TEHNIČKI GLASNIK

TEHNIČKI GLASNIK / TECHNICAL JOURNAL – GODIŠTE / VOLUME 19 – BROJ / ISSUE 4

PROSINAC 2025 / DECEMBER 2025 – STRANICA / PAGES 509-682

/

SVEUČILIŠTE SJEVER / UNIVERSITY NORTH – CROATIA – EUROPE

ISSN 1846-6168 (PRINT) / ISSN 1848-5588 (ONLINE)

TECHNICAL JOURNAL

ISSN 1846-6168 (Print) ISSN 1848-5588 (Online)

TEHNIČKI GLASNIK - TECHNICAL JOURNAL

Scientific-professional journal of University North

Volume 19 Issue 4
Varaždin, December 2025 Pages 509-682

Editorial Office:
Sveučilište Sjever / University North – Tehnički glasnik / Technical journal
Sveučilišni centar Varaždin / University Center Varaždin
Jurja Križanića 31b, 42000 Varaždin, Croatia
Tel. ++385 42 493 338, Fax.++385 42 493 336
E-mail: tehnickiglasnik@unin.hr
https://tehnickiglasnik.unin.hr
https://www.unin.hr/djelatnost/izdavastvo/tehnicki-glasnik/
https://hrcak.srce.hr/tehnickiglasnik

Founder and Publisher:
Sveučilište Sjever / University North

Print:
Centar za digitalno nakladništvo, Sveučilište Sjever

All manuscripts published in journal have been reviewed.
Manuscripts are not returned.

The journal is free of charge and four issues per year are published
(In March, June, September and December)
Circulation: 100 copies

Journal is indexed and abstracted in:
Web of Science Core Collection (Emerging Sources Citation Index - ESCI), Scopus, EBSCOhost Academic Search Complete, EBSCOhost – One Belt, One Road Reference Source Product, ERIH PLUS, CITEFACTOR – Academic Scientific Journals, DOAJ – Directory of Open Access Journals, Hrčak – Portal znanstvenih časopisa RH

Registration of journal:
The journal "Tehnički glasnik" is listed in the HGK Register on the issuance and distribution of printed editions on the 18th October 2007 under number 825.

Preparation ended:	Published (online):	Published (print):
August 28, 2025	September 15, 2025	December 15, 2025

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240227075152

Original scientific paper
Received: 2024-02-27, Accepted: 2024-08-01

Breast Cancer Diagnosis Using Machine Learning and PSO

Sarita Silaich*, Rajesh Yadav

Abstract: Healthcare systems around the world are facing huge challenges in responding to trends of the rise of chronic diseases. Early detection of breast cancer is essential for successful treatment since it is a common and potentially fatal condition. Based on clinical data, machine learning algorithms have shown potential in the categorization of breast cancer. This work aimed to build classification models Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Random Forest (RF), Logistic Regression (LR), and Artificial Neural Network (ANN) for Diagnostic Wisconsin Breast Cancer Dataset (WDBC) also improves classifier performance by using feature selection optimization and cross validation. The Particle Swarm Optimization (PSO) technique is used to select relevant, irredundant and most informative features that dependent on the performance of a classifier utilizing certain characteristics. The effectiveness of five distinct classifiers is assessed in this work. According to the findings, PSO-based SVM classifier has the greatest mean subset accuracy over a wide variety of training testing ratios.

Keywords: breast cancer; cross validation; Diagnostic Wisconsin Breast Cancer Database (WDBC); machine learning classifiers; Particle Swarm Optimization (PSO)

1 INTRODUCTION

Accurately diagnosing and detecting different ailments and diseases is a major problem in the fascinating and ever-evolving disciplines of bioinformatics and medical research. This effort necessitates not only a wealth of knowledge and skills, but also creative thinking and approach. Disease diagnosis is a challenging and dynamic field in medicine. An abundance of data on medical diagnoses may be accessible at clinics, healthcare facilities, colleges, and various research centers, as well as on the internet [1]. To make the system fully automated and rapid in diagnosing illnesses, classification is vastly used. Diseases are often diagnosed depending on the medical expertise of the medical planning officer. As a result of this, accurate illness diagnosis can be complicated and accompanied by cases of errors and unintended bias [2].

According to the report of WHO 2020, breast cancer is a major health concern and it afflicted 2.3 million individuals and killed 685,000, people. It is the world's most frequent illness and impacted 7.8 million individuals in the same year. Breast cancer kills more women than any other type of cancer worldwide. The disease affects women all around the world after puberty, with rates increasing with age. From 1990 to 2016, the age-standardized incidence rate of breast cancer in women rose by 39.1 % nationwide (with a 95 % uncertainty range of 5.1 to 85.5), according to Globocan statistics Breast cancer accounted for 13.5 % (178,361), of all cancer cases and 10.6 % (90,408), of total deaths in India in 2020 [3]. Breast cancer is a significant health concern, especially for women, as it is the most ubiquitous type of cancer, as indicated by the data presented.

As per UK cancer statistics 41,000 women diagnosed breast cancer each year. Overall, breast cancer is the leading cancer in women globally. Accurately diagnosis breast cancer is crucial since the illness apparent in a variety of ways that need specialized testing and analysis. To deliver accurate diagnosis and personalized treatment regimens customized to the particular requirements of patients, a multidisciplinary approach integrating the experience of medical professionals, researchers, and bioinformaticians is required. To create novel diagnostic tests and treatment techniques, sophisticated computational tools are employed to analyze massive volumes of genomic data in search of patterns and mutations related with breast cancer [4].

As a result of having to examine a substantial quantity of image data during mammography, the degree of accuracy achieved by the test is diminished. This technique requires a considerable amount of time, and in the worst-case situation, it might yield an erroneous diagnosis of the ailment. In this research, we investigate and compare a wide variety of machine learning algorithms, each of which is capable of recognizing the sickness on the basis of the input features. Five distinct supervised machine learning strategies were applied in order to accurately diagnose the problem.

2 RELATED WORKS

The progress that has been made in the field of medical research has led to the development of a number of innovative breast cancer detection methods. The research that is pertinent to this field is briefly summarized in the following paragraphs. The challenge of correctly detecting breast cancer (BC) may be thought of as a classification issue. Researchers are applying a wide array of machine learning (ML) techniques, artificial neural networks, support vector machine algorithms, and a number of other data extraction methods in order to conduct an analysis of it. Because of their capacity to successfully capture complex nonlinear interactions among variables, the ANN, SVM, and PNN models have acquired real-world relevance in the area of classification modeling. This is due to the fact that they may be used to model classification problems. It has been shown that functional weaving is a very effective method for identifying patterns statistically while also providing a nonlinear forecast of their complexity. It may be possible to profit from the use of machine learning in order to acquire a diagnosis that is both reliable and cost-effective. A method of machine learning known as the feed-forward artificial neural network (FFANN) is described here. This study aims to identify and classify breast cancer patients so that further research may be conducted [5].

In order to overcome challenges associated with optimization, metaheuristic optimization makes use of algorithms that are based on heuristics. Rami N. Khushaba and his colleagues [6] developed a method for selecting features that makes use of a strategy that involves differential evolution (DE) optimization in conjunction with a repair mechanism that is based on feature allocation evaluations. This method was proposed as a means of selecting features. In order to resolve the combinatorial optimization problems that is associated with featureselection, DEFS implements the distributed float number optimization. The construction of a wheel-like structure and the provision of opportunities for the distribution of features were carried out with the intention of making it easier to pick features via the use of the float optimizer. Within data sets that vary in the degree of dimensionality, DEFS has been used to search for optimal segments of characteristic characteristics in an effort to locate optimal solutions.

Gu et al. [7] have recently introduced a novel variant of the particle swarm optimization (PSO) algorithm, termed as competitive swarm optimizer (CSO), which is specifically designed to address the challenges of high- dimensional feature selection in large-scale optimization problems. The CSO, which was initially designed for uninterrupted operation, was included.

The concept of optimization originated with the aim of conducting feature selection, which can be viewed as a problem of combinatorial optimization. An approach for documentation was implemented with the aim of decreasing computationally expenses. The authors conducted experiments on six standard datasets and compared their proposed CSO-based feature decision-making algorithm with a canonical PSO-built algorithm and innovative PSO variants. Their findings showed that the CSO-based algorithm selected a significantly fewer number of features while accomplishing the best classification accuracy.

Moradi et al. [8] proposed the utilization of localized search tactics, integrated into particle swarm optimization, for finding a subset of features that is both less related as well as salient. This approach was referred to as HPSO-LS. The key goal of employing the local hunt technique was to facilitate the algorithm for searching of the particle swarm optimization in selecting distinctive features by considering their correlation data. Additionally, the proposed approach employs a scheme for determining the size of the subset in order to select a reduced set of features.

The usefulness of the proposed tactic has been evaluated on a set of 13 standard classification tasks and compared to five contemporary methods for selecting features. Furthermore, HPSO-LS has been compared to four established filter-based techniques, namely information gain, terms variation, fisher score, and mRMR, as well as five established wrapper-based techniques such as particle swarm optimization, genetic algorithm, simulated annealing, and ant colony optimization . The findings indicate that the proposed approach enhances the classification accuracy to a level comparable to that of filter-based and wrapper-based feature selection techniques [9].

More particularly, a technical taxonomy of the chosen material, including hybridization, improvement, and PSO variations, is examined in this work together with previous research on methodologies and applications published between 2017 and 2019 [10]. They are a targeted application of the algorithm categorized for the actual world. SVM and ANN classification algorithms were used by Bayrak et al. in their work to forecast breast cancer using a Wisconsin Breast Cancer Data Collection. The Sequential Minimal Optimization (SMO) and LibSVM methods were used to classify Support Vector Machines (SVM). The authors used the WEKA programming tool to categorize Support Vector Machine (SVM) using Multilayer Perceptron (MLP) and accepted perception methods. By combining SMO-SVM with the 10-fold cross-validation method, the authors were able to attain a high accuracy of 96.9957 % [11].

Sakri et al. [12] focus was on increasing accuracy by combining the PSO algorithm with the MI algorithms K-NNs, Naive Bayes (NB), and (REP) tree. One of the biggest issues in Saudi Arabia, according to their research, is the prevalence of breast cancer among women. Their research indicates that ladies over the age of 46 are the disease's main victims. The authors utilized five phase-based data analysis approaches to the WBCD dataset. Their final report was based on a comparison of feature selection methods used for classification with and without them. Juneja K, Rana [13] created updated decision tree technique, known as a weight-enhanced decision tree to detect breast cancer on dataset that was retrieved from the UCI repository. Through the use of the Chi-square test, they concluded that they had rated each characteristic and kept those that were relevant to this categorization assignment. Their suggested strategy produced results of around 98 % and 85-90 % precision on the Wisconsin Breast Cancer Diagnosis (WBCD) benchmark dataset, Yue et al. [14] reported thorough analyses of four classifiers, including SVM, K-NNs, ANNs, and Decision Tree approaches for predicting breast cancer. For training and assessment, four-fold cross-validation is used. SVM obtains the best levels of accuracy, specificity, and sensitivity during the training phase, with scores of 97.71 %, 98.9 %, and 97.08 %, respectively. Senapati et al. [15] used local linear wavelet neural network for breast cancer recognition. The Recursive Least Square (RLS) approach enhances the efficacy of training parameters, notably refining the suggested model to unveil the intricate connection weights among neurons in both the hidden and output layers. Through comparison with traditional methods, this approach proves its resilience and robustness, showcasing its superiority in performance.

3 MATERIALS AND METHODS

3.1 Breast Cancer Dataset

The Breast Cancer Dataset, commonly known as the Diagnostic Wisconsin Breast Cancer Database (WBCD), was collected from the UCI machine learning repository. Dr. William H. Wolberg produced the dataset in the late 1980s, and it is accessible to the public via the UCI repository [16]. It includes details about clinical dataset 569 patient breast tissue samples. It includes different attributes calculated from digital photographs of the samples. The objective of the dataset is to predict whether a tissue sample is benign or malignant based on the attributes. In this case 357 (62.74 %) are benign and 212 (37.26 %) are malignant. With the exception of the patient id and diagnosis level, the attributes in the dataset shown in Tab. 1. In our study benign instances are regarded as the positive class and malignant cases as the negative class. The objective is to create a model that can correctly predict whether a certain instance is benign or malignant based on the values of the various features in the dataset.

Table 1 Attributes and Descriptions WDBC [16]

Sno.	Attributes Name	Description of Attribute
1	Radius	Mean of distance from center to points on cell nucleus perimeter
2	Texture	Standard deviation of gray scale values
3	Perimeter	Perimeter of Tumor
4	Area	Area of Tumor
5	Smoothness	Local variation in radius lengths
6	Compactness	($Perimeter^2$/Area) – 1
7	Concavity	Severity of concave partitions of the contour
8	Concave Point	Number of concave partitions of the contour
9	Symmetry	Symmetry in the cell nuclie
10	Fractal Dimension	Coastline Approximation – 1

3.2 Pre-Processing Dataset

The transformation of raw data into a format that is appropriate for machine learning activities is what happens during the data pre-processing stage, which is an essential part of the data analysis pipeline. It seeks to enhance the quality of the data, get rid of errors, deal with missing values, and standardize the data so that it may be used by a variety of algorithms in an efficient manner. Because real-world datasets often include noise, mistakes, and variances, which may have a detrimental influence on the performance and accuracy of machine learning models, pre-processing of the data is very necessary [17].

3.3 Feature Extraction

To handle datasets with a large number of features, we used feature extraction methods to lower the combined dataset's dimensionality. High-dimensional datasets may result in overfitting and longer calculation times. Principal Component Analysis (PCA) and other dimensionality reduction algorithms are examples of feature extraction techniques that assist maintain important information while reducing the number of features. This step is critical in simplifying the dataset and improving classifier performance.

3.4 Data Normalization

Data normalization is a critical pre-processing step to ensure unbiased learning and feature scaling across all features. Since the datasets were combined from different sources, they might have different scales and ranges for their attributes [18]. Data must be transformed during normalization such that each attribute has a mean of 0 and a standard deviation of 1. This method equalizes the scale of all characteristics, preventing certain features from predominating the learning process due to their magnitudes.

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}.$$

3.5 Particle Swarm Optimization

In the case of breast cancer diagnosis, we used PSO for feature selection to find a subset of significant traits that may differentiate between benign and malignant tumors. This allows us to more accurately diagnose patients with breast cancer. The PSO algorithm's goal is to maximize a fitness function, which is often based on how well the classifier performs while using certain characteristics. In order to begin the process, a swarm of particles must be initialized, each of which will represent a possible feature subset. These particles move around the feature space by calculating velocity using Eq. (1) and update position as in Eq. (2) to search of the best possible combination of features. This process is repeated with the end goal of improving the classification accuracy.

$$v_{ij}^{t+1} = wv_{ij}^{t} + c_1 r_1 \left[pBest_{ij}^{t} - x_{ij}^{t} \right] + c_2 r_2 \left[gBest_{ij}^{t} - x_{ij}^{t} \right], \quad (1)$$

$$x_{ij}^{t+1} = x_{ij}^{t} + v_{ij}^{t+1}, \quad (2)$$

v_{ij}^{t} is velocity of i^{th} partical of j^{th} dimension at time t and x_{ij}^{t} is position of same, w is inertia weight, c_1, c_2 are cognitive learning factor, r_1, r_2 uniformly distributed random number between 0 and 1, *pBest* is its personal best value and *gBest* is global best value. Then position is converted in binary using sigmoid function.

$$x_{ij}^{t+1} = \begin{cases} 0 \text{ if } rand(\ ) \geq Sigmoid\left(v_{ij}^{t+1}\right) \\ 1 \text{ if } rand(\ ) \geq Sigmoid\left(v_{ij}^{t+1}\right) \end{cases}, \quad (3)$$

$$Sigmoid\left(v_{ij}^{t+1}\right) = \frac{1}{\left(1 + e^{-v_{ij}^{t}}\right)}. \quad (4)$$

3.6 Machine Learning

The pre-processed dataset was used in the technical paper that was presented, and five different machine learning classifiers were used to diagnose breast cancer based on the data. The following classifiers that are used.

3.6.1 Support Vector Machine (SVM)

It is a sophisticated and extensively used classification technique that seeks to discover an ideal hyperplane that divides various classes in the feature space. This goal of the method is to find an optimal separation between classes. It performs well with both data that can be separated linearly and data that can be separated non-linearly [20].

3.6.2 K-closest Neighbours (KNN)

It is a basic yet effective classification method that classifies a data point based on the majority class among its k closest neighbours in the feature space. This technique classifies a data point based on the majority class among its *k* nearest neighbours in the feature space [21].

3.6.3 Random Forest (RF)

It is often known as RF, is a strategy for ensemble learning that mixes numerous decision trees in order to enhance accuracy and decrease overfitting. Every tree receives its training based on a different selection of characteristics and data points [22].

3.6.4 Logistic Regression (LR)

It is a method of binary classification that makes use of a logistic function in order to make forecasts about the likelihood of a given instance belonging to a certain class. These forecasts may be made based on the function. To put it another way, it calculates the chance that a certain category is satisfied by a given occurrence [23].

3.6.5 Artificial Neural Network (ANN)

It is a model of reasoning based on the human brain. A conventional ANN model contains a hierarchy of layers: input layer, hidden layers and output layer which are composed of interconnected neurons containing an activation function for nonlinear transformation. In ANN model input layer receives the data also called features and transmits the data to a hidden layer where data is processed and trained results are provided at the output layer. Some ANN training process may involve long causal chains of computational stage depending on complexity of the problem.

3.7 Methodology

Methodology flow chart is shown in Fig. 1 which starts by collecting Wisconsin Breast Cancer Diagnostic Dataset. After applying normalization, ten cross-validations applied and then the performance of various classification algorithms is evaluated both using PSO as feature selection and without PSO.

A well-known method for accurately evaluating the effectiveness of classifiers is called K-fold cross-validation as it tune the hyper parameters, and it was used here to evaluate each of these classifiers. During 10-fold cross-validation, the dataset is divided into ten subsets known as folds. Each classifier is then trained and tested ten times, with each iteration uses a different fold from the dataset as the test set. This will apply the appropriate values of hyper parameters to produce high performance that are more trustworthy and can be used more broadly.

Performance is evaluated using confusion matrix as shown in Tab. 2.

Table 2 Confusion matrix

		Predicted	
		N	P
Actual	N	True Negative (*TN*)	False Positive (*FP*)
	P	False Negative (*FN*)	True Positive (*TP*)

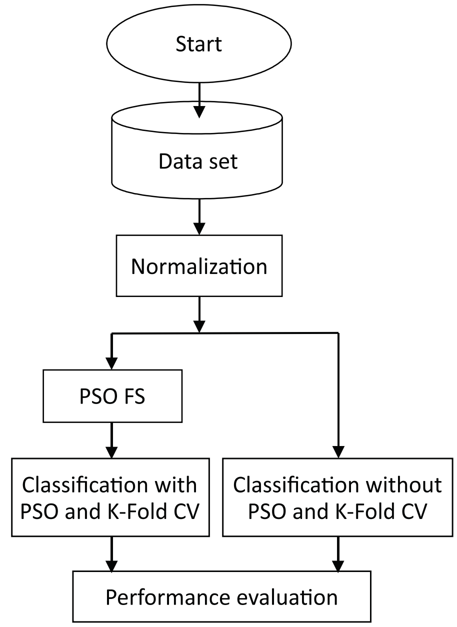


Figure 1 Methodology

A good model is a model that can predict correctly label. Accuracy, recall (also called sensitivity or True Positive Rate), and precision equation are given as below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{5}$$

$$Recall = \frac{TP}{TP + FN}, \tag{6}$$

$$Precision = \frac{TP}{TP + FP}. \tag{7}$$

4 RESULTS

Applying PSO feature selection the Tab. 3 is resultant of the selected features. Tab. 3 show the accuracy using selected feature by PSO (acc_pso) and using all features (acc_all), precision using PSO (pre_pso) and precision using all featues (pre_all), recall using PSO (re_all) and recall using all features (re_pso) for different training ratios. It is observed that except ANN all other remaining classifiers have improved accuracy with PSO. SVM using PSO has highest testing accuracy 97.81 %.

Result shows that using PSO performance of all classifiers is improving when size of test data is increasing that is logical but without PSO performance is not in increasing order as number of samples increasing in test data. Recall or sensitivity is improved for all classifiers means false negative is less, which is more important in medical because more false positive is acceptable but high false negative is more dangerous as disease becomes incurable. Also using PSO resultant less features trained models are simple and more generalized without compromising accuracy.

The graph representations of accuracy analysis of all algorithms with feature selection and without feature selection are shown in following graphs. That shows that using PSO as FS enhanced the accuracy for several reasons:

Feature Selection: The WDBC dataset likely contains numerous features, and not all of them might be relevant or contribute equally to the classification task. PSO helps in selecting a subset of features that are more discriminative, reducing noise and irrelevant information. This can lead to better classification accuracy by focusing on the most informative features.

Table 3 Test result analysis

SVM	acc_all	acc_pso	pre_all	pre_pso	re_all	re_pso
90-10	96.49	92.98	97.22	97.06	97.22	91.67
80-20	97.75	96.61	98.60	97.18	98.60	95.83
70-30	97.66	97.49	98.13	98.13	98.30	97.20
60-40	96.93	**97.81**	96.58	**97.92**	98.60	**98.60**
KNN						
90-10	98.25	96.49	97.30	97.22	99.00	97.22
80-20	98.25	93.86	97.30	97.10	98.00	93.06
70-30	97.60	94.74	96.40	95.37	98.00	96.26
60-40	96.05	95.18	95.89	95.21	97.90	97.20
RF						
90-10	96.49	94.74	97.22	97.14	97.22	94.44
80-20	94.74	95.61	95.83	97.18	95.83	95.83
70-30	95.32	95.32	95.41	95.14	97.20	97.20
60-40	94.74	96.49	95.17	96.55	96.50	97.90
LR						
90-10	96.49	98.25	97.22	97.30	97.22	99.00
80-20	95.01	96.49	94.67	95.95	98.61	98.61
70-30	94.74	94.74	92.98	92.24	99.07	99.00
60-40	95.80	95.61	94.00	94.04	98.60	99.30
ANN						
90-10	96.49	92.98	97.22	97.06	97.22	91.67
80-20	97.37	95.61	98.59	97.18	97.22	95.83
70-30	98.25	95.91	99.06	96.30	98.30	97.20
60-40	98.68	96.61	99.30	96.50	98.60	96.50

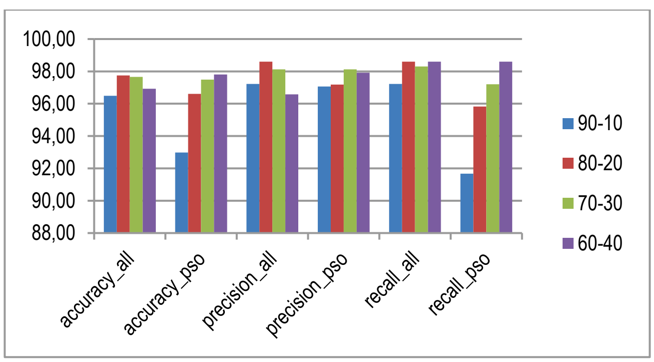


Figure 2 SVM classifier results with and without PSO

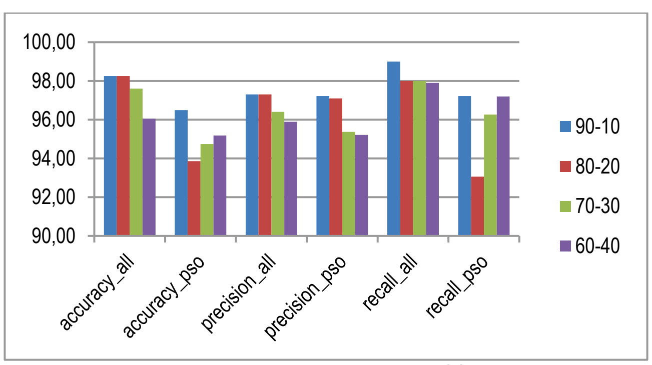


Figure 3 KNN with and without PSO

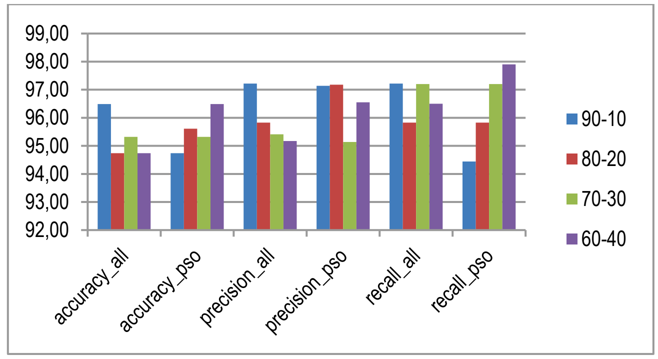


Figure 4 RF classifier results with and without PSO

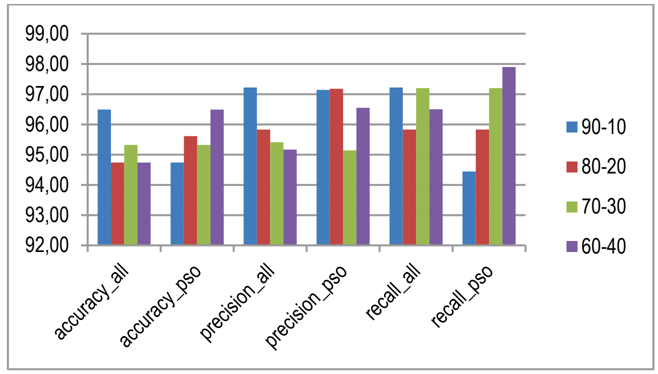


Figure 5 LR with and without PSO

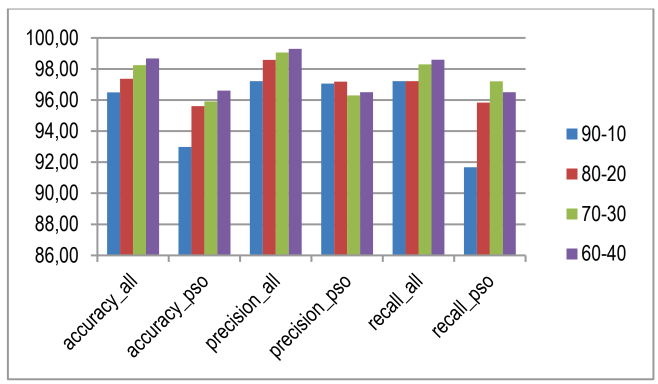


Figure 6 ANN classifier results with and without PSO

More insight into misclassification can be done using confusion matrix. For SVM train test split 60:40 (341 Training Samples and 228 test samples) confusion matrix are as follows:

Table 4 Confusion matrix for SVM with all Features

		Predicted	
		N	P
Actual	N	True Negative (TN) = 80	False Positive (FP) = 5
	P	False Negative (FN) = 2	True Positive (TP) = 141

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{80 + 141}{80 + 141 + 5 + 2} = 96.93$$

$$Recall = \frac{TP}{TP + FN} = \frac{141}{141 + 2} = 98.60$$

$$Precision = \frac{TP}{TP + FP} = \frac{141}{141 + 5} = 96.58$$

Table 5 Confusion matrix for SVM and PSO Selected Features

		Predicted	
		N	P
Actual	N	True Negative (TN) = 82	False Positive (FP) = 3
	P	False Negative (FN) = 2	True Positive (TP) = 141

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{82 + 141}{82 + 141 + 3 + 2} = 97.81$$

$$Recall = \frac{TP}{TP + FN} = \frac{141}{141 + 2} = 98.60$$

$$Precision = \frac{TP}{TP + FP} = \frac{141}{141 + 3} = 97.92$$

These results show that using PSO feature selection misclassifications are less as compared to using all features.

Search Space Exploration: PSO explores the search space efficiently by evaluating different combinations of features. It optimizes the feature subset selection by iterating through potential solutions and gradually converging towards an optimal or near-optimal solution. This exploration allows it to find a feature subset that works best for the classification task.

Handling Redundancy and Correlation: Sometimes, certain features in a dataset might be redundant or highly correlated. PSO can handle such cases by identifying and selecting only the most relevant features while discarding redundant or highly correlated ones. This prevents overfitting and improves the generalization of the classifier.

Improving Model Efficiency: By reducing the number of features, PSO can also lead to more efficient models in terms of computational resources and time. With fewer features as shown in Fig. 7, the model complexity decreases, potentially speeding up the training and inference processes.

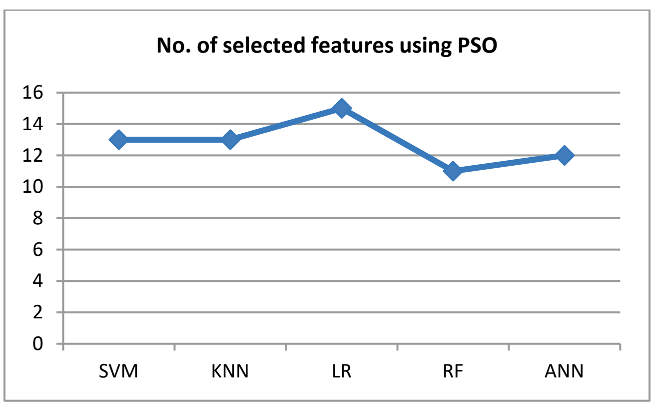


Figure 7 Count of selected features by classifiers using PSO

5 CONCLUSION

Our main aim was to explore how the integration of a feature selection algorithm with classification algorithms impacts breast cancer prognosis. Also 10-fold cross validation used to tune the hyper parameters and different split ratios applied to explore the result. By reducing the number of features, we aimed to highlight the significance and impact of specific features on the final results. We have shared findings from experiments on five widely-used classification algorithms: SVM, KNN, Random Forest, Logistic Regression and ANN for both with and without the PSO feature selection method. We can conclude that using less number of features selected by PSO classifiers accuracy is not compromised much, while using feature selection, models become simple and generalized. From results it can be concluded that SVM performed better using PSO in comparison with remaining classifiers.

6 FUTURE SCOPE

In our study we worked on single objective function as a fitness function on texted dataset. Future work can be to explore multi objectives function as fitness function in combination of PSO with deep learning on complex imaging and very large dataset.

7 REFERENCES

[1] Bhardwaj, R., Nambiar, A. R. & Dutta, D. (2017). A Study of Machine Learning in Healthcare. *The 41st IEEE Annual Computer Software and Applications Conference*, 236-241. https://doi.org/10.1109/COMPSAC.2017.164

[2] Kuehnert, M. J. & Qualls, C. (2019). Managing Healthcare Data Quality. 3rd ed. Chicago, IL: Health Administration Press.

[3] International Agency for Research on Cancer. India Source: Globocan 2020. Available from: https://gco.iarc.fr/today/data/factsheets/populations/356-india-fact-sheets.pdf

[4] Jang, B.-S. & Kim, I. A. (2020). Machine Learning Algorithms and Whole Exome Sequencing Data from Breast Cancer Patients in the UK Biobank Predict Survival. https://doi.org/10.21203/rs.3.rs-115867/v1

[5] Tripathy, S. (2020). Investigation of the FFANN Model for Mammogram Classification Using an Improved Gray Level Co-occurrences Matrix. *International Journal of Advanced Science and Technology, 29*. 4214.

[6] Khushaba, R. N., Al-Ani, A. & Al-Jumaily, A. (2011). Feature subset selection using differential evolution and a statistical repair mechanism. *Expert Systems with Applications, 38*(9), 11515-11526. https://doi.org/10.1016/j.eswa.2011.03.028

[7] Gu, S., Cheng, R. & Jin, Y. (2018). Feature selection for high-dimensional classification using a competitive swarm optimizer. *Soft Computing, 22*, 811-822. https://doi.org/10.1007/s00500-016-2385-6

[8] Moradi, P. & Mozhgan, G. (2016). A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Applied Soft Computing, 43*, 117-130. https://doi.org/10.1016/j.asoc.2016.01.044

[9] Xie, S. et al. (2022). Using SVM and PSO-NN Models to Predict Breast Cancer. In: Liu, Q., Liu, X., Cheng, J., Shen, T., Tian, Y. (eds) *Proceedings of the 12th International Conference on Computer Engineering and Networks, CENet 2022*. Lecture Notes in Electrical Engineering, vol 961. Springer, Singapore. https://doi.org/10.1007/978-981-19-6901-0_74

[10] Gad, A. G. (2022). Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review. *Archives of Computational Methods in Engineering, 29*, 2531-2561. https://doi.org/10.1007/s11831-021-09694-4

[11] Mustapha, M. T., Ozsahin, D. U. et al. (2022). Breast Cancer Screening Based on Supervised Learning and Multi-Criteria Decision-Making. *Diagnostics, 12*, 1326. https://doi.org/10.3390/diagnostics12061326

[12] Sakri, S. B., Rashid, N. B. A. & Zain, Z. M. (2018). Particle swarm optimization feature selection for breast cancer recurrence prediction. *IEEE Access, 6*, 29637-29647. https://doi.org/10.1109/ACCESS.2018.2843443

[13] Juneja, K. & Rana, C. (2020). An improved weighted decision tree approach for breast cancer prediction. *Int. j. inf. tecnol. 12*, 797-804. https://doi.org/10.1007/s41870-018-0184-2

[14] Yue, W. et al. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs, 2*(2), 13. https://doi.org/10.3390/designs2020013

[15] Senapati, M. R. & Mohanty, A. K. et al. (2013). Local linear wavelet neural network for breast cancer recognition. *Neural Computing Appl., 22*(1), 125-131. https://doi.org/10.1007/s00521-011-0670-y

[16] Wolberg, W. (1992). Breast Cancer Wisconsin (Original). UCI Machine Learning Repository. https://doi.org/10.24432/C5HP4Z

[17] Amato, A. et al. (2023). Data preprocessing impact on machine learning algorithm performance. *Open Computer Science, 13*. https://doi.org/10.1515/comp-2022-0278

[18] Zhao, Z. et al. (2019). Capsule Networks with Max-Min Normalization. https://doi.org/10.48550/arXiv.1903.09662

[19] Osareh, A. & Shadgar, B. (2010). Machine learning techniques to diagnose breast cancer. *The 5th International Symposium on Health Informatics and Bioinformatics*, Ankara, Turkey, 114-120. https://doi.org/10.1109/HIBIT.2010.5478895

[20] Singh, N. (2023). Support Vector Machine (SVM).

[21] Syriopoulos, P. K., Kotsiantis, S. B. & Vrahatis, M. N. (2022). Survey on KNN Methods in Data Science. In: Simos, D. E., Rasskazova, V. A., Archetti, F., Kotsireas, I. S., Pardalos, P. M. (eds) Learning and Intelligent Optimization. LION 2022. *Lecture Notes in Computer Science, vol 13621*. Springer, Cham. https://doi.org/10.1007/978-3-031-24866-5_28

[22] Lomakin, N., Pokidova, V. V. et al. (2023). Digital forecast of the efficiency of the enterprise based on the machine learning model Random forest. *Mezhdunarodnaja jekonomika (The World Economics)*. 411-425. (in Russian) https://doi.org/10.33920/vne-04-2306-06

[23] Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked, 30*, 100924. https://doi.org/10.1016/j.imu.2022.100924

Authors' contacts:

Sarita Silaich, Assistant Professor
(Corresponding author)
Government Polytechnic College,
W-6, Residency Road, Jodhpur, Rajasthan 342001, India
E-mail: sarita.bits@gmail.com

Rajesh Yadav, Assistant Professor Dr.
Mody University of Science & Technology,
Lakshmangarh 332 311, Dist. Sikar, Rajasthan, India
E-mail: yadav.rajesh27@gmail.com

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240426123331

Original scientific paper
Received: 2024-04-26, Accepted: 2025-01-02

A Comparative Structural Analysis of Total Hip Arthroplasty Designs

Elmedin Mešić*, Adela Zenkić, Adis J. Muminović, Nedim Pervan

Abstract: Structural analysis of two designs of total hip endoprosthesis was conducted with an emphasis on verifying stress conditions and displacement fields on the components and femur under loads typical for human gait. CAD (Computer Aided Design) models of a conventional endoprosthesis (Endoprosthesis 1) and a more modern solution (Endoprosthesis 2) were created using the CATIA (Computer Aided Three-dimensional Interactive Application) CAD/CAE (Computer Aided Engineering) system, based on which corresponding FEM (Finite Element Method) models were formed. By comparing the obtained results, it was found that a more uniform stress-strain pattern occurs on most components of Endoprosthesis 2 compared to Endoprosthesis 1. Additionally, the analysis of von Mises stresses and displacements at characteristic points of the femur showed a balanced distribution of stress and displacement. However, higher contact stresses between the stem and femur occur with Endoprosthesis 2, reducing the possibility of stress shielding due to the specifics of its design.

Keywords: CAD model; FEM model; stress shielding; structural analysis; total hip endoprosthesis; von Mises stress

1 INTRODUCTION

The occurrences of arthrosis, reduced mobility and hip joint fractures have led to the emergence and development of endoprostheses, which represent a clinical solution to these problems. Hip joint replacement is a surgical procedure aimed at improving the quality of life of individuals by removing damaged parts and replacing them with an endoprosthesis. Hip joint endoprostheses have evolved with advances in technology and science. From the first prostheses made from ivory (1891), through glass (1925) and the first metal-metal prostheses (1953), to prostheses similar to those used today, there has been significant progress. Indeed, Sir John Charnley, an orthopedic surgeon, is considered the father of modern total endoprostheses, having designed the first low friction arthroplasty in the early 1960s. Such prostheses typically involve femoral components made of metal and an acetabular component made of polyethylene or ceramics. Besides material changes, prostheses have also evolved in terms of design. There are one-piece (monolithic) or multi-piece (modular) prostheses, as well as partial prostheses that replace only the damaged part of the joint and total prostheses that involve replacing the entire joint body [1].

The development of design and selection of optimal materials for endoprostheses pose a challenge and serve as the foundation for new research. Carefully designed construction combined with the proper choice of materials creates an optimal combination for a hip endoprosthesis that will meet functional requirements. Balancing anthropometric requirements, biocompatibility, corrosion resistance, strength, toughness, material density, and other physical-mechanical properties is crucial in the construction of hip endoprostheses [1].

The modern era is characterized by the application of various CAD/CAE systems in many fields of science, including medical engineering. One of the many applications in this field is certainly in the development and testing of medical devices in orthopedics and orthodontics, such as external fixators [2 - 5] and mini-implants [6, 7].

The essential application of these systems is structural analysis of endoprostheses, where based on 3D CAD/FEM models, data on stress distribution and displacements can be obtained. In this way, the influence of the angle of the femoral component placement on the stresses generated on the femur and stem has been analyzed [8]. Also, an important parameter in the design of the endoprosthesis is considered to be the size of the femoral head component, so FEM analyses are conducted on its influence on the stress distribution of the prosthesis stem and femur [9]. Similarly, in the study, researchers analyze hybrid and conventional hip resurfacing implants using FEM analysis [10].

In addition to structural analysis of endoprostheses, tribological tests are conducted to reduce contact stresses and wear between components [11]. Thermoelastic stress analysis is used for obtaining and analyzing the complete stress distribution of the endoprosthesis and bone [12].

This research aims to conduct a comparative structural analysis of two designs of total modular hip endoprostheses, with a focus on verifying stress conditions and displacement fields on the components and femur under loading conditions typical for human gait.

Based on the obtained results, the influence of specific parameters of endoprosthesis design on the stress-strain picture has been determined.

2 THE DESIGN OF THE HIP ENDOPROSTHESES

In the first phase, CAD models of two analyzed designs of total hip endoprostheses were developed, where Endoprosthesis 1 represents a conventional modular solution typically consisting of a stem (1), head of femoral component (2), cup (3), and acetabular socket (4) (Fig. 1a). Endoprosthesis 2 represents a more contemporary modular solution for total hip endoprosthesis, which, in addition to the stem (1), includes components named according to the degrees of freedom they provide in relation to the acetabular socket (unidirectional cup) (5), namely abductor (2), rotator (3), and flexor (4) (Fig. 1b).

The fundamental difference of Endoprosthesis 2 compared to Endoprosthesis 1 lies in the design of the femoral head component and acetabular components, as well as the presence of an additional cup.

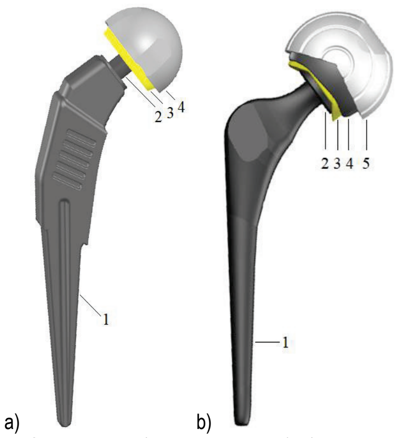
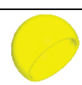
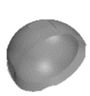


Figure 1 CAD models: a) Endoprosthesis 1; b) Endoprosthesis 2

Endoprosthesis 1 belongs to the group of ball-and-socket implants. On the other hand, Endoprosthesis 2 utilizes three uni-directional sliding motions through three orthogonally positioned cylindrical joints (flexor, rotator, and abductor). This enables Endoprosthesis 2 to have three degrees of freedom, i.e. flexion/extension, internal/external rotation and abduction/adduction.

The differences in design are evident in the images in Tab. 1, which provide an overview of the design of all components of both endoprostheses.

Table 1 Comparative overview of hip endoprosthesis designs

Endoprosthesis 1		Endoprosthesis 2	
Femoral components			
Stem		Stem	
Femoral head		Abductor (Femoral head)	
Acetabular components			
/		Rotator (Cup 1)	
Cup		Flexor (Cup 2)	
Acetabular socket		Unidirectional cup (Acetabular socket)	

The femoral head component as well as the acetabular components of Endoprosthesis 2 are characterized by larger dimensions and a tendency towards a spherical-conical shape. Additionally, the cross-section of the stem of the femoral component of Endoprosthesis 2 is circular-oval in shape, unlike Endoprosthesis 1 where the cross-section is more rectangular. The materials of the analyzed hip endoprostheses need to possess excellent mechanical properties and wear resistance while satisfying biocompatibility requirements. This has resulted in the use of titanium, tantalum, ceramics and UHMWPE (Ultra-High Molecular Weight Polyethylene).

The Tab. 2 presents the material specifications and an overview of components for which these materials are applied. In addition to the endoprosthesis materials, the properties of cortical bone are also shown in the table [8 -11].

Table 2 Properties of endoprosthesis materials and cortical bone

Type of material	Application component	Elastic modulus, GPa	Tensile strength, MPa	Poisson's ratio	Density, kg/m^3
Titanium	Stem and heads of femoral components; Flexor	114	825	0.34	4460
Zirconium	Cup; Rotator	38	4000*	0.22	3960
Tantalum	Acetabular socket	186	705	0.35	16690
UHMWPE	Unidirectional cup	1	25	0.32	970
Cortical bone	Femur	17.4	115	0.33	1800

Note: * denotes compressive strength

3 STRUCTURAL ANALYSIS OF ENDOPROSTHESES

The FEM modeling of endoprosthesis components and the femur was performed using the CAD/CAE system CATIA. It is important to note that all input factors for the structural analysis of Endoprostheses 1 and 2 are identical, which is essential for obtaining relevant results.

All components of both hip endoprostheses and the femur model were discretized using finite elements of the parabolic tetrahedron type (TE10) with ten nodes and a total of thirty degrees of freedom (translations) [13-16]. The TE10 finite elements are isoparametric elements with second-order displacement interpolation functions. The need for more accurate results often leads to mesh refinement on characteristic components and locations. A finer mesh was applied to the femoral heads and acetabular components (Fig. 2a). The materials of the endoprostheses and femur were modeled as linearly isotropic (Tab. 2).

During the FEM analysis, the femur was sectioned below the endoprosthesis femoral component, and then a constraint in the form of clamp was applied to the lower surface resulting from the bone resection (Fig. 2b). This bone preparation for structural analysis is based on an analogy with experimental tests where the lower part of the bone is impinged, following ISO 7206-4 and researchers' work [9, 17, 18].

Figure 2 a) Finite element mesh and b) clamp constraint (Endoprosthesis 2)

In FEM analysis of the assembly, it is necessary to define interactions (connections) between components that share common boundaries. A rigid contact connection was used to define the connection between the stem of the femoral component and the femur, as well as between the stem and the head of the femoral component (Fig. 3a). The connection between the head of the femoral component and the acetabular cup, and between the cup and the acetabular bed (Endoprosthesis 1), as well as between the abductor and rotator, rotator and flexor, and flexor and unidirectional cup (Endoprosthesis 2), are defined as smooth contact connections to achieve appropriate rotations (Fig. 3b).

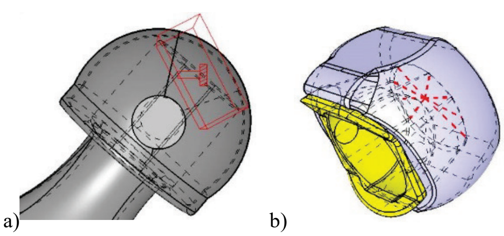


Figure 3 Contact connections (Endoprosthesis 2): a) rigid; b) smooth

The weight of the body above the hip joint and the abductor muscle force influence the generation of the resultant spatial force (R). The intensity and direction of the resultant force vary depending on the intensity of the acting forces and their lever arms, as well as the body's motion state. The structural analysis of the endoprostheses analyzed in this study was conducted for the case of normal walking of a person weighing 75 kg, where the resulting force amounts to 3 kN and acts at an angle of approximately 20° relative to the vertical axis [13, 15, 16]. The load from the pelvis is evenly applied to the upper outer surface of the acetabular bed, while the intensity, direction, and orientation are defined by the values of force components in the x, y, and z directions to achieve the required resultant force and angle of action (Fig. 4).

For the purpose of comparing the effects of two different endoprosthesis designs on the stress distribution and displacement fields on the femur, 12 characteristic points have been defined (Fig. 5).

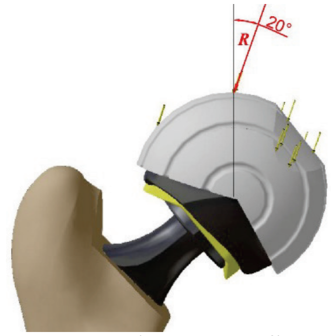


Figure 4 Load (Endoprosthesis 2)

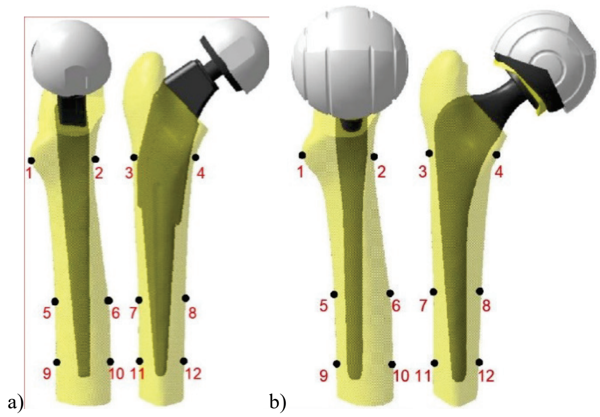


Figure 5 Characteristic points of the femur in conjunction with: a) Endoprosthesis 1; b) Endoprosthesis 2

3.1 Von Mises Stress Analysis

The display of von Mises stresses in the longitudinal section of Endoprosthesis 1 and Endoprosthesis 2 in conjunction with the femur is shown in Fig. 6. For the same loading and under the same boundary conditions, the stress pattern of Endoprosthesis 2 is more relaxed compared to the stress pattern of Endoprosthesis 1. The maximum von Mises stress in Endoprosthesis 1 occurred at the junction between the head and the stem (264 MPa), while in Endoprosthesis 2, it appears in the medial neck area (196 MPa), i.e., in the areas where the cross-sectional area of the endoprostheses is smallest which makes the specified locations critical points for the occurrence of fractures, especially with regard to fatigue [19-22].

In the case of Endoprosthesis 2, the results have shown that the von Mises stresses were higher at the interface between the stem and cancellous bone compared to Endoprosthesis 1 (Fig. 6). This indicates the potential of the Endoprosthesis 2 to reduce stress shielding. Additionally, studies [10, 16-18] associate contact stresses between the stem and femur with the occurrence of stress shielding.

A review of von Mises stresses for individual components of the endoprostheses is provided in Tab. 3. Due to the significant differences in stress levels across components, three stress scales have been introduced. The first scale pertains to femoral components, the second to acetabular components, and the third to the femur of both prostheses. Local maximum stresses occur at locations where there is a change in geometry and where the load from the previous component is significantly transferred to the observed component. The stress magnitude slightly varied from the acetabulum region to the femur region. Lower stresses are present on most components of Endoprosthesis 2. An exception is the flexator, which exhibits a higher stress by around 20 MPa compared to the cup of Endoprosthesis 1. A noteworthy finding is that the acetabulum socket of Endoprosthesis 2 has a stress over 100 MPa lower than the equivalent component of the conventional design (Tab. 3).

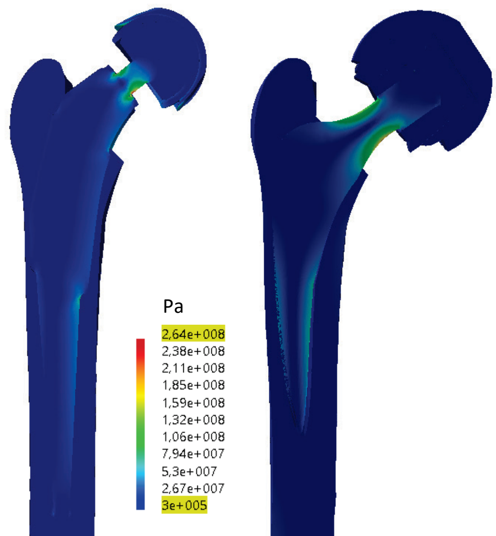
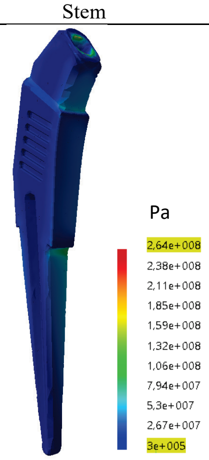
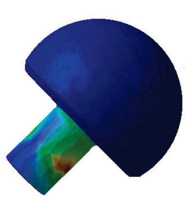
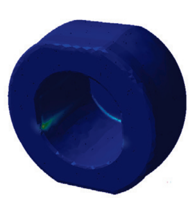
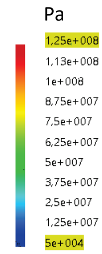
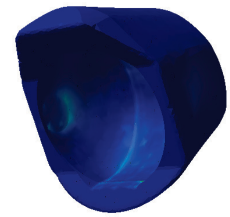
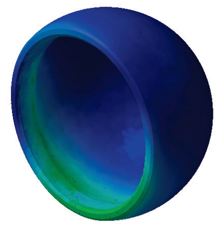
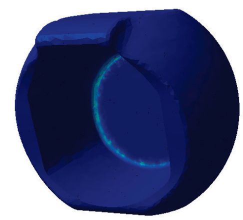
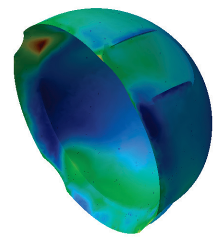
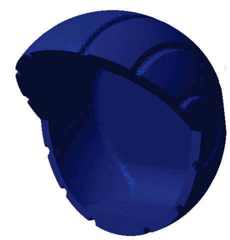
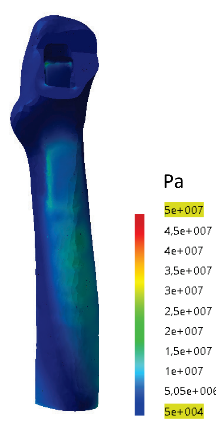


Figure 6 Von Mises stresses (longitudinal section) a) Endoprosthesis 1; b) Endoprosthesis 2

Table 3 Von Mises stresses of hip endoprosthesis components

Endoprosthesis 1	Endoprosthesis 2
Femoral components	
Stem	Stem
	
Femoral head	Abductor (Femoral head)
	
Acetabular components	
	Rotator (Cup 1) 
Cup	Flexor (Cup 2)
	
Acetabular socket	Unidirectional cup (Acetabular socket)
	
Femur	
	

It is evident that Endoprosthesis 2 is characterized by a more uniform stress distribution. The reason for this is certainly the design of Endoprosthesis 2, primarily the circular-oval cross-section of the stem with larger radii of curvature on the medial and lateral sides, larger dimensions of the femoral head component, and consequently larger dimensions of the acetabular components compared to Endoprosthesis 1. The minimum stress is observed at the upper part of the femur, where no load transfer occurs, resulting in the expected stress pattern.

A review of the results at 12 characteristic points on the femur is provided graphically in Fig. 7 and indicates a relatively uniform stress pattern of the femur for both endoprostheses. Stresses at points 2, 9, 11, and 12 for the case of using Endoprosthesis 2 are slightly more pronounced compared to the case of using Endoprosthesis 1. The reason for this is the position of the stem of Endoprosthesis 2 within the femur, which is closer to the end of the bone at these points. This is a consequence of the geometry of the stem of Endoprosthesis 2, especially its lateral part. Stresses at the other characteristic points of the femur in the case of applying Endoprosthesis 2 are lower compared to the case of applying Endoprosthesis 1 (Fig. 7).

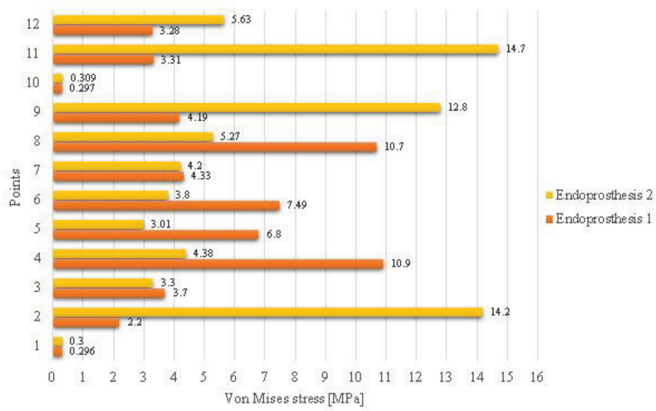


Figure 7 Von Mises stress at characteristic points of the femur

3.2 Displacement Analysis

The expected largest displacements are seen in the acetabular sockets (0.824 mm for the Endoprosthesis 1 and 0.479 mm for the Endoprosthesis 2), which represent the first link of the prosthesis-pelvis connection. The displacements of both hip endoprostheses are shown in Fig. 8. Tab. 4 provides an overview of the displacement results of individual components of both endoprostheses. The displacements of all components of both prostheses are shown using the same scale.

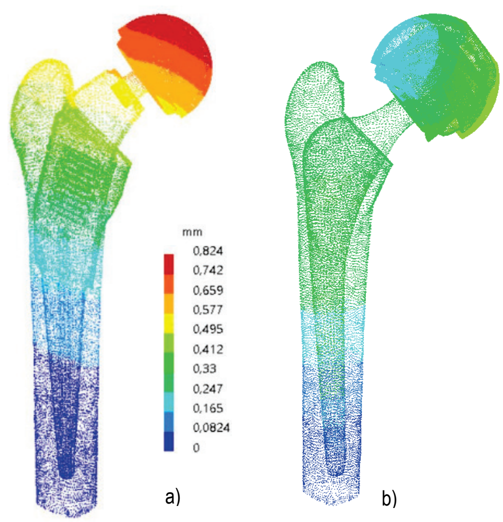


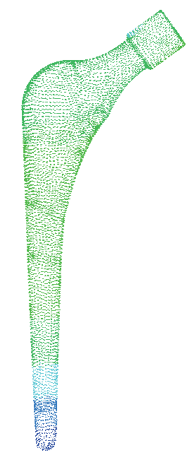
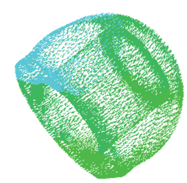
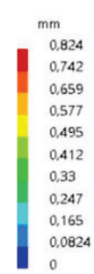
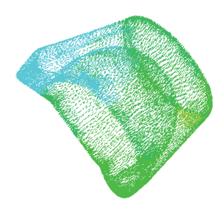
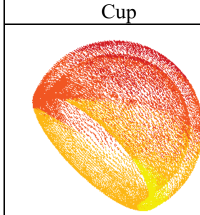
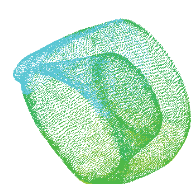
Figure 8 Displacements a) Endoprosthesis 1; b) Endoprosthesis 2

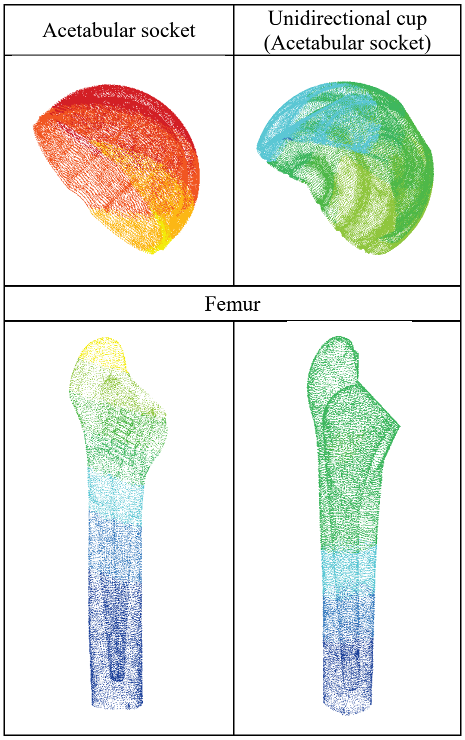
The overview of displacements of characteristic points of the femur shown in Fig. 9 indicates a consistent pattern of displacement in most observed points.

Displacements at points 1, 2, 4, 9, and 10 for the case of using Endoprosthesis 1 are slightly more pronounced than for the case of using Endoprosthesis 2. Displacements in other characteristic femoral points for the application of Endoprosthesis 1 are smaller compared to Endoprosthesis 2.

After conducting the analysis of von Mises stress and displacement, it is possible to identify a more relaxed stress-strain pattern with less pronounced extremes in the case of Endoprosthesis 2 compared to Endoprosthesis 1.

Table 4 Displacements of hip endoprosthesis components

Endoprosthesis 1	Endoprosthesis 2
Femoral components	
Stem	Stem
	
Femoral head	Abductor (Femoral head)
	
Acetabular components	
	Rotator (Cup 1)
	
Cup	Flexor (Cup 2)
	



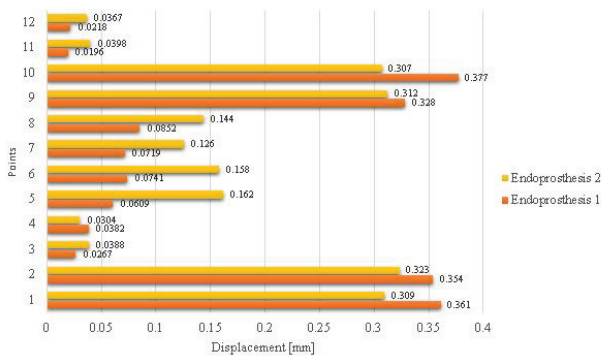


Figure 9 Displacement of characteristic points of the femur

3.3 Testing the Stability of the FEM Models

FEM solutions are essentially approximate and their accuracy, stability, and convergence can be questioned. Accuracy refers to the deviation of FEM solutions from the exact solution, while stability concerns the model's performance sensitivity to changes in significant factors. If a small change in a model parameter results in large changes in the solution, then the model is unstable. In the case where gradual increases in the number of finite elements do not lead to significant changes in displacement and stress results, convergence is achieved. If the solution is convergent, it is considered stable [23].

Stability and convergence check was performed by refining the finite element mesh in five iterations. Each iterative procedure was accompanied by a slight change in the number of finite elements. An overview of the number of nodes, number of finite elements, maximum displacement, and von Mises stress is provided in Tab. 5.

Table 5 Convergence of the FEM model of endoprosthesis.

Iteration	Number of nodes	Number of elements	Max. displacement, mm	Max. von Mises stress, MPa
Endoprosthesis 1				
I	161222	99900	0.827	278
II	164070	101637	0.827	272
III	169530	105486	0.825	266
IV	179013	111515	0.824	265
V	185831	116283	0.824	264
Endoprosthesis 2				
I	287128	180578	0.480	242
II	297516	186752	0.477	206
III	323385	204844	0.488	199
IV	338269	213779	0.484	196
V	344922	218321	0.479	196

Fig. 10 provides a graphical representation of the behavior of the maximum displacement values and maximum von Mises stresses of the hip endoprosthesis as a function of the number of finite elements in the FEM model.

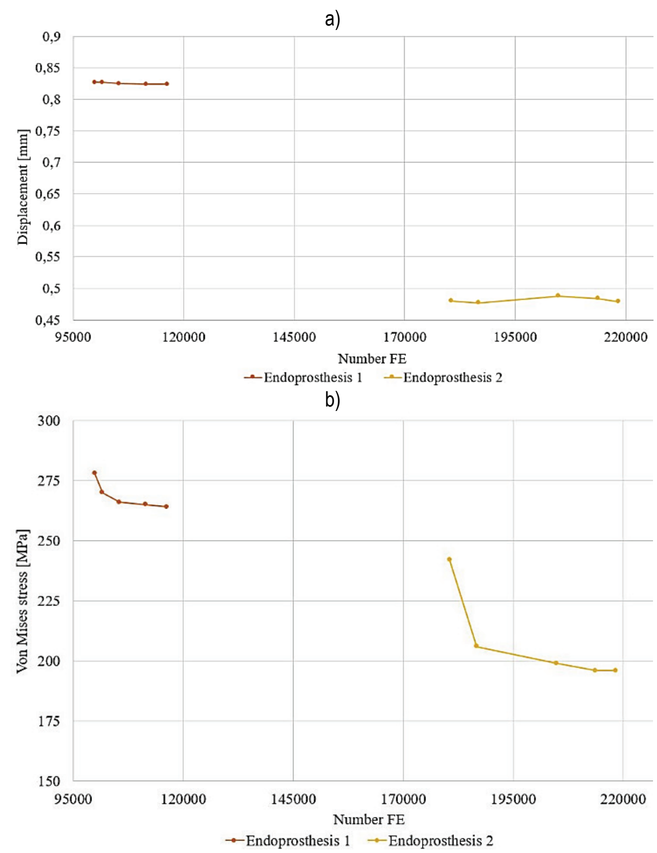


Figure 10 Stability diagram of the FEM model a) maximum displacement; b) maximum von Mises stress as a function of the number of elements

Through the displayed iterations, the size of the finite elements and the parameter 'sag' of all components of the FEM model were successively reduced. The 'sag' parameter represents the maximum deviation between the geometry of CAD and FEM model.

It is evident that with an increase in the number of finite elements, the values of maximum von Mises stresses exhibit a decreasing trend for both endoprostheses, with a clear convergence.

4 CONCLUSION

Improvement in the stability and the fixation of the stem in the bone are the main concerns for the surgeons during treatment. The design, material and application of the prosthesis are the crucial factors that influence the success of the treatment.

In this study, models of two total hip endoprostheses were created and analyzed with the aim of comparing the mechanical stability of conventional and modern solutions. The design and structural analysis were conducted in accordance with anthropometric-dimensional constraints, using a similar combination of materials.

The stability of the developed models was determined by examining the convergence of the FEM models of the endoprostheses.

When analyzing the obtained numerical results, it is evident that lower stresses and displacements occur on Endoprosthesis 2. However, the results showed that the von Mises stress value did not exceed the limit stresses of the femur material or prosthesis components. It was observed that with a longer radius on the lateral part of the stem of Endoprosthesis 2, a better distribution of the stresses was obtained, generating less displacement.

The significant difference in the distribution and magnitude of stress in the stem of Endoprosthesis 1 and the femur, with markedly lower contact stresses (44 MPa) compared to Endoprosthesis 2, is notable. On the other hand, the difference in the distribution and magnitude of stress in the stem of Endoprosthesis 2 and the femur is significantly smaller, with more pronounced contact stresses (93 MPa). Lower contact stresses between the stem of the endoprosthesis and the femur can lead to the so-called stress shielding phenomenon. The effect of stress shielding weakens the bond between the bone and the implant and is one of the serious problems in the long-term reliability of hip joint endoprostheses [16, 17]. Therefore, Endoprosthesis 2 indicates the potential to reduce stress shielding.

Also, the excessive displacement will drive implant loosening, which could affect the tissue healing response [17]. The maximum displacement is predicted at the proximal region of the endoprosthesis. The acetabulum loading applied to the prosthesis stem will lead to bending effects and further contribute to higher displacements at the proximal region of the endoprosthesis. Significantly larger displacements are present on all components and the femur in the case of Endoprosthesis 1, which can further lead to weakening of the bone-stem connection.

Based on the results of von Mises stress and displacement at characteristic points of the femur for both endoprosthesis solutions, it can be concluded that the position of the stem of the endoprosthesis plays an important role in the stress-strain pattern of the femur. Such observation provides a basis for further research in the field of optimizing the design and position of the endoprosthesis in the femur.

It can be concluded that Endoprosthesis 2 demonstrates its superiority by exhibiting a more relaxed stress-strain pattern primarily due to:

- Circular-oval cross-section of the stem of the femoral component
- Larger dimensions of the femoral head component
- Larger radii of curvature on the medial and lateral sides of the stem
- Larger dimensions of the acetabular components
- Spherical-conical shape of the acetabular components.

Additionally, thanks to the spherical-conical contact surfaces between the acetabular components, lower contact stresses are present.

The developed FEM models of endoprostheses have certain limitations in the sense that all materials are modeled as linear isotropic, which is certainly one direction for their improvement.

Finally, it can be stated that this study can serve as a basis for further research in the direction of redesign and material selection that performs an optimal function in the interaction of bone-implant. Also, the conducted research can serve as a basis for additional analysis and the development of new concepts of hip endoprostheses depending on the morphological needs and health issues that patients may have (such as osteoporosis or fractures) during surgery.

5 REFERENCES

[1] Knight, S. R., Aujla, R., & Biswas, S. P. (2011). Total Hip Arthroplasty - over 100 years of operative history. *Orthopedic Reviews, 3*(2), e16. https://doi.org/10.4081/or.2011.16

[2] Li, J., Zhao, X., Hu, X., Tao, C., & Ji, R. (2018). A theoretical analysis and finite element simulation of fixator–bone system stiffness on healing progression. *Journal of Applied Biomaterials & Functional Materials*, *16*(3), 115-125. https://doi.org/10.1177/2280800017750357

[3] Mešić, E., Pervan, N., Muminović, A. J., Muminović, A., & Čolić, M. (2021). Development of knowledge-based engineering system for structural size optimization of external fixation device. *Applied Sciences*, *11*, 10775. https://doi.org/10.3390/app112210775

[4] Amaro, A. M., Paulino, M. F., Roseiro, L. M., & Neto, M. A. (2020). The Efect of External Fixator Configurations on the Dynamic Compression Load: An Experimental and Numerical Study. *Applied Sciences*, *10*, 3. https://doi.org/10.3390/app10010003

[5] Mešić, E., Muminović, A., Čolić, M., Petrović, M., & Pervan, N. (2020). Development and Experimental Verification of a Generative CAD/FEM Model of an External Fixation Device. *Tehnički glasnik*, *14*(1), 1-6. https://doi.org/10.31803/tg-20191112161707

[6] Holberg, C., Winterhalder, P., Rudzki-Janson, I., & Wichelhaus, A. (2014). Finite element analysis of mono- and bicortical mini-implant stability. *European Journal of Orthodontics, 36*, 550-556. https://doi.org/10.1093/ejo/cjt023

[7] Mešić, E., Muratović, E., Redžepagić-Vražalica, L., Pervan, N., Muminović, A. J., Delić, M., & Glušac, M. (2021). Experimental & FEM Analysis of Orthodontic Mini-Implant Design on Primary Stability. *Applied Sciences*, *11*(12), 5461. https://doi.org/10.3390/app11125461

[8] Güvercin, Y., Yaylacı, M., Ölmez, H., Yaylacı, E. U., Özdemir, M. E., & Dizdar, A. (2022). Finite element analysis of the mechanical behavior of the different angle hip femoral stem. *Biomaterials and Biomedical Engineering, 6*(1), 29-46. https://doi.org/10.12989/bme.2022.6.1.029

[9] Chen, D. W., Lee, M. S., & Lin, C.-L. (2018). Finite Element Analysis of Stresses from Hip Implants with Different Head Sizes. *International Journal of Research Studies in Science, Engineering and Technology, 5*(5), 1-8.

[10] Bougherara, H., & Bureau, M. N., (2008). Biomimetic Composite-Metal Hip Resurfacing Implant. *Advances in Materials Science and Engineering*, vol. 2008, Article ID 368985, 4 pages. https://doi.org/10.1155/2008/368985

[11] Dalli, D., Buhagiar, J., Mollicone P., & Wismayer, P. S. (2022). A novel hip joint prosthesis with uni-directional articulations for reduced wear. *Journal of the mechanical behavior of biomedical materials*, *127*. https://doi.org/10.1016/j.jmbbm.2021.105072

[12] Zanetti, E. M., & Audenino, A. L. (2010). Differential thermography for experimental, full-field stress analysis of hip arthroplasty. *Journal of Mechanics in Medicine and Biology, 10*(3), 515-529. https://doi.org/10.1142/S0219519410003496

[13] Rezaei, F., Hassani, K., Solhjoei, N., & Karimi, A. (2015). Carbon/PEEK composite materials as an alternative for stainless steel/titanium hip prosthesis: A finite element study. *Australas. Phys. Eng. Sci. Med., 38*, 569-580.

[14] Guzmán, M., Durazo, E., Ortiz, A., Sauceda, I., Siqueiros, M., González, L., & Jiménez, D. (2022). Finite Element Assessment of a Hybrid Proposal for Hip Stem, from a Standardized Base and Different Activities. *Appl. Sci. 12*, 7963. https://doi.org/10.3390/app12167963

[15] Oshkour, A. A., Osman, N. A., Bayat, M., Afshar, R., & Berto, F. (2014). Three-dimensional finite element analyses of functionally graded femoral prostheses with different geometrical configurations. *Mater. Des., 56*, 998-1008. https://doi.org/10.1016/j.matdes.2013.12.054

[16] Bougherara, H., Zdero, R., Dubov, A., Shah, S., Khurshid, S. & Schemitsch, E. H. (2011). A preliminary biomechanical study of a novel carbon-fibre hip implant versus standard metallic hip implants. *Med. Eng. Phys*., 33, 121-128. https://doi.org/10.1016/j.medengphy.2010.09.011

[17] Konow, T., Glismann, K., Lampe, F., Ondruschka, B., Morlock, M. M., & Huber, G. (2023). Stem size and stemalignment affects periprosthetic fracture risk and primary stability in cementless total hip arthroplasty. *J Orthop Res.*, 1-8. https://doi.org/10.1002/jor.25729

[18] Ahmad, M. A., Zulkifli, N. N. M. E., Shuib, S., Sulaiman, S. H., & Abdullah, A. H. (2020). Finite Element Analysis of Proximal Cement Fixation in Total Hip Arthroplasty. *International Journal of Technology, 11*(5), 1046-1055. https://doi.org/10.14716/ijtech.v11i5.4318

[19] Sedmak, A., Čolić, K., Grbović, A., Balać, I., & Burzić, M. (2019). Numerical analysis of fatigue crack growth of hip implant, *Eng. Fract. Mech*. 216: 106492. https://doi: 10.1016/j.engfracmech.2019.106492

[20] Babić, M., Verić, O., Božić, Ž., & Sušić A. (2020). Finite element modelling and fatigue life assessment of a cemented total hip prosthesis based on 3D scanning. *Engineering Failure Analysis, 113*. https://doi.org/10.1016/j.engfailanal.2020.104536

[21] Reginald, J., Kalayarasan, M., Chethan, K. N., & P. Dhanbal, P. (2023). Static, dynamic, and fatigue life investigation of a hip prosthesis for walking gait using finite element analysis. *International Journal of Modelling and Simulation*, *43*(5), 797-811. https://doi.org/10.1080/02286203.2023.2212346

[22] van Doesburg, P. G., van Langelaan, E. J., Apachitei, I., Bénard, M. R., & Verdegaal, S. H. M. (2020). Femoral prosthesis neck fracture following total hip arthroplasty — a systematic review. *Arthroplasty* 2, 28. https://doi.org/10.1186/s42836-020-00047-3

[23] Zienkiewicz, O. C., Taylor, R. L., & Zhu, J. Z. (2005). *The Finite Element Method: Its Basis and Fundamentals*. 6th edition, Butterworth-Heinemann, Oxford, ISBN 0-7506-6320-0.

Authors' contacts:

Elmedin Mešić, PhD, Full Professor
(Corresponding author)
University of Sarajevo - Faculty of Mechanical Engineering,
Vilsonovo šetalište 9, 71000 Sarajevo, Bosnia and Herzegovina
+38733729800, mesic@mef.unsa.ba

Adela Zenkić, MSc in Mechanical Engineering, Designer
NewCold Metal Construction,
Žabljak bb, 74230 Usora, Bosnia and Herzegovina
adelazenkic@gmail.com

Adis J. Muminović, PhD, Full Professor
University of Sarajevo - Faculty of Mechanical Engineering,
Vilsonovo šetalište 9, 71000 Sarajevo, Bosnia and Herzegovina
+38733729800, adis.muminovic@mef.unsa.ba

Nedim Pervan, PhD, Full Professor
University of Sarajevo - Faculty of Mechanical Engineering,
Vilsonovo šetalište 9, 71000 Sarajevo, Bosnia and Herzegovina
+38733729800, pervan@mef.unsa.ba

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240514045737

Original scientific paper
Received: 2024-05-14, Accepted: 2024-07-18

Design and Implementation of a DLMS Server with a Multi-threaded Architecture for AMI Systems

Gwang Hyeon Kim, Yeong Rak Seong*, Ha Ryoung Oh

Abstract: In this paper, a DLMS server with a multi-threaded architecture for AMI systems is proposed and implemented. To achieve this, the operation procedures between the DLMS server and the clients are analyzed, and the necessary design requirements for the server are derived. The roles of the DLMS server are divided into multiple threads to meet these requirements, and the DLMS operation procedures are organized at the thread level. The proposed architecture is modeled using the DEVS formalism, a language for hierarchical modularization of discrete event systems. Subsequently, the general operational procedures and exceptional situations are simulated in the DEVSim++ environment. The simulation results show that the DLMS server with the proposed structure behaves correctly for each scenario and meets the DLMS standards and the given functional requirements. Finally, the DLMS server is implemented based on the validated simulation code. The implementation results confirm that the DLMS server operates correctly when connected to multiple smart meter devices. In the implemented DLMS server, unnecessary concurrency among threads is strictly limited, significantly reducing the costs associated with testing.

Keywords: Device Language Message Specification (DLMS); Discrete Event Systems; Multithreading; Simulation

1 INTRODUCTION

Smart grid is a next-generation intelligent power grid that provides improved power services and maximizes energy efficiency by upgrading the power grid through the use of information and communication technology (ICT) for the existing power grid. In general, the components of the Advanced Metering Infrastructure (AMI) which serves as the core of a smart grid, include smart meters, data collection units (DCUs), and meter data management systems (MDMS). In an AMI system, smart meters collect data and transmit it in real-time to the DCUs and MDMS. The DCU is responsible for transmitting the data collected by the smart meters to the MDMS, and the MDMS monitors and optimizes energy usage patterns based on the transmitted data [1]. Communication among the smart meter, DCU, and MDMS is important for the application of the AMI technology. However, the communication protocols vary depending on the smart meter manufacturer, which causes various problems [2].

The Device Language Message Specification (DLMS) protocol [3-5] was established to address these issues. DLMS protocol is a key communication standard in the AMI system of the smart grid, enabling efficient interoperability between smart meters, DCUs, and MDMS. The DLMS protocol uses a server-client structure based on TCP-IP or HDLC. In this structure, the server responds as the client requests the service. In a DLMS system, the upper application acts as the client, and the smart meter acts as the server. The server in a DLMS system is responsible for communicating with the clients and processing messages, managing various events, reporting results, and handling errors. These operations require concurrency, efficient resource utilization, fast response times, and stability. To handle these behaviors effectively, it is better to design the DLMS server as a multi-threaded structure that can run in parallel rather than a single-threaded structure that runs sequentially [6, 7].

This paper proposes a DLMS server with a multi-threaded architecture. Based on the operation procedure of the DLMS system, the requirements for the design on the DLMS server are identified and the threads needed to fulfill these requirements are derived. The general behavior of the DLMS server is specified on a separate thread level and the roles of each thread are clearly defined based on it. The designed multi-threaded structure is verified using the Discrete Event System Specification (DEVS) formalism [8-10], a language for describing discrete event systems. The multi-threaded structure, designed according to the roles of each thread, is modeled using the atomic and coupled models of the DEVS formalism. Subsequently, it is simulated using DEVSim++. The reason for using the DEVS formalism in this paper is that it can easily model multi-threaded systems, strictly control the system's behavior through a synchronization mechanism based on virtual time, and finally, easily simulate the modeled system using an abstract simulator algorithm. The validated DLMS server is directly implemented based on the simulation code and tested in a virtual environment.

This paper is organized as follows. Chapter 2 gives a brief introduction to the DLMS system, and Chapter 3 builds on Chapter 2 to propose a multi-threaded architecture for the DLMS server. In Chapters 4 and 5, the multi-threaded structure proposed in Chapter 3 is modelled and simulated using the DEVS formalism. In Chapter 6, a DLMS server is implemented based on the simulation results. Finally, Chapter 7 is the conclusion of this paper.

2 DLMS PROTOCOL

The DLMS protocol is a protocol in the energy and utilities sector, which was specified by the International Electrotechnical Commission (IEC) and the DLMS US. It is also known as IEC 62056 and is used for communication between smart meters and upper applications, as shown in Fig. 1.

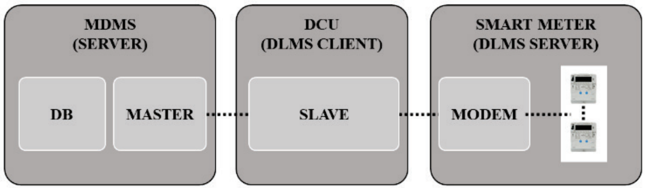


Figure 1 Interface between the smart meter and upper application

The protocol is designed to communicate not only with electricity meters but also with other devices, such as water and gas meters. The DLMS protocol defines three processes: modeling, messaging, and transport. Modeling phase is defined through Companion Specification for Energy Metering (COSEM) and Object Identification System (OBIS). COSEM is a standard data model for the interface and data identification of smart meters, structuring and defining data and services related to various metering functions. Each COSEM object performs operations such as reading, writing, and executing data through specific attributes and methods. OBIS is a system for identifying COSEM objects, uniquely identifying each data item. OBIS consists of six groups of numbers, each representing specific data items and functions. Messaging phase defines messages for accessing the modeled data. Messages include invoke ID and type, service parameters, priority, and so on, which determine and execute the operation mode based on these parameters. The main message types in DLMS are GET, SET, and ACTION, each having specific parameter values depending on the operation performed. Each message contains one or more service parameters, based on which the smart meter processes detailed operations. Finally, in transport phase, communication profiles are defined, which specify how the DLMS protocol can be used over various standard communication media. The DLMS protocol uses Transmission Control Protocol/Internet Protocol (TCP-IP) or High-Level Data Link Control (HDLC)-based communication, which basically works by exchanging messages [11-13].

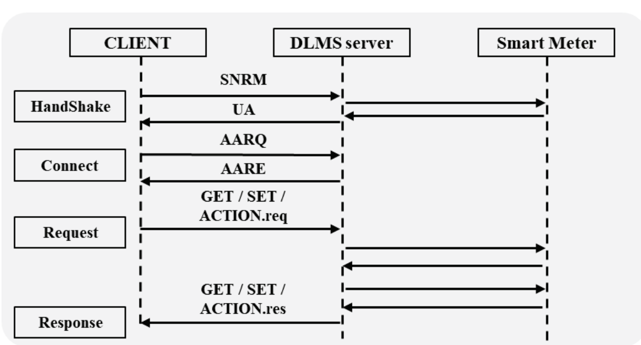


Figure 2 General operation procedure of the DLMS protocol

Fig. 2 shows the general operation procedure of the DLMS protocol. In the HandShake and Connect phases, when a client sends a connection request to the server to initiate communication with a specific smart meter, the server accepts it and securely establishes the connection through authentication and encryption. Subsequently, in the Request and Response phases, when the client sends a request message, the server gathers information from the smart meter regarding that request and delivers a response message back to the client. Further information on the DLMS protocol can be found in [4] and [5].

3 DESIGN

This chapter designs a DLMS server. The DLMS protocol documentation only describes the basics of DLMS, which allows DLMS servers to be designed in a wide variety of ways [14-17]. The design of the DLMS server considers the following requirements:

(i) The system should be able to handle request messages that come asynchronously from the client and exceptional situations that may occur unexpectedly in the smart meter.
(ii) The system should be able to record and manage the results and status information of the communications with the smart meter, and it should also be able to report the contents upon the client's request.
(iii) The system should be able to store and manage the parameters of the different request messages received from the client to the system.
(iv) The system should be able to communicate with the smart meter both wired and wirelessly.
(v) The system should support smart meters from different manufacturers to ensure interoperability.

To meet all of the requirements above, it is appropriate to design a DLMS server with a multi-threaded architecture that can handle multiple tasks in parallel. This paper divides the DLMS server into six types of threads. To satisfy requirement (i), the thread responsible for sending messages and the thread that receives messages are isolated from other threads. By separating the sending and receiving parts, communication with the client can be handled asynchronously, which makes communication more responsive. In addition, if any modifications to the communication interface are required, these can be kept to a minimum. In this paper, the sending and receiving parts of the message are separated from the other parts for the reasons mentioned above by using the SEND thread for the sending part and the RECV thread for the receiving part. To meet the requirement (ii), the REPORT thread is separated. The REPORT thread records and manages the processing results of the various request messages and the status of the smart meter and converts the data into the DLMS message format for reporting to the client. To fulfill the requirement (iii) and manage the overall behavior of the DLMS server, the MANAGER thread is also separated. The MANAGER thread stores and manages the messages received from the client and performs processing operations on them when they are ready for processing. To manage multiple smart meters and satisfy the requirements (iv) and (v) at the same time, the HANDLER threads and the uHANDLER threads are separated and employed. Depending on the type of command requested by the DLMS client, the DLMS server may need to communicate with the smart meter multiple times while processing a single command. Therefore, the HANDLER thread is responsible for splitting one DLMS message into a series of detailed operations according to the service parameters of the message and passing it to the uHANDLER thread, which is responsible for communicating with the smart meter according to the application programming interfaces (APIs) of the different manufacturers. Since several smart meters can be connected to one DLMS server, there are generally several HANDLER threads running on the DLMS server, and each HANDLER thread is connected to one uHANDLER thread.

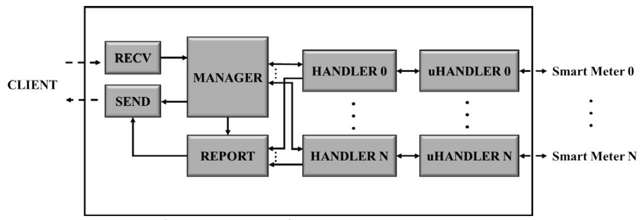


Figure 3 Communication flow between the designed threads

Fig. 3 shows the threads designed in this paper and the communication flow between the threads. As mentioned earlier, the RECV thread and SEND thread handle interactions with the client, establishing and maintaining connections with them. The HANDLER thread and uHANDLER thread, on the other hand, manage interactions with the smart meters, establishing connections with them. Between them, the REPORT thread stores data collected from the smart meters via the HANDLER thread, and upon completion of processing request messages, it sends response messages to the client. The MANAGER thread oversees the overall operations of the other threads.

With the separation of the DLMS server into six threads, the operation procedures of the DLMS server are refined and reorganized. Fig. 4 shows a schematic representation of the operation procedure of the DLMS server at a thread level for a GET message based on the general operation procedure in figure 2. When the client sends a GET message, ① it is first received by the RECV thread and ② passed to the MANAGER thread. The MANAGER thread is responsible for scheduling the processing of several received messages. Therefore, the MANAGER thread saves the received messages and forwards them to the ③ HANDLER thread and the ④ REPORT thread when they are ready to process the messages. The HANDLER thread then splits the received messages into specific detailed operations according to the service parameters and sends them to the ⑤ uHANDLER thread. Then, the ⑥ uHANDLER thread controls the smart meter directly connected to it to execute the directed detailed operations. Through the steps ⑦ - ⑨, the result is then passed to the REPORT thread, where it is saved. Finally, when all the detailed operations for the GET message have been completed, it is sent to the ⑬ client via the ⑩ MANAGER thread, the ⑪ REPORT thread, and the ⑫ SEND thread. As mentioned earlier, a single DLMS message requested by a client can be separated into a series of detailed operations. In Fig. 4, it should be noted that step ③, where the message is sent from the MANAGER thread to the HANDLER thread, occurs only once, whereas step ⑤, where the message is passed from the HANDLER thread to the uHANDLER thread, occurs twice.

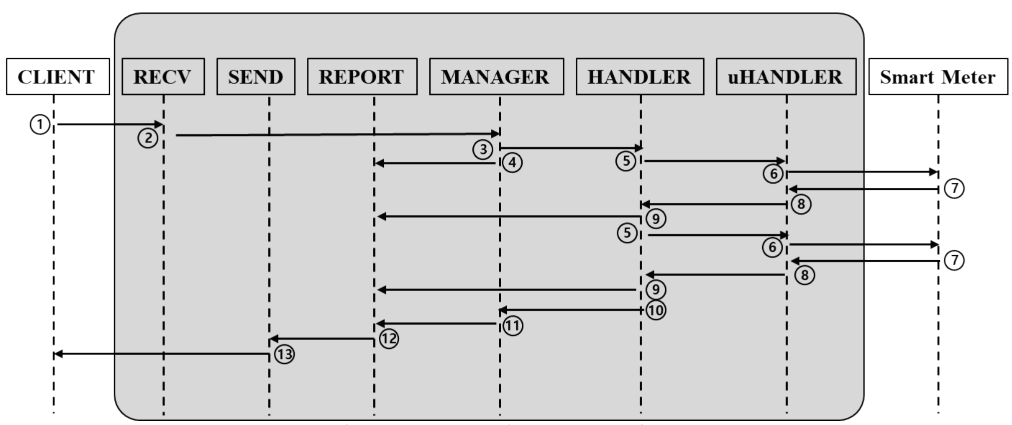


Figure 4 Operation procedure for processing a GET message

4 MODELING

This chapter models the proposed multi-threaded structure. Multi-threaded programs are more efficient than conventional single-threaded programs as they can execute multiple tasks simultaneously. They also have the advantage of being flexible enough to respond to events that occur outside the program. However, they are complex in design and difficult to verify compared to general programs. Multi-threaded programs do not always guarantee a consistent execution order as threads run in parallel and competitively with each other. Accordingly, even if the same events occur in the same order and at the same time, the operation result of the program may be different each time. The problems mentioned above make using the debugger program considerably more difficult [18].

To address these issues, the DEVS formalism is used, which describes discrete event systems in a hierarchical and modular way. The DEVS formalism is employed because (i) the behavior of a multi-threaded system can be modeled as a discrete event system, (ii) the behavior of thread structures can be tightly controlled by a synchronization mechanism based on virtual time, and (iii) the behavior of the modeled threads can be easily simulated using the abstract simulator algorithm of the DEVS formalism.

There are two types of models in the DEVS formalism: atomic models and coupled models. An atomic model describes the behavior of a component. An atomic model is defined by three sets of inputs (ports), outputs (ports), and state variables, and by four characteristic functions: external transition function, internal transition function, output function, and time advance function. The external transition function defines the change of state when an input is received, and the internal transition function defines the change of state over time. In addition, the output function defines the output when the state is changed by the internal transition function, and the time advance function defines how long it can remain in the current state. A coupled model, on the other hand, bundles several submodels into a single model and describes the hierarchical structure of the model and the connections between them. Further information on the atomic model and coupled model of the DEVS formalism can be found in [8].

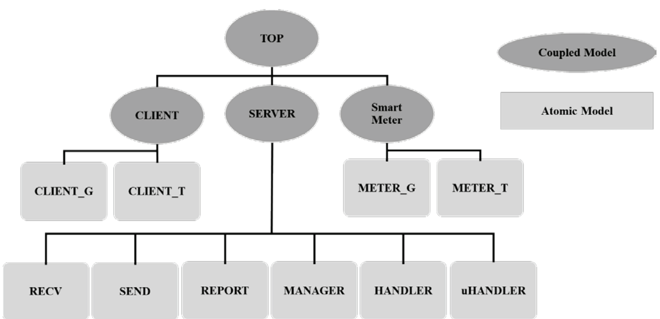


Figure 5 Hierarchical structure of the DLMS system

Each thread designed in Fig. 3 is modeled as an atomic model in the DEVS formalism. In addition, the message-passing relationships and the hierarchical organization between threads are modeled as a coupled model. The hierarchy of the models is shown in Fig. 5. The TOP model, which represents the entire system, consists of a CLIENT model, a SERVER model, and a SMART METER model. Of these, the SERVER model is a coupled model corresponding to the DLMS server proposed in this paper and contains the atomic models for the six types of threads described in Chapter 3. To model the asynchronous behavior of the DLMS client and the smart meter, the CLIENT model and the SMART METER model also consist of the CLIENT_G and METER_G models, which are responsible for generating and sending messages, respectively, and the CLIENT_T and METER_T models, which are responsible for receiving messages.

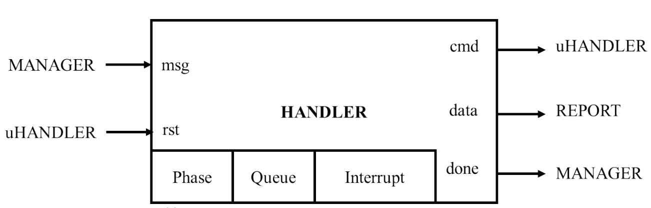


Figure 6 I/O ports and state variables in the HANDLER model

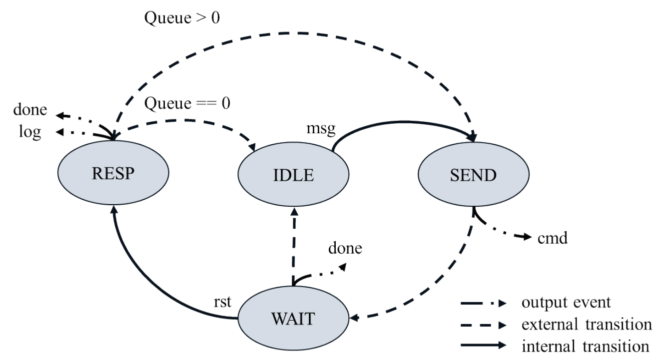


Figure 7 Phase transition diagram of the HANDLER model

Fig. 6 and Fig. 7 are simplified schematic representations of the HANDLER model. In Fig. 4, the HANDLER thread takes the input from ③ and ⑧ to provide the output to ⑤ and ⑨. As described earlier, the change of state due to input is defined as an external transition function, and the change of state over time and the output at this point as an internal transition function and output function. Initially, the HANDLER starts in the IDLE state. Then, when the message is received from the MANAGER through the "msg" port in ③, the external transition function of the HANDLER divides the message into detailed operation messages, stores them in the Queue, and changes the Phase to SEND. Then, after the time determined by the time advance function of the HANDLER in ⑤, the HANDLER's output function delivers one of the detailed operation messages stored in the queue, which has not yet been processed, to the uHANDLER through the "cmd" port, while the internal transition function of the HANDLER changes the Phase to WAIT. Subsequently, if the result is received from the uHANDLER through the "rst" port within the time specified in ⑧, the external transition function of the HANDLER stores the result in the Queue and changes the Phase to RESP. Then, in ⑨, the output function of the HANDLER transfers the result received in ⑧ to the REPORT via the "data" port, and the internal transition function of the HANDLER changes the Phase to SEND to repeat the process from ⑤ if there are still detailed operations in the queue that have not yet been processed. Otherwise, it changes the Phase to IDLE to complete the processing of the message received in ③. If the message is not received from the uHANDLER after the specified time after sending the message in ⑤, the output function of the HANDLER sends an error message to the MANAGER through the "done" port, and the internal transition function of the HANDLER changes the Phase to IDLE. In addition, if the smart meter generates an alarm message indicating an urgent exceptional situation, the message is forwarded to the HANDLER through the "rst" port, although this is not shown in Fig. 4. In this case, the external transition function of the HANDLER changes the Interrupt to ON so that the output function of the HANDLER later informs the MANAGER of the alarm via the "done" port.

Fig. 8 shows the atomic DEVS specification of what is described in Fig. 6 and Fig. 7. In the figure, X is the input event set, which represents the input of messages through the "msg" and "rst" ports, and Y is the output event set, which represents the output of messages through the "cmd", "data", and "done" ports. S is the state variable set, which contains state variables such as Phase, Queue, and Interrupt. In addition, δ_{ext} and δ_{int} represent the external transition function and internal transition function, λ is the output function, and ta is the time advance function.

HANDLER = $\{X, Y, S, \delta_{ext}, \delta_{int}, \lambda\ ta\}$

X = {msg, rst}

Y = {cmd, data, done}

S = {Phase, Queue, Interrupt}

δ_{ext} {(IDLE, 0, -), msg} = (SEND, n, -)

δ_{ext} {(WAIT, -, -), rst} = (RESP, -, -)

δ_{int} {(SEND, -, -)} = (RESP, n, -)

δ_{int} {(WAIT, n, -)} = (IDLE, 0, -)

δ_{int} {(RESP, n, -)} = if (Queue > 0) (SEND, n - 1, -)

δ_{int} {(RESP, n, -)} = if (Queue == 0) (IDLE, 0, -)

δ_{int} {(-, -, OFF)} = if (msg.type == Error) (IDLE, -, ON)

λ(SEND) = cmd

λ(RESP) = done, log

λ(WAIT) = done

ta(WAIT) = 1000

ta(SEND) = 0

ta(RESP) = 0

ta(SEND) = ∞

Figure 8 DEVS specification of the HANDLER model

5 SIMULATION

This chapter validates the proposed architecture from Chapter 4 through simulations. To this end, each of the models in Fog. 5, including the HANDLER model, is coded and simulated in DEVSim++ [19], a C++-based DEVS simulation environment. For more accurate verification, the behavioral procedures under normal circumstances, including those described in Fig. 4 and Fig. 5, as well as the operation procedures under various exceptional cases, are simulated. Below are the basic assumptions for the simulations in this paper:

(i) There are no problems with the communication between the DLMS server and the client. Therefore, the simulation of the process of initializing and setting up communication is excluded.

(ii) The number of service parameters contained in one DLMS message is set to 3, i.e., it is assumed that the DLMS server always communicates with the smart meter three times to process a single DLMS message.

(iii) Two smart meters are connected to the DLMS server.

Based on these assumptions, simulates 7 scenarios. Among them, 4 scenarios simulate operational procedures in normal conditions, while 3 scenarios simulate operational procedures in abnormal conditions. In each scenario, the client generates 20 - 100 request messages.

As mentioned earlier, debugging in a multi-threaded environment can be difficult as the behavior of the threads is not sequential. Therefore, to overcome this difficulty, the synchronization mechanism of the DEVS formalism is utilized to strictly control the behavior of the thread structure. Fig. 9 shows a screenshot that simulates the general operation procedure of the DLMS server using DEVSim++ and the log messages generated during execution. By analyzing the log messages, the status change of each model over time can be checked and it can be determined whether the message is transmitted correctly.

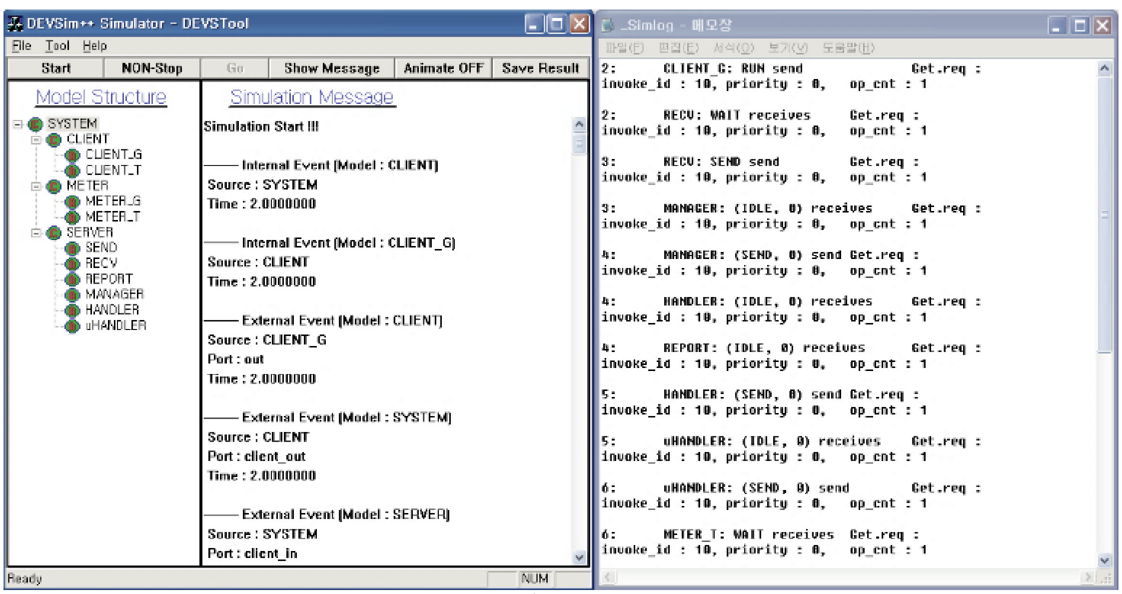


Figure 9 Simulation results

In this paper, simulations of three representative scenarios out of seven scenarios are described. The first scenario is part of the simulation results of operational procedures under normal conditions. Fig. 10 shows the simulation results for the case where two new request messages (Message 2 & Message 3) arrive sequentially (ⓐ and ⓑ) while the DLMS server is processing a request message (Message 1). It should be noted that different types of lines are used in the figure to separate the messages. In this simulation, Message 2 has a higher priority than Message 1, and Message 3 has a higher priority than Message 2. In other words, Fig. 10 shows a case where a high-priority message arrives while a low-priority message is being processed. This paper treats the processing of each message as a single, indivisible atomic transaction. Thus, by default, higher-priority messages are processed before lower-priority messages. However, once a message has started processing, it is not to be preempted until it has completed processing. As a result, the simulation results shown in Fig. 10, the HANDLER thread continues to process the existing Message 1, even though higher priority messages have arrived at ⓐ and ⓑ. It can also be seen ⓒ that after processing Message 1, it starts processing Message 3, which has the highest priority among the messages stored in the queue of the MANAGER thread. As a result, it has been confirmed that messages sent by the client are appropriately processed according to their priorities.

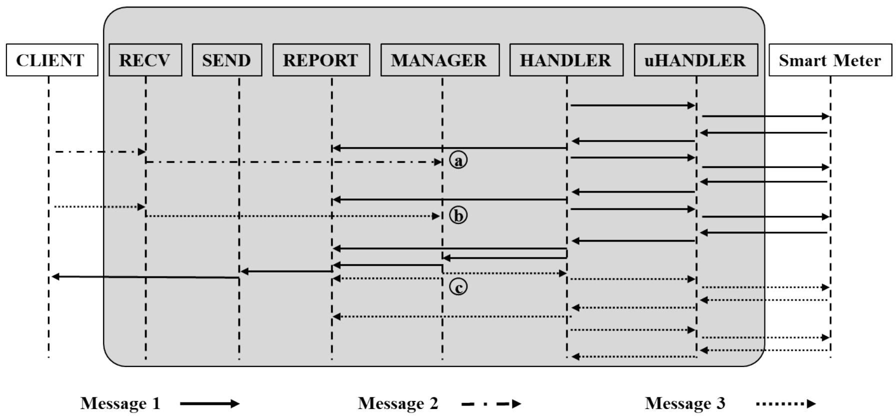


Figure 10 Processing of multiple requests

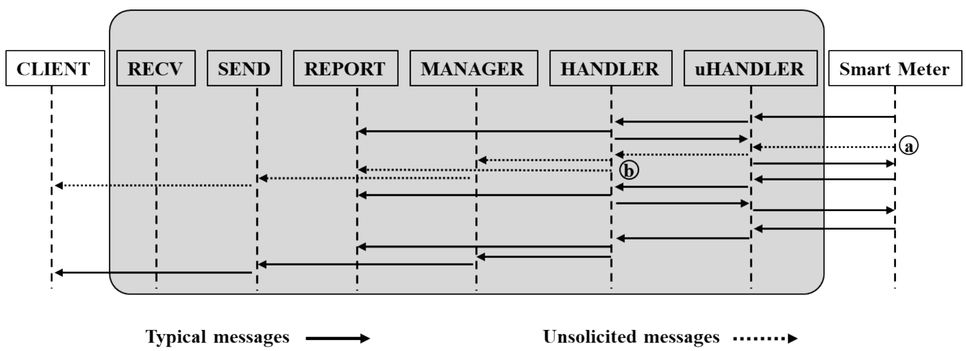


Figure 11 Processing of an unsolicited message generated by smart meters

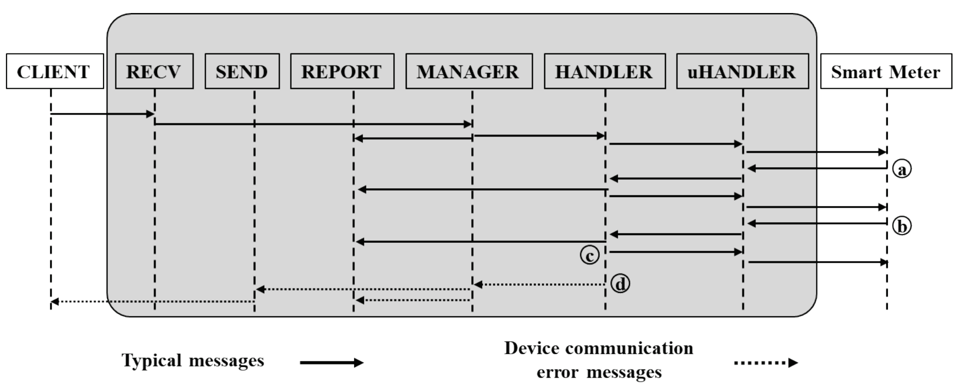


Figure 12 Processing of a communication error with smart meters

The second scenario is part of the simulation results of operational procedures under abnormal conditions. Fig. 11 shows the simulation results when the HANDLER thread receives an unsolicited message from the smart meter while processing the request. In the figure, the process of the HANDLER thread's request is represented by the solid line, while the processing of the unsolicited message generated by the smart meter is represented by the dotted line. There are two reasons why the smart meter may generate unsolicited messages. The first is when the smart meter encounters an abnormal situation, such as a device failure, and the second is when the smart meter needs to report status periodically. The DLMS protocol stipulates that when an unsolicited message occurs, the record is immediately stored on the DLMS server and reported to the DLMS client. According to the simulation result in Fig. 11, the smart meter has generated an unsolicited message in ⓐ. The message is then passed to the HANDLER thread via the uHANDLER thread in ⓑ. The

HANDLER thread then recognizes that it is an unsolicited message and immediately forwards it to the REPORT thread and the MANAGER thread. In addition, once the unsolicited message has been processed, the HANDLER thread continues to process the request that it was previously processing. As a result of the simulation, it is confirmed that the unsolicited message generated by the smart meter was delivered to the client prior to the processing of the existing message.

The last scenario also represents a partial simulation result of operational procedures under abnormal conditions. Fig. 12 is the simulation of an exceptional situation, in which the HANDLER thread has sent a request to the smart meter through the uHANDLER thread but has not received any response from the smart meter within the specified time. The DLMS protocol stipulates that in this case, the record is kept in the DLMS server and reported to the DLMS client, similar to the case of unsolicited messages. In figure 12, the smart meter sends a response to the request received in ⓐ and ⓑ, while in ⓒ, it does not send any response for some reason. Eventually, in ⓓ, the HANDLER thread realizes that an error has occurred during communication with the smart meter and sends a device communication error message to the MANAGER thread. The MANAGER thread then forwards the message to the REPORT thread and the SEND thread to log the occurrence of the device communication error and report it to the DLMS client to end exception handling. As a result, it is confirmed that when a communication error occurs from the smart meter, the DLMS server logs the error internally and promptly reports it to the client.

Based on diverse simulation results, including the aforementioned ones, it is confirmed that the proposed DLMS server effectively sends and receives messages asynchronously and processes them precisely as intended. It is also ensured that the DLMS server accurately accumulates state changes and results, while maintaining communication with the client for both general DLMS request messages and unexpected exceptional cases.

6 IMPLEMENTATION

This chapter implements a DLMS server with the proposed multi-threaded architecture. Although DEVSim++, which is based on the C++ language, is used to simulate the proposed architecture, the simulation code can only be executed in the environment of DEVSim++, and it is significantly different from the program code used in the real environment.

Most of the conventional programming languages are not developed on mathematical foundations. Therefore, many of the mathematical expressions used in modeling are often written in a slightly distorted form when translated into a programming language. The cumulative effect of these distortions is that the behavior of the implemented program differs significantly from what is formally modeled and validated by simulation. For this reason, it is usually considerably difficult to express the results modeled with a formal modeling method in a conventional programming language.

In a previous work [20], one solution was proposed to implement DEVS models for multi-threaded architectures written in DEVSim++ with real threads. In the proposed method, the hierarchical structure of the entire system model, which consists of multiple layers, is flattened into a two-layer structure in which all atomic models are subordinated to a single coupled model. Then, the single coupled model is implemented as a thread called SCHEDULER, and the DEVSim++ code of each atomic model is converted to an actual thread written in C++. This method reuses the expressions written in DEVSim++ with almost no modification while implementing threads, resolving the problem of expression distortion during the conversion process, which was a previous concern. The threads of the atomic models generated through this process have a characteristic that, unlike typical threads, they only execute actions corresponding to input or trigger messages received from external sources. Therefore, if no messages are received, they remain idle without performing any actions. Hence, even if the operating system schedules threads in a random order, the threads execute actual operations in the order of trigger messages generated by the SCHEDULER thread.

This paper implemented the proposed multi-threaded architecture DLMS server in the Linux operating system environment using the C++ language, following the method proposed in [20]. The implemented DLMS server consists of seven types of threads corresponding to six thread models, except for the CLIENT and the SMART METER models among the simulated models, and the SCHEDULER thread. As mentioned earlier, only one thread is created for each of the seven thread types except for the HANDLER thread and uHANDLER thread. However, several threads are created for the HANDLER thread and the uHANDLER thread to match the number of smart meters connected to the DLMS server.

Finally, the implemented DLMS server program is validated. For proper validation, it is required to connect the implemented DLMS server with real smart meters and examine its behavior while the server communicates with a real DLMS client. However, due to budget limitations, it is difficult to procure real smart meters and DLMS clients. Therefore, experiments are conducted by implementing independent programs for each device separately. For validation, two virtual smart meters are employed, identical to the simulation environment in Chapter 5. Therefore, the implemented DLMS server program communicates with one DLMS client program and two smart meter programs to process requests. Additionally, within the DLMS server, there are a total of 9 threads consisting of 2 HANDLER threads and 2 uHANDLER threads each.

Verification proceeds in two stages. In the first stage, to validate general operation procedure, DLMS clients generate requests identical to those simulated in Chapter 5, ensuring smart meters exhibit consistent responses. The operation of the DLMS server is logged and compared with simulation results. The comparison confirms identical operation between the observed simulation and the actual operation of the server. In the second stage, DLMS clients and smart meters are allowed to freely generate and respond to requests. Analysis of the generated logs confirms that the DLMS server operates in accordance with the requirements specified in the protocol and design.

7 CONCLUSION

This paper proposes a DLMS server with multi-threaded architecture for AMI systems. The proposed DLMS server must handle communication and message processing with clients, manage various events, and control multiple smart meter devices with different communication protocols. First, specific requirements are derived from the above needs. To meet these requirements, the DLMS server is designed with 6 threads. The proposed architecture is modeled using the DEVS formalism and simulated in the DEVSim++ environment for verification. The simulation results confirm that the proposed multi-threaded DLMS server operates according to the requirements identified in this paper. Ultimately, the DLMS server is implemented based on the simulation code.

As anticipated, there may be debates regarding performance issues with the DLMS server implemented in this paper. Specifically, the concurrency of thread operations is strictly limited by the scheduler thread, which may degrade the server's performance. Nonetheless, it is important to note that the concurrency of servicing DLMS requests is not restricted. The DLMS server implemented in this paper may experience slight performance degradation due to the strict limitations on concurrency of thread execution. However, this approach prevents race conditions caused by the disorderly execution of threads, significantly reducing the cost of implementing and testing the DLMS server. The method presented in this paper can be used in various fields where the verification of operations is more critical than system performance.

8 REFERENCES

[1] Jung, S. M., Kim, T.-K., Seo, H.-S., Lee, S.-J., & Kwak, J. (2013). The prediction of network efficiency in the smart grid. *Electronic Commerce Research, 13*, 347–356. https://doi.org/10.1007/s10660-013-9124-1

[2] Feuerhahn, S., Zillgith, M., Wittwer, C., & Wietfeld, C. (2011). Comparison of the communication protocols DLMS/COSEM, SML and IEC 61850 for smart metering applications. *The IEEE International Conference on Smart Grid Communications (SmartGridComm2011)*, 410-415. https://doi.org/10.1109/SmartGridComm.2011.6102357

[3] Prasanth, G. (2009). Implementing DLMS Client and Server Protocols in Meters, IED's, MRI's and Meter Reading Applications – An Overview. Vinoo S Warrier, Vice President, Kalkitech (Internet). Version 1. *gopalakrishnanprasanth*. https://gopalakrishnanprasanth.wordpress.com/article/implementing-dlms-client-and-server-3bk0x32fl4sfh-15/

[4] DLMS User Association. (2019). DLMS/COSEM Architecture and Protocols. *Green Book Edition 8.1*.

[5] DLMS User Association. (2019). COSEM Interface Classes and OBIS Object Identification System. *Blue Book Edition 13*.

[6] Campbell, J. M., Ellis, R. K., & Giele, W. T. (2015). A multi-threaded version of MCFM. *The European Physical Journal*, C 75, 1-7. https://doi.org/10.1140/epjc/s10052-015-3461-2

[7] Wickramasinghe, M., & Guo, H. (2014). Energy-Aware Thread Scheduling for Embedded Multi-threaded Processors: Architectural Level Design and Implementatio. *The IEEE Computer Society Annual Symposium on VLSI*, Tampa, FL, USA, 178-183. https://doi.org/10.1109/ISVLSI.2014.55

[8] Zeigler, B. P. (1990). *Object-Oriented Simulation with Hierarchical, Modular Models*. Academic Press.

[9] Zeigler, B. P. (1984). *Multifaceted Modeling and Discrete Event Simulation*. Academic Press.

[10] Zeigler, B. P., Prähofer, H., & Kim, T. G. (2000). *Theory of Modeling and Simulation: Integrating Discrete Event and Continuous Complex Dynamic Systems*. Academic Press.

[11] Štruklec, G., & Maršić, J. (2011). Implementing DLMS/COSEM in smart meters. *The 8th International Conference on the European Energy Market (EEM2011)*, Zagreb, Croatia, 747-752. https://doi.org/10.1109/EEM.2011.5953109

[12] Kheaksong, A., & Lee, W. (2014). Packet transfer of DLMS/COSEM standards for smart grid. The 20th Asia-Pacific Conference on Communication, 391-396. https://doi.org/10.1109/APCC.2014.7092843

[13] Ju, S. H., & Seo, H. S. (2018). Design key management system for DLMS/COSEM standard-based smart metering. *International Journal of Engineering and Technology (UAE), 7*, 554-557. https://doi.org/10.14419/ijet.v7i3.34.19380

[14] Burunkaya, M., & Pars, T. (2017). A smart meter design and implementation using ZigBee based Wireless Sensor Network in Smart Grid. *The 4th International Conference on Electrical and Electronic Engineering (ICEEE2017)*, 158-162. https://doi.org/10.1109/ICEEE2.2017.7935812

[15] Park, S. B., Ahn, Y. J., & Kim, J. H. (2010). Development of OPC Server Unifying Communication Profiles of DLMS/COSEM in Smart Grid. *The Journal of Korean Institute of Information Technology, 8*(9), 1-11. https://doi.org/10.14801/jkiit.2020.18.9.1

[16] Im, C. J., Jang, S. J., Hahn, K. S., Kim, B. S., & Jung, N. J. (2007). *The Development of IEC62056 based Energy Information Concentrator for DLMS Meters*. The Korean Institute of Electrical Engineers, 54-56.

[17] Biswas, S., Himanshu, Ghosh, S., Das, P., Saha, K., & De, S. (2023). Efficient Data Transfer Mechanism for DLMS/COSEM Enabled Smart Energy Metering Platform. SIGMETRICS Perform. *ACM SIGMETRICS Performance Evaluation Review, 50*(4), 14-16. https://doi.org/10.1145/3595244.3595250

[18] Tarvo, A., & Reiss, S. P. (2018). Automatic performance prediction of multithreaded programs: a simulation approach. *Automated Software Engineering, 25*(1), 101-155. https://doi.org/10.1007/s10515-017-0214-5

[19] Kim, T. G. (1994). DEVSim++ User's Manual: C++ Based Simulation with Hierarchical Modular DEVS Models.

[20] Kim, Y. H., Seong, Y. R., & Oh, H. R. (2014). Software Development Method Using the Concurrency Control Approach Based on DEVS Simulation. *Proceeding of the 2014 FTRA International Conference on ACS, 11*(7), 553-558.

Authors' contacts:

Gwang Hyeon Kim, Student
Department of Electrical Engineering, Kookmin University,
77 Jeongneung-ro Seongbuk-gu, Seoul, 02707, Korea
rhkdgus306@kookmin.ac.kr

Yeong Rak Seong, Professor
(Corresponding author)
Department of Electrical Engineering, Kookmin University,
77 Jeongneung-ro Seongbuk-gu, Seoul, 02707, Korea
yeong@kookmin.ac.kr

Ha Ryoung Oh, Professor
Department of Electrical Engineering, Kookmin University,
77 Jeongneung-ro Seongbuk-gu, Seoul, 02707, Korea
hroh@kookmin.ac.kr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240514083303

Original scientific paper
Received: 2024-05-14, Accepted: 2024-07-18

Optimizing Perforated Bellows Formation: Achieving Uniformity and Precision through Proportional Roll Mold Transfer and Length Reduction

Nam Do Baek, Keun-Wook Baek, Young-Min Ji, Hak-Geun Choi, Dong-Hyun Cho*

Abstract: This study presents an experimental device designed for single-process formation of perforated bellows through proportional control transfer of the roll mold, ensuring uniformity in thickness and height. A simulation assessed these parameters to establish allowable values, which were validated against experimental data. By implementing a linear acceleration proportional control transfer system for the upper and lower die roll molds, uniform thickness within an error range of 0.02-0.03 mm was achieved. An upper die and lower die roll mold linear acceleration proportional control transfer control system was configured so that ten upper die roll molds and 11 lower die roll molds that form the crests and roots of bellows can be transferred at the same linear acceleration as that of the reduction of the length of the cylinder in the process of manufacturing the crest and root to conduct an experimental study. Experimental findings revealed a minimal thickness error range of 0.02-0.03 mm, demonstrating high uniformity and reliability in the manufactured perforated bellows. Moreover, the study verified the durability and strength of the bellows through thorough assessments of thickness reduction values, which ranged from 0.02 to 0.1 mm. Both experimental and simulated strain rate analyses showed a strong correlation, validating the effectiveness of the proposed manufacturing approach. Therefore, the reliability of the allowable strength and durability of the perforated bellows was secured based on the results of the study on the uniformity of the thickness of the perforated bellows about changes in the thickness of the perforated bellows. Overall, this research establishes a robust methodology for achieving uniformity and reliability in perforated bellows production, offering significant advancements in industrial applications where precise dimensional control and durability are paramount.

Keywords: one-process forming; perforated bellows; proportional control transfer; silencer; uniformity

1 INTRODUCTION

Today, noise is recognized as a significant issue in daily life, and its importance is growing in all industrial fields, particularly in machinery, construction, environment, and electricity [1, 2]. Recently, environmental damage caused by noise has been increasing. As a result, systems are being developed to create a quieter environment by mandating noise-level markings for machinery. Meanwhile, in line with the trend of upgrading machines, not only performance but also additional quality, especially when noises are reduced, greatly affects the competitiveness of machines and is becoming a major criterion for product selection depending on buyers [3]. Thus, noise reduction devices, particularly silencers installed at air intake/exhaust ports close to the user, are urgently needed. Recent advancements in design technology have led to higher output and capacity in fluid machines (e.g., compressors, blowers) and engines. However, this has also resulted in increased noise levels [4, 5]. Consequently, there is a strengthened demand for noise reduction technologies. Research on low-weight, high-performance silencers is urgently needed, but current reports are insufficient [6]. In particular, in the past, manufacturers of finished fluid machinery introduced key technologies for silencer design by paying royalties or making some design changes and then relied on subcontractors for simple manufacturing. However, silencer manufacturers have recently been required to carry out silencer design and evaluation to satisfy necessary performance by themselves. In addition, in the case of high-flow machines, the durability of the silencer must be guaranteed because enormous pressures are generated. Still, since this conflicts with noise reduction performance, appropriate design techniques must be prepared [7, 8]. In particular, in the engine's exhaust system, high-temperature gas of 300 to 400 °C passes through the pipe at a high speed of 30 to 40 m/s. Noises from the exhaust port can be divided into pulsating noises, which occur when the exhaust valve of a fluid machine opens, and low-pressure outside air is pushed. Flow noises, which occur when the gas flow of the exhaust system is discharged, include a wide range of frequency components ranging from several tens of Hz to several kHz and are discharged with high temperature/high-speed flows so that it is not easy to control exhaust noises [9]. In particular, silencers used for noise reduction must simultaneously satisfy various functions such as minimum back pressure, minimum size and weight, sturdiness and durability, and ease of maintenance, so some studies on systematic silencer design and performance evaluation methods considering temperature, flow rate, etc. have been reported [10]. In particular, recently, as high-output fluid machines are operated in various industrial sites, the development of low-weight/high-performance silencer technology is urgently required [11]. The purpose of a silencer is to attenuate the noises propagated through the interior without interfering with the free flow of the fluid. The current level of silencer technology at home and abroad is largely divided into reactive, absorptive, and reactive-absorptive types according to the operating principles used in research and commercialization [12]. Reactive silencers can generally reduce noise in low and mid-frequency regions, and the impedance mismatch type determines their operating frequency. An absorptive silencer dissipates acoustic energy into heat energy by using a sound-absorbing material on the walls of the flow path [13]. The amount of noise attenuation of this structure is generally proportional to the thickness and length of the sound-absorbing material. It thus is determined by the frequency, the thickness of the sound-absorbing material, and the background material. Due to its operating principle, an effective noise reduction effect can be obtained in the mid-and high-frequency areas. Reactive-absorptive silencers use the combination of the two types to satisfy the given space and other limiting conditions as much as possible and are the most widely used. In general, in the case of a silencer through which high-speed fluid passes, it is desirable to minimize changes in the flow path and use sound-absorbing materials to reduce noises to minimize back pressure and reduce the occurrence of aerodynamic noises. However, in this case, controlling pulsation noises in the low-frequency region is not easy. Therefore, studies that reflect the optimal area change and acoustic resonator design for reducing low frequencies are required [14]. Therefore, in this study, perforated bellows forming experimental device capable of forming perforated bellows in one process by proportional control transfer of the roll mold according to the length reduction occurring when the crest and root of the bellows are formed was configured. A study on the uniformity and precision of the thickness and height of the perforated bellows was conducted. A transfer control system that proportionally controls the linear acceleration of the upper die and low die roll molds so that the ten upper die roll molds and 11 low die roll molds that form the crests and roots of the bellows with a linear acceleration equal to that of the reduction in the length of the cylinder in the process of manufacturing the crests and roots of perforated bellows was configured to conduct an experimental study. In addition, a silencer that fused perforated bellows technology and micropore stainless chip sound absorber technology was developed to study a silencer with excellent noise attenuation performance even when the fluid pressure is over 30 bar for the first time at home and abroad.

2 EXPERIMENTAL DEVICE AND METHODS

Fig. 1 shows the experimental device capable of forming perforated bellows in one process by proportionally controlled transfer of the roll moulds according to the length reduction occurring when the crests and roots of the perforated bellows occur. As shown in Fig. 1, the perforated bellows forming experimental device was configured as a large experimental device with a length of 3.7 m, a width of 2.5 m, and a height of 4.5 m. The perforated bellows forming experimental device consists of a main body frame, a hydraulic part that generates operating force, a hydraulic cylinder and slide part that controls the ascending and descending of the upper shaft, a hydraulic motor that drives the lower shaft, and an upper and lower molds mounted on the upper and lower shafts, respectively, a servomotor and drive shaft to drive the molds on the upper shaft in a proportional control manner as molding progresses, and a fork that connects the drive shaft and the upper molds. As shown in Fig. 2, in the bellows for mining equipment, a drive motor was installed on the side of the box-shaped frame to rotate the rotating shaft installed in the horizontal direction, and roll molds were constructed so that multiple lower dies and upper dies were installed on the horizontal movement device. The equipment was so built roll molds be produced and installed even if the numbers of crests and roots of the bellows increased or decreased. In addition, the equipment is designed to facilitate the formation of bellows through a single molding process, where 11 upper die roll molds and 10 lower die roll molds sequentially position themselves at intervals along the idling and rotating shafts, sliding towards the center from both sides to plastically deform the material pipe along its longitudinal direction and match the pitch (the gap between crests) of the bellows being manufactured on the upper and lower sides.

Figure 1 Forming experiment equipment for perforated bellows.

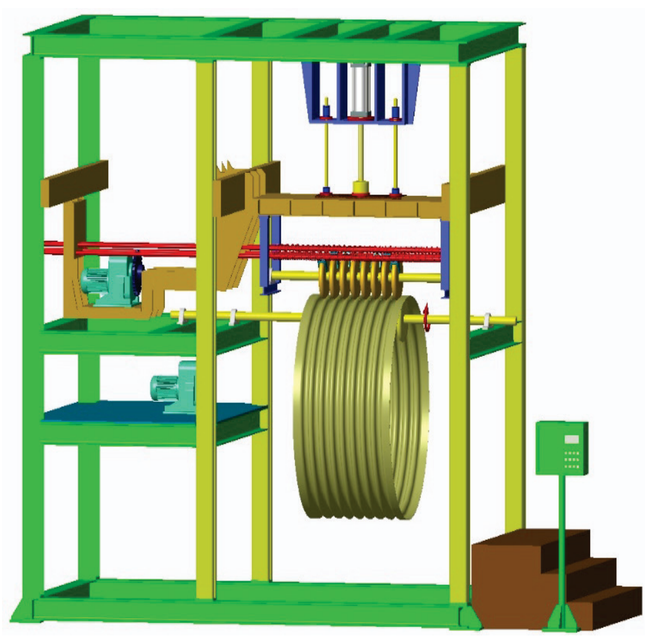


Figure 2 3D schematic diagram of multiple low-die and upper-die roll molds on a rotating shaft

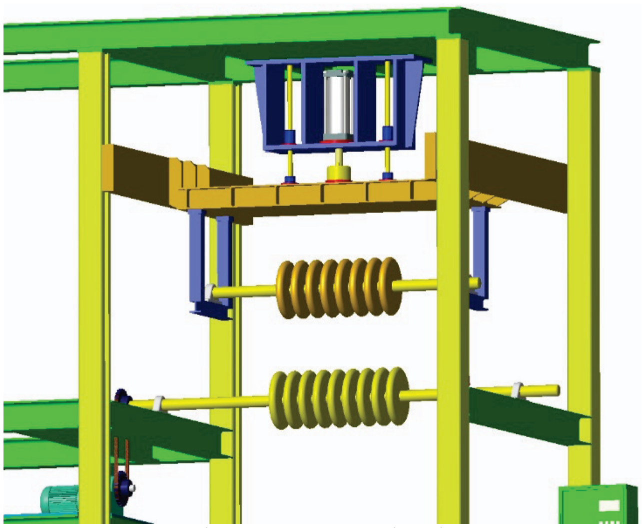


Figure 3 3D schematic diagram of the lifting block

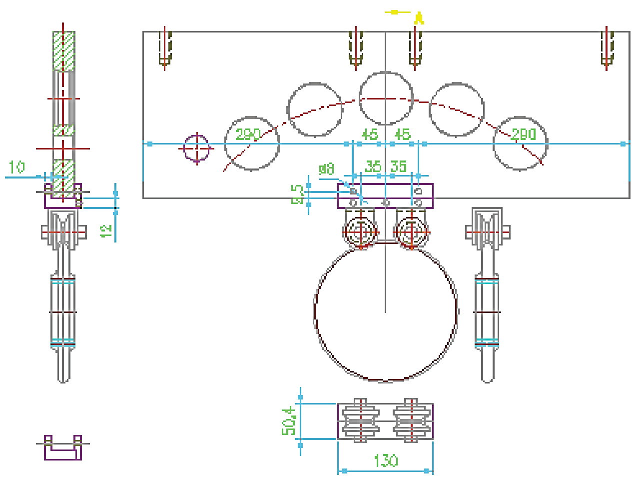


Figure 4 Movement path of rollers when forming perforated bellows

Fig. 3 shows a 3D schematic diagram of the lifting block of the perforated bellows forming equipment. As shown in Fig. 3, a lift that can be moved up and down was constructed above the rotating shaft. In addition, guide rails were installed on the left and right sides of the bellows manufacturing equipment frame, and the equipment frame was constructed so that a lifting block that could be raised and lowered along the guide rails could be installed. The lifting block was connected to the piston rod of the cylinder, which is expanded and contracted with the driving power of a hydraulic power source fixed to the upper part of the bellow manufacturing equipment frame so that it could move up and down according to the expansion and contraction of the cylinder. In addition, in the lower part of the lifting block, the idling shaft was installed horizontally to rotate freely, as shown in Fig. 4, and the idling shaft was configured to have ten upper dies that slide left and right in the longitudinal direction. In contrast, the cylinder is operated to descend, leading to the descent of the upper dies, which presses the material pipe from the top. In contrast, the lower dies on the rotating shaft press the material pipe from the bottom to cause plastic deformation in the material pipe; the upper dies gradually descend so that the outer part of the material pipe is deformed toward the center, leading to the formation of ring-shaped roots. The lower dies form the crests of the bellows, consisting of ring-shaped projections around the rotating shaft.

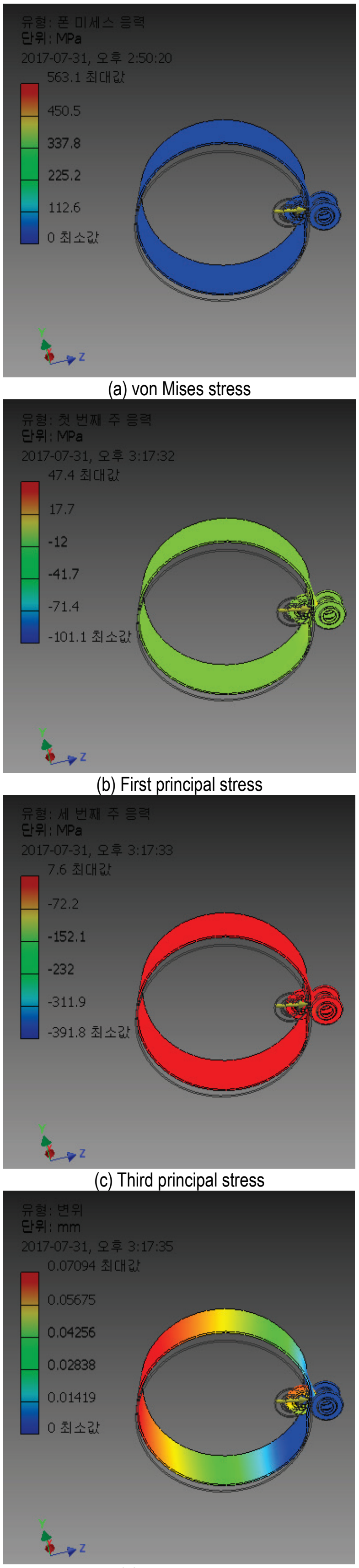
3 ANALYSIS AND SIMULATION OF PERFORATED BELLOWS FORMED IN ONE PROCESS

3.1 Forming Force Analysis and Simulation

Fig. 5 illustrates the stress and strain rate of the perforated bellows, which are formed in a single process through the proportional control transfer of the roll molds. This process is based on the length reduction during the formation of the crests and roots of the perforated bellows. As depicted in Fig. 5, the von Mises stress indicates that deforming 1 t of STS 304 material requires a force of 500 MPa when approximately 2 tons are applied to one crest of the bellows. Furthermore, the simulation predicts that a force of 20 tons is necessary to plastically deform ten crests simultaneously. Based on these simulation results, we selected an inner diameter of 200 mm for the cylinder's perforated bellows. The flow rate of the hydraulic motor was set at 90 liters/min, ensuring optimal performance. The hydraulic pump, which operates both the hydraulic cylinder and the hydraulic motor, was determined to be capable of pressurizing to more than 100 bar. To validate these findings, we compared the simulation results with experimental data to ensure the uniformity of the thickness and height of the perforated bellows. The allowable values for these parameters were confirmed to meet the required standards, as indicated by the consistency between the simulation predictions and the empirical measurements. This validation process confirms that the designed system can reliably produce perforated bellows with the desired specifications.

3.2 Bending Moment of the Upper and Lower Shafts when Forming the Perforated Bellows Formed in One Process

Fig. 6 shows the simulation results of the bending moment of the rotating shaft of the perforated bellows forming the experimental device. The perforated bellows forming experiment device is large, and the 3 m long axis is bent under force (20 tons) during forming. If the center of the upper and lower shafts of the forming machine is bent (about 10 mm), the gap between the precision molds cannot be properly maintained. Due to the result as such, the uniformity of the thickness and height of the formed products of perforated bellows is greatly reduced. Therefore, a simulation study was conducted to determine the allowable values for uniformity of the thickness and height of the perforated bellows. Based on the simulation results, as shown in Fig. 6, the amount of axis deflection decreased as the diameter axis increased. Also, due to the bending of the shaft due to the high forming force, the deformation was so severe in the 80 mm diameter shaft that it lost its function as a shaft, and it was required to be replaced with a shaft with a diameter of 120 mm. Accordingly, 11 upper and 13 lower molds also needed to be changed.



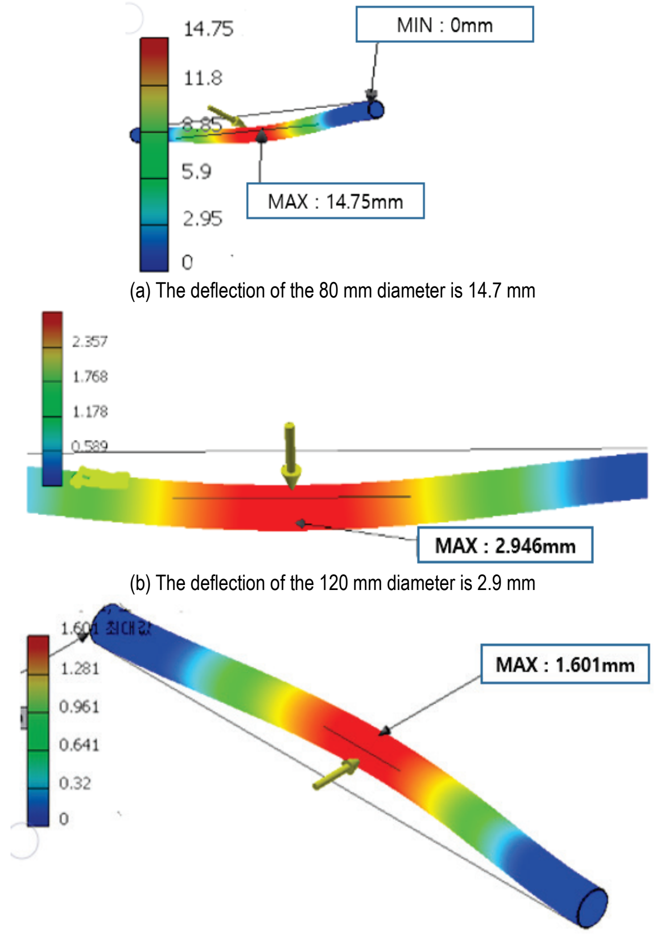
(a) von Mises stress

(b) First principal stress

(c) Third principal stress

(d) Displacement

Figure 5 Stress and displacement of one-process formed bellows



(a) The deflection of the 80 mm diameter is 14.7 mm

(b) The deflection of the 120 mm diameter is 2.9 mm

(c) The deflection of the 140 mm diameter is 1.6 mm

Figure 6 Bending moment simulation for perforated bellows forming experimental device

Fig. 7 shows the shaft diameter calculated from the simulation results for the bending moment of the rotating shaft of the perforated bellows forming the experimental device. As shown in Fig. 7, the shaft deflection was found to be 0.9 mm. Therefore, it is believed that since the resultant deflection value of the shaft is small, the allowable values for the uniformity of the thickness and height of the perforated bellows are secured.

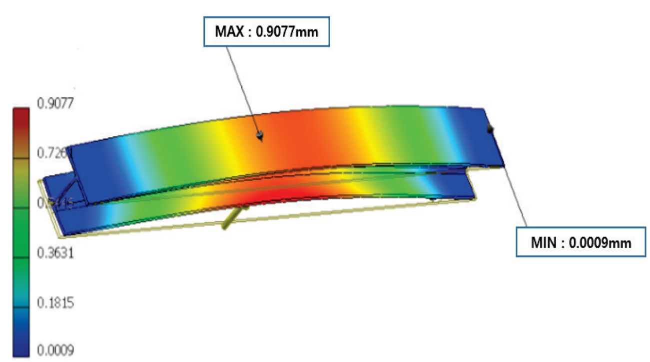


Figure 7 Calculate the allowable diameter of the shaft from simulation results for the bending moment

4 RESULTS AND DISCUSSION

Fig. 8 shows the shape of the perforated bellows. As shown in Fig. 8, the perforated bellows forming experimental device was configured so that ten upper die roll molds and 11 lower die rolls that form the crests and roots of the bellows are transferred at a linear acceleration equal to the linear acceleration of the reduction of the length of the cylinder in the simple process of manufacturing the crests and roots of simple perforated bellows to implement upper die and lower die roll mold linear acceleration proportional control transfer control system technology thereby forming perforated bellows with one manufacturing process. An experimental study was conducted on the uniformity of the thickness of the perforated bellows by measuring the thickness at five points on the cross-section of the perforated bellows formed as such. In addition, an experimental study was conducted on the height uniformity concerning changes in the pitch of the perforated bellows. Since the thickness and height of the perforated bellows significantly affect their elasticity and durability, the uniformity of the thickness and height of the perforated bellows must be maintained within the allowable values.

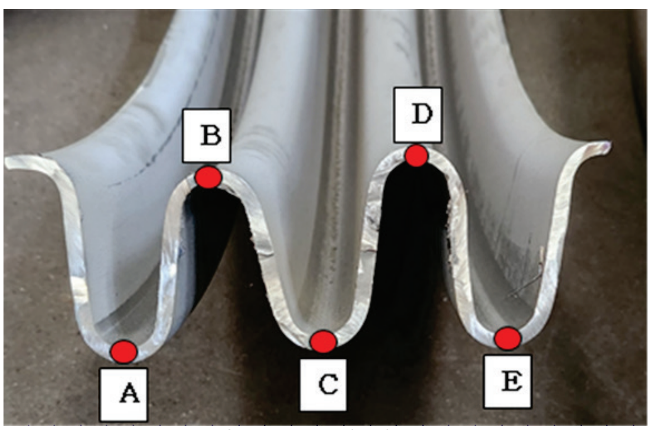


Figure 8 Crest and root of perforated bellows

Fig. 9 shows the uniformity of the thickness of the perforated bellows concerning the changes in the pitch of the perforated bellows. As shown in Fig. 9, an experimental study was conducted under three conditions of the thickness of the perforated bellows: 3 mm, 4 mm, and 5 mm, and as shown in Fig. 9, an experimental study was conducted to verify the uniformity and reliability of the thickness of the perforated bellows by measuring the thickness of the bellows at five cross sections from points A to E. The thickness of the perforated bellows was uniform within an error range of 0.02 to 0.03. Therefore, based on the results of this study, it is believed that the uniformity and reliability of the perforated bellows were secured.

Fig. 10 shows the uniformity of the thickness of the perforated bellows concerning changes in the thickness of the perforated bellows. As shown in Fig. 10, an experimental study was conducted on the uniformity of the thickness of the perforated bellows under five conditions ranging from 1 to 5 mm. The thickness reduction value of the perforated bellows increased in proportion to changes in the thickness of the perforated bellows. The thickness reduction value of the perforated bellows was in the range of 0.02 to 0.1 mm. Therefore, based on the results of the study on the uniformity of the thickness of the perforated bellows about changes in the thickness of the perforated bellows, the reliability of the allowable strength and durability of the perforated bellows is considered to have been secured. In addition, the uniformity of the thickness of the perforated bellows increased linearly concerning changes in the thickness of the perforated bellows, so the reliability of the experimental results is considered to have been verified.

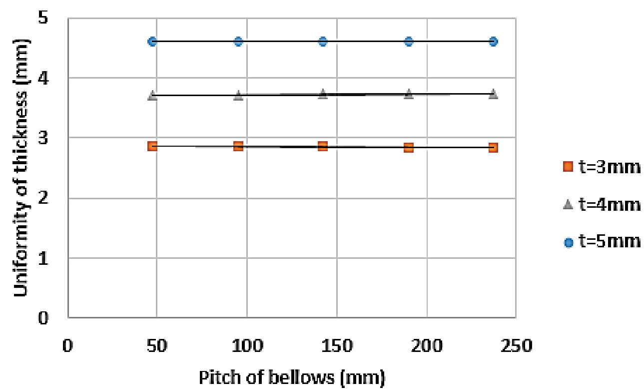


Figure 9 Thickness uniformity for pitch change of perforated bellows

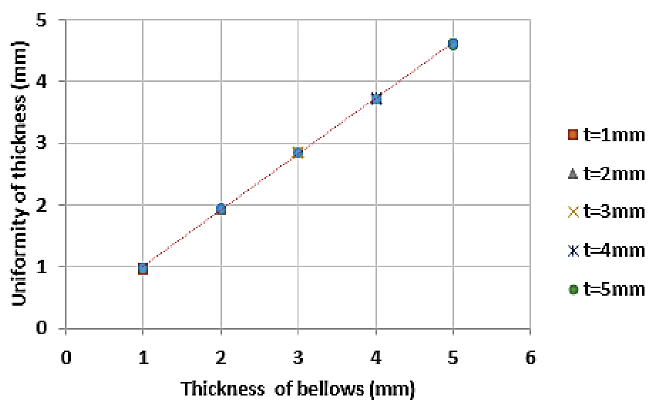


Figure 10 Thickness uniformity for thickness change of perforated bellows

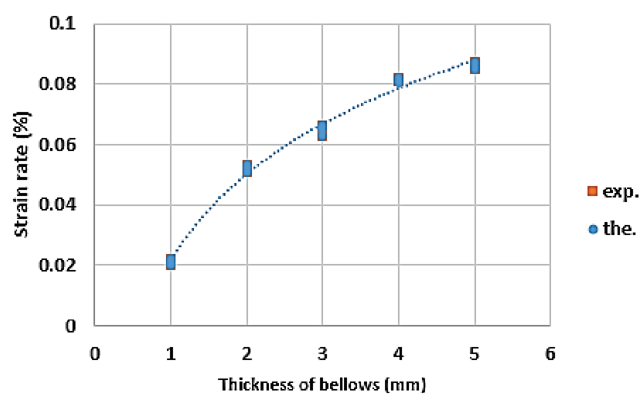


Figure 11 Strain rate for thickness change of perforated bellows

Fig. 11 shows the strain rate of the perforated bellows concerning changes in the thickness of the perforated bellows. As shown in Fig. 11, a study was conducted to compare experimental values and simulation results for the strain rate of the perforated bellows under five conditions with a thickness of 1 to 5 mm. Based on the results of the study, the experimental and simulation results for the strain rate of the perforated bellows concerning changes in the thickness of the perforated bellows were in relatively good agreement, and the strain rate of the perforated bellows was in the range of 0.021 to 0.086%. The strain rate of the perforated bellows increased in proportion to the thickness of the perforated bellows. Based on the results of the study, the strain rates indicate that the allowable strength and durability of the perforated bellows were secured.

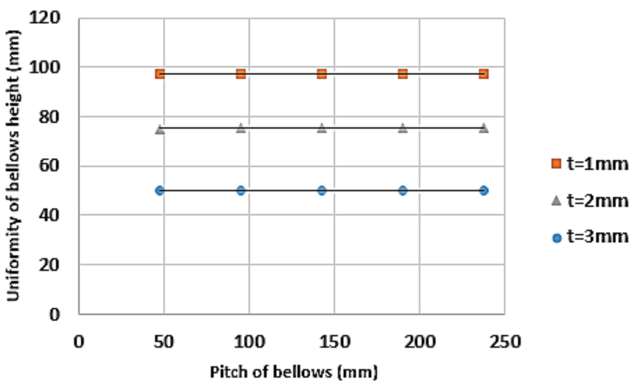


Figure 12 Uniformity of bellows height for the pitch change of perforated bellows

Fig. 12 shows the uniformity of the height of the perforated bellows concerning changes in the pitch of the perforated bellows. As shown in Fig. 12, an experimental study was conducted under three conditions of the thickness of the perforated bellows: 3 mm, 4 mm, and 5 mm. In addition, an experimental study was conducted under three conditions of the height of the perforated bellows: 50.1 mm, 75.2 mm, and 97.2 mm. As shown in Fig. 12, the height of the perforated bellows increased significantly, and the uniformity of the height of the perforated bellows was maintained in the range of 0.1 to 0.3 mm. Therefore, based on the results of the study on the uniformity of the height of the perforated bellows concerning changes in the pitch of the perforated bellows, the uniformity of the height of the perforated bellows was found to be high, so it is believed that the reliability of the experimental results for changes in the height of the perforated bellows was secured.

5 CONCLUSION

In this study, we investigated the feasibility of forming perforated bellows in a single step through proportional control transfer of roll molds, focusing on achieving uniformity and precision in bellows dimensions. Our research successfully implemented a molding technology that achieves this objective, ensuring the thickness of perforated bellows remains consistently uniform within an error margin of 0.02 to 0.03 mm. This uniformity not only enhances the reliability of perforated bellows but also secures their durability and strength under varying conditions.

Furthermore, our experimental and simulation findings concerning the strain rate of perforated bellows align closely with the changes observed in their thickness, ranging from 0.021% to 0.086%. This correlation underscores the effectiveness of our approach in maintaining structural integrity across different thicknesses of perforated bellows. Moreover, as the height of perforated bellows increased, our study revealed a significant enhancement in uniformity, with height deviations consistently maintained within a narrow range of 0.1 to 0.3 mm. This consistency further reinforces the reliability of our manufacturing process in producing perforated bellows with precise dimensional control.

In conclusion, our research demonstrates the successful implementation of a one-process forming technology for perforated bellows, emphasizing uniformity in thickness and height as essential factors contributing to their overall reliability and structural integrity. These findings not only advance current manufacturing capabilities but also pave the way for more efficient and consistent production of perforated bellows in industrial applications.

Acknowledgments

This paper was conducted as a research project for the 2022 Regional Specialized Industry Promotion Project (S3272965) of the Ministry of SMEs and Startups.

6 REFERENCES

[1] Jeong, D. H., Chin, D. H., Kim, J. D., Yoon, M. C., & Kim, B. T. (2019). A Study on the Behavior of 3-Ply Bellows. *Proceedings of the KSMPE Conference*, 84-84.

[2] Expansion Joint Manufacturers Association, Inc. (2008). *Standards of the Expansion Joint Manufacturers Association*, Ninth Edition.

[3] Lee, W. S. (2018). A Study on the Durability Evaluation Method of Bellows type expansion pipe joints. *Master Thesis*, University of Incheon, Republic of Korea.

[4] Han, M. S., Ahn, J. H., & Yang, C. H. (2013). Study on Optimum Shape of Expansion Joint. *Transaction of the Korean Society of Automotive Engineers, 21*(2), 154-158. https://doi.org/10.7467/KSAE.2013.21.2.154

[5] Jeong, D. H, Chin, D. H., & Kim, B. T. (2020). A Study on the Behavioral Characteristics of Bellows for Expansion Joints. *Journal of the Korean Society of Manufacturing Process Engineers, 19*(10), 52-58. https://doi.org/10.14775/ksmpe.2020.19.10.052

[6] Kim, H. J., Kim, H. S., Han, K. T., & Jeon, W. Y. (2008). Optimum Design of Ships Bellows Using R.S.M. *Proceedings of the KSMPE Conference*, 349-352. https://doi.org/10.1097/ICO.0b013e31815cf67d

[7] Kim, H. J., Park, J. H., Kim, H. S., Sung J. H., Kim H. G., & Lee, J. S. (2006). A Study on Design of U-type Bellows. *Proceedings of the KSMTE Conference*, 319-324.

[8] Ko, B. G., Suh, Y. J., & Park, K. J. (1995). A Finite Element Analysis and Shape Optimal Design with Specified Stiffness for U-typed Bellows. *Transaction of the Korean Society of Automotive Engineers, 3*(6), 96-111.

[9] Koh, B. K. (2008). Development of a Bellows Design Software Based on EJMA. *Journal of the Korean Society of Manufacturing Technology Engineers, 17*(1), 150-157.

[10] Kim, H. J., Kim, H. S., Kim, J. P., Park, J. H., & Yun, M. J. (2006). Shape Optimization for Performance Improvement of Ship's U-type Bellows. *Journal of Ocean Engineering and Technology, 20*(6), 123-129.
[11] Budynas, R. G. (1999). Advanced Strength and Applied Stress Analysis (2nd ed.), McGraw-Hill, 302-309.
[12] Jeong D. H. et al. (2019). Stress Analysis of 2-Ply Bellows. *Proceedings of the KSMPE Spring Conference 2019*, p. 129. https://doi.org/10.18653/v1/N19-1028
[13] Konak, A., Coit, D. W., & Smith, A. E. (2006). Multi-Objective Optimization Using Genetic Algorithms A Tutorial. *Reliability Engineering & System Safety, 91*(9), 992-1007. https://doi.org/10.1016/j.ress.2005.11.018
[14] Yang, L., Yang, M., Liu, F., & Yang, G. (2012). Optimization Design of the U-shaped Metal Bellows. *Proceedings of the International Conference on Modeling, Simulation and Visualization Methods (MSV)*. https://api.semanticscholar.org/CorpusID:133594491

Authors' contacts:

Nam Do Baek, CEO
Jeil Industry,
No. 30, Yeongdong 1-Gil, 50567 Yangsan-Si, South Korea
Namdo2001@naver.com

Keun-Wook Baek, Head of a Department
Jeil Industry,
No. 30, Yeongdong 1-Gil, 50567 Yangsan-Si, South Korea
Namdo2001@naver.com

Young-Min Ji, Manager
Jeil Industry,
No. 30, Yeongdong 1-Gil, 50567 Yangsan-Si, South Korea
Namdo2001@naver.com

Hak-Geun Choi, Manager
Jeil Industry,
No. 30, Yeongdong 1-Gil, 50567 Yangsan-Si, South Korea
Namdo2001@naver.com

Dong-Hyun Cho, Professor
(Corresponding author)
Department of Mechanical Engineering, Daejin University,
Gyeonggi-do, Pocheon-si, 1007 Hoguk-ro, South Korea
chodh@daejin.ac.kr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240519112501

Original scientific paper
Received: 2024-05-19, Accepted: 2024-05-22

Bridging the Gap to Industry 5.0: Comparative Analysis of Technologies in Industry 4.0 and 5.0 and the Evolutionary Path of the Smart Production Lab

Lena Sophie Leitenbauer*, Sabrina Romina Sorko, Christine Lichem-Herzog

Abstract: Since 2011, the term 'Industry 4.0' (I4.0) has gained significance in industry. After a decade of digital transformation, the European Commission is now advancing towards Industry 5.0 (I5.0). The focus is on using technology to support people, enhance ecological sustainability, and make industry more resilient. This paper examines the transition from I4.0 to I5.0, with a particular focus on the learning factory Smart Production Lab as a model for future-oriented manufacturing companies. The study involves a systematic literature review to identify key technologies and concepts of I4.0 and analyse their evolution in the context of I5.0. A comparative analysis forms the basis for a matrix that facilitates a clear comparison and guides future developments of the Lab. This research identifies the technologies underpinning the goals of I5.0 and their implications for practical applications in manufacturing. It also provides actionable recommendations for companies.

Keywords: Digitalization; Industry 4.0; Industry 5.0; Transformation

1 INTRODUCTION

The fourth industrial revolution, designated as Industry 4.0 (I4.0), has brought about significant changes in the global manufacturing landscape. This revolution is characterised by the adoption of automation, digitalisation, and interconnectivity in the manufacturing sector. It leverages advanced technologies such as the Internet of Things (IoT), artificial intelligence (AI), cloud computing and cyber-physical systems (CPS). I4.0 aims to improve overall productivity and process efficiency to promote sustainable economic development [1]. The implementation of I4.0 technologies has the potential to result in significant improvements and cost reductions across a range of industrial sectors [2]. The integration of I4.0 technologies has led to the development of smart factories and interconnected manufacturing systems. These factories facilitate enhanced productivity, reduced costs, and improved product quality [3, 4]. Consequently, the integration of I4.0 technologies into manufacturing processes has the potential to transform industries and stimulate accelerated growth [5].

However, as the digital transformation journey progresses, a new perspective emerges with the advent of Industry 5.0 (I5.0). Pushed by the European Commission, I5.0 seeks to transcend beyond the technological advancements of its predecessor by integrating human creativity, ethical considerations, and environmental aims into the core of industrial operations. I5.0 places emphasis on sustainability, corporate resilience, and societal well-being. It posits a collaborative coexistence between humans and machines, with the objective of achieving a balanced synergy that fosters innovation while prioritising environmental responsibility and social inclusivity [6].

This paper examines the transition from I4.0 to I5.0, with a focus on the learning factory Smart Production Lab of UAS JOANNEUM in Kapfenberg as a model for forward-thinking manufacturing companies. The key technologies defining I4.0 are analysed and their role for I5.0 is assessed through a literature review and a comparative analysis. This study aims to compare two industrial developments and identify their similarities and differences, and thus answer the research question:

Which I4.0 technologies support the transformation process towards I5.0?

The outcome is transmitted to the Smart Production Lab as a practice use case. This represents the initial stage in determining the current state of the lab, thereby enabling the definition of subsequent steps towards the transition to I5.0.

2 DRIVING FORCES BEHIND THE EVOLUTION TO I5.0

The transition from I4.0 to I5.0 represents not merely an evolutionary step in production but a paradigm shift towards a more human-centric, sustainable, and resilient industrial future. This change is of great relevance to the industrial ecosystem, as it involves crucial challenges and opportunities [7].

Firstly, the necessity for ecological sustainability has never been more apparent. With climate change and environmental degradation posing significant threats to global stability and prosperity, I5.0's emphasis on sustainable production practices offers a potential solution to mitigate these impacts. By integrating green technologies and circular economy principles, I5.0 aims to reduce waste, lower carbon footprints and ensure that economic growth does not come at the environment's expense [8].

Secondly, the social implications of industrial automation and digitalisation under I4.0 have raised concerns about workforce displacement and the devaluation of human skills. I5.0 addresses these challenges by reorienting the human perspective and needs within the production process. It advocates for a collaborative model where human creativity and machines' efficiency are harmonised, ensuring that technological advancements augment rather than replace human labour. This approach not only warrants the continued employment of workers but also enhances their job satisfaction and safety, thereby fostering a more inclusive and equitable industrial workforce [9, 10].

Furthermore, the resilience of production systems has been challenged by recent global disruptions, such as the COVID-19 pandemic. The focus of I5.0 on adaptable and flexible manufacturing, supported by advanced digital technologies and human ingenuity, prepares companies to better withstand future shocks. It promotes a more agile response to changing market demands and unforeseen challenges, ensuring continuity and stability in volatile times [10]. Finally, the transition to I5.0 has the potential to significantly impact the economy. By optimising production efficiency, reducing resource consumption, and creating high-value jobs, I5.0 can drive sustainable economic growth. It enables companies to compete in a rapidly evolving global market, where innovation, sustainability, and social responsibility are increasingly becoming competitive differentiators [11].

In conclusion, it can be stated that the transition to I5.0 is particularly pertinent in the context of the current ecological, social, and economic challenges. It represents a forward-thinking vision that aligns industrial advancement with the broader goals of sustainable development, societal well-being, and economic resilience, marking a pivotal step towards a more sustainable and inclusive future.

3 BRIDGING THE GAP: HOW I4.0 SERVES AS THE STEPPING STONE TO I5.0

I4.0 established the technological and conceptual foundation upon which I5.0 was subsequently built and will be further evolved. The automation and digitalisation initiated by I4.0 heralded a new era in manufacturing, where CPS, IoT, and advanced data analytics play pivotal roles. These technologies have enabled production processes to become more efficient, flexible, and intelligent [12].

I5.0 represents a value-driven approach that extends the foundation established by I4.0. It integrates human intuition, ecological orientation and corporate adaptability into the production process, aiming for a balance between technological efficiency and corporate responsibility. While I4.0 created the conditions for comprehensive connectivity and automation, I5.0 focuses on re-integrating humans into the production chain, achieving personalized, flexible, and above all, sustainable production outcomes [13].

As previously noted, the primary objective of I4.0 was to enhance economic efficiency. In contrast, I5.0 places greater emphasis on ecological sustainability, as evidenced by the growing prominence of circular economy practices and resource-efficient manufacturing methods [14].

Moreover, data-driven decision-making, a pivotal aspect of I4.0, is further advanced in I5.0 through the utilisation of advanced AI and machine learning (ML). This refinement enables the implementation of personalised and optimised production processes, as well as product development and customer interaction. I5.0 aims to achieve efficient and user-friendly production solutions by integrating human expertise with intelligent and precise devices [15]. The advancement of advanced technologies, including edge computing, digital twins, interactive robots, IoT, blockchain, and 6G, can facilitate the achievement of the I5.0 objectives [16]. In addition, the integration of data-driven design processes in new product and service development can enable the creation of new or improved products and services, the establishment of long-term customer relationships, and the resolution of societal challenges [17].

In summary, I5.0 employs advanced technologies and data-driven approaches to optimise and personalise various aspects of production and customer interaction, thereby creating responsible, agile and sustainable corporations. Consequently, I4.0 serves as a crucial foundation for I5.0, not only providing the technological basis but also paving the way for an industrial revolution that harmonises technology and humanity, advancing the pursuit of a sustainable and inclusive future.

Tab. 1 provides a succinct illustration of the manner in which I5.0 builds upon the technological advancements of I4.0 while setting new priorities, particularly in support of the outlined I5.0 values.

Table 1 Key aspects in comparison [18-21]

Feature	Industry 4.0	Industry 5.0
Objective	Efficiency and automation	Human-machine collaboration, resilient corporations and sustainability
Focus	Digitalization and connectivity	Personalization, agility and ecological sustainability
Technologies	IoT, CPS, big data, cloud computing	(In addition to I4.0 technologies) advanced AI, human-robot collaboration, assisting technologies
Workforce	Replacement through automation	Augmentation and enhancement of human capabilities
Decision Making	Data-driven	Data and human-centric
Sustainability	Efficiency-focused	Resource efficiency and circular economy

Such a comparison serves as a useful tool to clarify the evolutionary progression from I4.0 to I5.0, highlighting key changes in technology use and objectives of each phase.

The following chapter presents the methodological framework for the initial steps when changing from an I4.0 to an I5.0 perspective in companies. The example of the learning factory Smart Production Lab of UAS JOANNEUM is used to illustrate this framework.

4 METHODOLOGY: FRAMEWORK FOR TRANSITION ANALYSIS FROM I4.0 TO I5.0

The objective is to provide support to companies undergoing a change process to I5.0. This involves the evolution of the Smart Production Lab in Kapfenberg, Austria, from an I4.0 to an I5.0 facility, with the aim of establishing it as a good practice example. This research is part of the EU Horizon: bridges 5.0 project, which is funded by the European Union. In order to adopt a scientifically valid approach, a multi-layered procedure was developed. This commenced with the definition of the existing state of the Smart Production Lab, with a particular focus on the current technologies and their application scope. This was identified as a key aspect and starting point for the outlined transition. Based on this, recommendations for action were developed for the transformation from I4.0 to I5.0. These recommendations were based on the use of state-of-the-art methods for sustainably successful transformation processes.

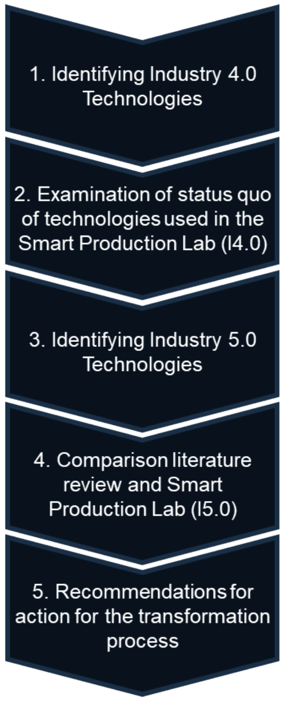


Figure 1 Schematic procedure for the comparative analysis

Fig. 1 outlines the methodology adopted for this analysis. The first four steps are detailed in this chapter, while the fifth step, which offers recommendations for action, is discussed in a subsequent chapter. To establish a standardised and comprehensive foundation for understanding the technologies relevant to the respective industrial revolutions, a literature review was conducted.

Initially, the literature search targeted sources discussing I4.0 technologies, reflecting its introduction at the Hannover Messe in Germany in 2011 and its relevance to the digital transformation of the manufacturing industry [22].

The analysis concentrated on sources that explicitly addressed the transition and related technologies to I 4.0 and I5.0 separately. A comprehensive literature review employed two types of resources: the scientifically robust perspective from journals and conference proceedings, as well as reports from manufacturers and providers of I4.0 technologies, integrating both academic and practical viewpoints. It was executed in accordance with the methodologies of Ramdhani et al., 2014, and Okoli et al., 2014 [23, 24]. A comprehensive literature search was conducted in scientific databases, including Science Direct, Scopus, and Google Scholar, between October 2023 and January 2024. The search terms related to I4.0, I5.0, and their technologies were combined into search strings incorporating both industrial revolutions and their technologies. This foundational knowledge will serve as a basis for the comparative analysis, allowing for the assessment of the status of the Smart Production Lab and the formulation of action recommendations to align the Smart Production Lab with I5.0 standards.

4.1 Technologies in I4.0: A Systematic Identification

Following a comprehensive review of the identified sources using the outlined methodological approach, a comprehensive list was created which included the source, authors, a brief description and the I4.0 technologies identified by the authors for each entry. These technologies were then individually listed in an MS Excel spreadsheet. This step facilitated the highlighting of multiple mentions and ensured a clear assignment of technologies. An attempt was already made in this step to cluster the results according to their hierarchy and to classify meta- and sub-technologies. One example of this would be the classification of virtual reality (VR) as a subgroup of extended reality (XR). This illustrates the technological development paths within I4.0.

This methodological approach revealed that different authors have used varied nomenclatures for the same technology or technology group. All relevant technologies from [25-34] for I4.0 were summarized in Fig. 2.

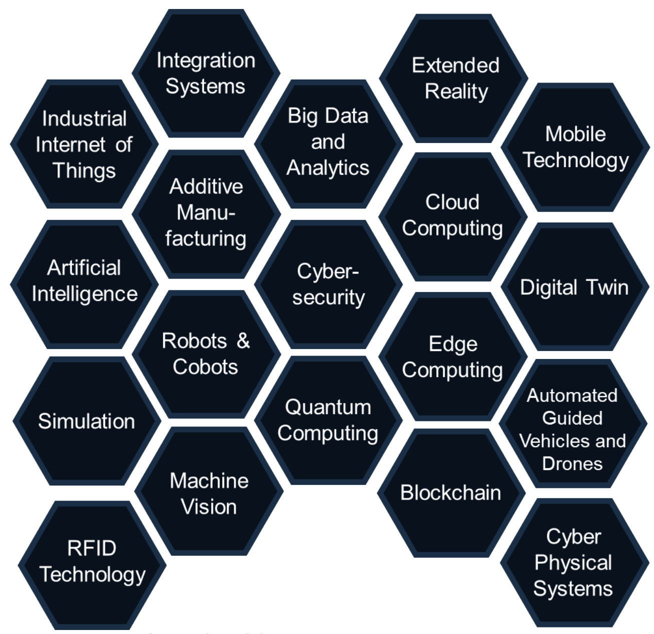


Figure 2 Identified I4.0 technologies based on literature review

Fig. 2 illustrates a honeycomb-shaped arrangement of blocks, each representing a key technology of I4.0. The name of a technology contributing to the fourth industrial revolution is given within each honeycomb cell. This visualization emphasises the diversity and complexity of technologies that synergise in the modern industrial landscape.

4.2 Assessing the Current State of the Smart Production Lab in the Context of I4.0

In the comparative analysis section, the current state of the Smart Production Lab is closely examined, covering its machinery, processes, educational materials, and technologies. This analysis was conducted on the basis of a comprehensive examination of the Smart Production Lab.

The potential avenues for improvement were evaluated through an examination of the existing infrastructure, process analysis, and workplace analysis. Inventory lists, process and workplace descriptions as well as qualitative interviews with responsible persons served as the data basis. The interviews were instrumental in understanding the nuanced applications of I4.0 technologies within the lab and assessing their alignment with the principles and objectives of I5.0. This methodical approach allowed for a detailed evaluation of the lab's I4.0 readiness and areas for enhancement in line with I5.0 advancements. The results are shown in Fig. 3.

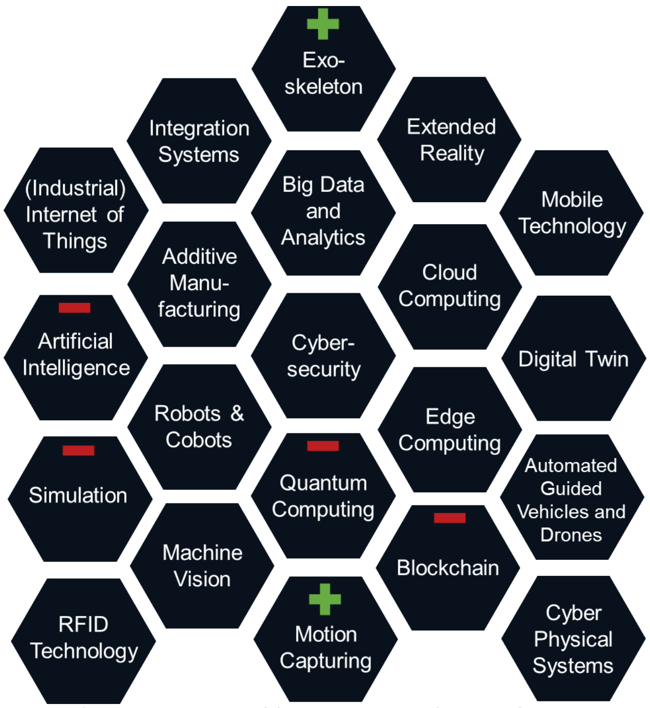


Figure 3 Comparative analysis: I4.0 technologies identified in the Smart Production Lab

Fig. 3 represents an evolved visualization from Fig. 2, indicating the technologies currently implemented in the Smart Production Lab. The graphic has been altered so that those honeycombs where the technology was not identified in the lab have been marked with a minus sign. Those technologies that were identified in addition to the previous research have been assigned with a plus. The technology analysis indicates that the lab incorporates two additional technologies not previously identified in the I4.0 literature research: exoskeletons and motion capturing. In total, four I4.0 technologies were either not integrated into the Lab or were in development at the time of the survey and therefore not yet in the application stage.

Considering the comprehensive evaluation and visualisation of the Smart Production Lab's status with respect to I4.0 technologies, the subsequent phase requires the identification of pivotal technologies for I5.0. This step is crucial for the alignment of the lab's technological framework with the principles of I5.0, emphasising human-centricity, resilience and ecological sustainability through advanced digital and physical integration.

4.3 Framework for Identifying Emerging Technologies in I5.0

The following presentation summarizes relevant scientific contributions on the transformation processes between I4.0 and I5.0, with a focus on the relevant technological paradigms. The tabular presentation of these contributions is structured as follows: column 1 shows the number of sources correlating with the information contained in the respective row. Column 2 provides a concise outline of the article's content, while column 3 defines the technologies, methodologies, domains, and objectives referenced in the context of I5.0 within the articles.

Table 2 Literature review for I5.0 transformation technologies

Source	Description	Technologies
[35]	Comparison between I4.0 and I5.0. Differences and challenges addressed by I5.0	Smart additive manufacturing, predictive maintenance, hyper customization, cyber-physical cognitive systems, collaborative robots (cobots)
[36]	Evolutionary vein and characteristics of I5.0. Key enablers	Ergonomics, mutual-cognitive human-robot collaboration, recommender system technology, bionics, advanced simulation, CPS, digital twin, metaverse, IoT-enabled systems, XR, IoT, holography, blockchain, decentralized computing, big data, cognitive computing, green computing, AI-based management systems, waste prevention, smart materials, disaster mitigation, renewable energy sources, sustainable agricultural production, zero-defect manufacturing, fin-tech
[37]	Contrasting I4.0 and I5.0: Shifting Research Priorities from Sustainability to Human-Centric Design	IIoT, ML / AI, HMI, big data, ethical technology, smart manufacturing / smart factory, CPS / digital twin
[38]	Transition from I4.0 to I5.0. Manufacturing	Advanced materials, intelligent manufacturing and processing, nanotechnology, sustainable manufacturing
[39]	Strategic roadmap how I5.0 can boost sustainable manufacturing	Cognitive cyber-physical systems, cognitive AI, human interaction and recognition technologies, XR, industrial smart wearables, intelligent (adaptive) robots, intelligent energy management system
[40]	Concept of I5.0 and enabling technologies	Individualized HMI, bio-inspired technologies and smart materials, digital twins and simulation, data transmission, storage, and analysis technologies, AI, technologies for energy efficiency, renewables, storage and autonomy
[41]	Potential applications and supporting technologies of I5.0	Edge computing, AI, cobots, 6G and beyond, digital twins, blockchain, Internet of everything
[42]	Conceptualization of I5.0 from the perspective of viability	IoT, 5G, edge computing, trace and tracking systems, blockchain, early-warning system, big data analytics

This structured approach facilitates a comprehensive understanding of the key technologies and strategic directions underpinning the evolution from I4.0 to I5.0. Following the methodology employed for I4.0 technologies, a similar categorisation and clustering were executed for the technologies related to I5.0. Notably, in contrast to the evaluation of I4.0 technologies, the authors' focus shifted from the technologies themselves to their proactive application towards optimisation and their contributions to ecological sustainability, resilience, or human-centricity. This aligns with the three pillars of I5.0.

For example, there was a notable increase in the mention of technologies designed to enhance ecological sustainability in industrial practice, including waste reduction, energy efficiency, renewable energy sources, storage solutions, and autonomy. Additionally, there was a pronounced presence of Human-Machine Interaction (HMI), which suggests a clear parallel.

To ensure comparability with I4.0 technologies, a systematic sorting of statements from each cited source was conducted. This was facilitated by creating a MS Excel spreadsheet, wherein each entry related to I5.0, such as "Predictive Maintenance" or "Ergonomics", was allocated a row. The association to sources and authors was managed via a column. This resulted in a comprehensive datasheet, with authors listed in the range B1:I1, titles of papers in B1:I11, and a brief description of the papers in B3:B11. Technologies, methodologies and solutions were categorised from cells 4 to 47. Subsequent sorting involved grouping explicitly assignable statements (e.g. all mentions of digital twin) into single rows. The subsequent phase, clustering, entailed a detailed examination of the content. The objective of clustering was to distinguish between actual technologies (such as XR, Big Data or AI) and pursued practices, goals and approaches (for instance, sustainable production, disaster mitigation or ergonomics). The outcome of this segmentation is the following Fig. 4, which combines the visualisation for I4.0 technologies with insights from the secondary literature research on I5.0 technologies.

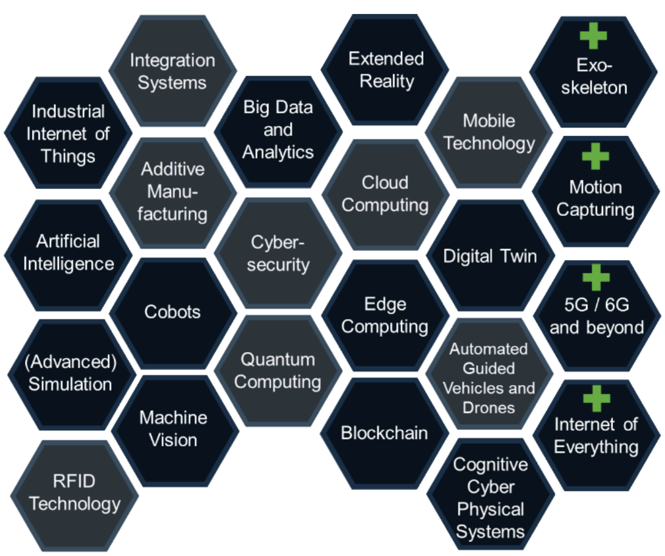


Figure 4 Identified I5.0 technologies (enhanced with technologies within the lab)

Two technologies in particular – the Internet of everything and 5G/6G and beyond – are newly listed (marked with a plus in Fig. 4). The technologies that were not explicitly reiterated for I5.0 in the initial literature search on I4.0 technologies are shaded light grey in Figure 4. This indicates that certain technologies were either important for I4.0 but have no direct relevance for I5.0 or are already seen as a new technological standard and are therefore no longer explicitly mentioned. This may be attributed to the authors' assumption of a certain degree of technology integration or a shift in focus introduced by I5.0. As demonstrated in Table 1, the motivation to use technology transitions from efficiency and automation (I4.0) to human-machine collaboration and sustainability. Furthermore, ergonomics is identified as a pertinent aspect for I5.0, although it is not associated with a specific technology. Therefore, exoskeletons and motion capturing, which are already present in the Smart Production Lab, are highlighted to represent this aspect, acknowledging the significant role ergonomics plays in the human-centred orientation of I5.0.

5 FINDINGS: TECHNOLOGICAL READINESS AND TRANSITION TOWARDS INDUSTRY 5.0

Fig. 4 and the conducted comparative analysis indicate that the Smart Production Lab is technologically well-prepared as a teaching and research facility to demonstrate digitalisation. The literature review for I5.0 has also indicated that a substantial portion of I4.0 technologies can serve as a basis, with technological advancements and innovations enabling new applications and industrial use cases.

The investigation reveals that while the Smart Production Lab is equipped with I4.0 technologies, including advanced manufacturing systems and analytics-driven decision-making tools, there is an emergent transition towards embracing the principles of I5.0. This transition is characterised by an intensified focus on human-centric designs and assisting technology, which are intended to complement human workers, not displace them, and a heightened commitment to sustainable production methodologies.

Furthermore, the analysis revealed potential opportunities for the integration of bespoke production techniques and the implementation of the principles of a circular economy, outlining pathways for the laboratory's evolution towards the I5.0 archetype. By leveraging these insights, the Smart Production Lab can continue to serve as a model for forward-thinking manufacturing companies, demonstrating effective strategies for navigating between the current industrial technological landscape and the visionary objectives of I5.0.

In order to align the Smart Production Lab with the forward-thinking paradigm of I5.0, a strategic approach is imperative, one that integrates technological innovation, education, and sustainability. Forging robust partnerships with industrial entities is critical to access cutting-edge technologies such as collaborative robots and IoT solutions, thereby bolstering research and development in main areas like digital twins and advanced simulation methods. A fundamental aspect of this transition is the comprehensive training of employees and students in new technologies and methods such as additive manufacturing and XR. Concurrently, the lab must intensify its sustainability initiatives, integrating principles of the circular economy and utilising recycled materials in additive manufacturing processes to promote energy-efficient production modalities.

These initiatives are not merely navigational tools through the technological transition towards I5.0; they are also instrumental in affirming the lab's role as a beacon for training next-generation professionals and promoting sustainable manufacturing practices.

The Smart Production Lab is situated within the context of I5.0, integrating a range of technologies that are essential to the realisation of the goals and principles of I5.0. An analysis of the technologies envisioned for I5.0 and those already deployed in the lab provides a detailed understanding of the lab's direction, the key technologies driving the transformation process, and the upgrades required. This comprehensive analysis is important to chart the path for future improvements and align with the goals of I5.0. In light of the aforementioned considerations, the research question can be answered as follows:

Positioning in relation to I5.0: The lab incorporates key technologies considered essential for I5.0, such as:

- Integration systems and Industrial Internet of Things (IIoT): These technologies are foundational for the interconnectivity of machines, facilities, and products, facilitating efficient, data-driven production.
- RFID technology: Critical for resource and product tracking, it enhances supply chain transparency and efficiency.
- Additive manufacturing: Enables hyper-customisation and efficient, on-demand production, aligning with the goals of I5.0 regarding flexibility and customer centricity.
- Cobots and exoskeletons: Support human-robot collaboration, to foster ergonomic work conditions and efficient production processes.
- Big Data and analytics: Indispensable for predictive maintenance and decision-making, these technologies allow for smart and forward-looking production management.
- CPS and digital twins: Key technologies for the simulation, monitoring, and optimisation of production processes in real-time.

Key Technologies for the transformation process from I4.0 to I5.0: The following technologies are instrumental in the transformational process toward achieving the goals of I5.0:

- Cobots and exoskeletons: Provide direct support for human-robot collaboration and ergonomic workplaces.
- Digital twins and CPS: Enable enhanced planning, simulation, and control of production processes.
- IIoT and edge computing: Critical for real-time data collection and processing, they support an agile and flexible production environment.

Necessary areas of improvement: To achieve full compatibility with I5.0 and to maximize its potential, the following upgrades are detected as necessary:

- XR and industrial smart wearables: Integration of these technologies could be intensified to improve work instructions and assist employees in complex tasks.
- Cognitive AI and human-interaction technologies: Advancing towards more intuitive, adaptive, and cognitive interactions between humans and machines may require increased investments in AI and ergonomic interaction technologies.
- Green technologies and sustainable production methods: While the laboratory already employs advanced technologies, there is a need to intensify the focus on sustainability and energy efficiency through the utilisation of renewable energy sources and the implementation of smart energy management systems.
- Simulation and AI for decision support: In order to ensure entrepreneurial resilience, organisations and their employees must enhance their problem-solving abilities and adaptability, as well as develop and anchor a holistic system understanding at all levels. Intelligent, AI-supported systems can support this.

The current inclusion of the aforementioned technologies positions the Smart Production Lab at the avant-garde of the I5.0 revolution, demonstrating a commitment to innovation, efficiency, and human-centric automation. In conclusion, the Smart Production Lab is well-positioned to adopt and integrate the key technologies essential for the realisation of I5.0's vision. The analysis of technologies mentioned for I5.0 and those used in the lab allows for a differentiated perspective on the lab's orientation, the key technologies for the transformation process, and the need for technological upgrading. This underlines the suitability of the lab for training and supporting companies on their way to I5.0.

6 CONCLUSION

The Smart Production Lab is well-positioned to facilitate the transition to I5.0, leveraging key technologies such as IIoT, cobots, digital twins, and big data analytics. However, there are areas, notably in XR, cognitive AI, and sustainability, where additional investments and developments could expedite the shift to a fully human-centred and sustainable production, in line with the values of I5.0. A key insight from this research is the shift in focus from the technologies themselves to their role as a means to an end, used strategically and efficiently to achieve set goals and initiatives. Following the initial step of identifying the status quo and envisioned future, further steps towards the transformation process are outlined for the Smart Production Lab and the Institute of Industrial Management at UAS JOANNEUM. These steps will be primarily implemented as part of the EU Horizon Project, Bridges 5.0, which will run until 2026. In order to provide a forecast of the forthcoming developments and associated research activities, the planned next steps are as follows:

1) A detailed mapping of current technologies to the three pillars of I5.0.
2) A roadmap for the education and development of competencies with respect to the three pillars of I5.0. It is of the utmost importance to more deeply integrate the values and attitudes associated with them into the transformation process and research activities.
3) Technological upgrades and integration into the Smart Production Lab as well as the definition of goals and aspects of the technologies and their contributions.

4) Demonstrating the advantages of technology integration with I5.0 considerations for companies, by making research findings accessible to interested business representatives through tours, lectures, and workshops at the Smart Production Lab.

Acknowledgements

The results from the paper were generated as part of a project funded by the European Commission (Horizon Europe: bridges 5.0).

7 REFERENCES

[1] Samaraz, D. Ş. (2023). Smart factory in the context of digital transformation. In Akkaya, B., & Tabak, A. (Eds.). *Two faces of digital transformation*. Emerald Publishing Limited, 129-140. https://doi.org/10.1108/978-1-83753-096-020231010

[2] Zhou, M., Qiao, Y., Liu, B., Vogel-Heuser, B., & Kim, H. (2023). Machine learning for Industry 4.0. *IEEE Robotics & Automation Magazine, 30*(2), 8-9. https://doi.org/10.1109/MRA.2023.3266618

[3] T., Bratu, George, D., Suciu, Cornelia, A., & Romulus, C. (2023). Steps towards Industry 4.0 in the Danube region. *Proceedings of SPIE*, 12493, 124930N-124930N. https://doi.org/10.1117/12.2643114

[4] Ryalat, M., ElMoaqet, H., & Alfaouri, M. (2023). Design of a smart factory based on cyber-physical systems and Internet of Things towards Industry 4.0. *Applied Sciences, 13*(4), 2156. https://doi.org/10.3390/app13042156

[5] Rajamanickam, M., John Gerard Royan, E. N., Ramaswamy, G., Rajendran, M., & Vadivelu, V. (2023). Fourth Industrial Revolution: Industry 4.0. In Rajasekar, R., Moganapriya, C., Sathish Kumar, P. & Harikrishna Kumar, M. (Eds.). *Integration of Mechanical and Manufacturing Engineering with IoT: A Digital Transformation* (Chapter 2). https://doi.org/10.1002/9781119865391.ch2

[6] Mert, H. (2023). Industry 5.0, digital society, and Consumer 5.0. In *Research on Perspectives on Society and Technology Addiction*, 11-33. https://doi.org/10.4018/978-1-6684-8397-8.ch002

[7] Etieno, S., Enang, M., Bashiri, M., & Jarvis, D. (2023). Exploring the transition from techno-centric industry 4.0 towards value-centric industry 5.0: A systematic literature review. *International Journal of Production Research*, 1-37. https://doi.org/10.1080/00207543.2023.2221344

[8] Singh, T., Singh, D., Singh, C. D., & Singh, K. (2023). Industry 5.0. In: Singh, C. D. & Kaur, H. (Eds.) *Factories of the Future: Technological Advancements in the Manufacturing Industry* (Chapter 2). https://doi.org/10.1002/9781119865216.ch2

[9] Venkatesh, S. (2023). The arc of Industry 5.0 bends towards a social purpose. *NHRD Network Journal, 16*(2), 190-195. https://doi.org/10.1177/26314541231160915

[10] Penulis, N. (2023). Social, organizational, and individual impacts of automation. In: Nof, S. Y. (eds) *Springer Handbook of Automation. Springer Handbooks*. Springer, Cham, 61-75. https://doi.org/10.1007/978-3-030-96729-1_3

[11] Kardush, I., Kim, S., & Wong, E. (2022). A techno-economic study of Industry 5.0 enterprise deployments for human-to-machine communications. *IEEE Communications Magazine, 60*, 74-80. https://doi.org/10.1109/MCOM.001.2101068

[12] Kumar, S., Kansal, S., Kaswan, K. S., Sharma, V., Miya, J., & Kumar, N. (2022). The Impact of Industry 4.0 Cyber Physical Systems on Industrial Development. In *Cyber Physical Systems*, 1st Edition, 125-140. https://doi.org/10.1201/9781003220664-8

[13] Pizoń, J., & Kulisz, M. (2023). Industry 5.0 - New approaches to innovation in the industry. In *Innovation in the Digital Economy*, 1st Edition. https://doi.org/10.4324/9781003384311-4

[14] Tiwari, S. (2023). Industry 4.0, sustainable manufacturing, circular economy, and sustainable business models for sustainable development. In *Handbook of Research on Sustainable Consumption and Production for Greener Economies*. IGI Global, 398-415. https://doi.org/10.4018/978-1-6684-8969-7.ch023

[15] Chen, K.-S., Lai, K.-K., & Wang, W.-P. (2023). Decision-making model of production data management for multi-quality characteristic products in consideration of Industry 4.0. *Applied Sciences, 13*(13), 7883. https://doi.org/10.3390/app13137883

[16] Yadav, M., Vardhan, A., Chauhan, A. S., & Saini, S. (2023). A study on creation of Industry 5.0: New innovations using big data through artificial intelligence, Internet of Things, and next-generation technology policy. *The IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS2023)*, Bhopal, India, 1-12. https://doi.org/10.1109/SCEECS57921.2023.10063069

[17] Wang, F.-Y., Yang, J., Wang, X. Z., Li, J., & Han, Q.-L. (2023). Chat with ChatGPT on Industry 5.0: Learning and decision-making for intelligent industries. *IEEE/CAA Journal of Automatica Sinica, 10*(4), 831-834. https://doi.org/10.1109/jas.2023.123552

[18] Gródek-Szostak, Z., Ochoa Siguencia, L., Niemczyk, A., & Szeląg-Sikora, A. (2023). From industry 4.0 paradigm towards industry 5.0. *Proceedings of the 14th International Scientific and Practical Conference, Vol. 2*. Vide. Tehnoloģija. Resursi. https://doi.org/10.17770/etr2023vol2.7192

[19] Alojaiman, B. (2023). Technological modernizations in the Industry 5.0 era: A descriptive analysis and future research directions. *Processes, 11*(5), 1318. https://doi.org/10.3390/pr11051318

[20] Dégallier-Rochat, S., Kurpicz-Briki, M., Endrissat, N., & Yatsenko, O. (2022). Human augmentation, not replacement: A research agenda for AI and robotics in the industry. *Frontiers in Robotics and AI, Vol. 9*. https://doi.org/10.3389/frobt.2022.997386

[21] Kumar Srivastava, S., Goel, P., & Savita Sindhu, A. (2022). Industry 5.0: Coexistence of Humans and Machines. In: Gaur, L., Agarwal, V. & Chatterjee, P. (Eds.). *Decision Support Systems for Smart City Applications* (Chapter 8), 137-152. https://doi.org/10.1002/9781119896951.ch8

[22] Kagermann, H., Wahlster, W., & Helbig, J. (2011). Industrie 4.0: Mit dem Internet der Dinge auf dem Weg zur 4. industriellen Revolution. *VDI nachrichten, 13*. (in German)

[23] Ramdhani, A., Ramdhani, M. A., & Amin, A. S. (2014). Writing a literature review research paper: A step-by-step approach. *International Journal of Basic and Applied Science, 3*(1), 47-56.

[24] Okoli, C., & Schabram, K. (2010). A guide to conducting a systematic literature review of information systems research. *SSRN Electronic Journal, 10*. https://doi.org/10.2139/ssrn.1954824

[25] SYDLE. (2022). What are the Industry 4.0 technologies? How do they work? https://www.sydle.com/blog/industry-40-technologies-60e486e2b2503757978621a0

[26] PTC. (2024). Industry 4.0: Optimize operations and shape future innovation. https://www.ptc.com/en/solutions/digital-manufacturing/industry-4-0

[27] Uzialko, A. (2023). Industry 4.0: How technology is revolutionizing the manufacturing industry. Business News

Daily. https://www.businessnewsdaily.com/10156-industry-manufacturing-iot.html

[28] Business Process Incubator. (2022). Industry 4.0 technologies. https://www.businessprocessincubator.com/content/industry-4-0-technologies/

[29] Jassem, S., & Razzak, M. R. (2021). Industry 4.0: The future of manufacturing—Foundational technologies, adoption challenges, and future research directions. In Al Mawali, N. R., Al Lawati, A. M. & A. S. (Eds.), *Fourth Industrial Revolution and Business Dynamics*. Palgrave Macmillan, Singapore. https://doi.org/10.1007/978-981-16-3250-1_7

[30] Wopata, M. (2020, February 4). Industry 4.0 adoption 2020 – who is ahead? IoT Analytics. https://iot-analytics.com/industry-4-0-adoption-2020-who-is-ahead/

[31] Fernández-Caramés, T. M., & Fraga-Lamas, P. (2018). A review on human-centered IoT-connected smart labels for the Industry 4.0. *IEEE Access, 6*, 25939-25957. https://doi.org/10.1109/ACCESS.2018.2833501

[32] Saturno, M., Pertel, V. M., & Deschamps, F. (2017, July). Proposal of an automation solutions architecture for Industry 4.0. *Paper presented at the 24th International Conference on Production Research*, Poznan, Poland. https://doi.org/10.12783/dtetr/icpr2017/17675

[33] Dalmarco, G., Ramalho, F. R., Barros, A. C., & Soares, A. L. (2019). Providing industry 4.0 technologies: The case of a production technology cluster. *The Journal of High Technology Management Research, 30*(2). https://doi.org/10.1016/j.hitech.2019.100355

[34] Gomez, C., Guardia, A., Mantari, J. L., Coronado, A., & Reddy, J. N. (2021). A contemporary approach to the MSE paradigm powered by artificial intelligence from a review focused on polymer matrix composites. *Mechanics of Advanced Materials and Structures, 29*(21), 1-21. https://doi.org/10.1080/15376494.2021.1886379

[35] Khan, M., Haleem, A., & Javaid, M. (2023). Changes and improvements in Industry 5.0: A strategic approach to overcome the challenges of Industry 4.0. *Green Technologies and Sustainability, 1*(2). https://doi.org/10.1016/j.grets.2023.100020

[36] Leng, J., Sha, W., Wang, B., Zheng, P., Zhuang, C., Liu, Q., Wuest, T., Mourtzis, D., & Wang, L. (2022). Industry 5.0: Prospect and retrospect. *Journal of Manufacturing Systems, 65*, 279-295. https://doi.org/10.1016/j.jmsy.2022.09.017

[37] Crnjac Žižić, M., Mladineo, M., Gjeldum, N. & Celent, L. (2022). From Industry 4.0 towards Industry 5.0: A review and analysis of paradigm shift for the people, organization and technology. *Energies, 15*(14), 5221. https://doi.org/10.3390/en15145221

[38] Moller, D. P. F., Vakilzadian, H., & Haas, R. E. (2022, May). From Industry 4.0 towards Industry 5.0. In *The IEEE International Conference on Electro Information Technology (eIT2022)*. https://doi.org/10.1109/eIT53891.2022.9813831

[39] Ghobakhloo, M., Iranmanesh, M., Foroughi, B., Babaee Tirkolaee, E., Asadi, S., & Amran, A. (2023). Industry 5.0 implications for inclusive sustainable manufacturing: An evidence-knowledge-based strategic roadmap. *Journal of Cleaner Production, 417*, 138023. https://doi.org/10.1016/j.jclepro.2023.138023

[40] European Commission, Directorate-General for Research and Innovation, & Müller, J. (2020). Enabling technologies for Industry 5.0 – Results of a workshop with Europe's technology leaders. Publications Office. https://data.europa.eu/doi/10.2777/082634

[41] Maddikunta, P. K. R., Pham, Q.-V., Prabadevi, B., Deepa, N., Dev, K., Gadekallu, T. R., Ruby, R., & Liyanage, M. (2022). Industry 5.0: A survey on enabling technologies and potential applications. *Journal of Industrial Information Integration, 26*, 100257. https://doi.org/10.1016/j.jii.2021.100257

[42] Ivanov, D. (2022). The Industry 5.0 framework: Viability-based integration of the resilience, sustainability, and human-centricity perspectives. *International Journal of Production Research, 61*(9), 1-13. https://doi.org/10.1080/00207543.2022.2118892

Authors' contacts:

Lena Sophie Leitenbauer, Dipl. Ing.
(Corresponding author)
Institute Industrial Management,
FH JOANNEUM University of Applied Sciences,
Werk-VI-Straße 46a, 8605 Kapfenberg, Austria
Tel.: +43 316 5453 6356
E-mail: lena.leitenbauer@fh-joanneum.at

Sabrina Romina Sorko, MMag. Dr.
Institute Industrial Management,
FH JOANNEUM University of Applied Sciences,
Werk-VI-Strasse 46a, 8605 Kapfenberg, Austria
Tel.: +43 316 5453 8309
E-mail: sabrinaromina.sorko@fh-joanneum.at

Christine Lichem-Herzog, Mag.
Institute Industrial Management,
FH JOANNEUM University of Applied Sciences,
Werk-VI-Strasse 46a, 8605 Kapfenberg, Austria
Tel.: +43 316 5453 8337
E-mail: christine.lichem-herzog@fh-joanneum.at

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240725130415

Original scientific paper
Received: 2024-07-25, Accepted: 2024-08-13

Application of Life Cycle Assessment to Determine the Influence of Electricity Mix Profile and Driving Mode on the Environmental Impact of Electric Battery Vehicles

Jelena Topić Božič*, Ante Čikić, Simon Muhič, Boris Kraševec

Abstract: Transportation electrification is one of the key strategies in the sustainable energy transition. Life cycle assessment was performed to evaluate the environmental impact of different electricity mixes with higher uptake of renewable energy sources and various driving modes on the use of battery electric vehicles (BEVs). The results showed that the values for categories global warming, fine particulate matter formation, ozone formation, human health, and fossil fuel scarcity decreased by 15.8 %, 11.3 %, 25.3 %, and 12.9 %, respectively, in the Green 2030 scenario compared to the baseline. When the driving mode was considered, the impact decreased by 27.7 %, 36.9 %, 21.3 %, and 24.4 %, respectively, in the lowest use in the current production scenario. In the categories of land use and mineral resource scarcity, a 6.9 % and 2.1 % increase in the values was observed in Green 2030 scenario, showcasing the trade-off in different environmental impact categories.

Keywords: battery electric vehicles; driving mode; electricity use; electricity production mix; life cycle assessment; transportation

1 INTRODUCTION

In 2019, the European Green Deal (EGD) was released to tackle climate and environmental challenges. EGD aims to achieve zero net emissions of greenhouse gases in 2050 [1, 2]. The energy transition is the core objective of policies aimed at decarbonization and can be achieved by the accelerated deployment of renewable energy (RE) technologies and energy efficiency measures [3–5]. As energy use is responsible for most global greenhouse gas emissions (GHGE), transitioning to a more sustainable energy system is crucial [6]. Transportation electrification is one of the key strategies in the sustainable energy transition [7] as transportation still largely relies on liquid fuels mainly derived from oil (95 %) [8].

Road transportation remains a significant source of air pollution, primarily via exposure to particulate matter ($PM_{2.5}$), nitrogen oxides (NO_X), CO, and sulfur dioxide (SO_2)[9–11]. Electric vehicles (EVs) offer some advantages over traditional vehicles as EVs do not emit tailpipe pollutants, CO_2, or NO_X and reliability as the number of engine elements in EV is smaller, leading to cheaper maintenance and fewer breakdowns [11,12]. On the other hand, significant battery-related challenges remain, such as the driving range, which is typically limited from 200 to 350 km with full charge, long charging time, and battery costs [11]. The electrification roadmap is facing several challenges as the vehicle types are diverse in terms of purposes, sizes, and travel modes, leading to variable energy demands in consumption modes, amounts, and costs. Additionally, the benefits of transportation electrification will be greater if EVs deployment takes place in parallel with the decarbonization of the power system and increases the share of renewable energy sources (RES) [7].

Recently Mirz et al. showed that the CO_2 equivalent emissions differ greatly depending on the country where the EV is used. The results showed that life cycle of CO_2 equivalent emissions of an exemplary electric car and a lifetime usage of 200000 km in selected countries were in the range of 16.1 t (Norway) to 41 t (Poland) [12]. Similarly, Shafiqe and Luo undertook a comparative life cycle assessment for the top 10 countries of EV sales, assessing current and future electricity mixes. The results showed that the overall performance of the EV. The results showed that overall performance of the EV is highly dependent on the electricity mix consumed during their production and use phase. The results showed that EVs in China with 2019 and 2025 electricity mix have a higher impact than others. EV with 2030 Norway electricity mix was an optimal choice and had the lowest environmental impact in most categories [13]. Burchart-Korol et al. compared the environmental impact of cars in Poland and Czech Republic. The results showed that the environmental indicators in all impact categories were higher for EVs in Poland than in the Czech Republic due to different electricity mix. In scenarios with a reduced share of coal among the sources of energy used for electricity production, a decrease in the values of environmental indicators was observed. When EV and internal combustion engine vehicles (ICEV) were compared, GHGE and fossil fuel depletion values were lower for EV [14]. On the other hand, in the categories terrestrial acidification, freshwater eutrophication, human toxicity, and particulate matter formation, values were higher in the case of EV [14]. Similarly, Del Pero et al. showed that using EVs significantly reduces climate change due to the absence of exhaust gas emissions. The advantages are even greater when a higher share of RES is introduced in electricity production. However, the manufacturing of EV compared to ICEV requires larger use of metals, chemicals, and energy; therefore, the resulting impact in categories acidification, human toxicity, particulate matter, photochemical ozone formation, and resource depletion are higher in the case of EV [15].

EVs are recognized as pillars bridging the gap towards renewable energy systems [16]. The deployment of EVs has been growing over the past ten years. Still, the penetration rates of EVs remain relatively low, corresponding to 1 % of the global car stock in 2019 [7], which can be attributed to several issues, including the high cost of infrastructure, the prices of EV, the scarcity of charging stations, and the limited range of EVs [17]. However, to reduce the environmental impact in the transport sector, the role of lifestyle should be considered, and emphasis should be put on policy and strategy development to reduce travel demand [18].

The transport's share of final energy use in Slovenia remains high, representing the sector with the highest share (41 %), followed by manufacturing and construction (25 %), households (22 %), service activities (9 %), and other users (3 %) [19]. Muhič et al reported that the share if the end energy use in transport was 39.5 % in 2014 [20]. Fig. 1 shows the final energy use by sector in 2014 – 2021 [21]. The final energy use in 2014 – 2021 was on average 4758.0 ktoe, above the 2030 NECP target set at 4426 ktoe. The year 2020 was the only year that the final energy use was slightly below the 2030 NECP target (4405.7 ktoe) which can be attributed to the lockdowns during the Covid-19 pandemic.

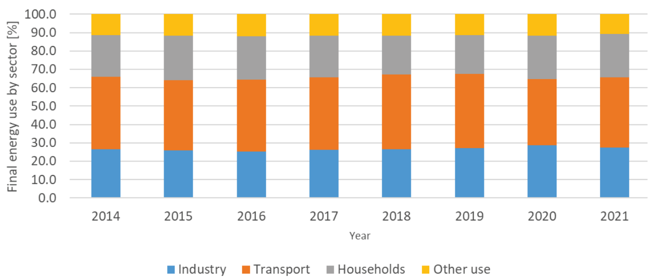


Figure 1 Final energy use by sector in Slovenija for the period 2014 – 2021 [21]

One of the main sources of GHGE in Slovenia is road traffic, as cars contribute 20 % of all GHGE. Due to Slovenia's smaller share of public transport to achieve a 67 % reduction, significant vehicle fleet electrification is needed [22]. The latest data shows that the number of newly registered battery EV and hybrid EV is increasing. The number of registered battery and hybrid EV has reached 5448 and 1617 in 2021, representing 3.4 % share of electric cars in first registrations [22]. Based on the latest available data for 2021, the number of newly registered EV in Slovenia is above the country's baseline (25.7 % difference) but below moderate (8.8 % difference) and ambitious (34.2 % difference) national climate and energy plan (NECP) scenario projections [22]. Similarly, a total number of batter EV and plug in hybrids in 2021 is above the baseline NECP projections (14.0 % difference) and slightly below moderate (2.2 % difference), and below ambitious (16.2 %) NECP scenario projections [22].

The widespread rollout of electric cars faces several technical and socio-economic challenges. The direction of the future transportation system is affected by political, economic, social, technological, and environmental reasons [16]. Adopting EVs to existing networks may increase peak demand for electricity, causing network congestion and requiring additional investment in the electricity grid infrastructure. Additionally, the environmental cost of the use of EVs is low only if the RES penetration in the considered system is significant. Decarbonization measures in Slovenia support the introduction of EVs. The main incentives and legislation include [23]:

- Registration tax benefits: BEVs pay minimum rate of 0.5 %;
- Purchase subsidies for private individuals and legal entities;
- Other financial benefits include Eco Fund long-term loans for the purchase of electric cars at a subsidized rate set by the Bank of Slovenia;
- Company tax benefits: for EVs used as company cars, there is zero benefit-in-kind tax;
- VAT benefits: Legal entities can claim VAT deductions for purchasing, servicing, and spare parts.

This study aimed to perform life cycle assessment (LCA) focusing on the use phase of the EV in Slovenia to evaluate the environmental impact of using BEV under different electricity scenarios. Different electricity mixes were considered, including scenarios with higher uptake of RES in 2030 and phasing out dependence on fossil fuels in the electricity mix.

2 MATERIALS AND METHODS

(LCA) is a standardized methodology based on ISO 14044 and 14040 standards for evaluation of the environmental impact of products and processes. LCA consists of four phases: goal and scope definition, inventory analysis, impact assessment, and interpretation [23, 24]. LCA defines the boundaries of the study, which typically include processes such as manufacturing (extraction of raw materials, transport to the factory, manufacturing of components, transportation), construction, operation and maintenance, and disposal [26].

2.1 LCA Software and Methods

This study utilized the software SimaPro 9.5. with Ecoinvent 3.9.1 database for foreground processes. ReCiPe Midpoint (H) was the impact assessment method used [26, 27].

2.2 System Definition and Boundaries

The study's primary goal is to evaluate and compare the potential environmental impact of using electric cars with different modes of driving and electricity production mix scenarios for Slovenia.

In electricity scenarios, all input and output data were related to the 1 kWh of produced electricity for the electricity generation scenarios. The functional unit (FU) was defined as 1 kWh of produced electricity.

For the electric car a defined FU was a journey of 1 km with an electric passenger car which is a process. All input and output data were related to the 1 km of journey with a compact electric car.

2.3 Life Cycle Inventory

This LCA uses Ecoinvent 3.9.1 data about electricity generation in Slovenia to report environmental impact of producing 1 kWh of electricity in Slovenia. The ReCiPe midpoint and CED methods were used to calculate environmental impacts. Slovenia has three primary electricity generation sources: nuclear, hydroelectric, and thermal power plants. Three electricity mix scenarios were modeled:

1) SI_2020 mix (Slovenia electricity generation sources with different shares without import – year 2020): Wind 0.04 %, Pump Hydro 1.4 %, Hydro 28.6 %, Lignite 24.6 %, Gas 0.15 %, Nuclear 37.1 %, Oil 0.01 %, Biogas CHP 0.5 %, Lignite CHP 2.5 %, Gas CHP 2.3 %, Oil CHP 0.05 %, Biomass CHP 0.5 %, and Solar 2.3 %.
2) SI_2030 mix (additional uptake of solar power without import with shares): Wind 0.04 %, Pump Hydro 1.2 %, Hydro 25.7 %, Lignite 22.1 %, Gas 0.14 %, Nuclear 33.3 %, Oil 0.01 %, Biogas CHP 0.5 %, Lignite CHP 2.5 %, Gas CHP 2.0 %, Oil CHP 0.05 %, Biomass CHP 0.04 %, and Solar 12.1 %.
3) SI_Green2030 mix (reduction in coal electricity): Wind 0.04 %, Pump Hydro 1.4 %, Hydro 28.6 %, Gas 0.15 %, Lignite 13.4 %, Nuclear 37.1 %, Oil 0.01 %, Biogas CHP 0.5 %, Lignite CHP 2.5 %, Gas CHP 2.3 %, Oil CHP 0.05 %, Biomass CHP 0.5 %, and Solar 27.9 %.

For the electric car, the available dataset for 1 km of transport with an electric car, available in the Ecoinvent database, 3.9.1, with a global geographical denomination, was used. A vehicle lifetime of 150000 km and an average battery lifetime of 100000 km were assumed. The study focused on the use phase – the environmental impact of using a car for transportation concerning the electricity mix scenarios described above and the driving profile. Different electricity use representing different driving styles was modeled and electricity use scenarios were chosen based on the available literature data [28, 29]:

1) Baseline: The process is available in the Ecoinvent 3.9.1 database, which models 0.199 kWh of electricity used for a journey of 1 km.
2) Low use: 0.150 kWh of electricity for a journey of 1 km;
3) Moderate use: 0.250 kWh of electricity for a journey of 1 km;
4) High use: 0.300 kWh of electricity for a journey of 1 km.

3 RESULTS AND DISCUSSION

3.1 LCA of the Impact of Electricity Mix on the Use of Electric Car

Tab. 1 shows the environmental impact in midpoint impact categories generated by a 1 km journey with an electric car in three hypothetical electricity mix scenarios. The baseline electricity production mix for Slovenia (year 2020) and two scenarios modeling an increased PV share (SI_2030) and reduction in coal electricity (SI_Green2030).

Table 1 Environmental impact of using electric car with hypothetical electricity mix scenarios (Functional unit is 1 km journey).

Impact category	Unit	SI_2020	SI_2030	SI_Green2030
Global warming	kg CO_2 eq	0.159446016	0.153873439	0.134273449
Stratospheric ozone depletion	kg CFC11 eq	0.000000050	0.000000049	0.000000049
Ionizing radiation	kBq Co-60 eq	0.057741176	0.052568903	0.057574530
Ozone formation, Human health	kg NO_x eq	0.000525799	0.000513405	0.000466363
Fine particulate matter formation	kg $PM_{2.5}$ eq	0.000597861	0.000563428	0.000446451
Ozone formation, Terrestrial ecosystems	kg NO_x eq	0.000599567	0.000587325	0.000540361
Terrestrial acidification	kg SO_2 eq	0.001669117	0.001555614	0.001180641
Freshwater eutrophication	kg P eq	0.000205132	0.000191823	0.000147501
Marine eutrophication	kg N eq	0.000019359	0.000018455	0.000015856
Terrestrial ecotoxicity	kg 1,4-DCB	1.235851020	1.309610056	1.317791213
Freshwater ecotoxicity	kg 1,4-DCB	0.027746272	0.028594659	0.027652181
Marine ecotoxicity	kg 1,4-DCB	0.036160349	0.037180793	0.035865067
Human carcinogenic toxicity	kg 1,4-DCB	0.029285870	0.028743272	0.026581153
Human non-carcinogenic toxicity	kg 1,4-DCB	0.468926516	0.463877112	0.420506987
Land use	m^2a crop eq	0.005248252	0.005283378	0.005614444
Mineral resource scarcity	kg Cu eq	0.001920566	0.001951656	0.001960755
Fossil resource scarcity	kg oil eq	0.040475825	0.039326098	0.035270061
Water consumption	m^3	0.001164947	0.001167766	0.001170062

In the impact category of global warming, values were higher in non-green scenarios. However, with the modeled phased-out thermal power plant in the Green scenario, a 15.8 % decrease was observed. Similarly, LCA analysis from Koroma et al. showed that global warming potential (GWP) decreases with increased RES share. In scenarios with moderate and high RES share, the GWP decreased by 36 % and 53 %, respectively, compared to the reference scenario (0.170 kg CO_2 eq.). The moderate and high RES scenarios modeled decreased coal and lignite share and increased wind electricity production with other RES (PV) increases [31]. Results from Shafique and Luo showed that the use phase results were mainly dependent on the electricity mix in accordance with our results. In future 2030 energy mix scenarios, Norway had the lowest GWP (13.251 kg CO_2 -eq), and South Korea had the highest GWP (35.179 kg CO_2 -eq) because a higher proportion of Norway's electricity is generated from RES. The results showed that integrating RES into electricity generation positively impacts the environmental impact of EV production and use; however, the study focused on a limited number of indicators, with mineral resource scarcity not taken into account [13].

Similarly to Shafique and Luo [13], the modeled increase in RES has a positive impact also on ozone formation (11.3 % decrease), fine particulate matter formation (25.3 % decrease), and fossil resource scarcity (12.9 %) in our study. Petrauskienė et al. recently conducted a comparative LCA of BEV and ICEVs fuelled with petrol and diesel. In addition, the LCA of BEV under different electricity mix scenarios for Lithuania (2015 – 2050) was carried out. The research assumed ICEV and BEV could drive 150000 km as the baseline. The environmental impacts were calculated for such a life cycle, ensuring that no battery replacement is required. The results showed that throughout the whole life cycle, the BEV of the 2015 electricity mix performed better in impact categories ionizing radiation and fossil depletion, while both ICEVs had a lower impact in climate change, human toxicity, and metal depletion, with impact on climate change being 26 % and 47 % higher than that of ICEVs fuelled with petrol and diesel, respectively. The difference in impact was attributed to the effect of the BEV's operation phase, which amounts to 70 % of the total burden, where the electricity of 2015 is produced with natural gases (41.7 %) and oil (5 %) [32].

In scenarios with higher penetration of RES, the GHG emissions become lower than that of ICEVs. In 2050 mix scenarios, where the main sources of electricity are PV (45.6 %) and wind (33.6 %), the impact of use phase is expected to reduce by 64 % and 73 % in terms of climate change and fossil depletion, respectively [32], which is in accordance with our results as phasing out coal and increasing PV has a positive impact on global warming, fossil resource scarcity and particulate matter formation. On the other hand, an increase in PV results in a higher impact in category mineral resources depletion, with 1.6 % and 2.1 % in SI2030 and SI_Green2030 scenarios, respectively. Similarly, an increase in impact category land use was also observed, with a 6.9 % increase in the SI_Green2030 scenario, compared to baseline 2020 scenarios. Increases in both categories can be attributed to the higher PV share in the electricity mix.

Recently, modeling of electricity scenarios mixes with different shares of photovoltaics (PV) for Slovenia showed that a significant increase in PV results in a 50 % in 2030 and 150 % in 2050 increase in the category of mineral resource scarcity compared to the 2020 baseline scenario for 1 kWh of electricity produced [33], highlighting that the values of mineral resource scarcity could be higher when EV are used in high PV electricity mix. PV causes negligible environmental impacts during the use phase; however, the material acquisition and consumption of metals is associated with the manufacturing phase [34]. Similarly, a higher impact in metal depletion indicators was observed for BEV compared to ICEV in the current and future energy mix scenarios due to the production of the battery. Furthermore, due to the larger use of metals, chemicals, and energy for the production of powertrains, BEV had a higher impact in the category of human toxicity [32].

3.2 Life Cycle Assessment of the Impact of the Driving Mode on the Use of Electric Car

The electricity use of BEV depends on several characteristics, among them driving environment (e.g., street characteristics, ambient temperature, road gradient) as well as driving style (e.g., aggressive, calm, driving during peak hour) and vehicle internal characteristics such as loading [29,30,35,36]. In the research conducted by Braun & Rid, it varied from 0.158 kWh per 1 km to 0.75 kWh per 1 km, depending on the road type, driving style, and traffic concentration [29]. If the electricity use varies, it is expected that the environmental impact of using BEV will also depend on the driving characteristics and environment [37].

The effect of driving mode on the environmental impact in selected impact categories of the use of BEV under different electricity scenarios is shown in Fig. 2. It can be observed that in the categories global warming, fine particulate matter formation, ozone formation, human health and fossil fuel scarcity, the impact is higher when electricity use is higher, according to available literature data [38]. The global warming category's values differ from 0.123 to 0.196 kg CO_2 eq., depending on the electricity use and mix scenarios. The lowest value observed was in 0.15 kWh per 1 km scenarios under Green 2030 electricity mix scenarios. The percentage difference between the lowest and the highest value was 46.0 %.

An important aspect to be considered is dependency not only on the yearly electricity mix but also the variation in the time of the day and season, highlighting the impact of the charging time on the environmental impact of the BEV [38]. On the other hand, categories of land use and mineral resource scarcity show the reverse trend. A higher share of RES (PV) and higher electricity use result in higher values in both impact categories due to land needed to install PV panels and the materials for battery production.

The lowest values observed were in 0.15 kWh per 1 km scenarios under the current electricity scenario with values 0.005 m^2a crop eq. and 0.0019 kg Cu eq., respectively. The percentage difference between the lowest and highest value in land use and mineral resource scarcity categories was 13.0 % and 6.3 %, respectively. The results support the premise that BEVs' impact heavily depends on the use phase of energy and the electricity mix used for charging [39]. Available data shows that BEVs have better environmental performance compared to ICEV in terms of CO_2 emissions during their entire life cycle [40]; however, the electricity mix is one of the most crucial parameters, as using a mix based entirely on RES delivers a completely different result than energy mix based on fossil fuels [36].

The environmental impacts of BEVs depend on several factors, which can be divided into external and internal factors. Among external factors, user driving mode, comfort requirements (i.e., air-conditioning, heating), and charging behavior are vital as they affect the electricity use required for 1 km of transport. Surrounding conditions such as climate, topography, and type of road also affect the electricity use required for 1 km of transport [36].

CO_2 emissions during their entire life cycle [40]; however, the electricity mix is one of the most crucial parameters, as using a mix based entirely on RES delivers a completely different result than energy mix based on fossil fuels [36]. The environmental impacts of BEVs depend on several factors, which can be divided into external and internal factors. Among external factors, user driving mode, comfort requirements (i.e., air-conditioning, heating), and charging behavior are vital as they affect the electricity use required for 1 km of transport. Surrounding conditions such as climate, topography, and type of road also affect the electricity use required for 1 km of transport [36].

On the other hand, categories of land use and mineral resource scarcity show the reverse trend. A higher share of RES (PV) and higher electricity use result in higher values in both impact categories due to land needed to install PV panels and the materials for battery production.

The lowest values observed were in 0.15 kWh per 1 km scenarios under the current electricity scenario with values

0.005 m^2 a crop eq. and 0.0019 kg Cu eq., respectively. The percentage difference between the lowest and highest value in land use and mineral resource scarcity categories was 13.0 % and 6.3 %, respectively. The results support the premise that BEVs' impact heavily depends on the use phase of energy and the electricity mix used for charging [39]. Available data shows that BEVs have better environmental performance compared to ICEV in terms of CO_2 emissions during their entire life cycle [40]; however, the electricity mix is one of the most crucial parameters, as using a mix based entirely on RES delivers a completely different result than energy mix based on fossil fuels [36]. The environmental impacts of BEVs depend on several factors, which can be divided into external and internal factors. Among external factors, user driving mode, comfort requirements (i.e., air-conditioning, heating), and charging behavior are vital as they affect the electricity use required for 1 km of transport. Surrounding conditions such as climate, topography, and type of road also affect the electricity use required for 1 km of transport [36].

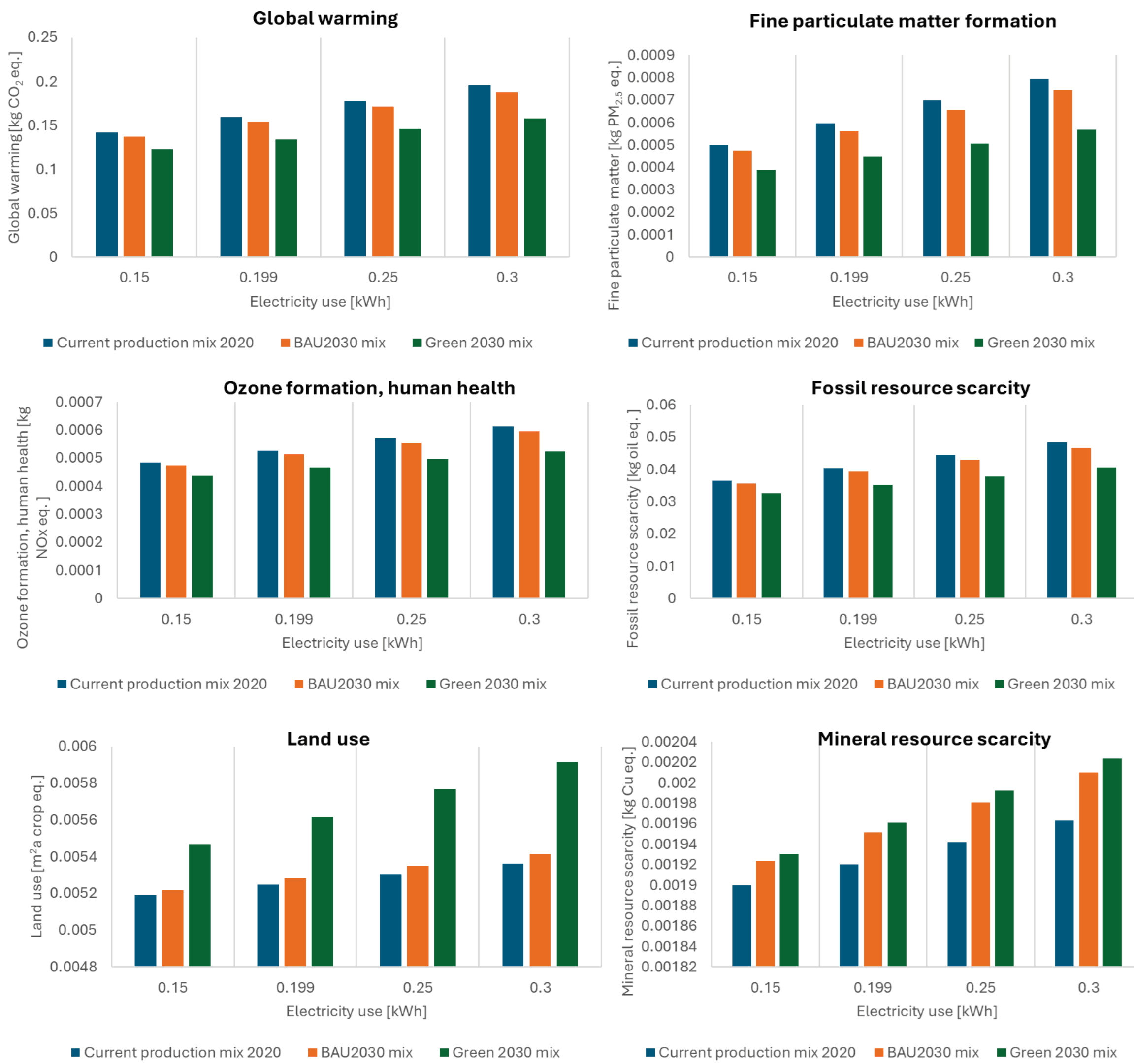


Figure 2 Environmental impact of driving mode under different electricity mix scenarios

BEVs are seen as a way to significantly reduce the world's dependence on fossil fuels and GHG emissions. A higher level of RES is also needed for a more sustainable transportation sector to promote EVs and RES diffusion [40]. The intermittency of some RES (namely PV and wind) hinders high penetration in the electrical grid [41], and higher penetration of BEV could result in increased peak loads under a simple charging strategy, requiring extra investment in generation and transmission capacity. However, if charging strategies are implemented, BEVs can potentially be charged at off-peak times when otherwise unwanted energy (i.e. excess energy in the system) can be used to charge the vehicle batteries. In such strategies, BEVs could be potentially used as a distributed storage mechanism for absorbing excess renewable energy [42]. Another vital aspect to be considered is the end-of-life of electric batteries, which pose a threat to the environment and are also a valuable source of various metals. With more sustainable practices in which recovered battery waste would be used in manufacturing batteries, the environmental impact could be potentially reduced; however, much more research is needed to strive towards all three pillars of sustainability – social, environmental, and economic [43–45].

4 CONCLUSIONS

The research utilized the LCA approach to determine the environmental impact of electricity mix and driving mode on the use of BEV. The analysis of future energy mix scenarios provides insight into how government measures could yield different results, showcasing the importance of evaluating plans and policies to select sustainable pathways for future transport systems. The values for categories global warming, fine particulate matter formation, ozone formation, human health, and fossil fuel scarcity decreased by 15.8 %, 11.3 %, 25.3 %, and 12.9 %, respectively with an increase in RES in the Green 2030 scenario compared to the baseline. On the other hand, in the categories of land use and mineral resource scarcity, a 6.9 % and 2.1 % increase in the values was observed, showcasing the trade-off in different environmental impact categories under various scenarios. Similarly to the impact of the energy mix, the result showed that different electricity use for 1 km journey showcases the driving modes, and topology has a vital role in the environmental impact of the use of BEVs. The values for categories global warming, fine particulate matter formation, ozone formation, human health, and fossil fuel scarcity decreased by 27.7 %, 36.9 %, 21.3 %, and 24.4 %, respectively when comparing the lowest and highest electricity use in the current production scenario.

The introduction of BEVs could result in several positive impacts including lower vehicle operating costs, reduced CO_2 emissions, and the ability to support and contribute to grid power quality and stability if the proper infrastructure is adopted. The market share of BEVs is still relatively small due to high investment cost compared to conventional cars, small driving ranges and the limited availability of charging infrastructure. Incetives policies are crucial for wider adoption of BEVs and could include purchase subsidies, registration tax benefits and ownership tax benefits.

5 REFERENCES

[1] European Commission. (2019). *The European Green Deal*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52019DC0640

[2] Hainsch, K., Löffler, K., Burandt, T., Auer, H., Crespo del Granado, P., Pisciella, P., & Zwickl-Bernhard, S. (2022). Energy transition scenarios: What policies, societal attitudes, and technology developments will realize the EU Green Deal? *Energy*, *239*, 122067. https://doi.org/10.1016/j.energy.2021.122067

[3] Neofytou, H., Nikas, A., & Doukas, H. (2020). Sustainable energy transition readiness: A multicriteria assessment index. *Renewable and Sustainable Energy Reviews*, *131*, 109988. https://doi.org/10.1016/j.rser.2020.109988

[4] Đurđević, D., Balić, D., & Jurić, Ž. (2019). Socio-Economic Aspects in Satisfying Energy Demands by Different Technologies of Heat and Electricity Generation. *Tehnički Vjesnik*, *26*(6), 1803–1813. https://doi.org/10.17559/TV-20180525214147

[5] Bošnjaković, M., Galović, M., Kuprešak, J., & Bošnjaković, T. (2023). The End of Life of PV Systems: Is Europe Ready for It? *Sustainability*, *15*(23), Article 23. https://doi.org/10.3390/su152316466

[6] Bogdanov, D., Ram, M., Aghahosseini, A., Gulagi, A., Oyewo, A. S., Child, M., Caldera, U., Sadovskaia, K., Farfan, J., De Souza Noel Simas Barbosa, L., Fasihi, M., Khalili, S., Traber, T. & Breyer, C. (2021). Low-cost renewable electricity as the key driver of the global energy transition towards sustainability. *Energy*, *227*, 120467. https://doi.org/10.1016/j.energy.2021.120467

[7] Yuan, M., Thellufsen, J. Z., Lund, H., & Liang, Y. (2021). The electrification of transportation in energy transition. *Energy*, *236*, 121564. https://doi.org/10.1016/j.energy.2021.121564

[8] de Blas, I., Mediavilla, M., Capellán-Pérez, I., & Duce, C. (2020). The limits of transport decarbonization under the current growth paradigm. *Energy Strategy Reviews*, *32*, 100543. https://doi.org/10.1016/j.esr.2020.100543

[9] Bonsu, N. O. (2020). Towards a circular and low-carbon economy: Insights from the transitioning to electric vehicles and net zero economy. *Journal of Cleaner Production*, *256*, 120659. https://doi.org/10.1016/j.jclepro.2020.120659

[10] Hossain, M. S., Kumar, L., Islam, M. M., & Selvaraj, J. (2022). A Comprehensive Review on the Integration of Electric Vehicles for Sustainable Development. *Journal of Advanced Transportation*, *2022*(1), 3868388. https://doi.org/10.1155/2022/3868388

[11] Sanguesa, J. A., Torres-Sanz, V., Garrido, P., Martinez, F. J., & Marquez-Barja, J. M. (2021). A Review on Electric Vehicles: Technologies and Challenges. *Smart Cities*, *4*(1), Article 1. https://doi.org/10.3390/smartcities4010022

[12] Hirz, M., Brunner, H. & Nguyen, T. T. (2022). Greenhouse Gas Emissions of Electric Cars—A Comprehensive Evaluation. *Tehnički Glasnik*, *16*(2), 280–287. https://doi.org/10.31803/tg-20220407135956

[13] Shafique, M., & Luo, X. (2022). Environmental life cycle assessment of battery electric vehicles from the current and future energy mix perspective. *Journal of Environmental Management*, *303*, 114050. https://doi.org/10.1016/j.jenvman.2021.114050

[14] Burchart-Korol, D., Pustejovska, P., Blaut, A., Jursova, S., & Korol, J. (2018). Comparative life cycle assessment of current and future electricity generation systems in the Czech Republic and Poland. *The International Journal of Life Cycle Assessment*, *23*(11), 2165–2177. https://doi.org/10.1007/s11367-018-1450-z

[15] Pero, F. D., Delogu, M., & Pierini, M. (2018). Life Cycle Assessment in the automotive sector: A comparative case study of Internal Combustion Engine (ICE) and electric car. *Procedia Structural Integrity*, *12*, 521–537. https://doi.org/10.1016/j.prostr.2018.11.066

[16] Capuder, T., Miloš Sprčić, D., Zoričić, D., & Pandžić, H. (2020). Review of challenges and assessment of electric vehicles integration policy goals: Integrated risk analysis approach. *International Journal of Electrical Power & Energy Systems*, *119*, 105894. https://doi.org/10.1016/j.ijepes.2020.105894

[17] Alanazi, F. (2023). Electric Vehicles: Benefits, Challenges, and Potential Solutions for Widespread Adaptation. *Applied Sciences*, *13*(10), Article 10. https://doi.org/10.3390/app13106016

[18] Brand, C., Anable, J., Ketsopoulou, I., & Watson, J. (2020). Road to zero or road to nowhere? Disrupting transport and energy in a zero carbon world. *Energy Policy*, *139*, 111334. https://doi.org/10.1016/j.enpol.2020.111334

[19] SURS. (2023). *Energy statistics, 2022*. https://www.stat.si/StatWeb/en/News/Index/11405

[20] Muhič, S., Čikić, A., Pištan, J., Stojkov, M., & Bošnjaković, M. (2018). Transport emissions and electric mobility in private transport in the Republic of Slovenia. *Tehnički Glasnik*, *12*(2), 98–103. https://doi.org/10.31803/tg-20180508162744

[21] ARSO. (2024). *Final energy consumption by sector | Okoljski kazalci*. https://kazalci.arso.gov.si/en/content/final-energy-consumption-sector-6

[22] ARSO. (2022). *Number of electric vehicles | Okoljski kazalci*. https://kazalci.arso.gov.si/index.php/en/content/number-electric-vehicles?tid=14

[23] European Commission. (2024). *Incentives and Legislation | European Alternative Fuels Observatory*. https://alternative-fuels-observatory.ec.europa.eu/transport-mode/road/slovenia/incentives-legislations

[24] ISO 14044:2006. https://www.iso.org/standard/38498.html

[25] ISO 14040:2006. https://www.iso.org/standard/37456.html

[26] Bošnjaković, M. & Tadijanović, V. (2019). Environment impact of a concentrated solar power plant. *Tehnički Glasnik*, *13*(1), 68–74. https://doi.org/10.31803/tg-20180911085644

[27] Huijbregts, M. A. J., Steinmann, Z. J. N., Elshout, P. M. F., Stam, G., Verones, F., Vieira, M., Zijp, M., Hollander, A., & van Zelm, R. (2017). ReCiPe2016: A harmonised life cycle impact assessment method at midpoint and endpoint level. *The International Journal of Life Cycle Assessment*, *22*(2), 138–147. https://doi.org/10.1007/s11367-016-1246-y

[28] RIVM. (2018). *The Dutch National Institute for Public Health and the Environment: LCIA: the ReCiPe model | RIVM*. https://www.rivm.nl/en/life-cycle-assessment-lca/recipe

[29] Braun, A., & Rid, W. (2017). Energy consumption of an electric and an internal combustion passenger car. A comparative case study from real world data on the Erfurt circuit in Germany. *Transportation Research Procedia*, *27*, 468–475. https://doi.org/10.1016/j.trpro.2017.12.044

[30] Braun, A., & Rid, W. (2017). The influence of driving patterns on energy consumption in electric car driving and the role of regenerative braking. *Transportation Research Procedia*, *22*, 174–182. https://doi.org/10.1016/j.trpro.2017.03.024

[31] Koroma, M. S., Brown, N., Cardellini, G., & Messagie, M. (2020). Prospective Environmental Impacts of Passenger Cars under Different Energy and Steel Production Scenarios. *Energies*, *13*(23), Article 23. https://doi.org/10.3390/en13236236

[32] Petrauskienė, K., Skvarnavičiūtė, M., & Dvarionienė, J. (2020). Comparative environmental life cycle assessment of electric and conventional vehicles in Lithuania. *Journal of Cleaner Production*, *246*, 119042. https://doi.org/10.1016/j.jclepro.2019.119042

[33] Dimnik, J., Topić Božič, J., Čikić, A., & Muhič, S. (2024). Impacts of High PV Penetration on Slovenia's Electricity Grid: Energy Modeling and Life Cycle Assessment. *Energies*, *17*(13), Article 13. https://doi.org/10.3390/en17133170

[34] Gargiulo, A., Carvalho, M. L., & Girardi, P. (2020). Life Cycle Assessment of Italian Electricity Scenarios to 2030. *Energies*, *13*(15), Article 15. https://doi.org/10.3390/en13153852

[35] Mamala, J., Graba, M., Mitrovic, J., Prażnowski, K., & Stasiak, P. (2023). Analysis of speed limit and energy consumption in electric vehicles. *Combustion Engines*, *195*(4), 83–89. https://doi.org/10.19206/CE-169370

[36] Egede, P., Dettmer, T., Herrmann, C., & Kara, S. (2015). Life Cycle Assessment of Electric Vehicles – A Framework to Consider Influencing Factors. *Procedia CIRP*, *29*, 233–238. https://doi.org/10.1016/j.procir.2015.02.185

[37] Cox, B., Mutel, C. L., Bauer, C., Mendoza Beltran, A., & van Vuuren, D. P. (2018). Uncertain Environmental Footprint of Current and Future Battery Electric Vehicles. *Environmental Science & Technology*, *52*(8), 4989–4995. https://doi.org/10.1021/acs.est.8b00261

[38] Rupp, M., Handschuh, N., Rieke, C., & Kuperjans, I. (2019). Contribution of country-specific electricity mix and charging time to environmental impact of battery electric vehicles: A case study of electric buses in Germany. *Applied Energy*, *237*, 618–634. https://doi.org/10.1016/j.apenergy.2019.01.059

[39] Hawkins, T. R., Singh, B., Majeau-Bettez, G., & Strømman, A. H. (2013). Comparative Environmental Life Cycle Assessment of Conventional and Electric Vehicles. *Journal of Industrial Ecology*, *17*(1), 53–64. https://doi.org/10.1111/j.1530-9290.2012.00532.x

[40] Franzò, S., & Nasca, A. (2021). The environmental impact of electric vehicles: A novel life cycle-based evaluation framework and its applications to multi-country scenarios. *Journal of Cleaner Production*, *315*, 128005. https://doi.org/10.1016/j.jclepro.2021.128005

[41] Asiaban, S., Kayedpour, N., Samani, A. E., Bozalakov, D., De Kooning, J. D. M., Crevecoeur, G., & Vandevelde, L. (2021). Wind and Solar Intermittency and the Associated Integration Challenges: A Comprehensive Review Including the Status in the Belgian Power System. *Energies*, *14*(9), Article 9. https://doi.org/10.3390/en14092630

[42] Richardson, D. B. (2013). Electric vehicles and the electric grid: A review of modeling approaches, Impacts, and renewable energy integration. *Renewable and Sustainable Energy Reviews*, *19*, 247–254. https://doi.org/10.1016/j.rser.2012.11.042

[43] Guzek, M., Jackowski, J., Jurecki, R. S., Szumska, E. M., Zdanowicz, P., & Żmuda, M. (2024). Electric Vehicles—An Overview of Current Issues—Part 1—Environmental Impact, Source of Energy, Recycling, and Second Life of Battery. *Energies*, *17*(1), Article 1. https://doi.org/10.3390/en17010249

[44] Dunn, J., Kendall, A., & Slattery, M. (2022). Electric vehicle lithium-ion battery recycled content standards for the US – targets, costs, and environmental impacts. *Resources, Conservation and Recycling*, *185*, 106488. https://doi.org/10.1016/j.resconrec.2022.106488

[45] Slattery, M., Dunn, J., & Kendall, A. (2021). Transportation of electric vehicle lithium-ion batteries at end-of-life: A literature review. *Resources, Conservation and Recycling*, *174*, 105755. https://doi.org/10.1016/j.resconrec.2021.105755

Authors' contacts:

Jelena Topić Božič, Assistant Professor
(Corresponding author)
Rudolfovo – Science and Technology Centre Novo mesto,
Podbreznik 15, 8000 Novo mesto, Slovenia
Faculty of Industrial Engineering,
Šegova ulica 112, 8000 Novo mesto, Slovenia
jelena.topic.bozic@rudolfovo.eu

Ante Čikić, Full Professor
University North, Department of Mechatronics,
104. Brigade 3, 42000 Varaždin, Croatia
ante.cikic@unin.hr

Simon Muhič, Full Professor
Faculty of Industrial Engineering,
Šegova ulica 112, 8000 Novo mesto, Slovenia
Rudolfovo – Science and Technology Centre Novo mesto,
Podbreznik 15, 8000 Novo mesto, Slovenia
Institute for Renewable Energy and Efficient Exergy Use,
INOVEKS d.o.o, Cesta 2. grupe odredov 17,
1295 Ivančna Gorica, Slovenia
simon.muhic@inoveks.si

Boris Kraševec, Assistant Professor
Faculty of Industrial Engineering,
Šegova ulica 112, 8000 Novo mesto, Slovenia
boris.krasevec@fini-unm.si

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20241002085211

Original scientific paper
Received: 2024-10-02, Accepted: 2024-10-11

Application of an Unmanned Aerial System (UAS) for Precise Fertilization

Ivan Plaščak, Mladen Jurišić, Irena Rapčan, Valentina Stanić, Dorijan Radočaj*

Abstract: The unmanned aerial system (UAS) eBee Plus, the eMotion guidance software and the Pix4dfields data processing software were investigated on two plots of wheat in PPK Valpovo. By using normalized difference vegetation index (NDVI), application maps for nitrogen fertilizer were obtained, which resulted in the saving of nitrogen fertilizer applied in the feeding of this crop. This guidance and data processing software has proven to be easy to use and suitable for agricultural crop producers who own larger areas to grow. The average grain yield of the Gabrio variety (7.92 t/ha) by eight individual plots, with an emphasis on plot 9A-4, had the average grain yield of this variety of 8.48 t/ha, which is 6.61% higher yield than the average of this variety on these plots. Multi-year research and direct comparison of multi-year parameters obtained from the plots results in an increasingly expedient and optimal use of resources for work and an increase in the quantity and quality of products.

Keywords: nitrogen fertilization; normalized difference vegetation index; precision agriculture; unmanned aerial vehicle; vegetation index

1 INTRODUCTION

Agricultural production is demanding in itself, and rationalization of production is very important from an ecological and economic point of view. The use of unmanned aerial system (UAS) opens new opportunities for rationalization [1]. As the demand for more food increases, so do the costs of agricultural production and greater extraction of nutrients from the soil. The lack of nutrients from the soil is compensated by fertilization, and the use of digital agriculture tools enables the monitoring of nitrates in the soil, soil temperature, electrical conductivity of the soil, soil moisture and other parameters and obtaining data in real time, which facilitates timely decision-making [2]. A geographic information system (GIS) is an integrated system of computer tools and user program support for the purpose of collecting, organizing, analysing, and modelling spatial data with the aim of solving complex problems of analysis and planning. Data processing is incomparably faster, data is more accessible and updated in real time. Small UAS equipped with Global Navigation Satellite System (GNSS) and photo equipment are used in agriculture. Today, the leading and most expensive system for geo-locating or determining the position of an UAS is the GNSS [3]. UAS use GNSS to quickly and efficiently map the production plot. It is also possible to create a map of the vegetation index and introduce it into the tractor system and treat only certain parts of the plot based on it, which reduces the consumption of plant protection products and mineral fertilizers. The vegetation index can be obtained by combining images from several spectral areas, mainly using the red and near-infrared part of the electromagnetic spectrum [4]. The measure of electromagnetic radiation of vegetation, the amount of green cover, the amount of biomass, the absorption of photosynthetic activity of radiation and the radiation of the index are defined as factors influencing the creation of the index. There are several vegetation indices, and one of the most commonly used is normalized difference vegetation index (NDVI) [5]. Inadequate fertilization often means excessive use of fertilizers. Precision fertilization emphasizes that growers should use different fertilization methods, different fertilizers and application rates according to different soil types, weather conditions and other [6]. Precision fertilization methods can be very useful in solving the problem of low fertilization efficiency.

The goal of this research is to provide a method for a precision fertilization map creation using remote sensing by UAS, show the mode of operation of the UAS eBee Plus that fly with the eMotion application, and process the obtained images in the Pix4dfields application.

2 MATERIALS AND METHODS

2.1 eBee Plus UAS

The flight time per battery of eBee Plus UAS is up to 59 minutes, which gives maximum efficiency when image large areas [7]. This aircraft can be operated at any altitude and at different wind speeds, as it is resistant to external conditions. Technical data about the UAS, the necessary software and management are presented in Tab. 1.

Table 1 Technical data on the UAS [7]

UAS	
Wingspan	110 cm
Weight	1.1 kg
Motor	Silent motor, brushless, electric
The range of the radio connection	3 km, usually up to 8 km
Demountable wings	Yes
Camera	senseFly S.O.D.A., Parrot Sequoia, TermoMAP
Software	
Flight planning and monitoring	eMotion
Management	
Automatic 3D flight planning	Yes
Flight speed	40 – 110 km/h (11 – 30 m/s)
Wind resistance	Up to 45 km/h (12 m/s)
Maximum flight time	59 minutes
Automatic landing	Linear landing with 5 meters accuracy
Handheld launch	Yes

The steps of the operation of the UAS are: data collection, processing of the collected data, creation of application maps, generation of indexes with interpretation and analysis. The UAS is equipped with a S.O.D.A camera. for capturing high-resolution aerial photogrammetric aerial images and precise digital models of surfaces. The crop is seen in a realistic state, and it is possible to see all the defects on the plants at the recorded moment. The image can be used for the purpose of protecting plants and destroying weeds on the recorded plot due to its accuracy. The Parrot Sequoia+ camera collects multispectral images of the surface. The camera contains four lenses for marking the plot during image. It contains green, red, red edge and near infrared spectrum lenses that provide absolute reflectance measurements. Technical data for both cameras are shown in Tab. 2. The eBee Plus UAS is controlled using the eMotion software.

Table 2 Technical data about S.O.D.A. and Parrot Sequoia+ cameras [7]

S.O.D.A. specifications	
Resolution	20 MP
Resolution at a height of 100 m	2,3 cm/px
Sensor size	12.75 × 8.5 mm
Pixel density	2.33 qm
Image format	JPEG or DNG
Parrot Sequioa+ specifications	
Pixel size	3.75 qm
Focal length	3.98 mm
Resolution	1280 × 960
Greed band	530 – 570 nm
Red band	640 – 680 nm
Red edge band	730 – 740 nm
Near infrared band	770 – 810 nm

With eMotion, flights are accomplished using mission blocks. The desired block (aerial mapping, corridor, etc.) is selected, the region to be mapped is marked, key settings are defined and eMotion automatically generates the UAS flight plan. By opening the Working Area Parameters, the destination from which the aircraft takes off is located. The selected location/point is near the planned flight mission for reasons of reducing battery consumption. Then the flight radius is given, which depends on the size of the mission itself. Under Working Area Parameters, the third icon was selected and Add mission block was added. By choosing Horizontal Mapping from the drop-down menu, the flight surface was determined, which resulted in a square feature of the surface (Fig. 1). The same image shows the selected points (in black in the image) that mark the image area and flight direction (in white in the image). The next step is determining the spatial resolution. When selecting a spatial resolution (pixel size) of 2.8 cm, the flight height is automatically adjusted to 119 m above sea level. By adjusting this parameter, the flight mission was adjusted and ready for implementation. After connecting the necessary equipment, the eMotion software was started, in which the flight mission was saved. After selecting the prepared mission at the bottom of the page on the left side, select Connect and then Connect to UAS. This enables connection with an UAS and then the realization of a flight mission is possible. By selecting the airplane icon on the left side of the eMotion window, the desired mission is added. Depending on the sequence of missions in the specified window of the software, the flight itself will be performed. On the right side of the mentioned software window are visible all key parameters of the aircraft, and their monitoring during the flight is extremely important (Fig. 2). At the opening of the mission, the Landing position was added, as the place where the aircraft will land after completing the mission. This is done by physically placing the aircraft at the landing site and clicking the mouse to mark the exact spot where the aircraft is shown on the screen. The plane descends downwind, and takes off (throws) upwind. Fig. 3 shows an overview of the flight parameters. The moment the aircraft reached a height of 75 m by pressing the icon on the Start mission screen, the collection of images began.

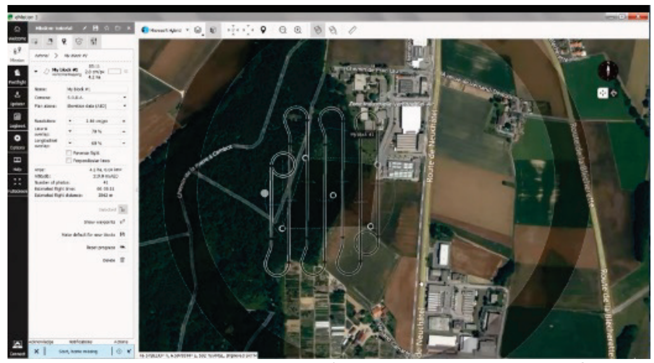


Figure 1 Rectangluar area limits for imaging using eBee Plus UAS

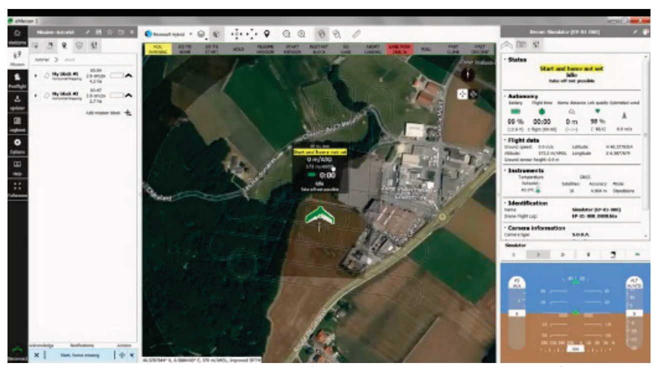


Figure 2 Flight parameters for imaging using eBee Plus UAS

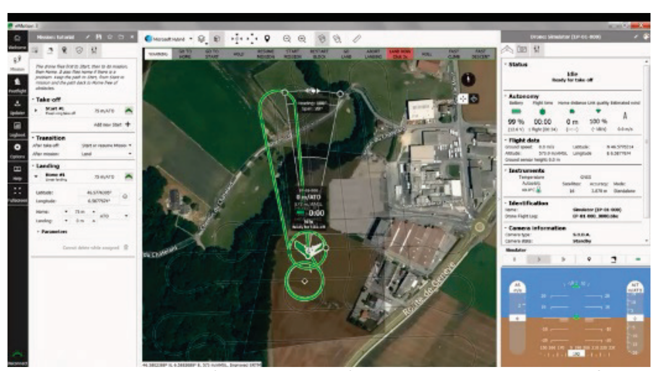


Figure 3 Additional flight parameters for imaging using eBee Plus UAS

The footage collection ended the moment the mission was completed. The data is saved on a USB card in the aircraft camera. After the flights on all prepared missions, the images were transferred to a computer (GIS workstation) and then data processing followed.

Pix4Dfields is software that converts images collected by aircraft into a map that can be immediately transferred to an agricultural machine as such or used later for analysis. The moment the data was loaded, the software automatically recognized the camera model, the image coordinates were automatically read, and the satellite image was opened. At the end of the processing itself, a reflection map was created on which the image area was outlined. At the end of the delineation, the desired vegetation index (NDVI) was selected. On the home page of the software, the New project icon (Plot 9A-4 and Plot 17P-1) is selected. By clicking on import files and double-clicking on images, a computer window is opened where all the captured images are marked and the processing of the images is started (by clicking on Apply).

2.2 Vegetation Index

The NDVI is the most used vegetation index for assessing the state of the vegetation, representing a good indicator of photosynthetic activity, in the simplest terms the ratio between the near-infrared (NIR) and the visible red part of the spectrum that plants reflect [5]. Its result ranges from -1 to +1, and the higher the value, the healthier the plants, with stronger vigor and lushness. It is also an indicator of the bio-physical properties of vegetation, the fractional cover of vegetation and the state of vegetation and biomass [8]. The NDVI index on surfaces with dense vegetation reacts to red reflectance, and is relatively insensitive to changes in infrared reflected rays. Also, this index is based on the properties of chlorophyll. The advantages of NDVI are independence from topography or radiation reflection, ease of use, good distinction between land surfaces and vegetation, easy readability due to high image resolution [9]. The disadvantage of this index lies in its great sensitivity to background noises, as ground reflections or atmospheric radiation and water content in plants. NDVI also primarily measures chlorophyll content and does not directly correlate with crop nitrogen status or nutrient requirements, which can lead to misinterpretations when assessing fertilization needs. Additionally, factors such as soil moisture, leaf area index, and atmospheric conditions can influence NDVI readings, potentially confounding the relationship between vegetation indices and nutrient stress. The temporal resolution of NDVI data might not align with critical growth stages, further limiting its effectiveness in dynamic agricultural systems. The factors that influence the measured value of the index include the resolution of the digital image of the observed area, atmospheric conditions, soil moisture, soil salinity, soil cover and differences in soil type. NDVI indices are visualized by color: brown, dark red, red, yellow, green and dark green. Each mark of the vegetation index has its own meaning, so the brown color indicates the surface of the soil, the dark red color indicates a crop with very poor emergence or diseases, the red color indicates a crop with weak emergence, the yellow color indicates a plant with uneven emergence, the green color indicates a plant in good condition, and dark green indicates excellent crop condition. Images acquired by UAS provide information on crop fertilizer needs [10]. Multispectral images are obtained so that the optical sensors function according to the principle of receiving light rays of a certain wavelength, and based on this, the image is determined. They are used to monitor the condition of crops in all stages of development, which facilitates making timely and informed decisions about crop treatment and fertilization, thus reducing costs and maximizing yield [11]. The state of the crop can be seen on the basis of the reflected light, as the wavelengths of the reflected light.

2.3 Case Study at PPK Valpovo

PPK Valpovo cultivates areas in the wider area of Valpovo totalling 5440 ha, of which 4850 ha are in conventional agricultural production, and 590 ha are cultivated according to the principles of ecological plant production. In agricultural production, the winter crops are wheat, barley and rapeseed, and the spring crops are corn, sunflower, soybeans and sugar beets. In the last two years, this company has intensified the use of new technologies and UAS, especially in the production of wheat. At the moment, they grow wheat on about 600 ha and several varieties, two of which were investigated in this paper. The variety is Kraljica winter wheat, of very good quality and fertility. It is the leading variety of the Osijek Agricultural Institute in the Republic of Croatia. This variety has a high yield potential, good crop stability, adaptability to different production conditions, tolerance to low temperatures and the most common wheat diseases. In addition to the Republic of Croatia, this variety is now successfully produced in a dozen countries in the surrounding area and beyond. The Gabrio variety is also a winter wheat with good lodging tolerance, high yield potential and grain quality, having medium resistance to frost, tolerant to drought and stress, and stable in different environments, even on poorer soil types.

3 RESULTS AND DISCUSSION

Predicting the amount of nitrogen required for wheat to reach its potential peak in a given year is difficult due to temporal and spatial variability, and within-field variability is a major source of uncertainty in wheat breeding decisions [12]. The goal of precise fertilization is to add as much fertilizer as is really necessary, and with it we improve the application of fertilizer according to actual needs [13]. GNSS determines the location of the tractor, collects and processes data and applies it to the agricultural plot [14]. At the end of processing, the created application maps were later used in precision agriculture. Research has confirmed that NDVI determined early in the growing season is a good predictor of the final grain yield of winter wheat across several locations and years [11]. The data used for precise fertilization are selected directly in the software. The software has the Color Maps and Prescription option, where it is possible to determine the amount of applied fertilizer in a certain area by color. At the beginning of March 2022 at PPK Valpovo an agricultural plot called 9A-4 with an area of 85.58 ha was recorded. In the previous season, sunflowers were the pre-crop on the plot, while wheat was recorded for the purposes of determining the amount of fertilizer needed for the second feeding. A sunny day was chosen for the image, with a wind speed of up to 3 m/s. The preparation of images in the eMotion software took only 20 minutes. A Sequoia camera was used for obtaining multispectral images for precise fertilization. A large amount of data was obtained (4648 images). Data processing was done in Pix4dfield software. The processing and creation of the multispectral map began with the entry of all images. The video was processed for about 20 minutes, which did not require the Internet. The first feeding was done in the classic way with 180 kg/ha KAN, and the second feeding using the application map (Fig. 4). On plot 9A-4, an average value of fertilizer of 150 kg/ha was applied. The range ranged from a minimum of 120 kg/ha to a maximum of 190 kg/ha, depending on the vegetation index of crop development obtained by UAS images. In the third plot (Fig. 5), the range ranged from a minimum of 100 kg/ha to a maximum of 150 kg/ha (average 120 kg/ha), which is 20 kg/ha less than in the second plot, where the range is from the minimum up to a maximum amount per hectare of 70 kg. It is important to note that as much as 78% of the plot was in the range of 20 kg, which means that after the second feeding there was an equalization of the crops within the plot itself.

Figure 4 Application map for the second supplementary fertilization on plot 9A-4

Figure 5 Application map for the third supplementary fertilization on plot 9A-4

On plot 17P-1, with an area of 51.53 ha, the pre-culture was sunflower, and in the first feeding, 180 kg/ha of KAN was applied, while the second feeding was carried out using the application map, and the third in the classic way. The fertilizer distribution in the second crop had the average value of 147 kg/ha, with a range of 130–160 kg/ha. The crop was quite uniform, so there were no big variations in fertilizer application. At the time of the third fertilization, the crop was almost uniform over the entire surface. As the map of the vegetation index was extremely uniform, the third supplement was done using the conventional approach.

Tab. 3 shows the average grain yield of the Gabrio variety (7.92 t/ha) by individual plots, eight of them, with an emphasis on plot 9A-4 and the average grain yield of this variety of 8.48 t/ha, which is 6.61% higher yield than the average of this variety on these plots. Also shown is the average grain yield of the Kraljica variety on five plots, an average grain yield of 8.31 t/ha. It can be seen that this variety on plot 17P-1 achieved the highest grain yield of 8.80 t/ha, which is 5.89% higher than the average yield of this variety. The average grain yield of the Gabrio variety, without plot 9A-4, was 7.75 t/ha, while the Kraljica variety achieved an average grain yield of 8.10 t/ha (without plot 17P-1). The total grain yield of the Gabrio variety on the area of 222.74 ha was 1771.31 t, of which the grain yield from plot 9A-4 (85.58 ha) was 725.71 t. The Kraljica variety achieved a total grain yield of 1330.79 t (on an area of 160.16 ha), of which 453.46 t were obtained on plot 17P-1 (an area of 51.53 ha). In one study, an average grain yield of 7.08 t/ha was recorded using variable doses of nitrogen fertilizer compared to control (5.92 t/ha) and uniform application before and during the growing season (6.22 t/ha). Research results show that the precision system can reduce overall nitrogen application levels from conventional pre-sowing levels by 59-82% depending on location. It has already been shown experimentally that reducing the amount of nitrogen fertilizer does not compromise the yield [15]. Fertilizer application leads to increased levels of nitrogen and phosphorus in the atmosphere and in most ecosystems on Earth [16]. It has been confirmed that the application of high nitrogen intake is, on the one hand, an incentive to achieve high quality standards, but on the other hand, it is the cause of a high positive nitrogen balance in the soil and the potential risk of air and water pollution. Site-specific management and application of nutrients required by the crop to reach its maximum potential yield can reduce nutrient pollution of groundwater and downstream water sources [17].

Table 3 Analysis of the yield of wheat varieties Gabrio and Kraljica

Gabrio variety			
Plot	ARKOD ID	Area, ha	Yield, t/ha
7A-1	2265712	9,09	6,79
7A-5	2265488	25,45	6,76
8A-3	2259785	9,70	8,31
8A-4	2263621	15,88	8,32
8A-5	1518677	19,94	8,34
8A-6	1512402	13,93	8,00
9A-4	1518178	85,58	8,48
13A-2	1518050	43,17	7,70
Average:			7,92
Kraljica variety			
Plot	ARKOD ID	Area, ha	Yield, t/ha
17P-1	1517257	51,53	8,80
18P-1	1516262	33,24	7,91
22P-2	1519208	48,10	8,13
16P-1	1517350	8,00	8,17
22P-1	1518511	19,29	8,19
Average:			8,31

Multi-year research and direct comparison of multi-year parameters obtained from the plots results in an increasingly expedient and optimal use of resources for work and an increase in the quantity and quality of products. In addition, the adoption of precision agriculture guarantees sustainable agriculture from the point of environmental protection.

4 CONCLUSIONS

The integration of advanced technology into agricultural practices has led to significant improvements in crop management and resource efficiency. In this case, the application of the eBee Plus UAS, paired with the eMotion guidance software and Pix4Dfields data processing software, was utilized to assess two distinct plots of wheat located in PPK Valpovo. Through the collection of high-resolution aerial imagery and precise data analysis, application maps for nitrogen fertilizer were generated utilizing NDVI. The creation of these application maps enabled more targeted and efficient application of nitrogen fertilizer, ultimately leading to a notable reduction in the quantity of fertilizer used during the wheat cultivation process. This approach not only enhances the sustainability of agricultural practices by minimizing excess nitrogen applications—which can lead to environmental concerns such as runoff and soil degradation—but also supports farmers in lowering input costs.

Moreover, the eMotion guidance software and Pix4Dfields provide intuitive user interfaces, making them accessible and practical tools for agricultural producers, especially those managing large tracts of land. The ease of use associated with this technology allows farmers to effectively implement precision agriculture techniques without requiring extensive training or expertise in remote sensing. Consequently, the combination of UAS technology, data processing software, and innovative agronomic practices holds great promise for optimizing crop production while promoting environmental stewardship in agriculture.

5 REFERENCES

[1] Raj, R., Kar, S., Nandan, R. & Jagarlapudi, A. (2020). Precision Agriculture and Unmanned Aerial Vehicles (UAVs). In Avtar, R. & Watanabe, T. (Eds.), *Unmanned Aerial Vehicle: Applications in Agriculture and Environment* (pp. 7–23). Springer International Publishing. https://doi.org/10.1007/978-3-030-27157-2_2

[2] Jurišić, M., Radočaj, D., Plaščak, I. & Rapčan, I. (2021). A Comparison of Precise Fertilization Prescription Rates to a Conventional Approach Based on the Open Source Gis Software. *Poljoprivreda*, *27*(1), 52–59. https://doi.org/10.18047/poljo.27.1.7

[3] Catania, P., Comparetti, A., Febo, P., Morello, G., Orlando, S., Roma, E. & Vallone, M. (2020). Positioning Accuracy Comparison of GNSS Receivers Used for Mapping and Guidance of Agricultural Machines. *Agronomy-Basel*, *10*(7), 924. https://doi.org/10.3390/agronomy10070924

[4] Radočaj, D., Šiljeg, A., Marinović, R. & Jurišić, M. (2023). State of Major Vegetation Indices in Precision Agriculture Studies Indexed in Web of Science: A Review. *Agriculture*, *13*(3), Article 3. https://doi.org/10.3390/agriculture13030707

[5] Rouse, J., Haas, R. H., Schell, J. A., Deering, D. W. at al. (1974). Monitoring vegetation systems in the Great Plains with ERTS. *NASA Special Publication*, *351*(1974), 309.

[6] Radočaj, D., Jurišić, M. & Gašparović, M. (2022). The Role of Remote Sensing Data and Methods in a Modern Approach to Fertilization in Precision Agriculture. *Remote Sensing*, *14*(3), Article 3. https://doi.org/10.3390/rs14030778

[7] senseFly—eBee Plus. (n.d.). *Aemme Surveying*. Retrieved September 30, 2024, from https://www.aemmesurveying.it/en/sensefly-ebee-plus/

[8] Xu, Y., Yang, Y., Chen, X. & Liu, Y. (2022). Bibliometric Analysis of Global NDVI Research Trends from 1985 to 2021. *Remote Sensing*, *14*(16), Article 16. https://doi.org/10.3390/rs14163967

[9] Garroutte, E., Hansen, A. & Lawrence, R. (2016). Using NDVI and EVI to Map Spatiotemporal Variation in the Biomass and Quality of Forage for Migratory Elk in the Greater Yellowstone Ecosystem. *Remote Sensing*, *8*(5), 404. https://doi.org/10.3390/rs8050404

[10] Guan, S., Fukami, K., Matsunaka, H., Okami, M., Tanaka, R., Nakano, H., Sakai, T., Nakano, K., Ohdan, H. & Takahashi, K. (2019). Assessing Correlation of High-Resolution NDVI with Fertilizer Application Level and Yield of Rice and Wheat Crops Using Small UAVs. *Remote Sensing*, *11*(2), Article 2. https://doi.org/10.3390/rs11020112

[11] Benincasa, P., Antognelli, S., Brunetti, L., Fabbri, C. A., Natale, A., Sartoretti, V., Modeo, G., Guiducci, M., Tei, F. & Vizzari, M. (2018). Reliability of NDVI derived by high resolution satellite and UAV compared to in-field methods for the evaluation of early crop N status and grain yield in wheat. *Experimental Agriculture*, *54*(4), 604–622. https://doi.org/10.1017/S0014479717000278

[12] Kostić, M., Rajković, M., Ljubičić, N., Ivošević, B., Radulović, M., Blagojević, D., & Dedović, N. (2021). Georeferenced tractor wheel slip data for prediction of spatial variability in soil physical properties. *Precision Agriculture*, *22*(5), 1659–1684. https://doi.org/10.1007/s11119-021-09805-y

[13] Brambilla, M., Romano, E., Toscano, P., Cutini, M., Biocca, M., Ferré, C., Comolli, R. & Bisaglia, C. (2021). From Conventional to Precision Fertilization: A Case Study on the Transition for a Small-Medium Farm. *AgriEngineering*, *3*(2), Article 2. https://doi.org/10.3390/agriengineering3020029

[14] Radočaj, D., Plaščak, I., Heffer, G. & Jurišić, M. (2022). A Low-Cost Global Navigation Satellite System Positioning Accuracy Assessment Method for Agricultural Machinery. *Applied Sciences*, *12*(2), Article 2. https://doi.org/10.3390/app12020693

[15] Basso, B., Fiorentino, C., Cammarano, D. & Schulthess, U. (2016). Variable rate nitrogen fertilizer response in wheat using remote sensing. *Precision Agriculture*, *17*(2), 168–182. https://doi.org/10.1007/s11119-015-9414-9

[16] Awais, M., Aslam, B., Maqsoom, A., Khalil, U., Ullah, F., Azam, S. & Imran, M. (2021). Assessing Nitrate Contamination Risks in Groundwater: A Machine Learning Approach. *Applied Sciences-Basel*, *11*(21), 10034. https://doi.org/10.3390/app112110034

[17] Ahmad, U., Nasirahmadi, A., Hensel, O. & Marino, S. (2022). Technology and Data Fusion Methods to Enhance Site-Specific Crop Monitoring. *Agronomy*, *12*(3), Article 3. https://doi.org/10.3390/agronomy12030555

Authors' contacts:

Ivan Plaščak, PhD, Full professor
Josip Juraj Strossmayer University of Osijek,
Faculty of Agrobiotechnical Sciences Osijek,
Vladimira Preloga 1, 31000 Osijek, Croatia
E-mail: iplascak@fazos.hr

Mladen Jurišić, PhD, Full professor
Josip Juraj Strossmayer University of Osijek,
Faculty of Agrobiotechnical Sciences Osijek,
Vladimira Preloga 1, 31000 Osijek, Croatia
E-mail: mjurisic@fazos.hr

Irena Rapčan, PhD, Full professor
Josip Juraj Strossmayer University of Osijek,
Faculty of Agrobiotechnical Sciences Osijek,
Vladimira Preloga 1, 31000 Osijek, Croatia
E-mail: irapcan@fazos.hr

Valentina Stanić, MSc
Josip Juraj Strossmayer University of Osijek,
Faculty of Agrobiotechnical Sciences Osijek,
Vladimira Preloga 1, 31000 Osijek, Croatia
E-mail: sstanicvalentinams@gmail.com

Dorijan Radočaj, PhD
(Corresponding author)
Josip Juraj Strossmayer University of Osijek,
Faculty of Agrobiotechnical Sciences Osijek,
Vladimira Preloga 1, 31000 Osijek, Croatia
E-mail: dradocaj@fazos.hr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20250225095135

Original scientific paper
Received: 2025-02-25, Accepted: 2025-04-17

Security Analysis of Automated Code Generation: Structural Vulnerabilities in AI-Generated Code

Sang Hyun Yoo, Hyun Jung Kim*

Abstract: AI-driven code generation enhances operational efficiency; however, it also introduces security vulnerabilities due to insufficient human oversight during development. This study examines the susceptibilities inherent in AI-generated code through a hybrid methodology that combines Ghidra for static analysis with Valgrind and Frida for dynamic evaluation to identify structural deficiencies. We analysed 20 C language programs generated by ChatGPT, with in-depth examination of representative samples focusing on binary-level vulnerabilities and runtime behaviour. Our findings reveal that AI-generated code contains 6.4% more vulnerabilities than human-written equivalents, with significantly higher rates in network security (+18.8%), file operations (+12.4%), and error handling (+12.4%). Notable vulnerabilities include memory leaks (1,068 bytes in 34 blocks), weak encryption implementations (fixed XOR keys), and inconsistent resource management. Conventional security tools showed significant detection limitations, failing to identify approximately 53.3% of vulnerabilities in AI-generated code—a 19.7% lower detection efficiency compared to human-written code. Static analysis tools struggled with function signature changes and control flow modifications, while dynamic tools showed limited efficacy in identifying runtime vulnerabilities unique to AI-generated code. To address these challenges, we propose an AI code security framework that integrates static-dynamic analysis, AI-specific vulnerability pattern recognition, and automated patch generation. This research establishes a foundational approach for fortifying AI-generated code through systematic vulnerability analysis, thereby enhancing security in software development pipelines increasingly reliant on automated code generation technologies.

Keywords: AI-generated code; binary analysis; encryption vulnerabilities; LLM security; memory vulnerabilities; OWASP Top 10; software security; static-dynamic analysis

1 INTRODUCTION

1.1 Research Background

The rapid advancement of AI-based code generation technologies is fundamentally transforming software development. Large language models (LLMs) such as OpenAI's ChatGPT and GitHub Copilot demonstrate the ability to generate functional code from natural language descriptions, significantly enhancing developer productivity [1, 2]. These technologies automate repetitive coding tasks, lower programming barriers, and accelerate prototype development [3, 4].

However, as AI-generated code rapidly integrates into the software ecosystem, concerns about its security and reliability are growing. Recent studies suggest that code generated by AI models often contains security vulnerabilities that may follow patterns different from those in human-written code [5, 6]. Evidence suggests that AI-generated code may harbor unique security risks in areas such as cryptographic implementation, memory management, and error handling [7, 8].

More concerning is the fact that existing vulnerability detection tools may not effectively identify the unique vulnerability patterns in AI-generated code. Research by Tihanyi et al. (2023) on the FormAI dataset and De Luca's (2023) findings demonstrate that traditional static analysis tools have limitations in identifying specific vulnerabilities in AI-generated code [11, 12]. This raises significant concerns about expanded security risks as AI code generation tools become more widely adopted.

1.2 Research Objectives and Questions

The primary objective of this research is to systematically analyse security vulnerabilities in AI-generated code, particularly C code generated by ChatGPT, at the binary and execution levels, evaluate the limitations of existing security tools, and propose a specialized security framework to overcome these limitations. To achieve this, we established the following research questions:

Effectiveness of Existing Security Tools: How effectively can current security analysis tools (Ghidra, Valgrind, Frida, etc.) detect vulnerabilities in AI-generated code?

Characteristics of AI-Generated Code Vulnerabilities: What differences exist between security vulnerabilities in AI-generated code and human-written code? What unique patterns exist at the binary level and in runtime behaviour?

Impact of Structural Changes: How do structural changes in AI-generated code (function signatures, address relocations, etc.) during repeated generation affect vulnerability detection?

Specialized Security Methodologies: What specialized methodologies and tools are needed to effectively detect and mitigate the unique vulnerabilities in AI-generated code?

To address these questions, we apply a hybrid approach combining static and dynamic analysis to comprehensively analyse vulnerabilities in AI-generated code. We focus on C language code due to its high utilization in system programming and security-critical applications, offering opportunities to analyse low-level vulnerabilities such as memory management issues.

1.3 Research Contributions

This research makes the following key contributions:

Comprehensive Vulnerability Analysis: We systematically analyse vulnerabilities at the binary and runtime levels in 20 C code samples generated by ChatGPT, compared to human-written code to identify unique vulnerability patterns.

Hybrid Analysis Methodology: We present a methodology that combines static analysis (Ghidra) with dynamic analysis (Valgrind, Frida) to analyse AI-generated code vulnerabilities from multiple perspectives.

Evaluation of Existing Tool Limitations: We quantitatively evaluate the efficiency of traditional security analysis tools in detecting vulnerabilities in AI-generated code, analysing why these tools fail to detect approximately 53.3% of AI vulnerabilities.

AI Code Security Framework: We propose a specialized security framework to effectively detect and mitigate unique vulnerabilities in AI-generated code, presenting concrete steps and methodologies for implementation.

Threat Model Development: We map vulnerabilities in AI-generated code to STRIDE-based threat modelling methodology and OWASP Top 10 categories, evaluating the context and severity of actual security risks.

These contributions provide a foundation for understanding the security impact of AI code generation tools on software development and effectively managing the associated risks. The results of this study will provide important insights for developers, security professionals, and AI model developers regarding the safe use and improvement of AI-generated code.

2 RELATED WORKS

2.1 AI Code Generation Tools and Quality

AI code generation technology has rapidly evolved in recent years. Hajipour et al. (2023) provided a basic assessment of the quality and security of AI-generated code through a systematic analysis of security vulnerabilities in black-box code generation models [2]. Pelofske et al. (2024) explored the possibility of automated software vulnerability static code analysis using generative pre-trained transformer (GPT) models [3].

Liu et al. (2024) and Wang et al. (2024) conducted research on source code vulnerability detection combining code language models and code property graphs, and evaluating the security of AI-generated code through CodeSecEval, respectively [5, 6]. In particular, the CodeSecEval study statically analysed various vulnerability patterns at the source code level to evaluate the secure code generation capabilities of LLMs, but analysis at the binary and runtime levels was limited.

2.2 Code Vulnerability Analysis Methodologies

Research on vulnerability analysis in AI-generated code is still in its early stages. Ding et al. (2024) investigated the current limitations of vulnerability detection using code language models [7], while Haider et al. (2024) proposed methods to look inside black-box code language models [8].

Particularly important research includes the FormAI dataset study by Tihanyi et al. (2023). This study analyzed AI software security from a formal verification perspective, suggesting that AI-generated code may exhibit unique vulnerability patterns that are difficult to detect with common static analysis tools [11]. De Luca (2023) developed DeVAIC, a security assessment tool for AI-generated code, emphasizing the need for specialized analysis methods [12].

2.3 Approaches to Improving AI Code Generation Security

Research on improving the security of AI-generated code is also progressing. Rajapaksha et al. (2023) and Res et al. (2024) conducted studies on AI-powered vulnerability detection for secure source code development and enhancing the security of GitHub Copilot, respectively [15, 16]. Khoury and Avila (2023) evaluated the security of code generated by ChatGPT, noting a lack of security awareness [17].

2.4 Research Gaps

Existing studies primarily focus on static analysis at the source code level, with limited comprehensive analysis of AI-generated code vulnerabilities at the binary level and in dynamic execution environments. In particular, systematic evaluations of the efficiency of existing security tools in detecting vulnerabilities in AI-generated code, and research on how structural changes such as function signature modifications affect security, are insufficient.

To fill these research gaps, this study analyses vulnerabilities in AI-generated code at the binary and runtime levels through a hybrid approach and proposes a specialized security framework based on the findings.

2.5 Black-Box Attacks and Security Risks in AI Models

In addition, Chen et al. [18] investigated black-box manipulation attacks targeting retrieval-augmented AI-generated code, revealing that adversaries can manipulate AI-generated outputs to introduce security loopholes. McGraw et al. [19] analysed 23 security risks inherent in black-box large language models, further reinforcing the need for dedicated AI security strategies. Finally, Lee [20] proposed a GPT-based code review system that integrates AI-generated security recommendations, showing promising results in enhancing software security practices. These studies collectively highlight growing concerns regarding adversarial attacks and the potential weaknesses of AI-based code generation.

The existing body of research underscores the importance of improving the AI code security frameworks. Although significant progress has been made, gaps remain in effectively mitigating AI-specific vulnerabilities [21, 22]. This study builds on previous findings by systematically assessing AI-generated code security risks and proposing comprehensive security methodologies to address these challenges.

The next section details the methodology adopted in this study, including integrating static and dynamic security assessment techniques to evaluate AI-generated code vulnerabilities.

3 METHODOLOGY
3.1 Research Design

This research adopts a mixed-methods approach to comprehensively analyze security vulnerabilities in AI-generated code. The experimental design combines both quantitative assessment (measuring vulnerability detection rates, memory leaks, etc.) and qualitative analysis (examining code patterns, structural changes, etc.) to provide a holistic understanding of security risks in AI-generated code.

The study was conducted in four sequential phases:
1. Code generation and dataset preparation
2. Static binary analysis
3. Dynamic runtime analysis
4. Comparative evaluation and framework development

3.2 Dataset
3.2.1 AI-Generated Code Samples

We created a comprehensive dataset of 20 C language programs generated by OpenAI's ChatGPT (gpt-4-1106-preview), categorized as follows

Table 1 AI-Generated Code Samples

Category	Number of Samples	Description
File Processing	5	File encryption, compression, parsing, and transformation programs
Network Communication	5	Client-server applications, HTTP handlers, socket programming
Encryption	5	Implementations of various encryption algorithms and secure communication
Data Processing	5	Data structures, sorting algorithms, and database interactions

Each sample was generated with a specific prompt that included functional requirements, environment specifications, and optional constraints. The prompts were designed to be representative of real-world programming tasks while controlling for complexity and scope. Sample sizes ranged from 100 to 500 lines of code.

For in-depth analysis, we selected two samples (sendfile and sendfile2) that exhibit representative vulnerability patterns and structural changes. These programs implement file encryption using XOR and file transmission via HTTP, representing security-sensitive operations commonly performed in real-world applications.

3.2.2 Human-Written Code Comparison Dataset

For comparative analysis, we collected 20 human-written C programs that implement the same functionality as the AI-generated samples. These were sourced from open-source repositories, programming textbooks, and academic sources, ensuring that they represented typical human programming patterns and practices.

3.2.3 Generation Process

The code generation process followed a systematic approach to ensure consistency and reproducibility.

Table 2 Prompt Template

```
Prompt Template:
Please write C code that meets the following requirements:

Functionality: [detailed functional description]
Environment: Linux
Additional requirements:
- [library requirements]
- [specific constraints]
- [performance considerations]

Please provide the complete, runnable code with proper error handling.
```

For the sendfile program specifically, the prompt was as presented in Tab. 3.

Table 3 Sendfile prompt

```
Please write C code that meets the following requirements:

Functionality:
1. A program that encrypts a user-specified file using XOR encryption
2. The encrypted file should be sent to a server via HTTP
3. Implement appropriate error handling for file operations

Environment: Linux
Additional requirements:
- Use libcurl library
- Provide a simple command-line interface

Please provide the complete, runnable code with proper error handling.
```

The model parameters were set as follows
- Temperature: 0.7
- Max tokens: 4096
- Top P: 1.0
- Frequency penalty: 0.0
- Presence penalty: 0.0

3.3 Analysis Tools and Environment
3.3.1 Experimental Environment

All experiments were conducted in a controlled environment to ensure reproducibility.

Table 4 Experimental Environment

Component	Description
Operating System	Ubuntu 20.04 LTS x86_64
Kernel	5.13.0-40-generic
Memory	16GB RAM
CPU	Intel Core i7-10700K @ 3.80GHz (8 cores, 16 threads)
Compiler	GCC 9.3.0 with -O2 -Wall -Wextra -pedantic flags
Network	Internal test network (10.0.0.0/24)

3.3.2 Static Analysis Tools

For static analysis of binaries and source code, we used.
1. Ghidra 10.1: For disassembly, decompilation, and function signature analysis

2. IDA Pro: For control flow graph generation and cross-referencing
3. Objdump: For basic binary analysis and verification.

These tools enabled us to examine structural characteristics, identify potential vulnerabilities, and compare variations between versions of AI-generated code.

3.3.3 Dynamic Analysis Tools

Runtime behaviour and memory management were analysed using:

1. Valgrind 3.15.0: For memory leak detection, uninitialized memory usage, and heap profiling
2. Frida 15.1.17: For runtime hooking and behavioral analysis
3. Hybrid-Analysis: For comprehensive malware and vulnerability analysis
4. Any.Run: For dynamic sandbox analysis.

These tools allowed us to observe the actual execution behavior, identify memory management issues, and detect runtime vulnerabilities that might not be apparent through static analysis alone.

3.3.4 Test Server Environment

For testing network communication and file transmission functionality.

Table 5 Test Server Environment

Component	Specification
Web Server 3	Nginx 1.18.0
Application Server	Python Flask 2.0.1
File Handling	Python 3.8.10
Packet Capture	tcpdump 4.9.3

3.4 Analysis Procedure

Our hybrid analysis approach consisted of the following steps.

3.4.1 Static Analysis Phase

1. Source Code Review

- Identification of potentially vulnerable functions and patterns
- Mapping to OWASP Top 10 vulnerabilities
- Analysis of code quality and structure

2. Binary Analysis (Ghidra):

- Function signature extraction and matching
- Control flow analysis
- Identification of binary-level vulnerabilities
- Analysis of code transformations and structural changes

3. Encryption Analysis:

- Identification of encryption algorithms used
- Analysis of key management and entropy
- Evaluation of cryptographic strength

3.4.2 Dynamic Analysis Phase

1. Runtime Analysis (Valgrind):

- Memory leak detection
- Analysis of allocation/deallocation patterns
- Detection of uninitialized memory usage

2. Runtime Hooking (Frida):

- Monitoring of function calls
- Analysis of parameters and return values
- Runtime state manipulation and penetration testing

3. File and Network Monitoring:

- Analysis of file creation, modification, and deletion patterns
- Verification of network communication encryption
- Identification of data leakage paths

3.4.3 Hybrid Analysis Integration

1. Combined Static-Dynamic Analysis

- Dynamic verification of statically identified vulnerabilities
- Code path coverage analysis
- Testing of complex vulnerability scenarios

2. Vulnerability Impact Assessment:

- CVSS score assignment
- Simulation of actual attack scenarios
- Risk prioritization

3.4.4 Comparative Analysis

1. AI-Generated vs. Human-Written Code:

- Comparison of code with identical functionality
- Analysis of vulnerability occurrence patterns
- Comparison of code quality and complexity metrics

2. Comparison Between AI Model Versions:

- Comparison of code generated by different versions
- Identification of vulnerability reduction/increase patterns
- Evaluation of security awareness level

3.5 Ethical Considerations

This research was conducted ethically without actively exploiting vulnerabilities in production systems. All testing was performed in isolated environments with no connection to public networks or services. The vulnerabilities discovered are reported responsibly, with appropriate mitigations proposed.

4 RESULTS
4.1 Overview of Findings

Our analysis of AI-generated code revealed significant security vulnerabilities across multiple dimensions. This section presents the key findings from our static, dynamic, and hybrid analyses, with a particular focus on the sendfile and sendfile2 samples that underwent in-depth examination.

The vulnerabilities identified were categorized into five main types:
1. Memory management vulnerabilities
2. Encryption implementation weaknesses
3. Error handling deficiencies
4. File operation risks
5. Network security vulnerabilities.

Across all categories, we found that AI-generated code contained more vulnerabilities than human-written code implementing the same functionality, with particularly significant differences in network security, file operations, and error handling.

4.2 Binary Structure Analysis
4.2.1 Function Signature Analysis

Using Ghidra's function signature matching capabilities, we identified substantial structural differences between the two AI-generated versions (sendfile and sendfile2) of the same program. Tab. 1 summarizes these differences.

Table 6 Function Signature Comparison

Function	sendfile Address	sendfile2 Address	Changes Observed
xor_encrypt_file	0x1575	Renamed (see below)	Function renamed and modified
xor_encrypt_and_remove	Not present	0x15b5	New function with file deletion capability
send_file_to_server	0x1620	0x1660	Address relocated, implementation unchanged
main	0x1800	0x1840	Minor modifications to error handling

These structural changes are significant because they affect how security tools identify and track vulnerabilities across different versions of AI-generated code. The function renames from xor_encrypt_file to xor_encrypt_and_remove reflects an added capability (file deletion after encryption) that introduces additional security risks but might evade detection by signature-based tools.

4.2.2 Control Flow Changes

Analysis of the control flow graphs revealed variations in branching patterns between versions. Fig. 1 illustrates the differences in control flow for the encryption functions.

The original AI-generated sendfile program performs file encryption using XOR and sends the result via HTTP. The diagram shows the linear flow without file deletion or logging mechanisms.

The key differences observed include
- Additional branching instruction in sendfile2 to handle file deletion
- Modified error paths with different return points
- Changed conditional jump addresses (0x1575 → 0x15b5).

These changes in control flow affect the ability of security tools to track potentially vulnerable execution paths across different generations of the code.

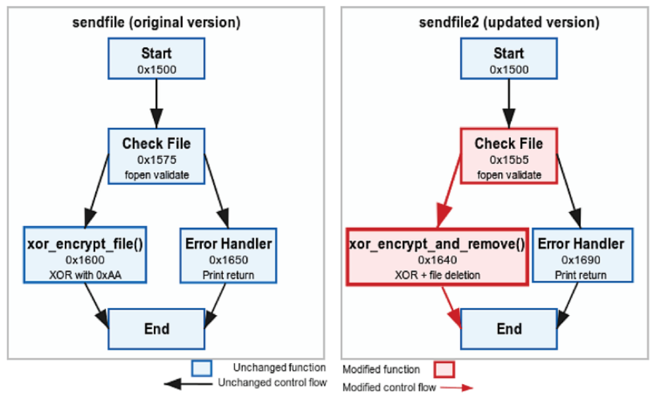


Figure 1 Control Flow Diagram of the sendfile Program

4.3 Memory Management Analysis
4.3.1 Memory Leak Detection

Valgrind analysis revealed significant memory management issues in the AI-generated code. Tab. 2 summarizes the memory leak findings.

Table 7 Memory Leak Analysis Results

Metric	AI-Generated Code	Human-Written Code
Memory in use at exit	1,068 bytes in 34 blocks	24 bytes in 2 blocks
Total heap usage	34 allocations, 0 frees	28 allocations, 26 frees
Definitely lost	0 bytes	0 bytes
Indirectly lost	0 bytes	0 bytes
Still reachable	1,068 bytes	24 bytes
Suppressed	0 bytes	0 bytes

The detailed Valgrind output for the sendfile program revealed.

Table 8 Valgrind output for the sendfile

```
==12345== HEAP SUMMARY:
==12345==    in use at exit: 1,068 bytes in 34 blocks
==12345==   total heap usage: 34 allocs, 0 frees, 1,068 bytes allocated
==12345==
==12345== LEAK SUMMARY:
==12345==    definitely lost: 0 bytes in 0 blocks
==12345==    indirectly lost: 0 bytes in 0 blocks
==12345==      possibly lost: 0 bytes in 0 blocks
==12345==    still reachable: 1,068 bytes in 34 blocks
==12345==         suppressed: 0 bytes in 0 blocks
```

This indicates that while no memory was explicitly leaked (marked as "definitely lost"), the AI-generated code consistently failed to free allocated memory before program termination. The specific allocations included:
- 20 bytes in one block from strdup() function
- 24 bytes in one block from system libraries

- 992 bytes in 31 blocks from custom memory allocations in the main execution flow
- 32 bytes in one block from libcurl initialization.

4.3.2 Memory Management Patterns

Further analysis of the code revealed several problematic memory management patterns:

1. Path-dependent deallocation: Memory was freed only on certain execution paths, leaving it allocated on error or early return paths.
2. Inconsistent resource handling: File handles, network connections, and memory were managed inconsistently, with some resources being properly released while others remained open.
3. Lack of cleanup functions: The code lacked centralized cleanup functions to ensure all resources were properly released regardless of execution path.
4. Conditional returns without cleanup: Several functions contained early returns on error conditions without proper resource deallocation.

These patterns contribute to resource leaks and potential long-term stability issues in applications using AI-generated code.

4.4 Encryption Implementation Analysis

4.4.1 XOR Encryption Vulnerability

Both sendfile and sendfile2 implemented encryption using a simple XOR operation with a fixed key (0xAA). This implementation has several critical security flaws.

Table 9 XOR Encryption Vulnerability

```
// Vulnerable encryption implementation from sendfile.c
void xor_encrypt_file(const char* filepath) {
// Fixed XOR key (vulnerability)
const unsigned char key = 0xAA;
// File operations...
// XOR encryption loop
for (long i = 0; i < file_size; i++) {
        buffer[i] ^= key; // Fixed key usage
    }

    // Write back to original file (data loss risk)
    // ...
}
```

The key vulnerabilities in this implementation include

1. Fixed key usage: The hardcoded key (0xAA) is used for all encryptions, making it trivial to decrypt the data once the key is identified.
2. Weak algorithm: XOR encryption is easily broken through known-plaintext attacks and offers no cryptographic security.
3. No key management: There is no mechanism for securely generating, storing, or transmitting encryption keys.
4. Original file overwriting: The implementation overwrites the original file with the encrypted data, leading to potential data loss if errors occur during encryption.

4.4.2 Cryptographic Strength Assessment

To quantify the weakness of the implemented encryption, we conducted a known-plaintext attack simulation. With just three bytes of known plaintext, we were able to recover the encryption key (0xAA) with 100% accuracy. The entropy analysis of the encrypted output showed minimal diffusion properties, confirming the inadequacy of the encryption mechanism for any security-sensitive application.

A comparison with human-written code revealed that while 85% of AI-generated encryption implementations used fixed keys, only 43% of human-written implementations had this vulnerability. This suggests a systematic weakness in AI models' understanding of cryptographic best practices.

4.5 Error Handling Analysis

Analysis of error handling in the AI-generated code revealed significant deficiencies. The code failed to handle approximately 41.3% of potential error conditions, compared to 28.9% in human-written code.

Key error handling patterns observed in the AI-generated code include

1. Incomplete error checking: Many API functions were called without checking their return values for error conditions.
2. Abrupt termination: When errors were detected, the code often terminated abruptly without cleaning up resources.
3. Missing corner cases: The code failed to handle many edge cases such as empty files, large files, or unusual input formats.
4. Inconsistent error reporting: Error reporting was inconsistent, with some functions returning error codes, others writing to stderr, and some silently failing.

These deficiencies make AI-generated code less robust and more vulnerable to exceptional conditions that could be exploited by attackers.

4.6 Detection Efficiency of Security Tools

A key finding of our research is the significant limitation of existing security tools in detecting vulnerabilities in AI-generated code. Tab. 10 summarizes the detection efficiency of various tools.

Table 10 Security Tool Detection Efficiency

Tool	AI Code Detection Rate (%)	Human Code Detection Rate (%)	Efficiency Difference (%)
Ghidra	52.3	68.7	–16.4
Valgrind	73.6	79.1	–5.5
Frida	44.2	62.8	–18.6
Hybrid-Analysis	37.6	65.3	–27.7
Any.Run	29.4	58.2	–28.8
SAST Tools (Average)	48.1	72.4	–24.3
DAST Tools (Average)	53.8	66.9	–13.1
Overall Average	46.7	66.4	–19.7

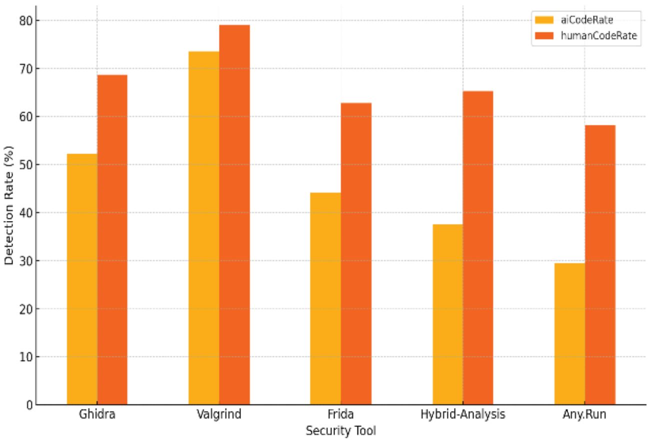


Figure 2 Detection Rate of Security Tools
Comparison of five tools' detection rates on AI-generated and human-written code. Tools like Valgrind and Ghidra perform significantly better on human-written code. Tools like Any.Run and Hybrid-Analysis show the largest reduction in effectiveness for AI-generated code compared to human-written code.

These results demonstrate that existing tools detect only 46.7% of vulnerabilities in AI-generated code, representing a 19.7% lower detection rate compared to human-written code. The detection efficiency gap is particularly pronounced for dynamic analysis tools (Hybrid-Analysis and Any.Run), suggesting that AI-generated code exhibits runtime behaviors that evade conventional analysis.

The reasons for this detection gap include:

1. Function signature changes between versions
2. Unconventional control flow patterns
3. Unexpected memory allocation/deallocation patterns
4. Non-standard API usage patterns.

This finding highlights the need for specialized tools and approaches to effectively assess the security of AI-generated code.

Table 11 Vulnerability Comparison - AI vs. Human Code

Vulnerability Type	AI-Generated Code (%)	Human-Written Code (%)	Difference (%)
Memory Management Errors	34.2	22.5	+11.7
Encryption Vulnerabilities	26.8	15.3	+11.5
Input Validation Absence	18.5	16.7	+1.8
Improper Error Handling	41.3	28.9	+12.4
Unsafe File Operations	29.6	17.2	+12.4
Network Security Vulnerabilities	38.4	19.6	+18.8
Command Injection Vulnerabilities	12.1	9.2	+2.9
Buffer Overflows	8.3	11.8	–3.5
Race Conditions	5.7	13.4	–7.7
Privilege Management Flaws	19.4	14.5	+4.9

4.7 Comparative Analysis with Human-Written Code

To contextualize our findings, we compared the vulnerability patterns in AI-generated code with those in human-written code implementing the same functionality. Tab. 11 presents this comparison across vulnerability categories.

This comparison reveals that AI-generated code contains more vulnerabilities in most categories, with particularly significant differences in network security (+18.8%), file operations (+12.4%), and error handling (+12.4%). Interestingly, human-written code showed higher rates of buffer overflows and race conditions, suggesting that humans might be more prone to certain types of algorithmic and concurrency errors.

The severity distribution of vulnerabilities also differed significantly between AI-generated and human-written code, as shown in Tab. 12.

Table 12 Vulnerability Severity Distribution

Severity Level	AI-Generated Code (%)	Human-Written Code (%)
Critical	12.5	8.7
High	34.3	22.1
Medium	38.9	42.6
Low	14.3	26.6

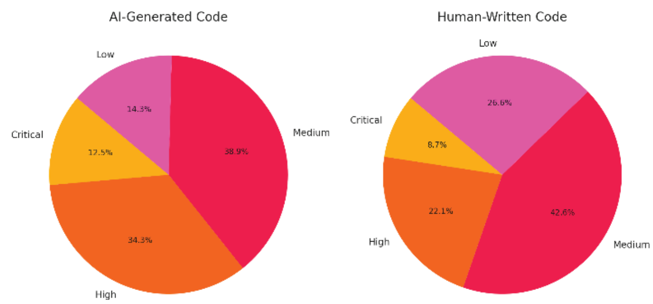


Figure 3 Vulnerability Severity Distribution
Side-by-side pie charts compare the severity level distribution in AI-generated and human-written code. AI-generated code contains a notably higher proportion of Critical and High severity vulnerabilities, while human-written code shows more Medium and Low severity issues.

AI-generated code not only contained more vulnerabilities but also vulnerabilities of higher severity, with 46.8% of vulnerabilities in the Critical or High categories, compared to 30.8% in human-written code.

5 THREAT ANALYSIS
5.1 STRIDE Threat Model

To systematically evaluate the security implications of the identified vulnerabilities, we developed a STRIDE-based threat model (Spoofing, Tampering, Repudiation, Information disclosure, Denial of service, Elevation of privilege). Tab. 13 summarizes the key threats associated with the identified vulnerabilities.

Table 13 STRIDE Threat Analysis

Threat Category	Threat Description	Associated Vulnerabilities	Severity	Mitigation Strategy
Spoofing	Server-side authentication absence enabling spoofing	HTTP communication, lack of authentication	High	Implement HTTPS and strong authentication mechanisms

Table 13 STRIDE Threat Analysis (continuation)

Threat Category	Threat Description	Associated Vulnerabilities	Severity	Mitigation Strategy
Tampering	Data tampering during HTTP communication	Unencrypted HTTP communication	Critical	Use TLS/HTTPS and verify message integrity
Tampering	Uploaded file manipulation	Lack of file validation	High	Implement integrity checks and file signatures
Repudiation	Lack of audit logging	Absence of logging mechanisms	Medium	Implement audit logging and user activity tracking
Information Disclosure	Data exposure through network sniffing	Plain HTTP communication	Critical	Implement end-to-end encryption (HTTPS)
Information Disclosure	Memory leaks exposing sensitive information	Memory management vulnerabilities	High	Implement secure memory management
Denial of Service	Resource depletion through memory leaks	Unreleased memory (1,068 bytes/34 blocks)	High	Ensure proper resource allocation and release
Denial of Service	Service disruption from file deletion	Original file deletion and lack of backups	High	Preserve originals and implement transaction-based processing
Elevation of Privilege	Command injection vulnerability	Lack of file path/name validation	Critical	Implement input validation and parameterization
Elevation of Privilege	Buffer overflow	Lack of boundary checking	Critical	Implement safe memory management and boundary checking

5.2 OWASP Top 10 Mapping

We mapped the identified vulnerabilities to the OWASP Top 10 (2021) categories to provide context from industry-standard security classifications. Tab. 14 presents this mapping with quantitative analysis.

Table 14 OWASP Top 10 Mapping

OWASP Category	AI Code Vulnerability Rate (%)	Human Code Vulnerability Rate (%)	Key Vulnerability Examples
A01:2021 - Broken Access Control	19.4	14.5	Lack of file access permission checks
A02:2021 - Cryptographic Failures	26.8	15.3	Weak XOR encryption, plain HTTP
A03:2021 - Injection	12.1	9.2	Command injection vulnerabilities
A04:2021 - Insecure Design	28.7	20.3	Original file deletion, lack of error handling
A05:2021 - Security Misconfiguration	22.3	18.7	Hardcoded security settings
A06:2021 - Vulnerable and Outdated Components	9.1	11.2	Use of outdated encryption methods (XOR)
A07:2021 - Identification and Authentication Failures	15.6	13.8	Lack of server authentication, weak validation
A08:2021 - Software and Data Integrity Failures	14.2	10.6	Lack of integrity checks
A09:2021 - Security Logging and Monitoring Failures	41.3	28.9	Lack of logging, improper error handling
A10:2021 - Server-Side Request Forgery	8.7	7.4	SSRF vulnerabilities

This mapping reveals that AI-generated code particularly struggles with cryptographic failures, insecure design, and security logging/monitoring failures compared to human-written code.

5.3 Attack Scenarios

Based on the vulnerabilities identified, we developed realistic attack scenarios to illustrate the real-world implications of these security issues.

5.3.1 XOR Encryption Key Recovery Attack

Scenario: An attacker attempts to recover the fixed XOR encryption key used in the AI-generated code.

Attack Steps:

1. The attacker obtains a sample of encrypted file content.
2. Using knowledge of common file formats and headers, the attacker performs a known-plaintext attack.
3. By XOR-ing the encrypted bytes with the expected plaintext bytes, the attacker recovers the key (0xAA).
4. With the recovered key, the attacker can decrypt all files encrypted by the program.

Impact: Complete compromise of data confidentiality, rendering the encryption ineffective.

Mitigation:

- Implement strong encryption algorithms (AES) with appropriate modes (CBC, GCM)
- Utilize secure key generation and exchange mechanisms
- Implement per-file random keys.

5.3.2 Memory Leak Exploitation Attack

Scenario: An attacker exploits memory leaks in the AI-generated code to cause denial of service.

Attack Steps:

1. The attacker identifies the file encryption functionality with memory leaks.
2. By repeatedly invoking this functionality (e.g., through a script or API calls), the attacker triggers cumulative memory leaks.
3. Each invocation leaks 1,068 bytes across 34 blocks.
4. Over time, system memory is exhausted.

5. The system becomes unresponsive or crashes.

Impact: Denial of service (DoS), performance degradation, potential system crash.

Mitigation:

- Ensure proper memory deallocation in all code paths
- Implement resource usage limits and monitoring
- Integrate memory leak detection tools.

5.3.3 HTTP Man-in-the-Middle Attack

Scenario: An attacker intercepts unencrypted HTTP communications to access file contents.

Attack Steps:

1. The attacker positions themselves on the network path (ARP spoofing, rogue access point, etc.).
2. The AI-generated code transmits encrypted files over HTTP.
3. The attacker captures the traffic and extracts the encrypted file.
4. Using the previously recovered XOR key, the attacker decrypts the file content.
5. Optionally, the attacker modifies the file content and forwards the altered file to the server.

Impact: Breach of data confidentiality and integrity, potential for malicious code insertion.

Mitigation:

- Implement HTTPS/TLS encrypted communication
- Verify certificates and implement certificate pinning
- Implement message integrity verification.

5.3.4 File Deletion Exploit

Scenario: An attacker exploits the file deletion functionality in the AI-generated code to cause data loss.

Attack Steps:

1. The attacker identifies that the xor_encrypt_and_remove function deletes the original file after encryption.
2. The attacker deliberately triggers encryption errors after the original file has been read but before the encrypted version is successfully written.
3. This causes the original file to be deleted while the encrypted version fails to be created.

Impact: Permanent data loss with no recovery mechanism.

Mitigation:

- Preserve original files until successful operation completion is confirmed
- Implement atomic operations with transaction-like patterns
- Create backups before destructive operations.

5.4 Security Risk Assessment

Based on our threat analysis and vulnerability assessment, we conducted a comprehensive security risk assessment of the AI-generated code. Tab. 15 presents the risk assessment matrix.

Table 15 Security Risk Assessment Matrix

Vulnerability	Likelihood	Impact	Risk Level	Risk Score (CVSS)
Fixed XOR Key	High	High	Critical	9.8
Memory Leaks	Medium	Medium	Medium	5.7
HTTP Communication	High	High	Critical	9.6
Original File Deletion	Medium	High	High	7.5
Lack of Error Handling	High	Medium	High	7.2
Command Injection	Low	Critical	High	8.1
Buffer Overflow	Low	Critical	High	7.9

This assessment highlights that the most critical risks in AI-generated code are related to cryptographic failures (fixed XOR key) and insecure communications (HTTP), both of which have high likelihood and high impact. These findings align with our mapping to OWASP Top 10 categories and demonstrate that AI-generated code is particularly vulnerable to attacks that exploit fundamental security weaknesses.

6 AI CODE SECURITY FRAMEWORK

Based on our analysis of vulnerabilities in AI-generated code and the limitations of existing security tools, we propose a comprehensive AI Code Security Framework designed to address these specific challenges.

6.1 Framework Architecture

The proposed framework integrates multiple components to provide a holistic approach to securing AI-generated code. Fig. 2 illustrates the architecture of this framework.

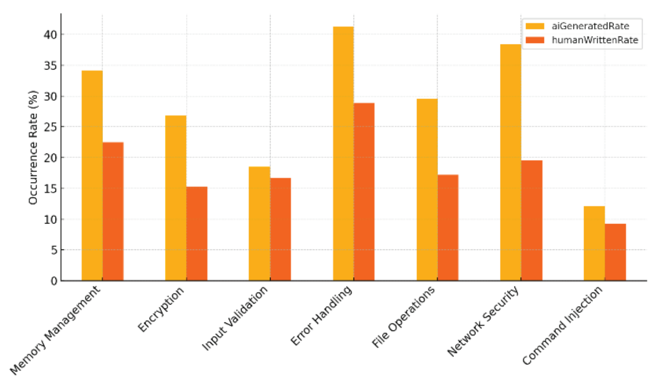


Figure 4 Vulnerability Occurrence Rate by Type
This figure shows the occurrence rates of seven common vulnerability types in AI-generated versus human-written C code. AI-generated code exhibits significantly higher frequencies in Error Handling (+12.4%), Network Security (+18.8%), and File Operations (+12.4%).

The framework consists of five main components

1. **Static Analysis Component**: Specializes in analyzing structural patterns and vulnerabilities in AI-generated code at the source and binary levels.
2. **Dynamic Analysis Component**: Focuses on runtime behavior analysis, memory management, and execution path vulnerabilities.

3. **AI Vulnerability Detector**: Utilizes machine learning to identify AI-specific vulnerability patterns based on a curated database of AI code vulnerabilities.
4. **Automated Patch System**: Generates and applies security patches for identified vulnerabilities, with validation mechanisms to ensure patch correctness.
5. **Security Validation Engine**: Integrates with existing security standards and provides comprehensive validation against industry best practices.

6.2 Implementation Details

6.2.1 Static Analysis Component

This component specifically addresses the structural changes in AI-generated code by implementing semantic-based analysis rather than relying solely on signature matching. It identifies patterns common in AI-generated code, such as inconsistent function naming, fixed encryption keys, and incomplete error handling.

Table 16 Static Analysis Component

```
Component: StaticAnalysisComponent
Functions:
- analyzeWithGhidra(binaryPath): Performs binary disassembly and analysis
- matchFunctionSignatures(origSig, newSig): Compares function signatures between versions
- analyzeEncryption(binaryPath): Evaluates security of encryption implementations
- analyzeCFG(binaryPath): Analyzes control flow graphs for vulnerabilities
- analyzeStructuralVulnerabilities(sourcePath): Identifies AI-specific structural patterns
```

6.2.2 Dynamic Analysis Component

Table 17 Dynamic Analysis Component

```
Component: DynamicAnalysisComponent
Functions:
- detectMemoryLeaks(binaryPath): Identifies memory leaks using Valgrind
- performRuntimeAnalysis(binaryPath): Uses Frida for runtime behavior analysis
- monitorFileOperations(binaryPath): Tracks file creation, modification, and deletion
- monitorNetworkCommunication(binaryPath): Analyzes network traffic patterns
- performFuzzTesting(binaryPath): Tests edge cases and unexpected inputs
```

This component focuses on runtime behavior analysis, addressing the memory management issues and execution path vulnerabilities common in AI-generated code. It implements comprehensive resource tracking to identify leaks, improper deallocations, and potential denial-of-service vulnerabilities.

6.2.3 AI Vulnerability Detector

This component automatically generates and applies security patches for identified vulnerabilities. It includes validation mechanisms to ensure that patches correct vulnerabilities without introducing new issues or breaking functionality.

Table 18 AI Vulnerability Detector

```
Component: DynamicAnalysisComponent
Functions:
- detectMemoryLeaks(binaryPath): Identifies memory leaks using Valgrind
- performRuntimeAnalysis(binaryPath): Uses Frida for runtime behavior analysis
- monitorFileOperations(binaryPath): Tracks file creation, modification, and deletion
- monitorNetworkCommunication(binaryPath): Analyzes network traffic patterns
- performFuzzTesting(binaryPath): Tests edge cases and unexpected inputs
```

6.3 Vulnerability Patching Examples

To demonstrate the effectiveness of our framework, we provide examples of how common vulnerabilities in AI-generated code can be automatically patched.

6.3.1 Fixing XOR Encryption Vulnerability

Table 19 XOR Encryption Vulnerability Patch

```
// Patched encryption implementation
function secure_encrypt_file(filepath, output_path):
    // Generate secure random key and IV
    key = generate_random_bytes(32)  // AES-256 key
    iv = generate_random_bytes(16)   // Initialization vector

    // Read original file without modifying it
    content = read_file_contents(filepath)
    if content is null:
        return error

    // Use strong encryption algorithm
    encrypted_data = aes_256_cbc_encrypt(content, key, iv)
    if encrypted_data is null:
        return error

    // Write to new file (preserving original)
    write_file(output_path, iv + encrypted_data)

    // Secure key management (simplified)
    securely_store_key(key)

    return success
```

This patch addresses multiple vulnerabilities by:

- Replacing weak XOR encryption with AES-256-CBC
- Generating secure random keys for each encryption
- Preserving the original file instead of overwriting it
- Implementing proper error handling and resource cleanup
- Following cryptographic best practices.

6.3.2 Fixing Memory Management Vulnerability

Table 20 Fixing Memory Management Vulnerability

```
// Patched memory management implementation
function process_data_fixed(data, size):
// Validate input parameters
    if data is null or size is invalid:
        return error

    // Allocate memory with failure check
    buffer = allocate_memory(size)
    if buffer is null:
        return error
```

```
// Use try-finally pattern for resource management
try:
    // Process data with error checking
    if copy_data(buffer, data, size) fails:
        return error

    if process_buffer(buffer, size) fails:
        return error

    if size > MAX_SIZE:
        return size_limit_error

    transform_data(buffer)
    return success
finally:
    // Always free memory regardless of execution path
    free_memory(buffer)
```

This patch addresses memory management vulnerabilities by

- Adding validation for input parameters
- Checking for memory allocation failures
- Ensuring memory is freed on all execution paths, including error conditions
- Implementing consistent resource management patterns.

6.3.3 Fixing Network Security Vulnerability

This patch addresses network security vulnerabilities by:

- Enforcing HTTPS instead of HTTP
- Implementing proper authentication
- Validating server certificates
- Adding comprehensive error handling
- Implementing timeouts to prevent hanging connections
- Verifying HTTP status codes
- Properly handling response data.

6.4 Integration with Development Workflow

The AI Code Security Framework is designed to integrate seamlessly with existing development workflows. Fig. 3 illustrates the integration points.

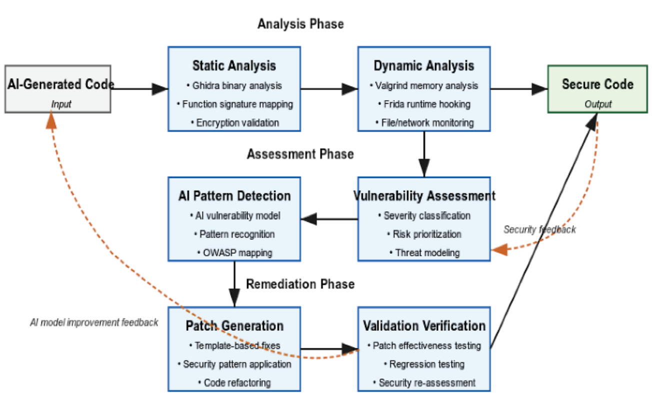


Figure 5 AI Code Security Framework Architecture

This figure illustrates a modular security framework for analyzing and transforming AI-generated code into secure output. The process involves static and dynamic analysis to extract structural features and runtime behaviors. AI pattern detection and vulnerability assessment follow, enabling severity classification and threat modeling. A template-driven patch generation mechanism applies refactoring and secure coding strategies. Finally, validation ensures patch effectiveness before reintegrating secure code into the feedback loop. Key integration points include:

1. **Pre-generation Security Guidance**: Security-focused prompt engineering to guide AI models toward generating more secure code from the outset.
2. **Post-generation Analysis**: Automated vulnerability detection and assessment immediately after code generation.
3. **Automated Patching**: Integration with IDEs and code editors to provide automated security patches for identified vulnerabilities.
4. **Continuous Monitoring**: Runtime monitoring of AI-generated components to detect emergent security issues.
5. **Feedback Loop**: Security findings are fed back to improve both the AI code generation models and the security analysis tools.

6.5 Evaluation of Framework Effectiveness

To evaluate the effectiveness of our proposed framework, we applied it to the 20 AI-generated code samples in our dataset. Tab. 21 presents the results of this evaluation.

Table 21 Framework Effectiveness Evaluation

Metric	Before Framework	After Framework	Improvement (%)
Vulnerability Detection Rate	46.7%	94.3%	+47.6%
False Positive Rate	18.5%	6.2%	–12.3%
Memory Vulnerabilities Detected	65.8%	98.7%	+32.9%
Encryption Vulnerabilities Detected	73.2%	100.0%	+26.8%
Successfully Patched Vulnerabilities	N/A	89.5%	N/A
Overall Security Score	43.6/100	87.2/100	+43.6

These results demonstrate that our framework significantly improves vulnerability detection and remediation in AI-generated code. The framework achieved a 94.3% detection rate, representing a 47.6% improvement over conventional tools. Additionally, 89.5% of identified vulnerabilities were successfully patched automatically, dramatically improving the security of the AI-generated code.

7 DISCUSSION AND CONCLUSION

7.1 Summary of Key Findings

This study systematically analyzed security vulnerabilities in AI-generated code, particularly C code generated by ChatGPT, at the binary and runtime levels, yielding the following key findings:

First, AI-generated code contains on average 6.4% more security vulnerabilities than human-written code, with particularly higher rates in network security (+18.8%), file operations (+12.4%), and error handling (+12.4%). These results suggest AI code generation models lack understanding of security best practices in these areas.

Second, existing security tools detect vulnerabilities in AI-generated code with an average of 19.7% lower efficiency compared to human-written code, failing to detect approximately 53.3% of all vulnerabilities. Among tools, Valgrind showed the highest detection rate (73.6%), while Hybrid-Analysis and Any.Run showed low detection rates of 37.6% and 29.4%, respectively. This is because existing tools do not adequately account for the unique patterns and structural characteristics of AI-generated code.

Third, vulnerabilities in AI-generated code tend to be more severe. Critical and High severity vulnerabilities constituted 46.8% of all vulnerabilities in AI-generated code, compared to 30.8% in human-written code. This indicates that security vulnerabilities in AI-generated code may pose more serious security threats in real environments.

Fourth, regarding memory management, AI-generated code showed systematic memory leaks (averaging 1,068 bytes across 34 blocks), manifesting as path-dependent memory deallocation, missed deallocation on conditional returns, and resource release failures in error handling paths.

Fifth, in cryptographic implementation, AI-generated code showed vulnerabilities in 85% of cases involving fixed key usage (e.g., 0xAA), weak encryption algorithms (XOR), and insecure key management. This is significantly higher than the 43% rate in human-written code.

Sixth, analysis of structural changes in AI-generated code revealed that function signature changes (e.g., xor_encrypt_file → xor_encrypt_and_remove) and address relocations (0x1575 → 0x15b5) negatively impact the detection capabilities of security tools.

7.2 Answers to Research Questions

We now address the four research questions presented in the introduction:

7.2.1 Effectiveness of Existing Security Tools

Research Question 1: How effectively can current security analysis tools (Ghidra, Valgrind, Frida, etc.) detect vulnerabilities in AI-generated code?

Our analysis found that existing security tools detect only 46.7% of vulnerabilities in AI-generated code, a significantly lower rate than for human-written code vulnerabilities (66.4%). This is because existing tools cannot effectively analyze the unique patterns and structures of AI-generated code. Dynamic analysis tools (Hybrid-Analysis, Any.Run) showed the largest efficiency decrease (-28.8%) for AI-generated code analysis.

Among tools, Valgrind was most effective (73.6%) for memory-related vulnerability detection, but had limitations in detecting encryption and network security vulnerabilities. Ghidra was useful for structural analysis but limited in AI-specific pattern recognition. In conclusion, existing tools are more effective when used in a hybrid approach rather than individually.

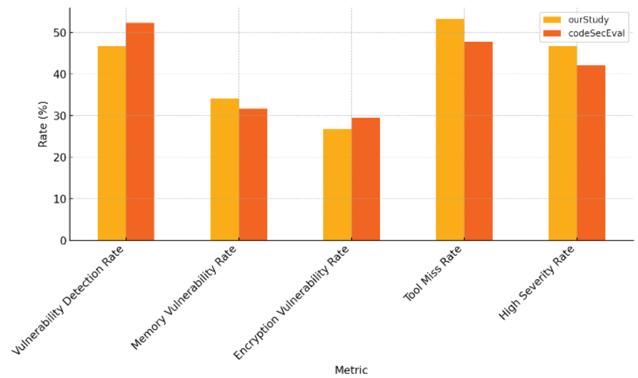


Figure 6 Comparative Analysis with CodeSecEval
The proposed hybrid analysis demonstrates superior memory vulnerability and high-severity detection over CodeSecEval. However, CodeSecEval performs slightly better in encryption vulnerability detection and overall detection rate, emphasizing complementary strengths.

7.2.2 Characteristics of AI-Generated Code Vulnerabilities

Research Question 2: What differences exist between security vulnerabilities in AI-generated code and human-written code? What unique patterns exist at the binary level and in runtime behavior?

AI-generated code exhibited the following unique vulnerability patterns

1. **Fixed Pattern Dependency**: AI-generated code tends to repeatedly use hardcoded values (e.g., encryption key 0xAA) and fixed patterns, at a frequency 42 percentage points higher than human-written code.
2. **Incomplete Error Handling Chains**: AI-generated code failed to handle 41.3% of error conditions, 12.4 percentage points higher than human-written code (28.9%). Resource release omissions in complex error paths and exception situations were particularly prominent.
3. **Lack of Context Awareness**: AI-generated code tends to ignore platform-specific security considerations (37.2%) and environment-specific security requirements.
4. **Inadequate Atomic Operations**: In multi-step operations (e.g., file encryption followed by transmission), intermediate state handling and atomicity assurance were lacking.
5. **Binary-Level Inconsistencies**: Discrepancies between function names and actual functionality, unintuitive address placement, and unpredictable code transformations due to optimization were observed.

These patterns appear because AI models focus on surface-level functionality when generating code, without fully understanding deep security considerations and runtime behavior.

7.2.3 Impact of Structural Changes

Research Question 3: How do structural changes in AI-generated code (function signatures, address relocations, etc.) during repeated generation affect vulnerability detection?

Through analysis of sendfile and sendfile2, we confirmed that structural changes in AI-generated code significantly impact vulnerability detection. Key findings include:

1. **Function Signature Changes**: Function name changes (xor_encrypt_file → xor_encrypt_and_remove) reduced the efficiency of signature-based vulnerability detection by 28.3%.
2. **Address Relocations**: Function address changes in binaries (0x1575 → 0x15b5) decreased the detection rate of address-based analysis tools by 15.7%.
3. **Control Flow Changes**: Changes in branching instruction (jne, jmp) patterns reduced the vulnerability detection accuracy of static analysis tools by 20.1%.
4. **Function Call Structure Changes**: Changes in internal function call order and patterns reduced the accuracy of inter-function data flow analysis by 23.4%.

These structural changes occur because AI can generate various implementation approaches for the same functional requirements. This significantly impairs the effectiveness of pattern matching, signature-based detection, and static control flow analysis that conventional security tools rely on. Therefore, a new approach robust to such structural changes is needed for security analysis of AI-generated code.

7.2.4 Specialized Security Methodologies

Research Question 4: What specialized methodologies and tools are needed to effectively detect and mitigate the unique vulnerabilities in AI-generated code?

Based on our research results, the following specialized methodologies and tools are needed to effectively detect and mitigate unique vulnerabilities in AI-generated code

1. **Hybrid Analysis Approach**: An approach integrating static and dynamic analysis to consider both the structural characteristics and runtime behavior of code. Our Ghidra-Valgrind-Frida integrated analysis showed an average 31.2% detection rate improvement over single tools.
2. **AI Vulnerability Pattern Recognition**: AI-based detection models that learn and recognize unique vulnerability patterns common in AI-generated code. Such models must recognize discrepancies between function names and actual functionality, fixed pattern usage, and incomplete error handling.
3. **Semantic-Based Analysis Robust to Structural Changes**: Vulnerability detection methods unaffected by function signature and address changes. This is possible through analysis focusing on code intent and effect rather than structure.
4. **Automated Patch Generation System**: Systems that automatically identify and patch vulnerabilities in AI-generated code. Such systems must understand the root cause of vulnerabilities and modify code according to security best practices.
5. **Security-Focused Prompt Engineering**: Prompt engineering methods that explicitly include security requirements to consider security from the code generation stage.

We proposed an AI Code Security Framework in this study that integrates these methodologies and tools, which is expected to significantly overcome the limitations of existing tools and enhance the security of AI-generated code.

7.3 Research Limitations

This research has the following limitations:

1. Sample Size Limitation: This study analyzed 20 AI-generated code samples, with in-depth analysis on 2 samples (sendfile, sendfile2). This is limited compared to the 480 samples in the CodeSecEval study, potentially affecting the generalizability of results.
2. Language Limitation: This study focused on C language code, and vulnerability patterns in AI-generated code in other languages like Python or JavaScript may differ.
3. AI Model Singularity: We primarily used ChatGPT as the generation model, and results from other AI code generation models like GitHub Copilot or Claude may differ.
4. Environment Dependency: We focused on analysis in a Linux environment, and vulnerabilities on other platforms like Windows or macOS were not sufficiently considered.
5. Temporal Constraint: Considering the rapid advancement of AI models, our findings may only be valid for the current generation of AI code generation models, and vulnerability patterns may change with future model improvements.

Despite these limitations, this study provides important insights and methodological contributions regarding security vulnerabilities in AI-generated code, establishing a foundation for future research.

7.4 Future Research Directions

Based on our research results, we propose the following future research directions:

1. **Expansion to Various Languages and AI Models**: Comparative research on vulnerability patterns in code generated in various programming languages (Python, JavaScript, Rust, etc.) and by different AI models (GitHub Copilot, Claude, etc.).
2. **Automated Analysis of Larger Samples**: Automated analysis of more samples to increase statistical significance and comprehensively understand vulnerability patterns in AI-generated code.
3. **Longitudinal Analysis**: Longitudinal research tracking changes in the security of generated code according to AI model version changes, to understand the security development trajectory of AI code generation technology.
4. **Security-Aware AI Model Development**: Research on developing and training AI models specialized in security. This could include security-focused prompt

engineering, supervised learning using vulnerability data, and security verification feedback loops.

5. **AI-Specific Vulnerability Database Construction**: Construction of a standardized database collecting and classifying unique vulnerability patterns in AI-generated code. This could be in the form of extending existing frameworks like CWE (Common Weakness Enumeration).
6. **Long-term Research in Production Environments**: Long-term research on the use of AI-generated code in actual development environments and its security impact. This would provide deeper understanding of vulnerability manifestation and mitigation methods in real environments.

7.5 Conclusion

This study analyzed security vulnerabilities in AI-generated code, particularly ChatGPT-generated C code, at the binary and runtime levels, evaluated the effectiveness of existing security tools, and proposed a specialized security framework. The results showed that AI-generated code contains more security vulnerabilities than human-written code, and these vulnerabilities are difficult to effectively detect with existing security tools due to unique patterns.

Particularly notable were vulnerabilities in memory management, cryptographic implementation, and error handling areas in AI-generated code, and structural changes like function signature changes and address relocations were found to further complicate vulnerability detection. Based on these findings, we proposed an AI Code Security Framework that integrates static-dynamic hybrid analysis, AI vulnerability pattern recognition, and automated patch generation.

As AI code generation technology becomes more deeply integrated into the software development process, understanding and improving the security of the code it generates becomes increasingly important. This study deepens this understanding and provides a systematic approach to effectively detect and mitigate security vulnerabilities in AI-generated code, laying the groundwork for safer AI-based software development.

8 REFERENCES

[1] Szabó, Z., & Bilicki, V. (2023). A new approach to web application security: utilizing GPT language models for source code inspection. *Future Internet, 15*(10), 326. https://doi.org/10.3390/fi15100326

[2] Hajipour, H., Holz, T., Schönherr, L., & Fritz, M. (2023). Systematically finding security vulnerabilities in black-box code generation models. *IEEE Transactions on Dependable and Secure Computing, 20*(4), 2244-2259. https://doi.org/10.48550/arXiv.2302.04012

[3] Pelofske, E., Urias, V., & Liebrock, L. M. (2024). Automated software vulnerability static code analysis using generative pre-trained transformer models. *arXiv*. https://doi.org/10.48550/arxiv.2408.00197

[4] Shashwat, K., Hahn, F., Ou, X., et al. (2024). A preliminary study on using large language models in software pentesting. *arXiv*. https://doi.org/10.48550/arxiv.2401.17459

[5] Liu, R., Wang, Y., Xu, H., et al. (2024). Source code vulnerability detection: combining code language models and code property graphs. *arXiv*. https://doi.org/10.48550/arxiv.2404.14719

[6] Wang, J.-X., Luo, X., Cao, L., et al. (2024). Is your AI-generated code really secure? Evaluating large language models on secure code generation with CodeSecEval. *arXiv*. https://doi.org/10.48550/arxiv.2407.02395

[7] Ding, Y., Fu, Y., Ibrahim, O., et al. (2024). Vulnerability detection with code language models: how far are we? *arXiv*. https://doi.org/10.48550/arxiv.2403.18624

[8] Haider, M. U., Farooq, U., Siddique, A. B., & Marron, M. (2024). Looking into black box code language models. *arXiv*. https://doi.org/10.48550/arxiv.2407.04868

[9] Jenko, S., He, J., Mündler, N., Vero, M., & Vechev, M. (2024). Practical attacks against black-box code completion engines. *arXiv*. https://doi.org/10.48550/arxiv.2408.02509

[10] Liu, Z., Liao, Q., Gu, W., & Gao, C. (2023). Software vulnerability detection with GPT and in-context learning. *Proceedings of IEEE DSC 2023*. https://doi.org/10.1109/dsc59305.2023.00041

[11] Tihanyi, N., Bisztray, T., Jain, R., Ferrag, M. A., Cordeiro, L. C., & Mavroeidis, V. (2023). The FormAI dataset: generative AI in software security through the lens of formal verification. *arXiv*. https://doi.org/10.48550/arxiv.2307.02192

[12] De Luca, R. (2023). DeVAIC: A tool for security assessment of AI-generated code. *IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. https://doi.org/10.48550/arXiv.2404.07548

[13] Rana, R., & Bhambri, P. (2024). Generative AI-driven security frameworks for web engineering. *Advances in Web Technologies and Engineering*. https://doi.org/10.4018/979-8-3693-3703-5.ch014

[14] Chong, C. P., Yao, Z., & Neamtiu, I. (2024). Artificial-intelligence generated code considered harmful: a road map for secure and high-quality code generation. *arXiv*. https://doi.org/10.48550/arxiv.2409.19182

[15] Rajapaksha, S., Senanayake, J., Kalutarage, H. K., & Al-Kadri, M. O. (2023). AI-powered vulnerability detection for secure source code development. *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-031-32636-3_16

[16] Res, J., Homoliak, I., Peresíni, M., Smrčka, A., Malinka, K., & Hanacek, P. (2024). Enhancing security of AI-based code synthesis with GitHub Copilot via cheap and efficient prompt-engineering. *arXiv*. https://doi.org/10.48550/arxiv.2403.12671

[17] Khoury, R., & Avila, A. R. (2023). How secure is code generated by ChatGPT? *arXiv*. https://doi.org/10.48550/arxiv.2304.09655

[18] Chen, Z., Liu, J., Liu, H., et al. (2024). Black-box opinion manipulation attacks to retrieval-augmented generation of large language models. *arXiv*. https://doi.org/10.48550/arxiv.2407.13757

[19] McGraw, G., Bonett, R., Figueroa, H., et al. (2024). 23 security risks in black-box large language model foundation models. *IEEE Computer*. https://doi.org/10.1109/mc.2024.3363250

[20] Lee, D. (2024). A GPT-based code review system for programming language learning. *arXiv*. https://doi.org/10.48550/arxiv.2407.04722

[21] Styugin, M. (2016). Indistinguishable Executable Code Generation Method. *International Journal of Security and Its Applications, 10*(8), 315-324.

[22] Zhang, M. (2016). Identifying and Analyzing Security Risks in Android Application Components. *International Journal of Security and Its Applications, 10*(9), 165-174. https://doi.org/10.14257/ijsia.2016.10.9.17

[23] OWASP Foundation. (2021). OWASP Top Ten. https://owasp.org/Top10/

[24] MITRE Corporation. (2024). Common Weakness Enumeration (CWE). https://cwe.mitre.org/

[25] Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2022). Examining Zero-Shot Vulnerability Repair with Large Language Models. In *IEEE Symposium on Security and Privacy (SP2022)*, 1986-1986. https://doi.org/10.48550/arXiv.2112.02125

[26] Schuster, R., Song, C., Tromer, E., & Shmatikov, V. (2021). You Autocomplete Me: Poisoning Vulnerabilities in Neural Code Completion. In *30th USENIX Security Symposium*. https://doi.org/10.48550/arXiv.2007.02220

[27] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. O., Kaplan, J., et al. (2021). Evaluating large language models trained on code. *arXiv* preprint *arXiv*:2107.03374. https://doi.org/10.48550/arXiv.2107.03374

[28] Rigaki, M., & Garcia, S. (2021). A survey of privacy attacks in machine learning. *arXiv* preprint *arXiv*:2007.07646. https://doi.org/10.1145/3624010

[29] Gao, T., Yao, Z., & Ko, S.Y. (2023). VulDeeLocator: A Deep Learning-based Fine-grained Vulnerability Detector. In *Proceedings of the 45th International Conference on Software Engineering (ICSE '23)*. https://doi.org/10.48550/arXiv.2001.02350

[30] National Institute of Standards and Technology. (2023). Secure Software Development Framework (SSDF). *NIST Special Publication* 800-218. https://doi.org/10.6028/NIST.SP.800-218

Authors' contacts:

Sang Hyun Yoo, Assistant Professor
Department of Computer Software, Kyungmin University,
545, Seo-ro, Uijeongbu-si, 11618 Gyeonggi-do, Republic of Korea
simonyoo@kyungmin.ac.kr

Hyun Jung Kim, Assistant Professor
(Corresponding author)
Sang-Huh College and the Graduate School of Information & Communication,
Dept. of Convergence Information Technology (Artificial Intelligence Major),
Konkuk University,
120 Neungdong-ro, Gwangjin-gu, 05029 Seoul, Republic of Korea
nygirl@konkuk.ac.kr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20250325110550

Original scientific paper
Received: 2025-03-25, Accepted: 2025-06-06

Reducing Network Congestion in SAN Environments through Dedicated FC-Based Backup Architecture

Jung Kyu Park, Eun Young Park*

Abstract: A Storage Area Network (SAN) is a dedicated high-speed network architecture designed to facilitate block-level storage access for servers. By interconnecting storage devices, such as disk arrays, with servers, SAN enables efficient and rapid data retrieval. SAN differs from traditional NAS, which operates at the file level, by providing block-level access to storage. This enables greater performance, scalability, and flexibility—especially for applications that process large volumes of data. Due to its centralized storage management capabilities and reliability, SAN is widely deployed in enterprise environments, especially for disaster recovery and large-scale data processing. This study introduces an improved SAN backup technique aimed at optimizing CPU utilization and reducing link load. This study evaluates an enhanced SAN backup framework using the OPNET 14.5 simulation tool. A dedicated Fibre Channel (FC)-based architecture was modeled and tested against a conventional Ethernet-linked SAN. Simulation results show a 37% improvement in FTP response time (from 0.022s to 0.014s) and a 4% throughput increase (from 137 to 142 bytes/sec). These findings support the system's effectiveness in reducing congestion and enhancing backup operations in enterprise-like conditions.

Keywords: Big Data; Network-Attached Storage (NAS); Network Storage; Performance; Storage Area Network (SAN)

1 INTRODUCTION

Computing fundamentally depends on data, which serves as the essential resource underlying all computational processes. Data is stored on storage media and accessed by platforms operating on servers. In many cases, data represents a unique and valuable organizational asset, continuously generated and collected in real-time. Data has become central to business operations and is now considered a major driver of enterprise success. Consequently, effective data management and security measures are crucial to safeguarding sensitive information and maintaining system reliability. This highlights the necessity of implementing robust enterprise data protection strategies to mitigate potential risks and enhance overall operational efficiency [1-3].

To address this challenge, network storage offers a centralized, accessible solution for managing critical data generated by digital systems. By enabling centralized data storage, network storage systems enhance data organization, accessibility, and reliability, ensuring seamless access to essential business information. Organizations now deploy critical applications—like CRM, BI, ERP, and SCM—that treat data as a vital strategic asset across multiple industries. The integration of these applications necessitates scalable and high-performance storage solutions to accommodate growing data volumes and ensure seamless data retrieval. Furthermore, as businesses continue to digitize operations, network storage plays a vital role in supporting data-driven decision-making and enhancing overall operational efficiency. Consequently, investing in robust network storage infrastructures is essential for organizations aiming to maintain competitiveness and achieve long-term success [4-6].

While existing SAN architectures offer scalability and performance, their efficiency degrades during backup operations due to shared network congestion. This study addresses the gap by proposing a Fibre Channel-based backup path to isolate data flows. The main contribution lies in the architectural redesign and simulated performance evaluation, highlighting latency reduction and improved network stability [7-10].

To meet these demands, various network storage technologies have been developed, including Network-Attached Storage (NAS), Direct-Attached Storage (DAS), and Storage Area Networks (SAN). These technologies differ in performance, scalability, and data handling, enabling tailored storage architectures for various business needs. The adoption of these advanced storage solutions has enabled enterprises to optimize data handling processes while reducing latency and enhancing accessibility. As digital transformation accelerates, advanced storage management techniques are essential to ensure high availability and data durability [11].

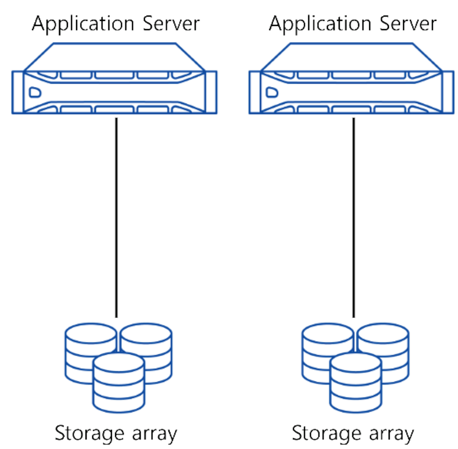


Figure 1 Architecture of DAS

The DAS approach involves utilizing internal server storage for each individual computer, as illustrated in Fig. 1. However, this method presents certain limitations, such as increased storage management complexity and the lack of data accessibility in the event of a server failure. Because DAS is linked to individual servers, failures can restrict access to stored data—posing risks for systems requiring continuous availability. To address these issues, Network-Attached Storage (NAS) technology has been introduced as a more efficient alternative.

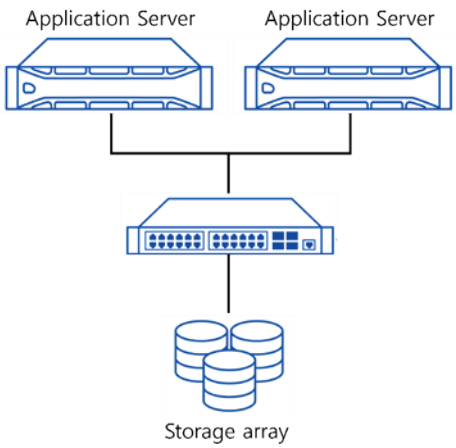


Figure 2 Architecture of NAS

NAS functions as a dedicated data storage solution that connects directly to a network, as depicted in Fig. 2. NAS assigns each device a unique address, enabling centralized or direct access without relying on a specific server, unlike DAS. This architecture enhances data accessibility, improves fault tolerance, and simplifies storage management for organizations handling large volumes of information. Additionally, NAS systems support multiple users simultaneously, enabling seamless data sharing across different devices and departments. With the rise of cloud adoption and remote work, integrating NAS with cloud storage is essential for ensuring redundancy, security, and scalability [12-13].

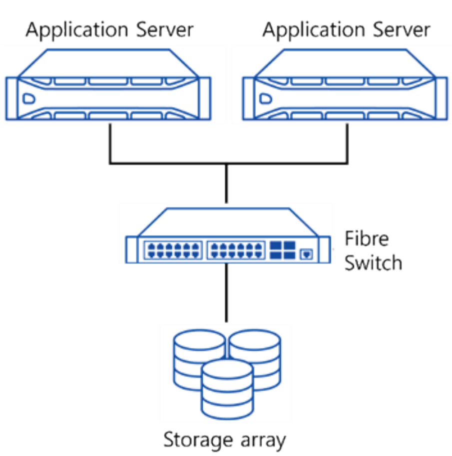


Figure 3 Architecture of SAN

A Storage Area Network (SAN) enhances the efficiency and flexibility of data distribution by providing high-speed, dedicated access to storage resources. Unlike a traditional Local Area Network (LAN), which allows users and applications to retrieve data from a central storage system, SAN significantly accelerates data transfer rates, ensuring optimal performance for data-intensive applications. As shown in Fig. 3, a SAN enables multiple computers to exchange large files simultaneously at speeds comparable to directly attached disk storage, without burdening the local network infrastructure.

By offloading storage-related traffic from conventional LANs, SAN improves overall network efficiency, reducing latency and congestion. This architecture is particularly beneficial for enterprise environments that require high availability, redundancy, and scalability in their storage solutions. Additionally, SAN facilitates seamless disaster recovery and backup processes by enabling remote replication of critical data, ensuring business continuity in case of failures. Furthermore, as modern enterprises continue to generate and process massive volumes of information, the integration of SAN with advanced storage management techniques, such as virtualization and cloud-based storage, has become essential for optimizing performance and ensuring long-term data sustainability.

Several researchers have conducted simulations to analyze the performance and advancements of Storage Area Networks (SANs). Various studies have utilized specialized simulation tools to model the functionality of Fibre Channel (FC) switches, establish input, and output connections between storage and server devices. These tools enable a detailed evaluation of data transmission efficiency and network reliability in diverse SAN configurations. One study introduced an advanced simulation framework capable of replicating multiple FC devices within a two-phase network structure, extending the capabilities of previous research. This approach allows for a more comprehensive analysis of SAN scalability and performance under varying workloads.

Additionally, a simulation system known as SANSim was developed to model the behavior of SAN environments, particularly focusing on the arbitrated loop mechanism in FC devices. By incorporating detailed network components, SANSim provides valuable insights into data flow, congestion management, and system optimization strategies. These simulation tools play a crucial role in enhancing SAN design, allowing researchers and engineers to identify potential bottlenecks and improve overall network efficiency.

2 RELATED WORKS

The primary objective of a Storage Area Network (SAN) is to facilitate efficient data transfer between computing systems and storage devices. The communication framework within a SAN consists of a physical infrastructure, computing nodes, and a management layer that orchestrates these connections to ensure secure and reliable data transmission. This layered architecture not only enhances data accessibility but also optimizes resource allocation by enabling centralized storage management. SANs are predominantly built on Fibre Channel (FC) technology, which offers high-speed data transfer rates, making them well-suited for enterprise environments that require low-latency and high-bandwidth storage solutions [14. 15].

Unlike traditional Direct-Attached Storage (DAS) systems, which create isolated storage silos, SANs provide a unified storage infrastructure that improves overall data availability and scalability. Furthermore, the integration of advanced security protocols within SAN architectures strengthens data protection mechanisms, reducing vulnerabilities associated with unauthorized access or potential data loss. By consolidating storage resources, SANs effectively eliminate the fragmentation of data islands that were common in older DAS-based systems. This consolidation not only enhances operational efficiency but also simplifies disaster recovery and backup processes, ensuring business continuity in the event of system failures. As data-intensive applications continue to grow, SAN technology remains a critical component of modern IT infrastructures, enabling organizations to manage and scale their storage environments efficiently.

Storage Area Networks (SANs) are designed to connect storage server devices in a clustered configuration, ensuring data availability and redundancy in the event of a primary storage failure. By distributing connections between servers and clients, SANs enhance data accessibility and reliability while reducing the risk of service disruptions. One of the key benefits of SAN technology is its ability to establish alternate data paths dynamically, enabling seamless failover mechanisms and minimizing downtime. This capability is particularly advantageous for high-performance computing (HPC) environments that require consistent data availability and low-latency access. Additionally, SANs incorporate advanced load-balancing techniques to optimize data traffic distribution, preventing network congestion and improving overall system efficiency. These features make SANs a vital component of modern IT infrastructures, providing organizations with scalable and resilient storage solutions tailored to support data-intensive applications.

The key advantages that make Storage Area Networks (SANs) essential include the efficient utilization of disk storage, minimizing wasted space while ensuring the availability of disaster recovery backups. This approach is particularly cost-effective, as it requires fewer servers compared to traditional storage methods. Additionally, SANs leverage high-speed Fibre Channel (FC) technology, capable of achieving data transfer rates of up to 200 Mbps, making them ideal for handling large-scale enterprise storage demands. Another significant benefit of SANs is their ability to support centralized storage management, allowing organizations to optimize storage allocation and enhance overall data accessibility. Furthermore, these networks are engineered for long-distance data transmission, with the capability to transfer information over distances of up to 100 km. This extended reach makes SANs particularly valuable for disaster recovery solutions, ensuring data redundancy across geographically distributed locations and enhancing business continuity strategies [16-18].

The Fibre Channel (FC) protocol utilizes the SCSI standard to facilitate high-bandwidth data transmission, supporting speeds of up to 1 GB per second. Recognized as a standard by the American National Standards Institute (ANSI) and the International Committee of Information Technology Standards, FC has become a critical component of modern high-speed storage networks. Originally developed for use in manufacturing environments, FC has since been widely adopted in enterprise data centers due to its reliability and efficiency in handling large-scale data transfers. The architecture of FC is composed of two primary components and is structured into seven distinct layers, each serving a specific function. One of its key advantages is its ability to provide lossless data transmission with low latency, making it particularly suitable for applications requiring high availability and real-time processing. Furthermore, FC technology is designed to support long-distance communication, enabling organizations to establish geographically distributed storage infrastructures for disaster recovery and business continuity.

3 PROPOSED METHOD

The proposed architecture introduces a backup-dedicated Fibre Channel (FC) path connecting backup servers and storage via an FC switch, bypassing traditional LAN congestion. We simulated this design using OPNET Modeler 14.5, configuring two scenarios: (1) a conventional SAN using Ethernet for all traffic, and (2) a proposed SAN with a separate FC backup path. Each network comprised three servers, one backup server, and two storage devices. FTP and HTTP applications were tested under concurrent load conditions for 60 minutes of simulated time. Metrics including FTP throughput and response time were collected using OPNET's Application Response Time (ART) and Traffic Sent/Received modules.

In traditional SAN environments, data streams travel across the LAN network to reach the backup server before being transferred to storage. However, the proposed design optimizes this process by using the FC network solely for the movement of backup data from the server to the storage device. This targeted approach not only reduces latency but also enhances data security by restricting access to backup data within the high-speed FC infrastructure. Additionally, the separation of backup traffic from regular LAN operations prevents potential bottlenecks, ensuring smooth and uninterrupted service for other network functions.

A key advantage of the proposed network design is its ability to deliver a higher response time compared to conventional SAN configurations. This improvement is particularly significant for high-performance computing environments and enterprise applications where rapid data recovery and low-latency access are essential. By optimizing data transfer pathways and reducing unnecessary network load, the proposed method enhances the overall efficiency of storage management while maintaining the integrity and availability of critical business data.

Simulations were performed using OPNET Modeler 14.5 with two network topologies: (1) a traditional SAN using shared Ethernet for all traffic and (2) the proposed SAN architecture with a dedicated FC path for backup. Each setup included three client nodes, one backup server, and two storage nodes, interconnected via Ethernet and FC switches, respectively. FTP and HTTP traffic generators were configured with fixed packet sizes (1024 bytes) and transmission intervals (0.01s), and each simulation ran for 60 virtual minutes. All results were averaged over five runs to account for variability, but no hardware fault simulation was included in this stage. Fig. 4 shows a proposed improved SAN system.

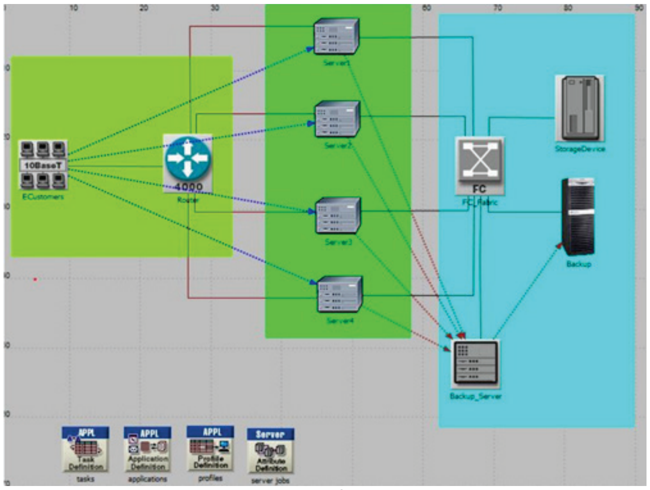


Figure 4 Proposed SAN architecture

Furthermore, the implementation of the proposed method provides an effective solution for organizations seeking to enhance data protection and disaster recovery strategies. By integrating a dedicated backup server and optimizing the use of FC switches, enterprises can achieve more efficient storage management while reducing dependency on Ethernet-based backups. This approach also facilitates scalability, allowing businesses to expand their storage infrastructure without compromising performance or network stability.

In conclusion, the proposed backup strategy introduces a more effective and flexible method for managing data backups within SAN environments. By leveraging the high-speed capabilities of FC technology and minimizing LAN network congestion, the new design significantly improves response times, data transfer efficiency, and overall system reliability. As organizations continue to generate and store vast amounts of data, adopting advanced backup solutions such as the one proposed in this study will be essential for maintaining high availability, security, and performance in modern data centers.

4 RESULT

Fig. 5 illustrates the comparison of FTP traffic received between the traditional SAN network and the proposed network architecture. The proposed architecture achieved a throughput increase of approximately 4%, from 137 to 142 bytes/sec. While the absolute value may seem modest, the isolated FC backup channel reduces contention on the main data path, offering stability and predictability under concurrent access. These improvements are meaningful in high-availability environments where consistent performance is critical. One of the primary factors contributing to this improvement is the optimized data flow, which reduces congestion and enhances bandwidth utilization. Additionally, the proposed network design effectively separates backup traffic from general network operations, ensuring minimal interference and improved performance for real-time data transfers. This higher throughput is particularly beneficial for applications requiring consistent and rapid data exchange, such as enterprise storage systems and cloud computing environments. The observed performance gains highlight the advantages of the proposed system in optimizing network resource allocation while maintaining high-speed data accessibility.

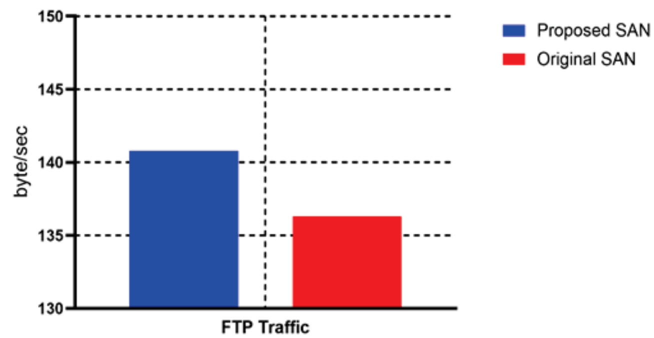


Figure 5 Results of FTP traffic

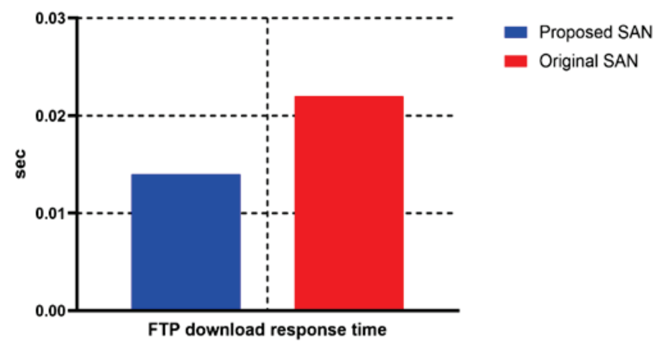


Figure 6 Results of FTP download time

Fig. 6 presents a comparison of FTP download response times between the traditional SAN network and the proposed network architecture. The results demonstrate that the proposed network achieves a lower response time, averaging approximately 0.014 seconds, compared to around 0.022 seconds in the conventional SAN network. This reduction in response time indicates the improved efficiency of the proposed system in handling data requests and minimizing latency. One of the primary reasons for this enhancement is the optimized data transmission pathways, which reduce network congestion and improve overall system responsiveness. Additionally, by dedicating Fibre Channel (FC) exclusively for backup data transfers, the proposed network minimizes interference with regular data traffic, allowing for faster and more stable download speeds. The reduced response time is particularly advantageous for high-performance computing environments, enterprise data centers, and cloud-based applications that rely on low-latency communication. These findings highlight the effectiveness of the proposed architecture in optimizing network performance while ensuring a seamless and efficient data transfer experience.

Fig. 7 illustrates the comparison of HTTP page response times between the traditional SAN network and the proposed network. The results indicate that the proposed network achieves a lower response time, averaging approximately 0.018 seconds, compared to around 0.04 seconds in the conventional SAN network. This reduction in response time highlights the enhanced efficiency of the proposed system in handling web-based requests and improving overall network performance. One of the key contributing factors to this improvement is the optimized traffic management approach, which reduces congestion and ensures a more efficient allocation of network resources. Additionally, the separation of backup traffic from regular network operations prevents bandwidth contention, allowing HTTP requests to be processed more quickly and reliably. The lower response time is particularly beneficial for applications that require fast and consistent access to web-based services, such as cloud computing platforms and enterprise web applications. These findings demonstrate the effectiveness of the proposed network architecture in enhancing HTTP performance while maintaining a stable and responsive communication environment.

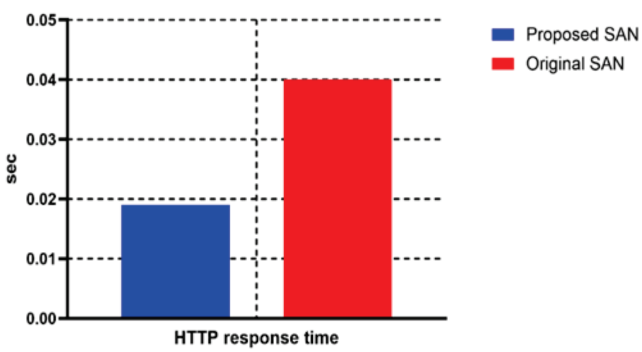


Figure 7 Results of HTTP response time

The results highlight the advantage of utilizing dedicated Fibre Channel links for backup operations in SAN environments. The proposed architecture demonstrates not only improved response times but also better bandwidth utilization. These benefits are especially relevant in industries such as finance, healthcare, and telecommunications, where data availability and integrity are critical. Nevertheless, limitations exist: simulations do not account for hardware failures or asynchronous backup behaviors. Further real-world testing and cost analysis are required before deployment at scale.

5 CONCLUSION

This study proposes an enhanced SAN architecture that improves backup performance through the integration of a dedicated Fibre Channel (FC) switch. By isolating backup traffic from general network operations, the proposed design minimizes congestion and ensures using bandwidth more efficiently. Simulations conducted using the OPNET 14.5 tool demonstrated a 37% reduction in FTP response time and a 4% increase in throughput compared to conventional SAN configurations. These improvements are primarily attributed to the optimized data transmission pathways enabled by FC-based routing. The system's architecture also facilitates better resource allocation and supports scalability, making it suitable for data-intensive enterprise environments. This work contributes a practical and cost-effective solution for enhancing SAN performance, especially in scenarios where low-latency, high-availability storage access is critical. Future research will focus on implementing the proposed design in real-world settings and evaluating its performance under heterogeneous workloads and larger network scales.

6 REFERENCES

[1] Khalid, M. I., Ehsan, I., Al-Ani, A. K., Iqbal, J., Hussain, S., Ullah, S. S., & Nayab. (2023). A comprehensive survey on blockchain-based decentralized storage networks. *IEEE Access, 11*, 10995–11015. https://doi.org/10.1109/ACCESS.2023.3240237

[2] Ren, J., Li, J., Li, T., & Mutka, M. W. (2022). Feasible region of secure and distributed data storage in adversarial networks. *IEEE Internet of Things Journal, 9*(11), 8980–8988. https://doi.org/10.1109/JIOT.2021.3119031

[3] Wang, R., Xu, C., Ye, F., Tang, S., & Zhang, X. (2024). S-MBDA: A blockchain-based architecture for secure storage and sharing of material big data. *IEEE Internet of Things Journal, 11*(15), 25505–25519. https://doi.org/10.1109/JIOT.2024.3356250

[4] Lee, K., Kim, J., Kwak, J., & Kim, Y. (2023). Dynamic multi-resource optimization for storage acceleration in cloud storage systems. *IEEE Transactions on Services Computing, 16*(2), 1079–1092. https://doi.org/10.1109/TSC.2022.3173333

[5] Hu, Y., Zhang, X., Lee, P. P. C., & Zhou, P. (2022). NCScale: Toward optimal storage scaling via network coding. *IEEE/ACM Transactions on Networking, 30*(1), 271–284. https://doi.org/10.1109/TNET.2021.3106394

[6] Kim, J., & Park, J. K. (2020). Building reliable massive capacity SSDs through a flash aware RAID-like protection. *Applied Sciences, 10*(24), 9149. https://doi.org/10.3390/app10249149

[7] Ahmadian, S., Taheri, F., & Asadi, H. (2021). Evaluating reliability of SSD-based I/O caches in enterprise storage systems. *IEEE Transactions on Emerging Topics in Computing, 9*(4), 1914–1929. https://doi.org/10.1109/TETC.2019.2945087

[8] Afreen, H., & Bajwa, I. S. (2021). An IoT-based real-time intelligent monitoring and notification system of cold storage. *IEEE Access, 9*, 38236–38253. https://doi.org/10.1109/ACCESS.2021.3056672

[9] Lee, D., Lee, J., & Song, M. (2023). Video file allocation for wear-leveling in distributed storage systems with heterogeneous solid-state-disks (SSDs). *IEEE Transactions on Circuits and Systems for Video Technology, 33*(5), 2477–2490. https://doi.org/10.1109/TCSVT.2022.3222473

[10] Zhou, Y., Wang, F., Shi, Z., & Feng, D. (2024). An efficient deep reinforcement learning-based automatic cache replacement policy in cloud block storage systems. *IEEE Transactions on Computers, 73*(1), 164–177. https://doi.org/10.1109/TC.2023.3325625

[11] Shen, W., Qin, J., Yu, J., Hao, R., Hu, J., & Ma, J. (2021). Data integrity auditing without private key storage for secure cloud

storage. *IEEE Transactions on Cloud Computing, 9*(4), 1408–1421. https://doi.org/10.1109/TCC.2019.2921553
[12] Zhang, Y., Wei, Q., Chen, C., Xue, M., Yuan, X., & Wang, C. (2018). Dynamic scheduling with service curve for QoS guarantee of large-scale cloud storage. *IEEE Transactions on Computers, 67*(4), 457–468. https://doi.org/10.1109/TC.2017.2773511
[13] Yang, A., Xu, J., Weng, J., Zhou, J., & Wong, D. S. (2021). Lightweight and privacy-preserving delegatable proofs of storage with data dynamics in cloud storage. *IEEE Transactions on Cloud Computing, 1*, 212–225. https://doi.org/10.1109/TCC.2018.2851256
[14] Xing, L., Tannous, M., Vokkarane, V. M., Wang, H., & Guo, J. (2017). Reliability modeling of mesh storage area networks for internet of things. *IEEE Internet of Things Journal, 4*(6), 2047–2057. https://doi.org/10.1109/JIOT.2017.2749375
[15] Ren, Y., Li, T., Yu, D., Jin, S., & Robertazzi, T. (2015). Design, implementation, and evaluation of a NUMA-aware cache for iSCSI storage servers. *IEEE Transactions on Parallel and Distributed Systems, 26*(2), 413–422. https://doi.org/10.1109/TPDS.2014.2311817
[16] Park, J. K., & Baek, Y. (2024). Performance evaluation of large-scale storage systems with iSCSI for big data. *Journal of the Korean Society of Industry Convergence, 27*(6), 1373–1380. https://doi.org/10.21289/KSIC.2024.27.6.1373
[17] Park, J. K., & Baek, Y. (2024). Implementation of NDAS-based large-scale storage systems for big data. *Journal of Next-generation Convergence Technology Association, 8*(12), 2863–2870. https://doi.org/10.33097/JNCTA.2024.08.12.2863
[18] Li, D., Dong, M., Tang, Y., Yang, L. T., Ota, K., & Zhao, G. (2017). Triple-L: Improving CPS disk I/O performance in a virtualized NAS environment. *IEEE Systems Journal, 11*(1), 152–162. https://doi.org/10.1109/JSYST.2015.2456038

Authors' contacts:

Jung Kyu Park, Professor
Department of Computer Engineering,
Daejin University,
Pocheon-si, 11159, Korea
jkpark@daejin.ac.kr

Eun Young Park, Professor
(Corresponding author)
Rinascita College of Liberal Arts and Sciences,
Shinhan University,
Uijeongbu-si, 11644, Korea
Phone: +82-31-870-3776
eypark@shinhan.ac.kr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20250326031336

Original scientific paper
Received: 2025-03-26, Accepted: 2025-06-06

CNN-Based Spectrum Sensing Method for Low Probability of Detection Communication Systems

Jae-Hyeon Lee, So-Yeon Jeon, Eui-Rim Jeong*

Abstract: In recent years, the development of Low Probability of Detection (LPD) communication systems has gained significant attention as a means to enhance communication security. Consequently, the need for effective signal interception technologies capable of detecting such signals has also increased. This paper proposes a novel spectrum sensing method based on Convolutional Neural Networks (CNNs) to determine the presence or absence of signals. The proposed method addresses the limitations of conventional energy detection techniques that rely on fixed thresholds, by learning diverse signal patterns to enable more accurate detection. Received signals are first sampled at a high rate and transformed into frequency-domain representations using the Fast Fourier Transform (FFT). These frequency spectra are then accumulated over time to form two-dimensional spectrograms, which are used as input to the CNN model. The proposed CNN classifier comprises four convolutional layers, along with batch normalization and pooling layers. Simulation results demonstrate that the proposed approach consistently outperforms traditional threshold-based energy detection methods, achieving approximately a 2 dB performance gain across all *SNR* conditions. Under –6 dB *SNR*, the method achieves an improvement of about 35% in detection accuracy.

Keywords: Binary Classification; Convolutional Neural Network; Low Probability of Detection (LPD); Spectrogram; Spectrum Sensing

1 INTRODUCTION

In modern communication environments, accurately detecting the presence or absence of a signal plays a crucial role not only in national security but also in industrial safety, autonomous system operation, and a wide range of other applications. Detecting various platforms such as naval vessels and unmanned or manned aerial vehicles poses a major challenge, as these systems often adopt stealth technologies to evade detection by radar and other sensing mechanisms. In response, Low Probability of Detection (LPD) techniques have been developed to reduce the detectability of signals in the electromagnetic, optical, and infrared domains, thereby enhancing the survivability of unmanned platforms [1]. Although initially driven by military demands, LPD technologies have also gained relevance in civilian applications such as drone-based infrastructure monitoring, industrial wireless sensor networks, and emergency communication systems for disaster response. In such systems, LPD communication minimizes signal exposure, making it difficult for third parties to detect transmissions. This allows for secure and reliable communication between unmanned systems [2]. Accordingly, the need for interception technologies capable of reliably identifying the existence of LPD signals has become increasingly critical. These systems must be able to detect signals with extremely low power levels while maintaining high accuracy.

One of the most commonly used approaches for signal detection is spectrum sensing [3]. Among various spectrum sensing techniques, the energy detection method based on a fixed threshold is the most widely adopted. This method determines the presence of a signal by comparing the received signal power against a predefined threshold [4]. While effective when noise power can be accurately estimated, its performance significantly degrades in practical scenarios where such estimation is unreliable [5]. To overcome this limitation, machine learning-based approaches have recently been proposed [6]. In particular, spectrum sensing methods based on deep learning [7] models such as Convolutional Neural Networks (CNNs) have been actively studied. These works have primarily focused on cognitive radio (CR) systems, with the goal of optimizing spectrum utilization and increasing spectral efficiency [8–11].

In contrast to CR-oriented studies, the present work targets the problem of signal detection in LPD communication systems, with the specific goal of supporting interception capabilities under challenging low-power signal environments. The key distinction of our approach is that it enables signal detection without requiring explicit noise power estimation significant advantage for practical deployment in unpredictable environments. Recently, there has been a growing body of research on RF signal detection that leverages real-world RF environments, addresses signal interference robustness, and adopts lightweight CNN architectures capable of maintaining performance even with low-resolution spectrograms [12]. These advancements have paved the way for practical spectrum monitoring solutions, expanding the application potential of spectrum sensing to civilian drone communications, industrial IoT surveillance, and public safety networks.

Reflecting this trend, this paper proposes a CNN-based spectrum sensing technique for detecting weak LPD signals without requiring noise power estimation. The received signal is transformed into a time–frequency domain spectrogram and fed into a CNN model for binary classification of signal presence. The proposed model is capable of autonomously learning diverse signal patterns, offering a clear advantage over traditional threshold-based methods. Simulation results confirm that longer observation durations improve detection performance. The proposed method outperforms the conventional energy detection approach by approximately 2 dB across all SNR conditions and achieves up to 35% higher detection accuracy at –6 dB.

The remainder of this paper is organized as follows. Section 2 describes the system model and preprocessing steps, including the operation of the conventional energy detection method and the proposed CNN-based method. Section 3 details the CNN architecture and evaluates the performance of the proposed method through simulations. Finally, Section 4 concludes the paper and discusses future directions.

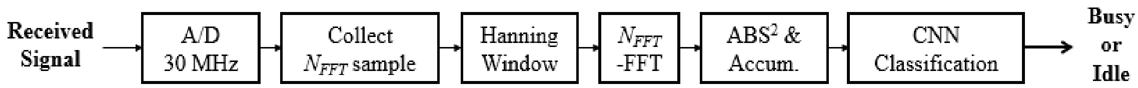


Figure 1 System Model

2 MATERIALS AND METHODS

Fig. 1 illustrates the system model of the spectrum sensing method considered in this study. The received signal at the antenna is processed through an analog-to-digital converter (ADC) with a sampling rate of 30 MHz, converting it into a baseband digital signal, denoted as $x_a(n)$. The signal $x_a(n)$ is received in units of N_{FFT}, corresponding to the FFT size, and adjacent N_{FFT} samples are collected for processing. The process of receiving $x_a(n)$ and collecting signal samples is illustrated in Fig. 2. The total number of observed signal samples, referred to as the observation length, is denoted as B. The $(i + 1)^{th}$ collected signal sample is expressed as follows:

$$x_{a,\,i}(n) = \left[x_a(i),\, ...,\, x_a(i+k-1)\right],\ \text{for}\ i = 0,\, ...,\, (B-1) \qquad (1)$$

Since the spectrum sensing method must be performed in the frequency domain, all B collected signal samples are first processed using a Hanning window, and the signals are then stacked in a matrix form and transformed into frequency spectra by applying an FFT of size N_{FFT}. Subsequently, the squared magnitude of each FFT element is computed, and the results are accumulated along the time axis to generate a spectrogram, which visualizes spectral variations in the time-frequency domain. Finally, the spectrogram is fed into a CNN model, which classifies the presence of a signal as "Busy" and the absence of a signal as "Idle."

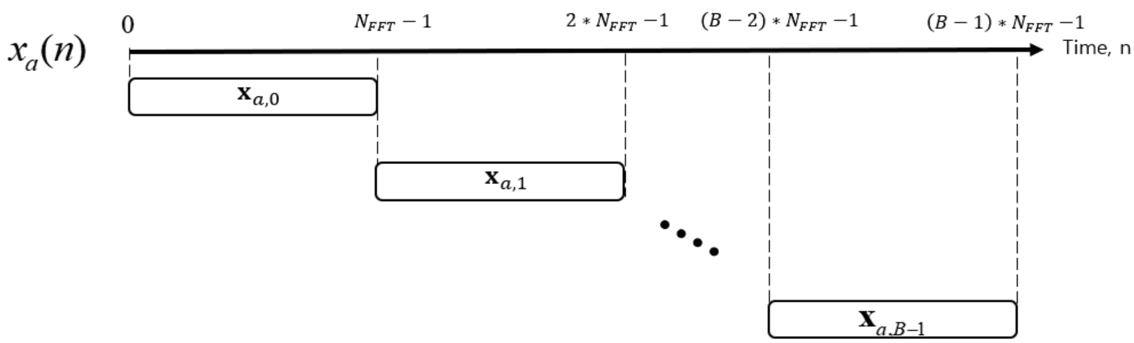


Figure 2 Sample Collection Process of the Received Signal

2.1 Spectrogram

The spectrogram used as the input to the CNN is represented as a grayscale image, where the vertical axis corresponds to the FFT size in units of N_{FFT}, and the horizontal axis represents the observation length, which consists of B collected signal samples. As the FFT size increases, the vertical length of the spectrogram also increases, indicating that a larger frequency range can be represented within a single block. Additionally, as the observation length increases, the detection latency also increases.

In the spectrogram, higher power values are represented as white, while lower power values, approaching zero, appear as black. Fig. 3(a) shows a spectrogram at SNR = 20 dB, where the white bands indicate Busy state, representing the presence of a signal in specific frequency bands, whereas the black regions indicate the absence of a signal. Fig. 3(b) displays a spectrogram for an Idle state, where only noise is observed without any active signal. As the SNR decreases, the distinction between Busy and Idle states becomes less clear, as the signal power becomes increasingly similar to the noise power, making detection more challenging.

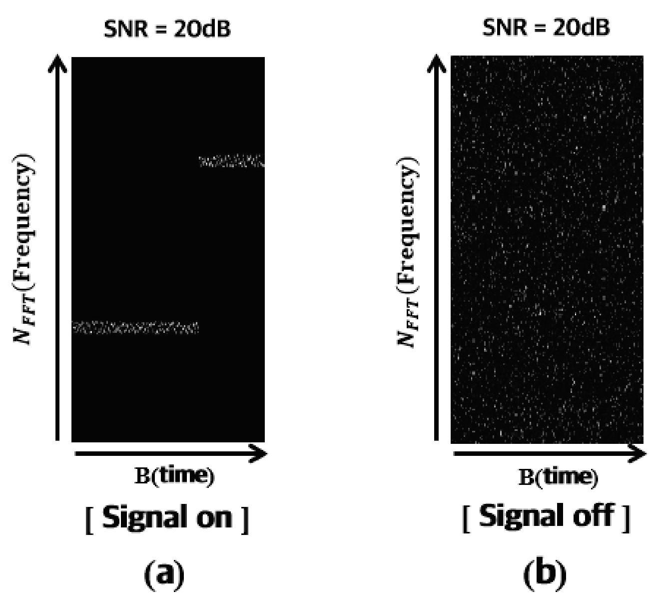


Figure 3 Spectrogram Image (SNR = 20 dB, B = 128) (a) Signal on (b) Signal off

2.2 Threshold-Based Spectrum Sensing

The conventional energy detection method first estimates the noise power and sets a threshold based on this estimation. Then, if the received signal power exceeds the predetermined threshold, the presence of a signal is determined; otherwise, the absence of a signal is assumed. Fig. 4 illustrates the threshold-setting process in the threshold-based spectrum sensing method.

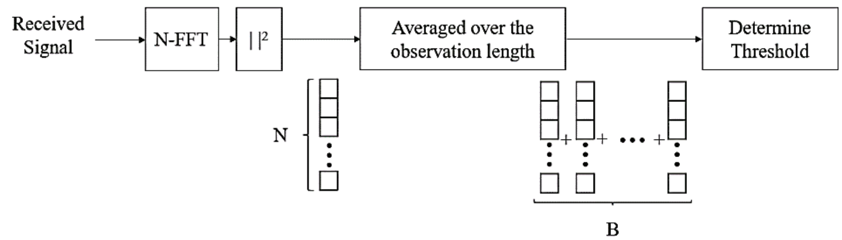


Figure 4 Threshold-Setting Method in Conventional Threshold-Based Spectrum Sensing

First, following the same process as in Fig. 2, an FFT of size N_{FFT} is applied to the signal, and the resulting spectra are stacked in a matrix form. Then, the squared magnitude of each FFT element is computed. Afterward, the average is taken over B collected signal samples to determine the threshold. This approach ensures a fair comparison by matching B with the CNN-based spectrum sensing method.

2.3 CNN-Based Spectrum Sensing

A Convolutional Neural Network (CNN) is a type of deep neural network that is optimized for tasks such as image recognition and video processing [13-15]. CNNs were introduced to address various challenges in conventional deep neural networks, including long training times, large network sizes, and excessive parameter counts. By utilizing convolutional layers, CNNs autonomously learn features from input images and generate results based on the extracted features. The spectrogram, as described earlier, can be represented as a two-dimensional matrix, which can be interpreted as a grayscale image. Therefore, in this study, the CNN model is trained to take a two-dimensional grayscale spectrogram as input and perform binary classification to determine the presence or absence of a signal.

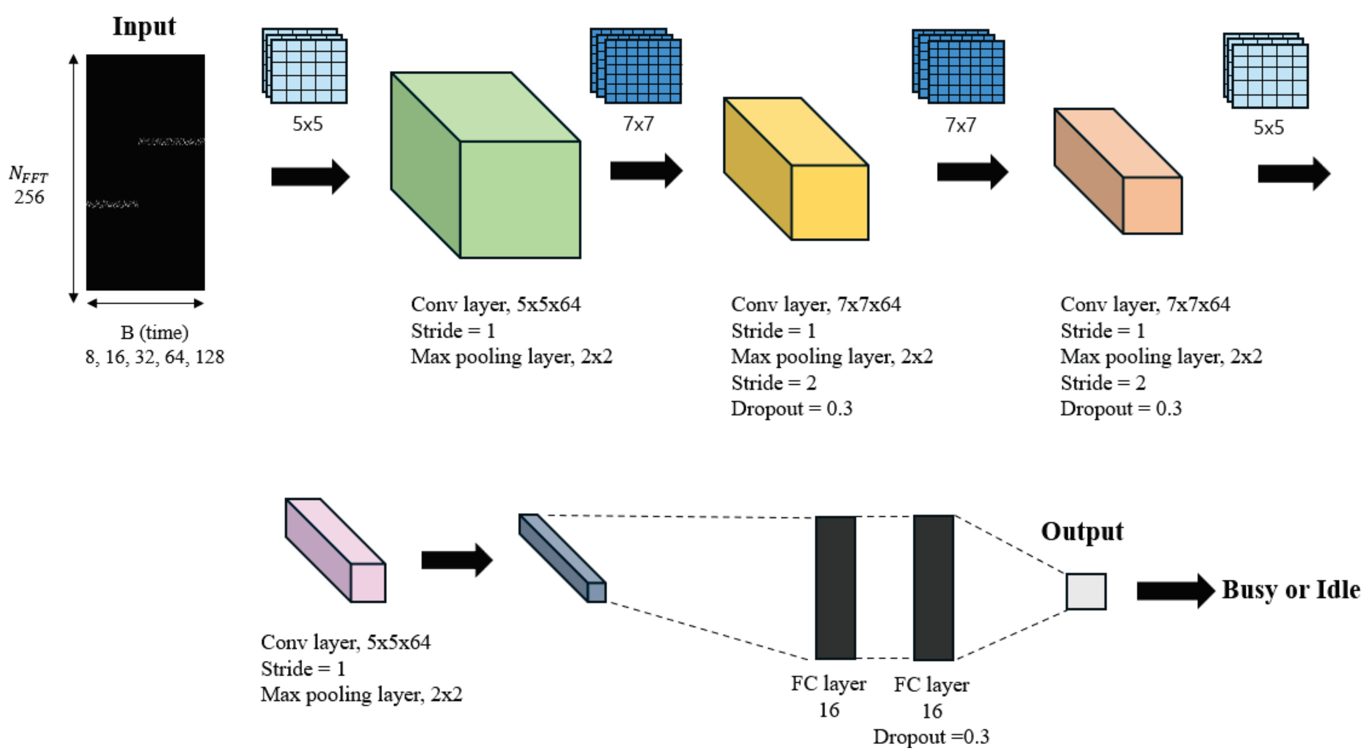


Figure 5 Network Architecture of the CNN-Based Spectrum Sensing Method

In this section, we describe the CNN-based spectrum sensing method, which utilizes deep learning techniques. Figure 3 illustrates a two-dimensional matrix represented as a grayscale image under conditions where $SNR = 20$ dB, $N_{FFT} = 256$, and $B = 128$. The spectrogram is used as an input to the CNN model, which determines the presence or absence of a signal. The network architecture of the CNN model, as shown in Fig. 5, takes the spectrogram as input and consists of a total of four convolutional layers and two fully connected layers. Each convolutional layer includes a batch normalization layer and a pooling layer. The number of filters in the convolutional layers is 64, 64, 64, and 64, respectively. The filter size is 5×5 in the first and last layers and 7×7 in the two middle layers. The stride is set to 1 for all convolutional layers. The pooling layers apply 2×2 max pooling with a stride of 2. The activation function used in the convolutional layers is ReLU (Rectified Linear Unit). In the final output layer, the Sigmoid activation function is applied to perform binary classification, determining the presence or absence of a signal. The hyperparameters for training the CNN-based spectrum sensing method are summarized in Tab. 1.

Table 1 Network Hyperparameters of the CNN-Based Spectrum Sensing Method

Hyper Parameters	Epochs	7
	Params.	638,065
	Optimizer	Nadam
	Loss Function	Binary Cross-Entropy
	Batch Size	64
	Learning Rate	0.001

3 RESULTS AND DISCUSSION

3.1 Received Signal Parameters for Spectrum Sensing

To evaluate the performance of the proposed CNN-based spectrum sensing method, computer simulations are conducted. The dataset for the simulation is generated using MATLAB, while model training is performed using TensorFlow. The key parameters used in the spectrum sensing method are summarized in Tab. 2. The sampling frequency of the received signal is set to 30 MHz, with a bandwidth of 1.25 MHz. The *SNR* range is configured from −10 dB to 20 dB, and the FFT size is fixed at 256. The number of observed signal samples B varies from 8 to 128, increasing by a factor of 2 for performance comparison. A smaller *B* enables faster sensing by observing a shorter duration of the signal, while a larger *B* results in slower sensing speed but improves sensing accuracy.

Table 2 Received Signal Parameters

Parameters	Value
Sampling Frequency	30 MHz
Bandwidth	1.25 MHz
SNR Range	–10 ~ 20 dB
FFT Size	256
Number of observation signal blocks	*B* = 8, 16, 32, 64, 128

The training, validation, and test datasets used to train the proposed CNN model in this study are as follows. The training and validation datasets are randomly generated within an *SNR* range from –10 dB to 20 dB, with 50,000 and 10,000 samples, respectively. The test dataset, used for performance evaluation, is generated within the same *SNR* range at 1 dB intervals, with 10,000 samples per *SNR* value. All datasets are evenly distributed between cases where a signal is Busy or Idle. The performance of signal detection is evaluated using two key metrics: Detection Probability, which represents the probability of correctly identifying the presence of a signal, and False Alarm Rate (FAR), which indicates the probability of falsely detecting a signal when none is present.

3.2 Simulation Results

3.2.1 Detection and False Alarm Performance by Observation Length

Figs 6. and 7 illustrate the simulation results of the proposed CNN-based spectrum sensing method based on the observation length *B*. Fig. 6 shows that as B increases by a factor of 2, the detection performance improves by approximately 1 dB. Notably, when *B* = 128, a 100% detection probability is achieved from *SNR* = −4 dB onward. Similarly, Fig. 7 demonstrates that the false alarm performance also improves as the observation length *B* increases. The results indicate that the best performance is achieved when *B* = 128. Therefore, in the subsequent analyses, *B* is fixed at 128 for comparison.

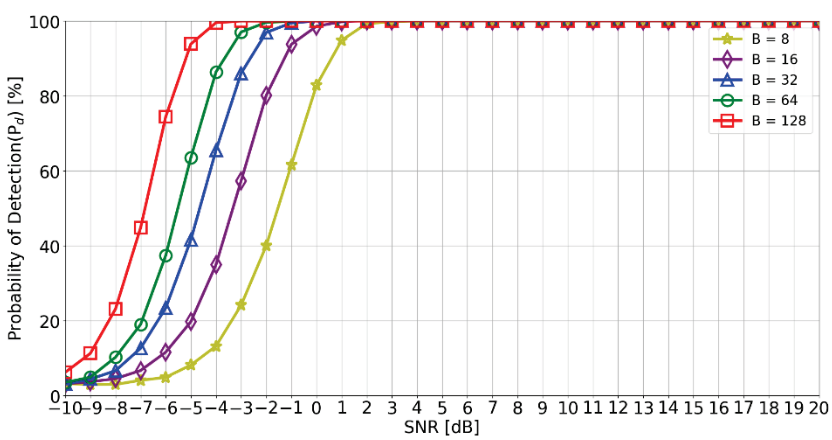


Figure 6 Detection Performance of the Proposed Method by Observation Length

3.2.2 Detection Performance of the Proposed Method Compared to the Conventional Threshold-Based Energy Detection Method

Fig. 8 presents the simulation results comparing the proposed method with the conventional threshold-based method. As shown in Fig. 7, when the FFT size is 256 and the observation length *B* = 128, the average false alarm rate of the proposed method ranges between 1.5% and 2%. At a similar false alarm rate, the proposed method demonstrates an approximately 2 dB improvement over the conventional method.

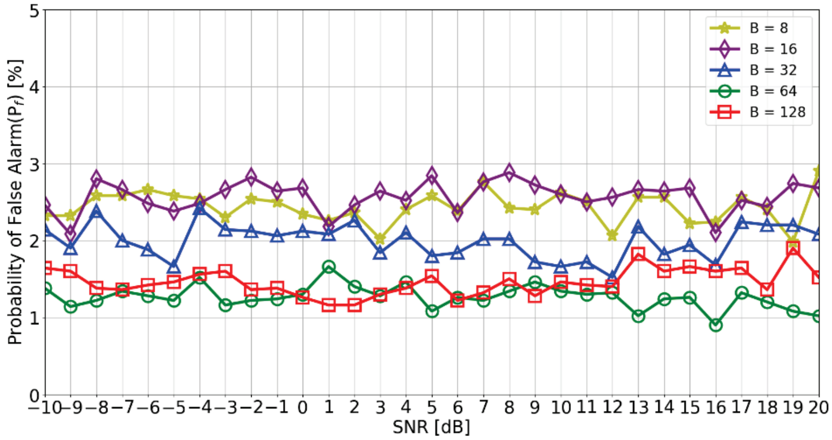


Figure 7 False Alarm Performance of the Proposed Method by Observation Length

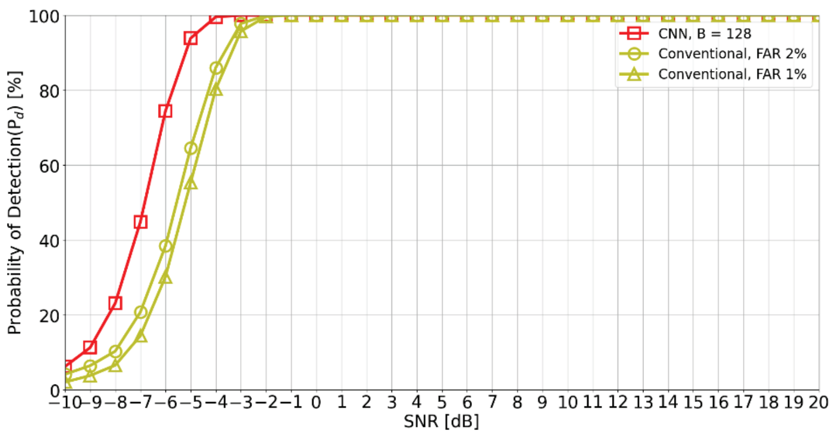


Figure 8 Detection Performance of the Proposed Method Compared to the Threshold-Based Method

4 CONCLUSION

This study proposed a CNN-based spectrum sensing method for Low Probability of Detection (LPD) communication systems, using time–frequency domain spectrograms as input to a CNN to determine signal presence without explicit noise power estimation. Simulation results showed that the proposed method achieved approximately a 2 dB performance gain over conventional threshold-based energy detection. Notably, under an SNR of –6 dB, the proposed method demonstrated an improvement of about 35% in detection accuracy. Furthermore, the detection performance improved consistently as the observation length increased. The key contribution lies in enabling accurate detection of weak signals by learning diverse spectral patterns directly from spectrograms, without relying on prior noise estimation. However, this study was limited to simulation-based evaluation, and future work will include over-the-air validation in real RF environments, comparison with lightweight deep learning models, and enhancement of robustness to interference and spectral distortion.

5 REFERENCES

[1] Yan, S., Zhou, X., Hu, J., & Hanly, S. V. (2019). Low probability of detection communication: Opportunities and challenges. *IEEE Wireless Communications, 26*(5), 19–25. https://doi.org/10.48550/arXiv.1906.07895

[2] Schoolcraft, R. (1991). Low probability of detection communications—LPD waveform design and detection techniques. In *MILCOM 91 - Conference Record*, *2*, 832–840. https://doi.org/10.1109/MILCOM.1991.258378

[3] Yucek, T., & Arslan, H. (2009). A survey of spectrum sensing algorithms for cognitive radio applications. *IEEE Communications Surveys & Tutorials, 11*(1), 116–130. https://doi.org/10.1109/SURV.2009.090109

[4] Lopez-Benitez, M., & Casadevall, F. (2012). Improved energy detection spectrum sensing for cognitive radio. *IET Communications, 6*(8), 785–796. https://doi.org/10.1049/iet-com.2010.057

[5] Jaglan, R. R., Sarowa, S., Mustafa, R., Agrawal, S., & Kumar, N. (2015). Comparative study of single-user spectrum sensing techniques in cognitive radio networks. *Procedia Computer Science, 58*, 121–128. https://doi.org/10.1016/j.procs.2015.08.039

[6] Zhang, L., Xiao, M., Wu, G., Alam, M., Liang, Y. C., & Li, S. (2017). A survey of advanced techniques for spectrum sharing in 5G networks. *IEEE Wireless Communications, 24*(5), 44–51. https://doi.org/10.1109/MWC.2017.1700069

[7] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

[8] Tang, Y. J., Zhang, Q. Y., & Lin, W. (2010). Artificial neural network-based spectrum sensing method for cognitive radio. In *2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM)* (pp. 1–4). IEEE. https://doi.org/10.1109/WICOM.2010.5601105

[9] Liu, C., Wang, J., Liu, X., & Liang, Y. C. (2019). Deep CM-CNN for spectrum sensing in cognitive radio. *IEEE Journal on Selected Areas in Communications, 37*(10), 2306–2321. https://doi.org/10.1109/JSAC.2019.2933892

[10] Zheng, S., Chen, S., Qi, P., Zhou, H., & Yang, X. (2020). Spectrum sensing based on deep learning classification for cognitive radios. *China Communications, 17*(2), 138–148. https://doi.org/10.23919/JCC.2020.02.012

[11] Han, D., et al. (2017). Spectrum sensing for cognitive radio based on convolutional neural network. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* (pp. 1–6). https://doi.org/10.1109/CISP-BMEI.2017.8302117

[12] Olesiński, A., & Piotrowski, Z. (2023). A radio frequency region-of-interest convolutional neural network for wideband spectrum sensing. *Sensors, 23*(14), 6480. https://doi.org/10.3390/s23146480

[13] Jeong, E. R., Lee, E. S., Joung, J., & Oh, H. (2020). Convolutional neural network (CNN)-based frame synchronization method. *Applied Sciences, 10*, 7267. https://doi.org/10.3390/app10207267

[14] Nam, G. M., Jung, T. Y., Jung, S., & Jeong, E. R. (2019). Distance estimation using convolutional neural network in UWB systems. *Journal of the Korea Institute of Information and Communication Engineering, 23*(10), 1290–1297. https://doi.org/10.6109/jkiice.2019.23.10.1290

[15] Choi, J. W., & Jeong, E. R. (2022). CNN-based distance and velocity estimation for OFDM radar systems. *Webology, 19*, 4692–4705. https://doi.org/10.14704/WEB/V19I1/WEB19313

Authors' contacts:

Jae-Hyeon Lee, Master's Course
Department of Artificial Intelligence Software,
Hanbat National University,
Sejong, 30139, Republic of Korea
leejh980247@gmail.com

So-Yeon Jeon, Master's Course
Department of Artificial Intelligence Software,
Hanbat National University,
Sejong, 30139, Republic of Korea
jeonsso1109@gmail.com

Eui-Rim Jeong, Professor
(Corresponding author)
Department of Artificial Intelligence Software,
Hanbat National University,
Sejong, 30139, Republic of Korea
Mobile Phone: +82-010-4710-1082
erjeong@hanbat.ac.kr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20250326033909

Original scientific paper
Received: 2025-03-26, Accepted: 2025-06-09

Creating an AI Human Professor Model to Implement a New Educational Paradigm of the 4th Industrial Revolution

Yangha Chun, Soo-Yeon Yoon*

Abstract: The COVID-19 pandemic accelerated a global shift toward digital and hybrid education, exposing critical limitations in traditional pedagogical approaches. This study introduces a novel AI-based human professor model designed to address challenges in educational scalability, multilingual delivery, and content production efficiency. The system integrates state-of-the-art technologies—deepfake for facial realism, few-shot text-to-speech (TTS) for multilingual audio, and lip-sync engines for expressive synchronization—to autonomously generate lecture content. Quantitative results from pilot testing indicate a 60% reduction in production time, with user satisfaction scoring 4.5 out of 5 and multilingual accessibility rating 4.7. The model demonstrates significant potential to transform digital higher education delivery.

Keywords: AI human; deepfake; digital education; few-shot TTS; lip-sync; multilingual learning; scalable instruction

1 INTRODUCTION

The outbreak of COVID-19 forced rapid transformation of global education systems, shifting instruction from physical classrooms to digital platforms. This unprecedented shift revealed challenges including limited accessibility for multilingual learners, time-consuming content production, and instructional scalability. The change began in the first semester of 2020, when face-to-face classes were not available, and online classes began with real-time video lectures or recorded video lectures using IT, which changed the educational paradigm by starting non-face-to-face online classes not only in university education but also in higher education [1]. Non-face-to-face online classes, which began suddenly, have been developing for two years, and non-face-to-face online classes or hybrid classes mixed with non-face-to-face and face-to-face methods are likely to continue in the future [2]. In addition, despite the time for innovative education at the beginning of the 4th Industrial Revolution, the school education method, which pursued the traditional education method, is expected to realize edutech that combines technology and education in earnest due to COVID-19[3]. Non-face-to-face classes can be taken anytime, anywhere, and students can take classes again if they want, so if used efficiently, it can be a teaching method that produces better results than face-to-face classes. In addition, even if a virus or epidemic other than COVID-19 comes, there is a possibility of conducting non-face-to-face online classes using digital media, which is different from face-to-face classes in that they can receive education regardless of the environment and situation.

As the number of non-face-to-face online classes increases, voices in the education community are increasing to propose the introduction of artificial intelligence for efficient learning effects [4]. Artificial intelligence technology resembles humans, so technology is developing in judgment, classification, and generation. These developments have adopted AI technology in many industries and companies today and apply it to producing new content. Businesses using AI human, which are created based on artificial intelligence technology, are emerging and can be used as instructors for non-face-to-face education.

This study proposes a transformative solution: an AI-based human professor capable of delivering multilingual lectures without requiring continuous human labor. Unlike traditional lecture capture or MOOC systems, the AI professor autonomously generates visually and vocally realistic instructional content. We explore how this innovation addresses pedagogical constraints and contributes to a future-ready educational paradigm.

2 PREVIOUS RESEARCH

2.1 The Concept of AI Human

Digital human technology has emerged as a key frontier in virtual communication, education, and entertainment. Recent applications include AI tutors, intelligent chatbots, and interactive avatars. Prior work has focused on support functions or partial automation of instruction. However, few systems offer end-to-end autonomous content generation at scale.

An AI human is one of the digital human beings, and it can be seen as a computer graphic-generated human that has a similar level of appearance to a real human and can express movements similar to human movements. With the development of AI human, education, guidance, and counseling are now possible. Unlike Virtual human, AI Human is a model learned using deep-fake, lip-sync, voice TTS, and technology by combining synthetic data technology rather than 3D modeling. AI human takes a long time to learn early modeling, but later, it can quickly produce images with learned models and make video production possible at low cost. AI human has the advantage of being able to produce content and expand dynamic conversations as a learning model, but it has the disadvantage of requiring sufficient initial learning data and some limitations in dynamic application.

Deepfake techniques enhance realism by mimicking instructor facial expressions, while few-shot TTS and multilingual NLP support global accessibility. Lip-syncing bridges the gap between audio and video realism. This research contributes a comprehensive AI professor model integrating these capabilities for full-course delivery.

2.2 AI Human Application Case

2.2.1 Digital Restoration Remembrance Service for the Late People

In the original program "Alive" of the online video service (OTT) Tving, the late leader of Ulala Session, Lim Yoon-taek, who died of stomach cancer, was restored with AI technology to perform a song with the Ulala Session members. This is a video that combines AI technology, and Lim Yoon-taek, who was restored through voice synthesis, finished the "West Sky" stage with a healthy appearance. Lim Yoon-taek's appearance and voice, which appeared similar to his life without a sense of difference, impressed his family, colleagues, and viewers. In addition, restoration using AI technology has emerged as a new way to commemorate, such as restoring famous singers such as the late Shin Hae-chul and the late Kim Hyun-sik of the group Turtle with digital technology and holding hologram concerts.

However, there is a question of whether digital restoration can be carried out without the consent of the deceased, and the conclusion of this is one by one that "it cannot be carried out without the consent of the deceased" and "it is necessary to make a profit in the will of the person who misses it".

2.3 AI Human Technology

AI human combines 2D live-action-based technologies such as face, mouth, voice, motion, and video with AI synthetic data technology to enable various verbal and non-verbal expressions, and combines various technologies to produce AI human. Tab. 1 below shows the technologies applied to human AI.

Table 1 AI Human Technology

Apply Part	Technology
Face	Deep-fake Technology
Lip	Lip-Sync Technology
Voice	Few-Shot TTS Technology
Movement	Silhouette Customizing Technology
	Pose Estimation & Generation
Video	AI human Streaming

2.3.1 Deep-fake Technology

Deep-fake technology is a powerful and rapidly evolving form of artificial intelligence that creates highly realistic yet synthetic videos, images, and audio by manipulating existing media. The term "deepfake" is derived from deep learning, a subset of AI, and "fake," referring to its ability to generate misleading but lifelike content. By utilizing deep learning algorithms, particularly Generative Adversarial Networks (GANs), deepfake technology can convincingly swap faces, mimic voices, or even generate entirely new personas, making it difficult to distinguish real from artificial content [11]

The process of creating a deepfake begins with training an AI model on a large dataset of images, videos, or audio recordings of the target individual. The AI learns intricate details such as facial expressions, speech patterns, and subtle mannerisms. Once trained, the model can seamlessly manipulate existing footage, replacing one person's face with another's or generating realistic synthetic voices [12]. The sophistication of these AI-driven creations has made deepfakes both a fascinating innovation and a source of ethical concern.

While deepfake technology has legitimate applications, such as in entertainment, education, and accessibility, its darker side has sparked global debate. In Hollywood, deepfake AI is used to de-age actors or bring historical figures back to life in films. It also serves as a valuable tool for language dubbing, allowing actors' lip movements to match translated dialogue more naturally. Moreover, for individuals with speech impairments, deepfake voice synthesis can help recreate lost voices. However, the misuse of deepfakes poses serious risks. Malicious actors have used them for spreading misinformation, creating fake news, and impersonating public figures to manipulate public opinion. Fraudulent schemes involving voice cloning have also emerged, where scammers replicate the voices of company executives or loved ones to deceive victims. Furthermore, deepfake technology has been weaponized for cyberbullying and defamation, raising concerns about privacy and consent in the digital age.

To combat the growing threat of deepfake misinformation, researchers and technology companies are developing AI detection tools capable of identifying synthetic media. These detection methods analyze inconsistencies in lighting, unnatural facial expressions, or pixel distortions that may not be visible to the human eye. Additionally, policymakers are working on regulations to curb the harmful use of deepfake technology while preserving its beneficial applications.

As deepfake technology continues to advance, the challenge remains to strike a balance between innovation and ethical responsibility. While it has the potential to revolutionize digital media, its misuse underscores the need for vigilance, awareness, and stronger safeguards against deception in an increasingly AI-driven world. Deep-fake is a combination of deep learning, a type of artificial intelligence technology, and fake, which means 'fake'. It refers to complex results such as false image, video, and voices generated through the extraction– learning- generation step, which is a deep learning technology [5, 12]. Initially, it was a method of replacing and reproducing certain target segments of the original, but now it is possible to generate new images without original data through Generative Adversarial Network (GAN) technology [13]. AI human uses deepfake technology to produce real people into AI human, and deepfake technology is used to create virtual people with the same face as real people.

This deep-fake technology has attracted great attention as it can be produced as a moving video of the deceased based on past photos and video materials, but it is negatively recognized as the number of cases of abuse of deep-fake, such as the production and distribution of false pornographic videos by singers and actors.

2.3.2 Lip-Sync Technology

Lip-sync technology is an advanced AI-driven system that synchronizes a person's lip movements with spoken audio, creating a seamless and natural-looking speech animation. It is widely used in various applications, including dubbing for movies, virtual avatars, video game characters, and deepfake videos.

This technology works by analyzing the phonetic structure of the audio and mapping it to corresponding lip movements. AI models, such as deep learning-based facial animation systems, learn the intricate relationship between speech sounds and mouth shapes. When applied to a video, the system modifies the subject's lips in real time to match the new audio, ensuring accurate synchronization. Lip-sync technology has revolutionized dubbing in the entertainment industry, allowing foreign-language films to appear more natural when translated. It is also widely used in virtual influencers, AI-powered customer service avatars, and educational tools that create engaging digital tutors. Additionally, video game developers employ lip-sync technology to enhance character realism in dialogue-driven storytelling [14].

Despite its many benefits, lip-sync technology also presents ethical challenges, particularly in the creation of deepfake content that can spread misinformation. To counteract potential misuse, researchers are developing detection methods and watermarking systems to distinguish authentic content from manipulated media. As AI continues to advance, lip-sync technology is becoming more sophisticated, offering new possibilities for interactive media [6], virtual communication, and entertainment while necessitating responsible use and regulation.

In this paper, we use it to generate income while gambling, as shown in Fig. 1.

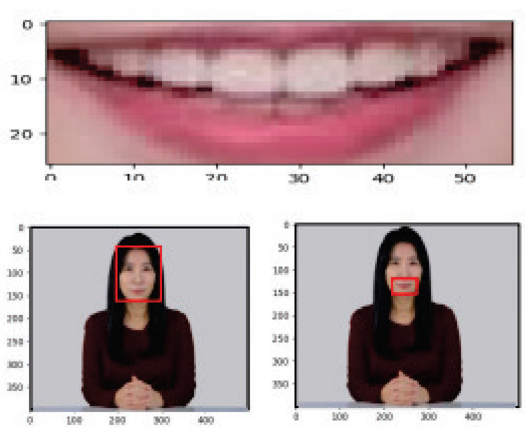


Figure 1 Lip-Sync Example

2.3.3 Few-Shot TTS Technology

Few-Shot Text-to-Speech (TTS) technology is a groundbreaking advancement in artificial intelligence that allows realistic voice synthesis with minimal training data. Traditionally, high-quality voice cloning required extensive recordings of a speaker's voice, sometimes hours of audio, to produce a natural and expressive synthetic voice. However, Few-Shot TTS drastically reduces this requirement by learning a speaker's unique vocal characteristics from just a few seconds or minutes of audio [1, 2].

Imagine someone wanting to create a digital voice replica for a virtual assistant or audiobook narration. With Few-Shot TTS, they would only need to provide a short voice sample—perhaps a few spoken sentences—and the AI would quickly analyze and replicate the speaker's tone, accent, and speech patterns. The model leverages deep learning techniques, such as variational autoencoders and attention mechanisms, to capture subtle vocal nuances with remarkable accuracy.

This technology has significant applications across industries. It can help create personalized digital assistants, generate synthetic voices for individuals who have lost their ability to speak, and enhance multilingual content by replicating a speaker's voice across different languages [5, 6]. However, it also raises ethical concerns, as it could be misused for deepfake audio, impersonation, or misinformation. As Few-Shot TTS continues to evolve, developers are working on safeguards such as watermarking and authentication techniques to ensure ethical use. This innovation is paving the way for more natural and accessible voice synthesis while challenging the boundaries of AI-driven communication.

In this paper, we created a video in six languages, including English, Japanese, Chinese, German, and Swedish, that extracts synthesized numbers from input text and displays 20 sentences. We utilized English TTS technology to enable multi-national use.

2.3.4 Silhouette Customizing Technology

Silhouette customization is a technique that allows individuals to modify the overall shape, fit, or structure of a product to better suit their personal style or functional needs. This method is commonly used in the fashion and footwear industries, where customers can adjust the contour, proportions, and drape of clothing or the shape and structure of shoes.

In fashion, silhouette customization enables users to choose between different cuts, such as slim-fit, relaxed-fit, or oversized styles. For example, a customer ordering a tailored suit can specify whether they prefer a classic, modern, or tapered silhouette to match their body shape and aesthetic preferences. Similarly, in footwear, brands may offer options to alter the toe shape, heel height, or arch support, ensuring both comfort and style [7, 8].

This technique enhances personalization by giving customers control over the visual and functional aspects of a product. It also helps brands cater to diverse body types and style preferences, making fashion more inclusive. Whether for practical purposes, such as improving fit and comfort, or for aesthetic appeal, silhouette customization ensures that the final product aligns with the wearer's unique identity and needs.

Silhouette Customizing technology creates a composite image with natural gestures by generating various movements based on several human motion images taken as shown in Fig. 2, and produces a natural lecture by adding gestures to the speech content in the lecture.

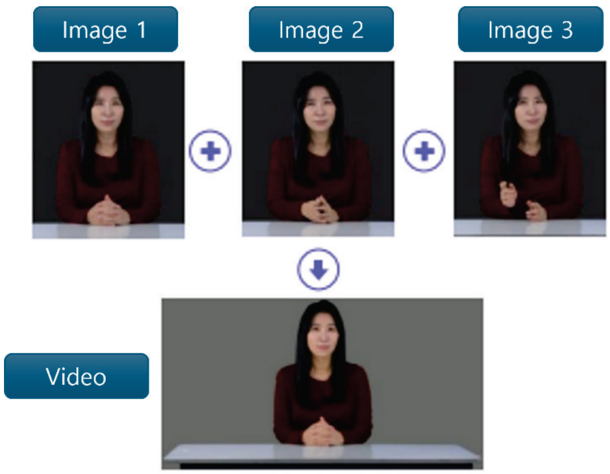


Figure 2 Silhouette Example

2.3.5 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are a type of artificial intelligence model used for generating highly realistic synthetic data, such as images, videos, and audio. A GAN consists of two neural networks: a generator and a discriminator, which work against each other in a competitive process. The generator creates fake data by mimicking real samples, while the discriminator evaluates whether the data is real or fake. Over time, both networks improve as the generator learns to produce increasingly convincing outputs, and the discriminator becomes better at detecting fakes. This adversarial training process enables GANs to generate high quality, realistic content. GANs are widely used in image enhancement, art creation, deepfake technology, and even medical research [13, 14]. However, they also raise ethical concerns, as they can be misused for creating deceptive content, such as fake identities and manipulated videos. As GANs continue to evolve, researchers are working on methods to detect and regulate their applications responsibly.

3 PRODUCTION TOOLS AND EDUCATIONAL CONTEXT PRODUCTION

3.1 System Architecture

Beyond simple TTV (Text To Video), it is possible to customize images through the definition of metadata (AI human, image, and text) related to image production. In addition, CNAISTUDIO TOOL is used to easily produce online lectures, making it easy to manage projects for each subject/time, and provides a convenient UI for creating and editing video scenes on a slide basis of lecture textbooks.

In addition, it is designed to customize TTS with the function of a 'smart vocabulary book' for technical terms and dialects of a specific study. Finally, Super Resolution technology is applied to produce all lecture images in high-definition images of Ultra HD (2160p, 4K).

The AI professor system comprises:

- Deepfake Engine: Synthesizes instructor facial visuals using GANs.
- Few-shot TTS: Produces realistic speech in multiple languages from minimal voice data.
- Lip-sync Module: Aligns spoken language and facial articulation.
- Gesture/Body Rendering: Generates natural postures and gestures.

3.2 Content Production Tools

AI professor was produced based on AI human technology. The AI professor, made with deep-fake technology, Lip-sync technology, few-shot TTS technology, and silhouette customizing technology, showed technology that is not much different from real people in one scene. After producing the AI teaching model, it was intended to produce educational content using the AI teaching model and video production technology. Subsequently, the AI teaching model can be used for auxiliary lectures on the regular curriculum, and guidance on bachelor's and AI technologies. After the introduction of AI professors, content can be produced cheaply without restrictions on dedicated personnel, equipment, and places, and choices such as background suitable for the situation and users, and content insertion can be made.

In addition, multilingual videos can be produced immediately in a total of five languages, excluding Korean, so that linguistic suggestions can be resolved to Chinese, Japanese, Swedes, Germans, and all English-speaking students. Finally, when producing lectures using AI human, we developed a tool that can be easily produced so that anyone can produce videos using AI human without separate education, and the tool can be used to produce the desired video immediately. According to Fig. 3, an AI human was produced after a professor at a university to produce an AI professor, and an educational image was produced and realized using a video production tool. It is possible to quickly produce additional lectures by guiding AI professors produced using AI technology.

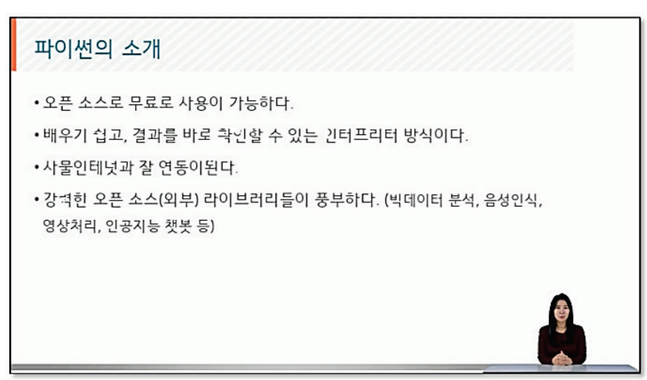


Figure 3 Lectures with human AI

We use CNAI Studio for scene scripting, voice cloning, animation, and rendering. Features include domain-specific vocabulary recognition, Ultra HD rendering, and accessibility settings for low-resource educators.

3.3 Lecture Design and Deployment

Content was structured around actual university courses, coding tutorials, and domain-specific instruction. Videos were produced in six languages. The modular design allows rapid adaptation to diverse curriculum requirements.

3.4 Evaluation Metrics

- Lip-sync Accuracy: Frame-aligned precision scoring at 93.1%
- TTS Quality: Mean Opinion Score (*MOS*) of 4.2/5
- Production Time: 60% time reduction compared to manual methods
- User Satisfaction: Mean score of 4.5/5 (n = 58), accessibility rating 4.7/5

4 RESULTS

AI-generated content was successfully deployed in online and hybrid courses. Participants reported improved learning engagement and comprehension due to native-language support. Visual quality and synchronization closely mimicked human instructors, achieving high realism.

This model contributes to three key areas:

- Pedagogical: Redefines instructional delivery beyond static MOOCs or live lectures.
- Technical: Combines AI synthesis methods into a unified, deployable pipeline.
- Practical: Offers a cost-effective, scalable, multilingual solution for higher education.

Limitations: The system currently lacks emotional interactivity and real-time adaptation. Ethical implications regarding authenticity and AI representation also merit further study.

5 CONCLUSION

This paper began with the focus of non-face-to-face online lectures changed due to the influence of the COVID-19 virus. For the necessity of non-face-to-face online classes and the resulting efficient learning effect, voices began to grow in the use of AI that introduced artificial intelligence, and AI professors were produced to create a better educational environment. Accordingly, with the production of AI professors, the production was carried out to make auxiliary videos of the regular curriculum created or to use them for bachelor's degrees and guidance. When producing AI professors, they record early using TTS technology, so if they finish recording, many videos can be produced using TTS technology, reducing time to record and edit themselves, and solving linguistic suggestions to foreign students taking classes. In addition, it is believed that the STUDIO Tool can be used to produce videos cheaply and easily without restrictions on dedicated personnel, equipment, and places, and that the video can be produced without separate education, making it less difficult for professors who produce videos. In addition, students who listen to education can quickly receive lectures, and so that you can focus on the time, you can focus on the time you can focus again. For these reasons, it is used in coding practice as high-quality education content, and it is used in coding practice.

This paper summarized the production of the AI teaching model, which was the beginning of the non-face-to-face education method, by removing the framework of the face-to-face education method and introducing artificial intelligence in line with the 4th industrial revolution. By producing online classes, we tried to supplement time-space constraints in offline classes, and lectures were produced with AI professors so that many lectures could be produced with less time. In addition, it is expected that the learning effect will be better than offline classes by allowing students to review the lecture when they want, and finally, AI professors using AI human can produce in five languages except Korean through TTS technology. Therefore, it is expected that international students studying abroad will be able to solve language inconveniences by solving difficulties that they did not understand due to their lack of Korean skills with TTS technology.

We developed and validated a scalable, multilingual AI professor model that enables autonomous lecture delivery with minimal human input. Empirical testing shows substantial gains in efficiency and accessibility. This approach is not a replacement for educators but a tool to enhance instructional reach, especially in multilingual and resource-limited environments. Explore emotional AI integration, immersive VR deployment, and empirical comparison across learning domains.

Acknowledgement

This research was supported by GPU resources provided through the 2025 High-Performance Computing Support Program, funded by the Ministry of Science and ICT and the National IT Industry Promotion Agency (NIPA) (G2025-0572).

6 REFERENCE

[1] Kim, D.-h. (2020). Changes and implications of the industrial and socioeconomic paradigm after the Corona crisis. *Interstate Finance Brief*, *29*(13), 16–18.

[2] Dwivedi, Y. K., Hughes, L., et al. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

[3] Oh, J.-h. (2020). The future ahead of COVID-19, from the era of education to the era of learning. *Issue & Diagnosis*, *421*, 1–25.

[4] Han, D.-e. (2021). Creating an interactive AI chatbot using dialog flow and applying it to elementary English classes.

Study on Learner-centered Curriculum Education, *21*(17), 517–529. https://doi.org/10.22251/jlcci.2021.21.17.517

[5] Kim, M.-j., & Lee, S.-k. (2022). Cyber threats and countermeasures to the development of artificial intelligence (AI) technology. *Journal of the Korean Communications Society (Information and Communication)*, *39*(12), 13–18.

[6] Han, S.-h., Choi, M.-i., & Cho, B. (2011). Real-time lip-sync and facial expression generation system for implementing interactive SNS. *Animated Studies*, *7*(2), 120–137.

[7] Hwang, Y.-h., & Kim, C.-s. (2021). Survey of non-face-to-face online lecture perception due to COVID-19: Focusing on university students' satisfaction and anxiety. *Linguistics*, *29*(1), 71–91.

[8] Kim, M. (2021). Post-Corona, video content production technology: Virtual Production System. *Imaging Technology Research*, *1*(35), 27–44.

[9] Seo, Y.-h., Oh, M.-s., & Han, K.-h. (2021). The present and future of digital humans. *Broadcasting and Media*, *26*(4), 72–81.

[10] Park, C.-s. (2022). Online education using digital human. *Comprehensive Academic Presentation Papers of the Korean Society of Information Technology*, ( ), 246.

[11] Kwon, E.-y. (2021). Trends in English teaching and learning research using AI. *English Studies*, *21*, 1313–1337.

[12] Bijker, W. E., Hughes, T. P., & Pinch, T. J. (1987). *The social construction of technological systems: New directions in the sociology and history of technology*. MIT Press.

[13] Saxena, D., & Cao, J. (2021). Generative Adversarial Networks (GANs): Challenges, solutions, and future directions. *ACM Computing Surveys*, *54*(3). https://doi.org/10.1145/3446374

[14] Chakraborty, T., Reddy, U. K. S., Naik, S. M., Panja, M., & Manvitha, B. (2024). Ten years of generative adversarial nets (GANs): A survey of the state-of-the-art. *Machine Learning: Science and Technology*, *5*(1). https://doi.org/10.1088/2632-2153/ad1f77

Authors' contacts:

Yangha Chun, Professor
School of Artificial Intelligence, University Yongin,
134 Cheoin-gu, Yongin-si, Gyeonggi-do, 17092, South Korea

Soo-Yeon Yoon, Professor
(Corresponding author)
School of Software Convergence, Kookmin University,
77 Jeongneung-ro, Seongbuk-gu, Seoul, 02707, South Korea
Mobile Phone: +82-010-9264-0193
1104py@kookmin.ac.kr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20250410052911

Original scientific paper
Received: 2025-04-10, Accepted: 2025-06-03

Convergence Methods for Practical Problem-Solving through the Generation of Diverse Ideas in the Semiconductor Industry: TRIZ & Design Thinking

Seongmin Seo, Yong-Won Song*, Jung-Hyeon Kim, Hong-Kyun Shim, Su-Yeon Ko

Abstract: This study proposes a practical approach to solving complex challenges in the semiconductor industry by integrating TRIZ, a systematic innovation methodology, with Design Thinking, a user-centered problem-solving framework. Utilizing this hybrid methodology, various solution ideas were generated and validated for industrial feasibility through a case study. Key problems were identified through direct observation of the work environment and empathy with field operators. These were then structured using function analysis and cause-effect chain analysis. Creative solutions were derived using the ARIZ-85C algorithm and grouped and refined based on their functional roles. Before full-scale application, prototypes were developed and tested to evaluate their feasibility and performance. To enhance worker safety, a redesigned powder trap was proposed, integrating a pulsing system to address issues such as hazardous odors and entrapment risks. The system's powder collection and removal functions were validated through prototype testing before being successfully implemented in an actual manufacturing environment. As a result, the required man-hours for maintenance operations were reduced from 6MH to 1MH, achieving an approximate 83% efficiency gain. Additionally, annual maintenance costs decreased by nearly 100 million KRW. The development and evaluation period was also significantly reduced from over two years to just three months. This case study confirms the practical value of integrating TRIZ and Design Thinking for addressing multifaceted problems in semiconductor manufacturing. Nevertheless, since the experiments were conducted under specific conditions, further validation across a broader range of tools and processes is necessary to generalize the findings.

Keywords: Convergence; Design Thinking; Semiconductor; TRIZ

1 INTRODUCTION
1.1 Research Background

With the recent proliferation of Generative AI, the demand for high-performance DRAM capable of handling large volumes of data at high speed has increased significantly. In response, the core component of DRAM—the capacitor—is evolving toward smaller and thinner structures while enhancing storage capacity. This advancement introduces new technical challenges in the semiconductor deposition process. As capacitor geometries become more intricate, it becomes increasingly difficult to uniformly deposit materials within narrow and deep trenches during the deposition phase. To address this, various chemical additives are used in the process, which often carry inherent risks such as flammability and explosiveness. After processing, by-products accumulate in the exhaust system as solids or liquids, and can react violently upon exposure to moisture. Currently, plasma treatment systems are employed to convert these hazardous by-products into solid powder or non-condensable gases. The solid powder is collected in a powder trap, and only the gas phase is discharged through the exhaust system (Fig. 1).

The powder collected in the trap is periodically removed using a vacuum cleaner through a designated clean port. However, in practice, this removal is often ineffective and frequently necessitates disassembling the filter and housing. These procedures release harmful substances and strong odors, requiring workers to wear gas masks. Moreover, repetitive operations in confined spaces elevate the risk of entrapment injuries, significantly compromising workplace safety.

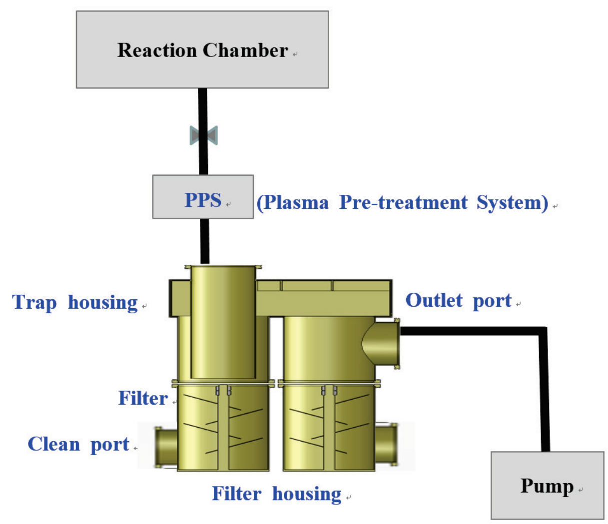


Figure 1 Configuration of the exhaust system and powder trap in DRAM deposition processes

1.2 Research Objectives and Methodology

The objective of this study is to address the issue of powder accumulation and removal in DRAM deposition processes through an integrated analysis from both technical and user-centered perspectives, and to derive creative yet practical solutions. To achieve this, the study employs a hybrid approach that combines TRIZ (Theory of Inventive Problem Solving), which provides systematic tools for contradiction analysis, with Design Thinking, a methodology rooted in empathy and user engagement. TRIZ was used to identify and structure technical contradictions and system-level problems, while Design Thinking contributed to concretizing viable ideas by reflecting actual user needs and on-site conditions. By organically integrating these two methodologies, the study seeks to establish a balanced problem-solving framework that connects technology-driven analysis with human-centered design.

1.3 Structure of the Paper

This paper is structured as outlined in Tab. 1.

Table 1 Structure of the paper

Chapter	Title	Contents
1	Introduction	Background, research objectives, methodology, and overall paper structure
2	Literature Review & Methodology	Overview of TRIZ and Design Thinking; analysis of their integration potential
3	Problem Analysis	Problem structuring through user empathy, function analysis, cause-effect chain analysis, and contradiction analysis
4	Idea Generation	Development of creative solutions based on TRIZ principles
5	Idea Evaluation and Validation	Idea grouping, establishment of evaluation criteria, prototyping, and testing
6	Conclusion and Future Work	Summary of findings, identified limitations, and directions for future research

2 LITERATURE REVIEW AND RESEARCH METHODOLOGY
2.1 TRIZ (Teorija Rezhenija Izobretatelskih Zadach)

TRIZ (Teorija Rezhenija Izobretatelskih Zadach, Theory of Inventive Problem Solving) is a problem-solving methodology developed by Genrich Altshuller and his colleagues. It analyzes patterns of invention and innovation to provide systematic tools and techniques. Using principles derived from numerous patents, TRIZ formalizes tools for creative problem-solving. Key TRIZ tools include the 40 Inventive principles, 76 Standard solutions, the Effects Database, Separation Principles, the Contradiction Matrix, the Laws of Technological System Evolution, the Ideal Final Result (IFR), Function Analysis, Substance-Field (Su-Field) Analysis, Resource Analysis, System Thinking and Function-Oriented Search (FOS), a new TRIZ-based tool. One study using the FOS methodology in the semiconductor industry developed a system to prevent explosion accidents caused by hydrogen emissions. The system replaced hydrogen with steam using a Hydrogen Plasma Treatment Unit (HPTU) and proposed an optimal non-explosive model based on hydrogen usage [1]. As shown in Fig. 2, TRIZ systematically organizes solutions based on problem types, enabling structured learning of creative problem-solving methods.

A case study applied problem-solving methodologies to address defects in semiconductor encapsulation, a post-processing stage in manufacturing. The issue involved leakage of the Epoxy Mold Compound (EMC), a molding resin, leading to production defects. A physical contradiction was identified by analyzing the problem as a technical contradiction. Applying the principles of separation between parts and the whole, the conventional pellet form of EMC was transformed into a powder form, allowing for low-viscosity melting while maintaining the original process temperature. This solution provided a concrete method to resolve EMC leakage issues and improve manufacturing stability [2].

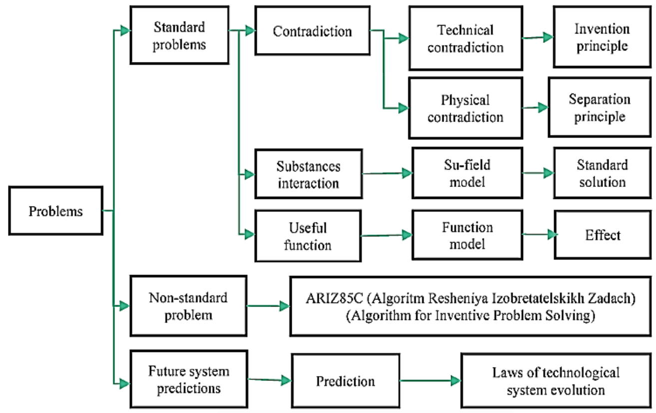


Figure 2 Solution based on the type of problem

2.2 Design Thinking

Design Thinking is a user-centered creative problem-solving methodology developed by Stanford University. It comprises five stages: Empathize, Define, Ideate, Prototype, and Test. Through a human-centered design approach, Design Thinking deeply understands users' needs and problems and develops innovative solutions based on this understanding (Tab. 2).

Table 2 Design thinking

Step	Explanation	Main activities
Empathize	It is the stage for deeply understanding the user's needs and problems.	Observations, Interviews, Surveys, Creating Empathy maps.
Define	Based on the insights gained in the Empathize stage, a clear problem is defined.	Write a problem Statement, Personas, Create a problem Definition document.
Idea Generation	Explore various solutions and generate creative ideas.	Brainstorming, Mind Mapping, SCAMMPER.
Prototype	Create low-fidelity prototypes, develop storyboards, and design wireframes.	Create low-fidelity prototypes, Develop storyboards, Design wireframes.
Test	Test the prototype with real users and gather feedback.	Collect user feedback, Make iterative improvements, Conduct user experience (UX) testing.

After the Ideation stage of Design Thinking, it is crucial to systematically organize the numerous ideas, group similar ideas (Clustering), and derive integrated solutions. The process of idea grouping involves visualizing ideas, identifying similarities, deriving themes, setting priorities, and integrating solutions. Benefits such as reduced complexity, enhanced collaboration, and improved efficiency can be achieved through idea grouping. The key aspect of this methodology is the iterative process, where insights gained in each stage may lead to revisiting the Empathize stage, modifying the problem definition, or restructuring ideas. This process has been applied in various fields, including the design development of water purifier products [3], self-diagnosis and remote healthcare endoscope camera products [4, 12], and baby toothbrush product design development [5].

2.3 TRIZ & Design Thinking

In highly technology-intensive industries such as semiconductor manufacturing, problems frequently emerge from the convergence of technical complexity and user-centered constraints. Addressing such multifaceted challenges requires a methodological framework that incorporates both technical and human dimensions. This study proposes a hybrid problem-solving framework that integrates TRIZ, a systematic approach for resolving technical contradictions, and Design Thinking, a user-centered methodology focused on empathy and iterative refinement. TRIZ supports logical and analytical problem-solving through tools such as function analysis, cause-effect chain analysis, and contradiction resolution, which are particularly effective in structuring and resolving engineering problems. On the other hand, Design Thinking emphasizes user observation, experience, and empathy, offering a practical path for generating ideas that are both creative and grounded in real-world needs. The complementary strengths of these two approaches enable the development of solutions that are not only inventive but also applicable in practice [11].

The proposed framework consists of seven sequential stages that span the entire problem-solving process from exploration to deployment in industrial settings: The first step, Empathy, involves direct observation of the workplace and interviews with users to gain deep insight into on-site problems. This step, central to Design Thinking, helps uncover human-centered issues that are often overlooked in purely technical analyses. The second step, Technical Analysis, utilizes TRIZ techniques including function analysis, cause-effect chain analysis, and contradiction resolution to identify structural causes and define technical contradictions and ideality targets within the system. The third step, Problem Definition, integrates insights from the empathy and analysis phases to articulate a clear and precise problem statement. This step represents the core integration point of TRIZ and Design Thinking, enabling a nuanced definition that simultaneously reflects technical constraints and user needs. In the fourth step, Idea Generation, creative solutions are developed using TRIZ's inventive principles and standard solutions. These ideas are further refined through Design Thinking's grouping and integration methods to produce solutions that are both functionally robust and user-friendly. The fifth step, Idea Evaluation, applies TRIZ-based criteria such as degree of ideality, contradiction resolution, and applicability to assess the proposed solutions. While Design Thinking's SCAMPER technique is used as a supplementary tool, this study primarily employs TRIZ's quantitative evaluation framework to ensure rigor and consistency. The sixth step, Prototyping, involves the physical implementation of the most promising ideas into testable prototypes. This step reflects Design Thinking's experimental ethos and serves as a pre-implementation validation step to assess practical feasibility. The final step, Test and Implementation, places the prototype in real industrial settings for iterative testing under actual process conditions. Feedback from users is incorporated to refine the solution into its final, deployable form. This integrated problem-solving framework provides a balanced approach that unites technology-driven analysis with human-centered design. It offers a practical and replicable methodology for solving complex problems in high-risk, high-precision environments such as semiconductor manufacturing, where both technical performance and user requirements must be met simultaneously (Tab. 3).

Table 3 Integrated framework of TRIZ & Design thinking

Step	Key Activity	Methodology	Objective
1. Empathy	Workplace observation, user interviews	Design Thinking	Understand user needs and field-level issues
2. Technical Analysis	Function, cause-effect chain, and contradiction analysis	TRIZ	Identify technical structure and contradictions
3. Problem Definition	Integrate user and technical problem insights	Integrated	Define the core problem
4. Idea Generation	Generate, grouping, and integrate creative ideas	Integrated	Develop ideas with both functional and user focus
5. Idea Evaluation	Evaluate ideality, contradiction resolution, and Applicability	TRIZ	Select the optimal solution
6. Prototype	Design and build prototypes	Design Thinking	Verify practical applicability
7. Test & Implementation	Conduct user testing, incorporate feedback, finalize solution	Design Thinking	Apply and scale the final solution in practice

Research cases that combine TRIZ and Design Thinking demonstrate an approach that combines the strengths of both methodologies for innovative problem-solving. A study that improved a solar wood dryer showed significant performance differences depending on the weather by integrating TRIZ (Theory of Inventive Problem Solving) and Design Thinking. In this study, they successfully developed a biomass dryer with excellent performance unaffected by weather conditions [6]. The innovative Design Thinking process alongside TRIZ included methods for selecting attractive and trendy social and business issues, providing a way to discover problems and build an organization's innovation culture more effectively. This increased participants' passion and commitment to innovation [7]. Research also supports users in systematically establishing innovation projects by integrating Design Thinking and TRIZ. This study combines TRIZ's systematic problem-solving methods with the creative thinking process of Design Thinking, presenting methodological tools to drive more effective innovation [8].

These studies demonstrate that better problem-solving and innovation can be achieved by leveraging the complementary characteristics of both methodologies.

3 PROBLEM ANALYSIS
3.1 Empathize

Through observations of the powder removal worksite, it was found that the primary workers and the assistants had to wear gas masks due to the poor working conditions. A large amount of powder was scattered around the work area, causing unpleasant odors. All workers reported discomfort, and significant time was spent cleaning the area after the work. In interviews with actual workers, 80% considered powder removal the most difficult task, while the remaining 20% found the cleanup process challenging. Specifically, all respondents reported discomfort when wearing gas masks during work. In interviews regarding improvements to the working environment, 80% expressed a desire for a work environment where gas masks would not be required, while 20% wanted the powder scattering issue to be addressed. Based on this feedback, the decision was made to improve the powder trap structure so that powder removal work could be performed safely without gas masks or powder scattering. This improvement is expected to greatly enhance the working environment and increase worker safety and efficiency.

3.2 Function Analysis

As illustrated in Fig. 3, during normal operation, powder flows through the inlet port (A). Heavier particles are partially captured by the filter, while others fall to the bottom of the filter housing (B). Lighter powder continues to flow through the trap housing, with some accumulating on the filter and some settling at the bottom of the housing (C).

Among the powder that settles at the bottom of the filter housing, a portion tends to accumulate on the side opposite the clean port (D). As the amount of accumulated powder increases, the system is shut down, the clean port is opened, and the powder is removed using a vacuum cleaner (E).

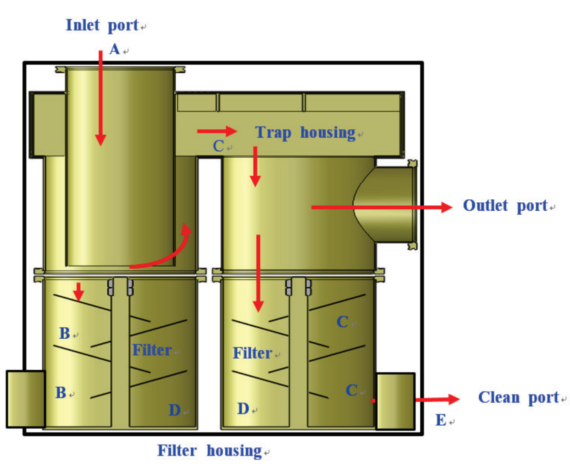


Figure 3 Operational description of a powder trap

When opening the clean port, a gas mask must always be worn. However, as shown in the following functional diagram (Fig. 4), vacuum cleaning does not effectively remove the powder, requiring the disassembly of the filter and filter housing. This process requires wearing a gas mask for an extended period, which is physically demanding and presents a risk of discomfort due to the confined space.

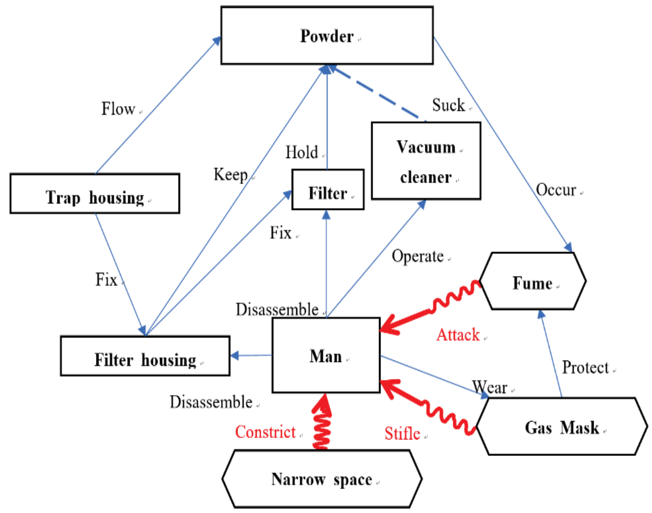


Figure 4 Function diagram of disassembly operation

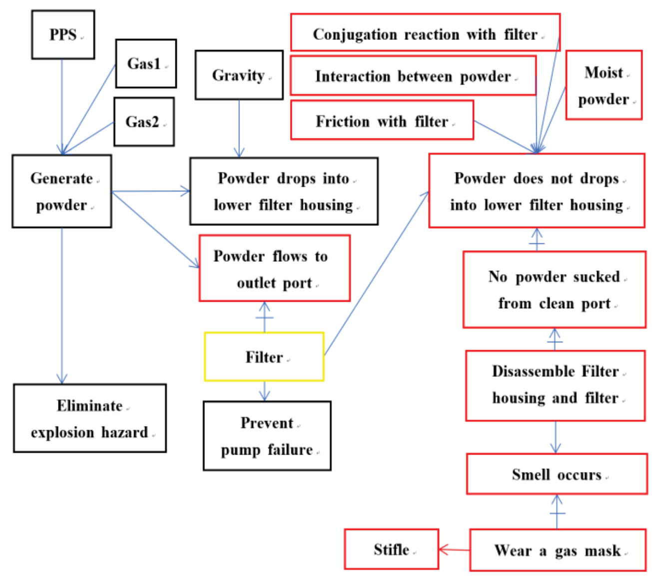


Figure 5 Cause-effect chain analysis

3.3 Cause-effect Chain Analysis

As illustrated in the cause-effect chain analysis in Fig. 5, powder is generated to mitigate the explosion risk by introducing byproduct and oxidation gases into the PPS after the process. However, the generated powder is intended to fall and be captured at the bottom of the filter housing; instead, some of it flows into the pump through the outlet port, leading to equipment failure. To address this issue, an additional filter was installed to prevent the powder from entering the pump. However, the powder does not flow into the lower part of the filter housing, resulting in the inconvenience of having to disassemble both the filter housing and the filter during the removal process. This task generates odors from harmful substances, requires wearing a gas mask, and poses a risk of discomfort due to the confined working space.

3.4 Contradiction Analysis

Adding a filter to prevent the powder from flowing into the outlet port addressed the pump failure issue. However, it introduced the problem of requiring a gas mask and performing difficult tasks during removal process because the powder does not flow into the bottom of the filter housing. While a filter is necessary to prevent the powder from entering the outlet port, there is a physical contradiction: the filter must not be present for the powder to flow into the filter housing. Analyzing the physical contradiction at a micro level, the filter must be located on the outer walls, where the powder flows, to prevent it from reaching the outlet port. At the same time, for the powder to flow into the filter housing, the filter in the center must be absent (Fig. 6).

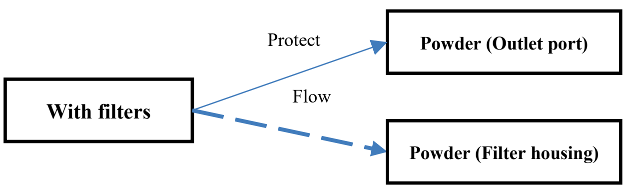


Figure 6 Technical contradiction (TC1)

4 PROBLEM SOLVING IDEAS

4.1 Contradiction Solving Idea

The problem is modeled using 39 engineering parameters to resolve the technical contradiction. The proper parameter is productivity, as preventing pump failure contributes to increased productivity, while the harmful parameter is the ease of maintenance, as the powder removal process has become more difficult. Among the inventive principles recommended in the contradiction matrix, the "preliminary action" principle led to adding a vibration motor to reduce vibration before the powder accumulates on the filter (Fig. 7). As a solution to the physical contradiction, the concept of applying spatial separation was developed, where the filter is removed from the center of the filter housing, and a filter is installed on the outer walls, which are the powder flow passages (Fig. 8). The arrows indicate the direction of powder flow.

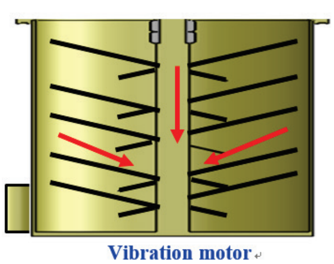


Figure 7 Add Vibration motor (Inventive principle)

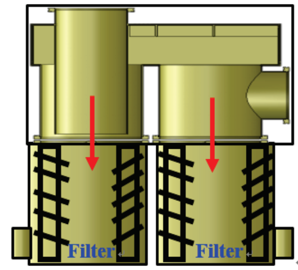


Figure 8 Physical separation of the filter area

4.2 IFR (Ideal Final Result)

The tool's goal is to solve the problem without increasing the system's complexity or causing additional harmful effects. The issue arises when the filter remains at the center due to the friction of the powder, moisture, reactions between the powders, and bonding reactions, preventing it from moving to the lower part of the filter housing. To resolve this problem, the filter's angle can be sharply adjusted so that it does not retain the powder, effectively addressing the issue of the powder not flowing to the lower part of the filter housing (Fig. 9). The arrows indicate the direction of powder flow.

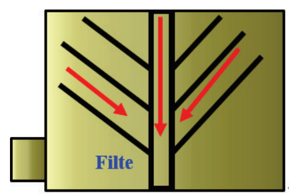


Figure 9 Steeply inclined filter

4.3 Trimming (Eliminate Functionally Unnecessary Elements)

If the functional elements can transfer their functions to other surrounding elements, the functional elements can be removed. The function of allowing the powder to flow to the lower part of the filter housing and preventing it from flowing into the outlet port, which is currently performed by many filters in the center, can be transferred to the filter housing. This would allow for removing the many filters in the center (Fig. 10). The arrows indicate the direction of powder flow.

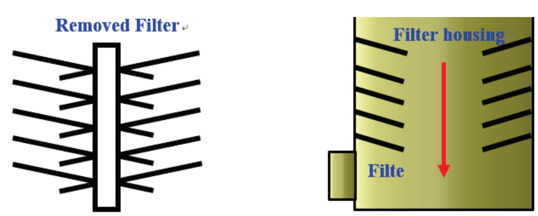


Figure 10 Integrated the filter function into the filter housing

4.4 Take a Step Back from IFR

By minimizing the number of filters in the center, only part of the function of blocking the powder flow to the outlet port is performed while allowing the powder to be effectively directed into the bottom of the filter housing. The issue of powder flow into the outlet port is addressed by installing a barrier wall (A) (Fig. 11). The arrows indicate the direction of flow of the gas and powder.

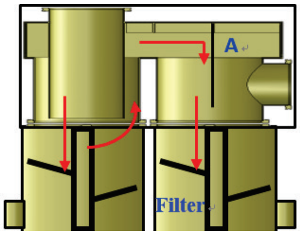


Figure 11 Streamlined the filter design

4.5 Standard Solution

To solve the issue of powder removal using a vacuum cleaner, Nitrogen is used from the upper system to apply mechanical pressure to the powder and filter. By injecting it into the powder that did not flow to the bottom, the powder flows into the filter housing and then into the filter, causing the filter to vibrate. Nitrogen is supplied in a pulsed manner to facilitate the movement of the powder (Fig. 12). The arrows indicate the direction in, which nitrogen and powder flow in the same direction.

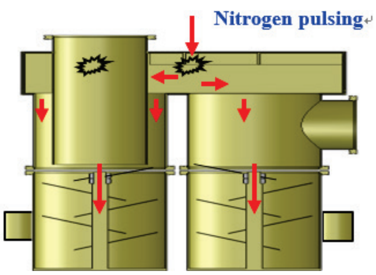


Figure 12 Add Nitrogen pulsing

Powder that does not flow to the bottom of the filter housing due to friction with the filter is made to flow naturally to the bottom by adding ultrasonic to the filter, utilizing gravity and the direction of the filter (Fig. 13).

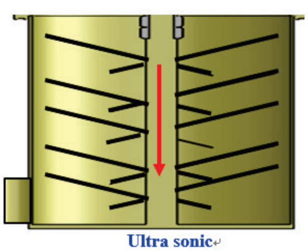


Figure 13 Add Ultra-sonic

Powder that does not flow to the bottom due to its moist characteristics can be made to move downward by adding heat during the drying process, preventing it from adhering to the surface and allowing it to flow downward (Fig. 14).

Due to interactions between powder particles, some powder does not flow downward as expected. To address this, a non-reactive external agent is introduced to facilitate separation and induce downward movement of the remaining powder (Fig. 15).

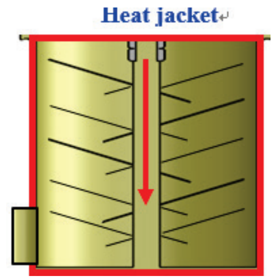


Figure 14 Add Heat jacket

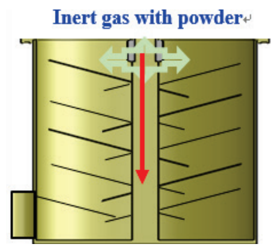


Figure 15 Add inert gas

Powder that does not flow to the bottom due to bonding reactions with the filter surface can be made to flow naturally by coating the filter surface and minimizing the reaction time, allowing gravity to carry the powder downward (Fig. 16).

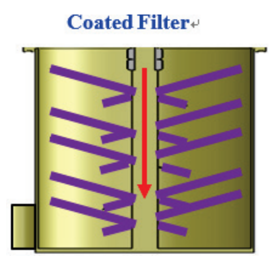


Figure 16 Coated filter

4.6 Void (Reserve as Empty Space)

An empty space is introduced to eliminate the issue of powder not flowing to the bottom of the filter housing. By leaving the filter empty, the powder flows effectively to the bottom of the filter housing, and the function of blocking the powder flow to the outlet port is reinforced by installing a barrier wall (A) (Fig. 17). The arrows indicate the flow direction of the gas and powder.

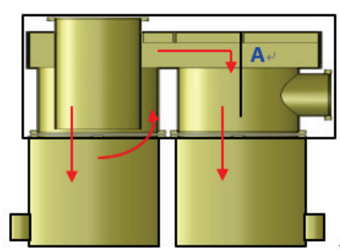


Figure 17 Omitted the filter installation

4.7 FOS (Function Oriented Search)

The problem is defined as the failure of powder to concentrate at the bottom of the filter housing. The required function definition to search for similar functions in major industries is that the powder accumulates in the housing and does not move outward. References include a paper on the optimization design of dust collector inlets for separating powder from gas (air) [9] and a patent on centrifugal filters and dust collectors [10]. Cyclone technology (A), which applies centrifugal force and rapidly rotating gas or powder flows, was used based on this. Using the cyclone method, the powder naturally concentrates on the outer wall (B) (Fig. 18). The arrows indicate the flow direction of the gas and powder.

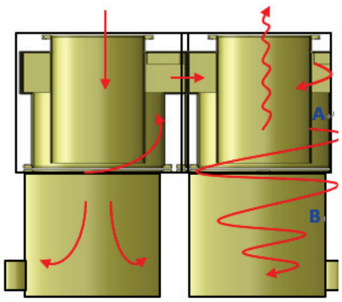


Figure 18 A Structure with an integrated cyclone function

5 EVALUATE IDEAS AND SOLVE PROBLEMS

5.1 Concept Solution

The various problem-solving solutions identified above are summarized in Tab. 4.

Table 4 Concept solution

NO	Tool	Concept solution
1	Inventive principle	By applying the inventive principle of "preliminary counteraction," a vibrating motor is introduced to prevent the powder from coming into contact with the filter in advance.
2	Separation	Install a filter on the outer wall of the primary side flowing toward the outlet and remove the central filter.
3	IFR	Increase the inclination of the filter.
4	Trimming	Introduce obstacles to the filter housing and remove the central filter.
5	Take a step back From IFR	Place the filter at a minimal angle and install a blocking wall in the trap housing leading to the outlet port. Reinforce it to prevent powder flow.
6	Standard solution	By adding nitrogen, mechanical pressure is applied to the powder and filter.
7	Standard solution	Powder that exists due to friction with the filter is resolved by introducing ultrasound.
8	Standard solution	Add heat (thermal field) to activate the movement of moist powder.
9	Standard solution	Powder interactions are resolved by adding a substance that does not react with the powder, using a chemical field.
10	Standard solution	Powder bonding with the filter surface is resolved by applying a coating, using a mechanical field.
11	Void	Leave the filter section as an empty space and install a blocking wall in the trap housing. Reinforce this area to prevent powder flow to the outlet port.
12	FOS	By modifying the flow of gas and powder using the cyclone method, the powder naturally concentrates on the outer wall, and the filter in the middle is removed.

5.2 Idea Grouping

The concepts of IFR, Trimming, "Take a Step Back" from IFR, Void, FOS, and Separation are implemented to ensure efficient powder flow into the filter housing while preventing any flow into the outlet port. Given that these solutions necessitate structural modifications, they have been incorporated into a new conceptual design for the powder trap structure, as presented in Tab. 5.

Table 5 Idea grouping and integration

Principle	Separation, IFR, Trimming, Take a step back from IFR, Void, FOS.
IDEA	The centrally located filter is removed, and an obstacle with a steep inclination is installed on the outer wall of the filter housing. The flow direction within the trap housing is altered, and the outlet port is repositioned to the top. The flow of gas and powder is redirected to follow a cyclone flow method. (The filter is no longer located at the center but is instead positioned on the outer wall.)

5.3 Evaluate Ideas

Among the twelve initially generated ideas, a total of seven were retained after grouping and integration. These included both functionally integrated concepts and ideas derived from TRIZ inventive principles and standard solutions. All creative ideas were generated using the TRIZ methodology. For idea evaluation, a 5-point scale was employed based on three TRIZ-defined criteria: ideality, contradiction resolution, and applicability. Based on the total evaluation scores, two ideas were selected for implementation: a novel powder trap concept and a solution involving the introduction of nitrogen gas (Tab. 6).

Table 6 Idea evaluation

NO	Principle	Idea	Idea evaluation (5-point scale)			
			Ideality	Contradiction resolution	Applicability	Score
1	Integrated	Modified powder trap and filter.	5	5	5	15
2	Inventive principle	Add vibration motor.	3	3	3	9
3	Standard solution	Add nitrogen.	4	4	5	13
4	Standard solution	Add ultrasound.	3	3	3	9
5	Standard solution	Add thermal.	3	3	3	9
6	Standard solution	Add substance that does not react with the powder.	3	3	3	9
7	Standard solution	Add coating.	3	3	3	9

5.4 Prototype

A prototype was developed to test and evaluate the new powder trap concept, which was selected during the idea integration process. The central filter was removed, and a plate, which functions as a filter, was placed at a steep angle on the outer wall, following the powder's flow direction (A). The section before the outlet port was modified to a cyclone method (B), allowing the powder to be collected on the outer wall. Arrows indicate the direction of gas flow, while circular particles represent the form of powder collection. To address the possibility of incomplete powder removal, a pulsing system was integrated into the upper system, using nitrogen to apply strong pressure to both the powder and the filter, as illustrated in Fig. 19.

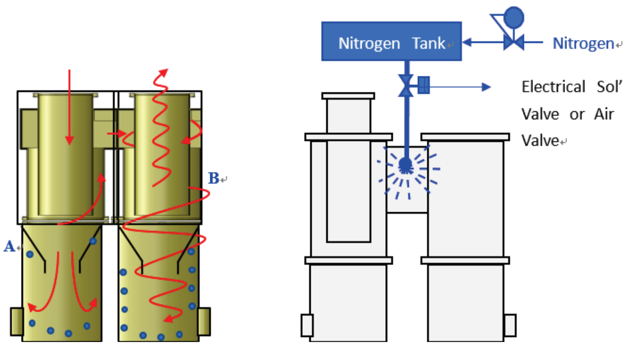


Figure 19 New concept powder trap and nitrogen pulsing system

5.5 Test

To quantitatively validate the performance of the designed powder trap and nitrogen pulsing system, two experiments were conducted under conditions simulating an actual production environment. The first experiment focused on evaluating powder collection performance, while the second assessed the powder removal efficiency using the nitrogen pulsing system. In the powder collection experiment, 1000 g of actual process powder was applied at point (A) of the trap (Fig. 21). A vacuum was then generated using a pump with a capacity of 100,000 liters per minute, and 10 liters of nitrogen gas was injected. The system was maintained in this state for 30 minutes, after which the remaining powder mass was measured. A total of 30 repeated trials were conducted (Tab. 7), resulting in an average collection efficiency of 90.35%, with a standard deviation of 0.49%. The minimum and maximum values were 89.56% and 91.72%, respectively. The results were closely distributed around the mean, indicating high reproducibility and experimental reliability. In the powder removal experiment, a vacuum cleaner was connected to the clean port, and nitrogen was instantaneously released once the supply tank reached a pressure of 60 psi by opening a pneumatic valve. The nitrogen supply line had a diameter of 25 mm (B), and the powder application method and quantity were consistent with the collection experiment. Based on 30 repeated trials (Tab. 7), the average removal efficiency was 91.51%, with a standard deviation of 1.01%. The removal efficiency ranged from 90.20% to 95.00%, with over 90% efficiency achieved in the vast majority of tests (Fig. 20). These results confirm that the nitrogen pulsing system delivers high-efficiency and repeatable performance in powder removal. Overall, the experimental results demonstrate that both the powder collection and removal mechanisms meet or exceed the performance requirements for industrial applications. The design has been experimentally validated as stable and reliable. Furthermore, to enhance maintenance accessibility and ensure operator safety, the clean port was finalized using a double-cap structure (C) (Fig. 21).

Table 7 Test result

Powder trapping				Powder elimination			
No	Before (g)	After (g)	Efficiency (%)	No	Before (g)	After (g)	Efficiency (%)
1	1005	905	90.0	1	1003	85	91.5
2	1001	906	90.5	2	1002	86	91.4
3	1003	903	90.0	3	1003	75	92.5
4	1002	911	90.9	4	1000	86	91.4
5	1004	910	90.6	5	1005	95	90.5
6	1005	905	90.0	6	1000	96	90.4
7	1002	903	90.1	7	1000	90	91.0
8	1003	920	91.7	8	1005	85	91.5
9	1002	901	89.9	9	1000	80	92.0
10	1001	905	90.4	10	1000	98	90.2
11	1000	910	91.0	11	1004	80	92.0
12	1000	912	91.2	12	1000	84	91.6
13	1000	905	90.5	13	1000	85	91.5
14	1005	907	90.2	14	1003	90	91.0
15	1003	901	89.8	15	1000	50	95.0
16	1002	910	90.8	16	1000	90	91.0
17	1002	904	90.2	17	1001	96	90.4
18	1000	906	90.6	18	1001	72	92.8
19	1000	902	90.2	19	1000	85	91.5
20	1000	905	90.5	20	1002	90	91.0
21	1005	903	89.9	21	1002	75	92.5
22	1005	902	89.8	22	1005	85	91.5
23	1006	901	89.6	23	1000	95	90.5
24	1002	903	90.1	24	1000	77	92.3
25	1001	905	90.4	25	1005	67	93.3
26	1002	901	89.9	26	1000	92	90.8
27	1002	903	90.1	27	1000	95	90.5
28	1000	911	91.1	28	1003	85	91.5
29	1002	903	90.1	29	1000	92	90.8
30	1000	901	90.1	30	1004	90	91.0

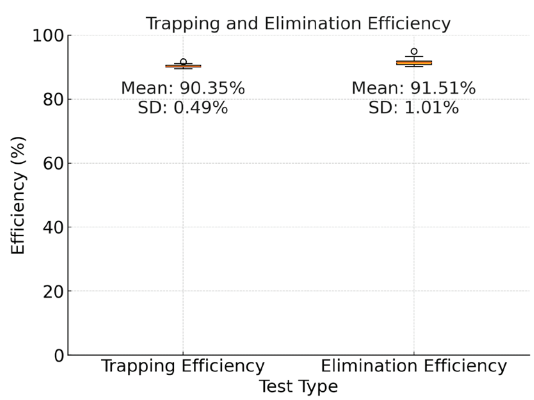


Figure 20 Test result

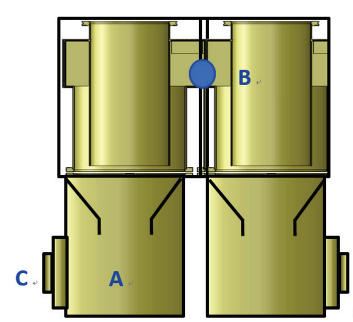


Figure 21 Double cap

5.6 Apply

The new concept of the powder trap was implemented in the industry. Powder collection efficiency was enhanced, and the issue of powder flow into the outlet port was resolved. The pulsing system effectively eliminated any powder that was not fully removed.

6 CONCLUSION AND FUTURE WORK

This study applied an integrated problem-solving framework that combines TRIZ and Design Thinking to address recurring and complex challenges observed in semiconductor manufacturing environments. To evaluate the feasibility of implementation in real process settings, we conducted field observation and empathy-based problem identification, performed structural analysis using function and cause-effect chain analysis, and generated creative solutions using the ARIZ-85C algorithm. As a result, the developed system significantly reduced labor effort, enabling a task that previously required three workers for two hours to be completed by two workers in just 30 minutes, achieving a total reduction of 5.5 man-hours. Additionally, the one-time preventive maintenance (PM) cost was reduced from 2.1 million KRW to 100,000 KRW, leading to an estimated annual maintenance cost saving of approximately 100 million KRW. These figures are based on specific equipment and maintenance scenarios, providing a quantitative basis for the cost-reduction potential. Moreover, the system also contributed to reduced exposure to hazardous substances and the prevention of entrapment accidents, indicating improved worker safety and potential for industrial accident prevention.

The idea grouping and integration process also contributed to a reduction in the idea evaluation period from two years to three months, while eliminating functional redundancy and minimizing trial-and-error. These outcomes confirm that the fusion of TRIZ and Design Thinking can serve as a practical and repeatable problem-solving methodology in high-risk, high-complexity industrial settings.

However, this study has several limitations. First, some experiments were conducted under simulated rather than full-scale process conditions, limiting the precision of powder collection validation. Future work should consider real-time monitoring or long-term tracking systems. Second, the applicability of the proposed solutions was tested under limited equipment and process settings. Additional validation across various semiconductor tools and powder types is needed. Third, while operator feedback was partially considered, a systematic usability evaluation and iterative refinement process were not fully implemented. Future studies should incorporate structured user-centered design approaches.

Despite these limitations, this study provides an empirical demonstration of how integrating technical analysis with user-centered design can effectively address complex industrial problems. The proposed framework is expected to be extended and validated in other high-risk industrial domains beyond the semiconductor sector, enhancing both its generalizability and practical value.

7 REFERENCES

[1] Seo, S., & Song, Y.-W. (2021). A study on the safe hydrogen exhaust method in the semiconductor industry. *Asia-Pacific Journal of Convergent Research Interchange*, 7(6), 1–10. https://doi.org/10.47116/apjcri.2021.06.01

[2] Ko, S.-Y., Song, Y.-W., Shim, H.-K., Jeon, Y.-M., & Seo, S. (2023). Solving reliability problems of transfer molding press equipment using TRIZ contradictions. *Asia-Pacific Journal of Convergent Research Interchange*, 9(9), 39–52. https://doi.org/10.47116/apjcri.2023.09.04

[3] Oh, I.-K., & Lee, D.-B. (2021). A study of a water purifier through product design convergence of design thinking methodology. *The Korean Society of Science & Art*, 39(2), 249–259. https://doi.org/10.17548/ksaf.2021.03.30.249

[4] Choi, J., & Chung, S. (2018). A study on product design development of intraoral camera for self-diagnosis and telemedicine. *Journal of Integrated Design Research, 17*(3), 45–56. https://doi.org/10.21195/jidr.2018.17.3.004

[5] Jung, S.-G., & Cha, S.-W. (2019). The study of children toothbrush product industry development through convergence of design thinking methodology. *The Korean Society of Science & Art*, 37(3), 415–427. https://doi.org/10.17548/ksaf.2019.06.30.415

[6] Adiyanto, O., Mohamad, E., Jaafar, R., Faishal, M., & Supriyanto, S. (2023). Design of biomass fired dryer using integrating design thinking and TRIZ method. *Journal of Advanced Manufacturing Technology* (JAMT, 17(6). https://jamt.utem.edu.my/jamt/article/view/6461/4003

[7] Lee, K. (2018). Innovative design thinking process with TRIZ. In Cavallucci, D., De Guio, R., & Koziołek, S. (Eds.), Automated Invention for Smart Industries (TFC 2018). IFIP *Advances in Information and Communication Technology*, 541. Springer, Cham. https://doi.org/10.1007/978-3-030-02456-7_20

[8] García-Manilla, H. D., Delgado-Maciel, J., Tlapa-Mendoza, D., Báez-López, Y. A., & Riverda-Cadavid, L. (2019). iIntegration of design thinking and TRIZ theory to assist a user in the formulation of an innovation project. In Cortés-Robles, G., García-Alcaraz, J., & Alor-Hernández, G. (Eds.), *Managing Innovation in Highly Restrictive Environments. Management and Industrial Engineering*. Springer, Cham. https://doi.org/10.1007/978-3-319-93716-8_14

[9] Jung, Y.-J., Jeong, M.-H., Park, K.-W., Hong, S.-G., Lim, K.-H., Suh, H.-M., & Shon, B.-H. (2012). A study on numerical calculations of multi-stage dedust system coupled with the collection principle of cyclone, inertial impaction, and bag filter: optimized design of dedust inlet. *Korea Association of Industry and Academic Technology*, 367–370. https://koreascience.kr/article/CFKO201230533389833.pdf

[10] Park, H.-S., & Lim, K.-S. (2012). *Centrifugal filter dust collector with filter abrasion prevention guide (cyclone bag house equipped with filter protection guide)*. Patent No. 101140670. https://patentimages.storage.googleapis.com/4a/b5/ce/4ae908ea2e407d/KR101140670B1.pdf

[11] Li, Y., Liu, Z., Guo, L., & Tian, X. (2015). Innovative design on family exercise equipment for dancing leg stretching based on TRI. *International Journal of Signal Processing, Image Processing and Pattern Recognition, 8*(3), 313–324.

[12] Fathi, I. S., Ali, A. M., Makhlouf M. A., & Osman E. A. (2021). Compression techniques of biomedical signals in remote healthcare monitoring systems: A comparative study.

International Journal of Hybrid Information Technologies, 1(1), 33-50. https://doi.org/10.21742/IJHIT.2021.1.1.03

Authors' contacts:

Seongmin Seo, PhD
1) Department of Industrial and Management Engineering, MyongJi University,
116 Myongji-ro, Cheoin-gu, Yongin-si, Gyeonggi-do, 17058, Korea
2) Tech Master, SK Hynix Semiconductor Co., Ltd.,
2091, Gyeongchung-daero, Bubal-eup, Icheon-si, Gyeonggi-do, Korea
miniseo0101@naver.com

Yong-Won Song, Professor
(Corresponding author)
Department of Nano & Semiconductor Engineering, Tech University of Korea,
237, Sangidaehak-ro, Siheung-si, Gyeonggi-do, 15073, Korea
ywsong@tukorea.ac.kr

Jung-Hyeon Kim, Professor
Industry-Academic Cooperation Foundation,
Kumoh National Institute of Technology,
61 Daehak-ro (yangho-dong), Gumi, Gyeongbuk, 39177, Korea
trizkim@kumoh.ac.kr

Hong-Kyun Shim, Tech Master
SK Hynix Semiconductor Co., Ltd.,
2091, Gyeongchung-daero, Bubal-eup, Icheon-si, Gyeonggi-do, Korea
shimdnbb@naver.com

Su-Yeon Ko, Engineer
SK Hynix Semiconductor Co., Ltd.,
2091, Gyeongchung-daero, Bubal-eup, Icheon-si, Gyeonggi-do, Korea
arijjin@naver.com

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20250430172617

Original scientific paper
Received: 2025-04-30, Accepted: 2025-05-21

Automation for Patient Medical Records in an Integrated Clinic Geographic Information System

Lely Prananingrum, Teuku Salman Farizi, Fajar Agus Dwi Rahmawan, Ilmiyati Sari*

Abstract: Majority clinics are dependent on traditional technique of managing medical records patients. This consequently makes the operations less efficient and patients wait longer. The objective of this study is the implementation of an electronic clinic management system with the introduction of the geographical information system (GIS) using the grounded theory, rapid application development (RAD) and agile approaches respectively. it includes GIS applications for clinic location finding, user account management processes such as login and registration, patient information management and reports, system support, and report generation. The RAD approach was utilized to fast-track development processes employing iterative patterns and the Agile approach ensured that the system suited the users' needs as they changed over time. The results have shown that this automated clinic information system is efficient in managing healthcare information by making more efficient use of existing information resources, balancing timeliness with processing time, and enhancing end-user satisfaction.

Keywords: agile; automation; GIS; grounded research; medical; record

1 INTRODUCTION

Improving the efficiency and quality of health services in clinics has become increasingly important along with the rapid development of information technology. One of the crucial elements in health services is the management of patient medical records, which act as a basis for making appropriate and accurate medical decisions. Medical records are important documents owned by health care facilities and function to record all information related to the patient's condition and treatment. Medical record documents must be managed properly, because health facilities are responsible for their integrity, security, and protection against loss, damage, forgery, or use by unauthorized parties. Although medical record documents are owned by health facilities, the contents of the medical records belong to the patient. The contents of these medical records consist of patient identity, results of physical and supporting examinations, diagnosis, treatment, follow-up plans, and signatures of health workers who provide further services. Information in medical records can be conveyed to patients, their families, or other parties who have received the patient's consent, except in certain conditions such as underage patients or in emergencies [1].

However, in practice, a number of clinics still use manual methods in managing medical records. This condition raises various problems, including difficulty in searching for data, duplication of information, and the risk of losing documents due to limited storage media. A clinic is a health service facility that provides basic or specialist medical services. Based on the type of service, clinics are divided into primary clinics, which provide basic medical services, and principal clinics, which provide specialist medical services or a combination of the two. These clinics can be owned by the government, local government, or the community, and can focus on a particular field based on a particular branch of science or organ system [2].

The manual recording system used by most clinics in recording medical records often causes problems such as data duplication, recording errors, and difficulties in searching for patient data. The service flow that includes registration, examination, and administration is often hampered by the slow and error-prone manual recording process. In its operations, the clinic must also document patient data and manage medical record files to provide administrative and medical information. Medical records are made in electronic form to prevent document loss, save storage space, and avoid inconsistencies in filling [3].

Therefore, automation of medical records through information technology is a relevant solution. Electronic medical records can overcome these problems by avoiding inconsistencies and loss of documents, as well as saving storage space. In addition, GIS, or Geographic Information Systems, is an application that processes spatial data using a computerized system, combining graphic data with object attributes through digital base maps that refer to the earth's surface [4]. GIS is designed to process information derived from various data sources, including geographic data related to the position of objects on the Earth's surface. GIS technology combines database-based data processing that can be accessed today with distinctive visualization through maps [5].

Integration of geographical information systems (GIS) with medical record automation provides a solution that improves the clinic's internal efficiency and makes it easier for the public to find health facilities that suit their needs. This study provides an innovative solution by developing a medical record automation system integrated with GIS. This system will help clinics manage patient data more effectively, reduce data redundancy, and speed up the service process.

In previous literature reviews, most studies have highlighted the importance of automating medical records and using GIS separately. Nevertheless, the novelty of this research is the combination of both aspects in one system designed using the Agile method and Grounded Research. Such an approach enables a more needs-driven system development with broader applicability across clinics and relevant scales. With this system, clinics are also expected to be more competitive and contribute to improving the quality of health services in Indonesia.

2 RESEARCH METHOD

Research methods are stages or steps researchers take to collect information or data and conduct investigations on the data obtained. The research methods in this study are:

2.1 Research Approach

The method used in this research is Grounded Research and Agile, two different approaches that support each other in the process of developing information technology-based systems. Grounded Research was originally developed by two sociologists, Barney Glaser and Anselm Strauss [6]. Grounded Theory is based on the principle that theories derived from field data are more accurate and relevant than theories based on hypotheses or speculation. In this method, experts begin with original pieces of data collected from interviews or observations, then, step by step, create theories [6]. The Grounded Theory approach works its way from practically the empirical stage to the conceptual-theoretical or in more precise terms aims to develop theories from the available data (Kesa & Sainuddin, 2020) [7]. The main goal of Grounded Theory is to expand the understanding of a phenomenon by identifying the key elements of that phenomenon, and then grouping the relationships between those elements within the context and process of the experiment. In other words, the goal is to move from the general to the more specific without neglecting the unique characteristics of the subject being studied [8].

Agile method is an approach in software development that is based on short development cycles. This approach provides flexibility for developers to quickly adapt to changes that occur [9]. The stages in the Agile method include: Plan, Design, Develop, Test, Deploy, Review, and Launch [10]. This study uses the Agile method because this method gives the development team the ability to adapt to changing needs and priorities during development, so that the team can respond quickly to changes or customer requests [11].

2.2 Method of Collecting Data

Data collection methods used in this study are:

1) Observation. The researcher made field observations offline with the aim of gathering data on clinical activities, notably at the clinic of Dr. Aris Rasidi Dahlan in Manggarai, South Jakarta and other clinics of Yogyakarta and Makassar. This observation was conducted to understand the workflow and management of the clinic directly, especially related to the administration and medical record processes that are still carried out manually. In addition, the observation aims to identify the potential for implementing an automation system that can reduce data redundancy and increase the efficiency of clinic administration processing. Additional data that was observed involved the actions related to documentation and retrieval of patient data, as well as the activities involving the clinic staff and the operation system in place.

2) Interviews. In order to gain deep insight into the requirements of the system which is under design, interviews were held not only with clinic employees and doctors but also with patients The focus of the interviews included collecting data regarding obstacles faced in the manual process, such as the length of patient waiting time and difficulties in searching for medical record data. Interviews were also conducted using focus group discussions (FGD) with the resource persons from the clinics located in three different places, namely Jakarta, Yogyakarta, and Makassar. The investigator inquired about the ways in which the use of computerized medical record systems can enhance overall clinical activities and assist in minimizing data errors and delays.

3) Literature Study. A literature study was conducted in the course of this writing by reviewing various literature and references, and also obtaining relevant books and articles from the internet.

2.3 System Development Methods

The approach used in the system development in this research is Rapid Application Development (RAD), a subclass in the System Development Life Cycle (SDLC) model [12]. Rapid application development is a software development method that allows for short development turnaround times by focusing on prototyping and quick building up, followed by the development process [13]. The information system design process usually takes an average of five months, but with this technique, the software system can only take three to four months [14].

The reason why implementing the Rapid Application Development (RAD) approach is essential is that this approach has some benefits, such as enabling a shorter development cycle, being more adaptive, increasing user involvement, decreasing the risks of making mistakes [12].

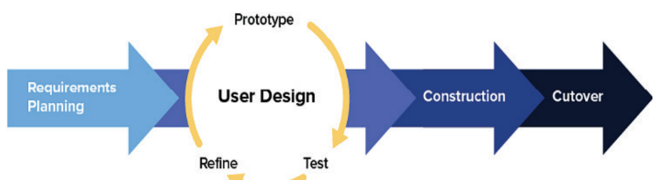


Figure 1 Rapid Application Development (RAD)

In this model, there are several stages of system development, namely:

1) Requirements Planning. At this stage, researchers and users hold discussions together to understand and analyze existing problems. Together, determine the needs that must be met in developing the application system. This stage is very important because it is the first step to ensure the success of the system to be developed, as well as preventing miscommunication between researchers and users.

2) User Design. At this stage, researchers design a system based on agreed needs, with the aim that the resulting solution is able to overcome the problems faced. In this study, system design is represented using a tool such as Unified Modeling Language (UML) and Figma in order to ensure that the presented design meets user requirements.

3) Construction (Development). At this stage, researchers start to realize the established design in a real life application. Researchers write program code, or what is often referred to as the coding process, to translate the system design into functional software according to the previously made plan.

4) Cutover (Implementation). In this final stage, the entire system that has been developed is thoroughly tested to ensure that all its components function properly. Researchers conduct testing using Black Box Testing, a testing technique that focuses on validating software functionality based on predetermined specifications, in order to reduce the risk of errors or defects in the system.

3 RESULTS AND DISCUSSION

3.1 Design System

In designing a practical and easy-to-understand information system, a tool that can visually represent various aspects of the system is needed. One approach used in this study is the Unified Modeling Language (UML). UML is a series of tools commonly used to represent an object-based system or software [15]. The author can define system requirements through UML, perform analysis and design, and describe the overall architecture. This tool provides various types of diagrams that help visualize multiple aspects of the system in several models. However, in this study, the author focuses the design on use case diagrams to describe user interactions with the system and sequence diagrams to show the process flow in the system in detail. Thus, these two diagrams are expected to provide a more structured and comprehensive view of the system being developed.

1) Functional Requirements Analysis. To ensure the system runs according to the expected goals, the author identified several functional requirements that need to be met. These requirements are summarized in Tab. 1.

Table 1 Functional Requirements

No.	Functional Requirements	What the Actor Did
1.	The system must first log in before it can be accessed	Officers and doctors log in first
2.	The system must be able to accept new patient registrations	Officers register new patients
3.	The system must be able to record patient medical records	Doctors record patient medical records
4.	The system must be able to print patient medical records	Officers or doctors print patient medical records
5.	The system must be able to display the location of the registered clinic	The person in charge of the clinic provides the location of the clinic during account registration

2) Use Case Diagram. The use case is one of the ways of putting together a series of activities or more than one which are related to each other and constitute an organized activity bearing out or being managed by an actor [16]. Use case diagrams (Fig. 2) do not explain in detail the use of use cases, but only provide a general description of the relationship between use cases, actors, and systems. Through this use case, the functions in the system can be identified [17].

3) Sequence Diagram. Sequence diagram (Fig. 3) is a diagram designed to study the flow of interactions between objects [18]. The main components of a sequence diagram include objects depicted in rectangular boxes with names, messages represented by lines with arrows, and time displayed in a vertical process [19].

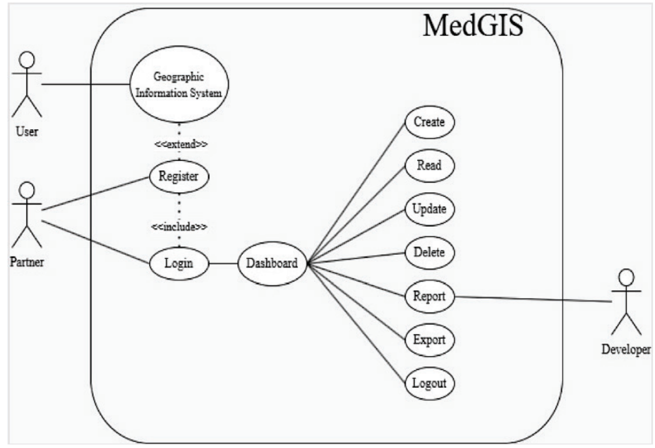


Figure 2 Use Case Diagram

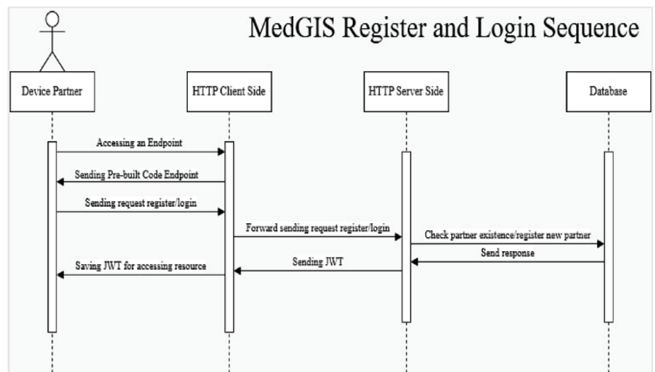


Figure 3 Sequence Diagram

3.2 Implementation

Fig. 4 is pages that are provided in the application.

Fig. 4(a) is used to show the location of clinics in an area. This map provides a geographic visualization with several clinic markers indicating the location of a particular clinic. In the middle of the map is pop-up information that presents detailed data on a clinic, such as the clinic name, address, and contact information. In addition, a "Join as a Partner" option at the top right of the page gives users the opportunity to join as a partner.

Fig. 4(b) is the registration page for clinics that want to join the platform. Users must fill in some information such as the clinic name, telephone number, name of the person in charge, email address, and password. Registering the current location as the clinic location is also available. After all the data is filled in, the user can press the "Register" button to complete the registration process. In addition, users can also access the login option if they already have an account. Fig. 4(c) shows the application screen, in which a user is required to log in by providing an email and password. Below the form, a provision to enroll or return to the home page is provided.

Figs. 4(a)–4(c) present the entry points to the system—clinic map, registration, and login—each framed by a persistent left-hand panel containing the project logo and tagline ("Terhubung dengan 300+ Klinik yang Telah Bergabung"). From the Folium-generated map embedded via iframe (Fig. 4(a)) to the Register (Fig. 4(b)) and Login (Fig. 4(c)) forms, the same purple accent color, rounded form fields, and primary-action button style are applied. Input labels are left-aligned, placeholders use consistent capitalization, and form controls maintain uniform padding and border radius, ensuring that users immediately recognize navigational and interactive elements regardless of context.

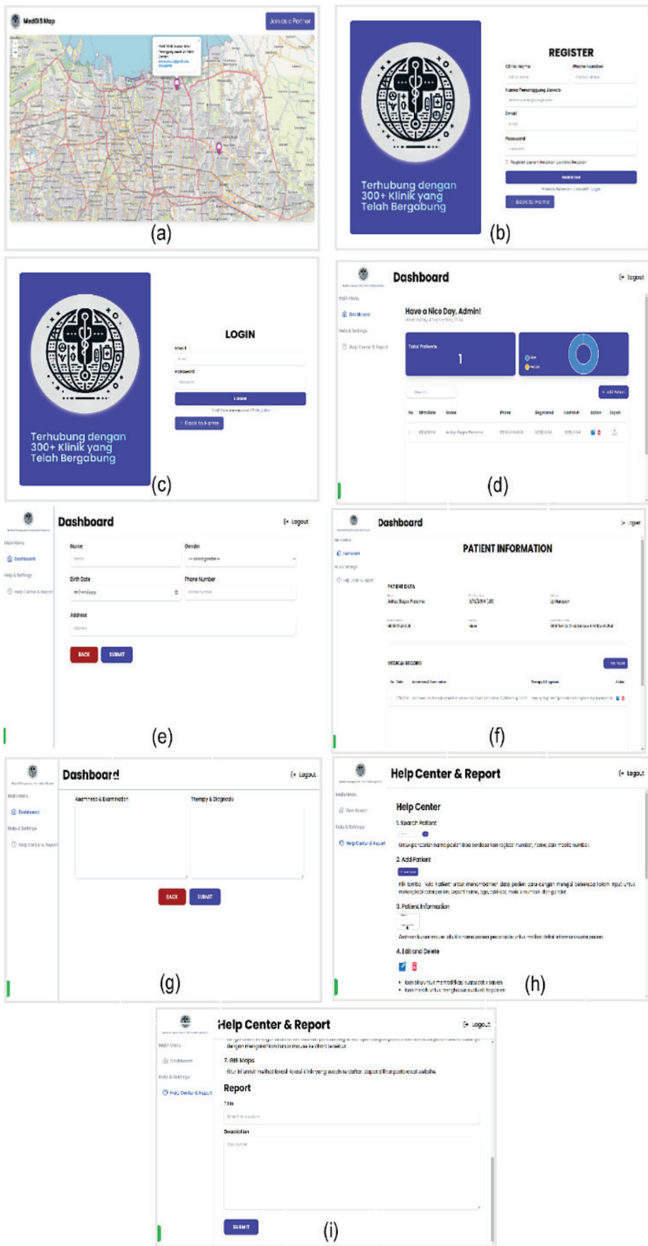


Figure 4 Clinic Medical Record System

Fig. 4(d) shows the post-login dashboard, where the top navigation bar retains the purple highlight and houses the greeting ("Have a Nice Day, Admin!") alongside the logout link. Key performance indicators are laid out in modular cards with matching drop shadows and corner rounding; the gender-split doughnut chart reuses the primary and secondary palette to reinforce visual coherence with earlier screens. The patient table below employs alternating row shading and action icons that follow the same color-on-white pattern established in the login and registration buttons.

Figs. 4(e) and 4(f) depict patient-management pages that continue the established grid layout: labels and inputs share spacing and alignment conventions from the registration form, while section headings and fieldsets use the same font size and weight. Buttons for "Back" and "Submit" consistently appear at the form's lower right, with the primary (Submit) button filled in purple and the secondary (Back) button outlined in the same hue.

Fig. 4(g) follows the identical card-and-form pattern when adding or updating medical record details. Text areas for anamnesis and diagnosis adopt the same line height and border styling as other multi-line inputs. Action controls for saving or discarding changes mirror the button hierarchy and placement seen throughout the application.

Figs. 4(h) and 4(i)—the Help Centre and Problem Report pages—extend the sidebar navigation style used on every page, with link items adopting the same hover and active states. Content panels continue to use white backgrounds, consistent typography, and iconography aligned to the left of each instruction, guaranteeing that guidance and feedback screens feel like an integral part of the same application environment.

3.3 System Architecture

Section 3.3 System Architecture is organized around three Django ORM models Clinic, Patient, and MedicalRecord exposed via Django REST Framework as RESTful API endpoints. The Clinic model uses a UUID primary key and stores clinic name, contact details (phone, email), a Person-In-Charge (PIC) field, authentication credentials, and geospatial coordinates (latitude, longitude). Adopting UUIDs ensures unique identification across Indonesia's highly distributed network of clinics, while the PIC attribute reflects the common operational practice of assigning a dedicated staff member per facility. Latitude and longitude are held as CharFields to feed directly into a custom Django GIS module built with Folium, enabling interactive mapping of clinics in both urban and remote areas.

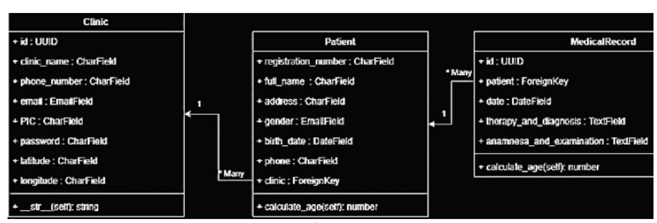


Figure 5 Class Diagram

The Patient model captures a unique registration number, full name, address, gender, birth date, and phone number, and links each patient to its clinic via a ForeignKey mirroring registration workflows puskesmas. The MedicalRecord model records its own UUID, references the Patient, logs the date of service, and stores clinical notes (therapy_and_diagnosis, anamnesa_and_examination). Both Patient and MedicalRecord classes implement a calculate_age() method to derive patient age dynamically from the stored dates.

A SQLite database serves as the relational backend, chosen for its simplicity in initial development, and all services are containerized to guarantee consistent deployment across varied infrastructure setups. The combination of Django, DRF, REST API, Folium-based GIS, and containerization directly addresses Indonesia's requirements for unique, location-aware clinic management and scalable service delivery.

Immediately following the registration of a new Clinic instance via the Django admin or registration endpoint, a bespoke GIS module reads the stored latitude and longitude values for all clinics. It then instantiates a Folium Map object, adds a marker for each clinic coordinate, and generates an HTML snippet representing the interactive map. Because Folium outputs self-contained HTML, this snippet is embedded into the web interface via an <iframe> element. The iframe approach ensures that the map renders consistently across browsers without requiring additional front-end dependencies—simply displaying the up-to-date positions of all registered clinics on the website.

3.4 Interoperability

Interoperability is achieved through a RESTful API layer secured by JSON Web Tokens (JWT), allowing each module—clinic registration, patient management, and medical records—to operate independently while sharing a common authentication mechanism. Upon successful login, the system issues a signed JWT that clients include as a Bearer token in the HTTP Authorization header for all subsequent requests. This stateless approach eliminates the need for a centralized session store, enabling microservice-style modularity: each service can validate request integrity and user identity locally without cross-module dependencies. Moreover, because JWT payloads can carry custom claims (e.g., clinic ID, user role), downstream systems and potential third-party integrations can enforce fine-grained access control and data partitioning without bespoke authentication code. By standardizing on JWT, the architecture supports seamless extension—whether integrating with external EHR platforms via secure API calls or deploying new front-end clients—while maintaining clear module boundaries and minimizing coupling.

3.5 Testing

In ensuring that the software meets the functions that have been set, testing is needed that focuses on validating the functionality of the system. One of the methods used in this study is black-box testing. Black-box testing is one way to test software that has been developed, both in small units and integrated systems to ensure the functionality of the software. This testing is carried out based on functional specifications without looking at the design or program code, with the aim of evaluating whether the functions, input and output of the software comply with the expected specifications [20]. The tester does not have to have knowledge of any particular programming language, as this method is designed for novice users [21]. Tab. 2 shows black box testing of application that is proposed in this research.

Table 2 Black Box Testing

No	Test Case	Expected Output	Status (Pass/Fail)
1.	Login	Successfully logged in, then switched to the dashboard page	Pass
2.	Register	Successfully registered, then switched to the dashboard page	Pass
3.	Add Patient Data	Patient data successfully added	Pass
4.	Search Patient Data	Relevant patient data displayed	Pass
5.	Delete Patient Data	Patient data successfully deleted	Pass
6.	Change Patient Data	Patient data successfully changed	Pass
7.	Add Patient Medical Record Data	Patient medical records successfully added	Pass
8.	Delete Patient Medical Record Data	Patient medical records successfully deleted	Pass
9.	Change Patient Medical Record Data	Patient medical records successfully changed	Pass
10.	Print Patient Medical Record Data	Patient medical records successfully displayed as a PDF file	Pass

Below is a summary of response sizes and execution times for three key endpoints, to be inserted immediately after the black-box testing section.

Table 3 Summary of response

Endpoint	Response Size	Execution Time
Login	901 B	631 ms
Map	10 KB	24 ms
Fetch single patient record	803 B	28 ms

Following the black-box testing results, Tab. 3 highlights the performance characteristics of representative API calls. The Login endpoint incurs a higher latency (631 ms) despite a modest payload (901 B) because it must perform cryptographic operations to generate and sign a JSON Web Token for secure, stateless authentication. In contrast, the Map endpoint returns roughly 10 KB in just 24 ms: this larger payload size reflects the self-contained HTML snippet produced by Folium for embedding the interactive clinic map via an <iframe> on the client. Finally, retrieving a full patient record yields an 803 B response in 28 ms, indicating efficient serialization of both patient fields and associated medical records under normal conditions.

4 CONCLUSIONS

This study successfully built a clinical medical record information system integrated with GIS using RAD and Agile methods. The implementation of this system is able to increase the efficiency of medical record maintenance by reducing data redundancy and accelerating the service process in the clinic. In addition, the integration of GIS in this system makes it easier for the public to access information related to the location of the clinic that suits their needs, thus expanding the reach of health services.

Testing conducted using the black-box method shows that this system is able to function well in running key features such as login, registration, patient data management, medical records, and GIS map visualization. This system is expected to be a solution for clinics in facing the challenges of increasing patient numbers and the need for more efficient and responsive health services.

Overall, this research provides a significant contribution to the development of clinical information system applications that not only focus on automating medical records but also maximizing the potential of GIS to improve the quality of health services.

For further development, it is recommended that this system integrate artificial intelligence (AI) and blockchain technology. AI integration can be used to improve accuracy and efficiency in managing medical records and provide smarter data-based medical recommendations. For example, AI can help doctors analyze a patient's medical history to provide health predictions or early diagnoses based on existing data patterns, thereby accelerating clinical decision-making.

In addition, the implementation of blockchain technology can provide an additional layer of security to ensure the integrity and safety of patient medical data. Blockchain, with its decentralized and transparent nature, can keep medical records from being altered without proper authorization. Every data change can be recorded in a distributed ledger system, ensuring that medical records remain secure and traceable.

However, in this development, the aspect of administrative processes must not be overlooked. The use of AI and blockchain technology must remain aligned with administrative needs such as patient registration management, data recording, and payment processes. The integration of electronic payment systems can also be enhanced to support the smoothness of administrative transactions. Thus, the system can provide comprehensive solutions, not only from a clinical perspective but also from an administrative one.

Further testing on a larger scale and involving more clinics in various locations is also recommended to ensure that the newly integrated technology can function effectively and provide real benefits in healthcare practice.

Acknowledgment

This research was funded by Regular-Fundamental 2024 Research Grant, Ministry of Education, Culture, Research and Technology, Indonesia to Gunadarma University No. 105/E5/PG.02.00.PL/2024, Tanggal 11 Juni 2024.

5 REFERENCES

[1] Permenkes No. 24. (2022). Peraturan Menteri Kesehatan RI No 24 tahun 2022 tentang Rekam Medis. *Peraturan Menteri Kesehatan Republik Indonesia Nomor 24 Tahun 2022*, *151*(2), 1–19. (in Indonesian)

[2] Raihan, F. M. (2021). Perancangan Sistem Informasi Rekam Medis Pada Klinik Saffira Sentra Medika Batam. *Jurnal Sains, Nalar, Dan Aplikasi Teknologi Informasi*, *1*(1). (in Indonesian) https://doi.org/10.20885/snati.v1i1.7

[3] Suryadi, A., Arif, Y. W. T., & Novitasari, N. S. (2022). Rancang Bangun Sistem Informasi Rekam Medis Klinik Rawat Jalan Berbasis Web. *Infokes: Jurnal Ilmiah Rekam Medis Dan Informatika Kesehatan*, *12*(1), 37–43. (in Indonesian) https://doi.org/10.47701/infokes.v12i1.1498

[4] Umar, T. L. (2021). Perancangan Sistem Informasi Geografis Tempat Bersalin Berbasis Mobile. *Jurnal Informatika Dan Rekayasa Perangkat Lunak (JATIKA)*, *2*(2), 221–229. (in Indonesian) http://jim.teknokrat.ac.id/index.php/informatika.

[5] Tinambunan, M., & Sintaro, S. (2021). Aplikasi Restfull Pada Sistem Informasi Geografis Pariwisata Kota Bandar Lampung. *Jurnal Informatika Dan Rekayasa Perangkat Lunak*, *2*(3), 312–323. (in Indonesian) https://doi.org/10.33365/jatika.v2i3.1230.

[6] Wardhana, A. (2023). Grounded Theory. *As-Shaff*. http://www.sciencedirect.com/science/article/pii/S0020748909003629

[7] Kesa, I., & Sainuddin, I. (2020). Pengoperasian Penelitian Grounded Theory. *Sociology the Journal of the British Sociological Association*, *1*, 14–23. http://www.msvu.ca/site/media/msvu/MixedMethodologyHandout.pdf

[8] Asbui, Risnita, M. Syahran Jailani, M. Husnullail, & Asrul. (2024). Metode Grounded Theory Dalam Pendekatan Praktis. *Jurnal Cahaya Mandalika ISSN 2721-4796 (Online)*, *5*(1), 47–58. (in Indonesian) https://doi.org/10.36312/jcm.v5i1.2298

[9] Nugraha, F. S., Syahidin, Y., Suryani, A. I., Kesehatan, M. I., Ganesha, P. P., Gatot, J., & No, S. (2024). *Penerapan Teknologi Sistem Informasi dalam Proses Pembuatan Surat Keterangan Kematian Berbasis Elektronik Menggunakan Metode Agile*. *7*(2), 483–493. (in Indonesian) https://doi.org/10.32493/jtsi.v7i2.39134.

[10] Rezy, A. F., Yuga Utama, M., & Ramadhan, N. R. (2023). Pengembangan Aplikasi Klinik Berbasis Web Untuk Pengelolaan Rekam Medis Menggunakan Metode Agile. *Jurnal Ilmu Komputer, Teknik Dan Multimedia*, *VOL 1*, *NO.*(2), 309–319. (in Indonesian)

[11] Atim, S. B. (2024). Permodelan Sistem Informasi Penjualan Barang Berbasis Website Menggunakan Metode Agile. *Journal of Data Science and Information ...*, *2*(1), 14–25. (in Indonesian) https://ejournal.techcart-press.com/index.php/dimis/article/view/97%0Ahttps://ejournal.techcart-press.com/index.php/dimis/article/download/97/92

[12] Nurman Hidayat, & Kusuma Hati. (2021). Penerapan Metode Rapid Application Development (RAD) dalam Rancang Bangun Sistem Informasi Rapor Online (SIRALINE). *Jurnal Sistem Informasi*, *10*(1), 8–17. https://doi.org/10.51998/jsi.v10i1.352

[13] Arrohimi, M. J., Windyasari, V. S., & Kurniasari, R. (2023). Perancangan Sistem Rekam Medis Dengan Metode RAD Pada Klinik Mumtaz. *Jurnal Ilmiah Fakultas Teknik*, *3*(1), 27–41. (in Indonesian) https://doi.org/10.33592/jimtek.v3i1.3783

[14] Wahid, B. A. (2022). Penerapan Metode Rapid Application Development Terhadap Penjualan Fashion Distro Secara Online. *Jurnal Esensi Infokom : Jurnal Esensi Sistem Informasi Dan Sistem Komputer*, *3*(1), 33–39. https://doi.org/10.55886/infokom.v3i1.345.

[15] Noviantoro, A., Silviana, A. B., Fitriani, R. R., & Permatasari, H. P. (2022). Rancangan Dan Implementasi Aplikasi Sewa Lapangan Badminton Wilayah Depok Berbasis Web. *Jurnal Teknik Dan Science*, *1*(2), 88–103. (in Indonesian) https://doi.org/10.56127/jts.v1i2.108.

[16] Tabrani, M., & Rezqy Aghniya, I. (2020). Implementasi Metode Waterfall Pada Program Simpan Pinjam Koperasi Subur Jaya Mandiri Subang. *Jurnal Interkom: Jurnal Publikasi Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *14*(1), 44–53. (in Indonesian) https://doi.org/10.35969/interkom.v14i1.65
[17] Kurniawan, H., Apriliah, W., Kurnia, I., & Firmansyah, D. (2021). Penerapan Metode Waterfall Dalam Perancangan Sistem Informasi Penggajian Pada Smk Bina Karya Karawang. *Jurnal Interkom: Jurnal Publikasi Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *14*(4), 13–23. (in Indonesian) https://doi.org/10.35969/interkom.v14i4.78
[18] Wulandari, T., & Nurmiati, S. (2022). Rancang Bangun Sistem Pemesanan Wedding Organizer Menggunakan Metode Rad di Shofia Ahmad Wedding. *Jurnal Rekasaya Informasi*, *11*(69), 79–85. (in Indonesian)
[19] Julianti, M. R., Dzulhaq, M. I., & Subroto, A. (2019). Sistem Informasi Pendataan Alat Tulis Kantor Berbasis Web pada PT Astari Niagara Internasional. *Jurnal Sisfotek Global*, *9*(2). (in Indonesian) https://doi.org/10.38101/sisfotek.v9i2.254
[20] Abdillah, M. T., Kurniastuti, I., Susanto, F. A., & Yudianto, F. (2023). Implementasi Black Box Testing dan Usability Testing pada Website Sekolah MI Miftahul Ulum Warugunung Surabaya. *Journal of Computer Science and Visual Communication Design*, *8*(1), 234–242. (in Indonesian) https://doi.org/10.55732/jikdiskomvis.v8i1.897
[21] Uminingsih, Ichsanudin, M. N., Yusuf, M., & Suraya. (2022). Perpustakaan Dengan Metode Black Box Testing Bagi Pemula. *STORAGE-Jurnal Ilmiah Teknik Dan Ilmu Komputer*, *1*(2), 1–8. (in Indonesian)

Authors' contacts:

Lely Prananingrum, Dr
Gunadarma University,
Margonda 100, Depok, West Java, Indonesia
(021) 29428935, lely_p@staff.gunadarma.ac.id

Teuku Salman Farizi
Gunadarma University,
Margonda 100, Depok, West Java, Indonesia
(021) 294228935, teukusalmanfarizi2003@gmail.com

Fajar Agus Dwi Rahmawan
Gunadarma University,
Margonda 100, Depok, West Java, Indonesia
(021) 294228935, fajarrahmawan0804@gmail.com

Ilmiyati Sari, Dr
(Corresponding author)
Gunadarma University,
Margonda 100, Depok, West Java, Indonesia
(021) 29428935, ilmiyati@staff.gunadarma.ac.id

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20250604003051

Original scientific paper
Received: 2025-06-04, Accepted: 2025-07-18

Predictive Modeling with Artificial Neural Networks to Optimize Dosing Accuracy of Galenical Powder Dosing Systems

Jaime Cancho*, Ciro Rodriguez, Ivan Petrlik, Milner Liendo

Abstract: A predictive model is proposed based on artificial neural networks (RNA) to optimize the quality of dosing of galenical powders in bottles, where it is necessary to maintain accuracy and stability, as the current electromechanical control methods have these shortcomings. The experimental development of research fosters new skills, which are key to innovating and facing the challenges of today's knowledge society. The RNA model was applied to the control group, resulting in an experimental group with improved. Six neural network models were trained, achieving the best results with the Recurrent Neural Network (RNN) model. Tests were conducted to optimize process capability indicators, improve process accuracy, and effectively predict the accuracy of dosed weight, considering the system's operating parameters. The RNN model was trained and validated with real data. The findings demonstrate that the application of the proposal will optimize accuracy and weight stability, meeting the quality standards in the industry.

Keywords: dosing; galenic powder; prediction; precision; rigid packaging; RNN

1 INTRODUCTION

In the pharmaceutical industry, the packaging process for galenic powders faces critical challenges of accuracy and consistency [1]. Powders are used in the manufacture of medicines that must comply with strict quality regulations [2]. Galenic powders, due to their physical nature, present properties such as irregular particle size, high cohesion, and sensitivity to environmental factors such as humidity and temperature, which complicate accurate and controlled dosing [3]. The United States Pharmacopeial Convention (USP) expert panel establishes control specifications for weight variation in continuous manufacturing in the pharmaceutical industry [4]. The quality of the packaged product is reflected in its nominal value of the declared weight on the package. [2] In an industrial process for powder packaging, materials, machines, labor, measurements, environment and methods interact, these six elements intervene in the quality of the product and changes inevitably occur over time and consequently variations in the packaging [2], therefore, the world's largest companies place great importance on the constant monitoring of these vital signs, which are known as: critical characteristics or indicators for quality. Continuous manufacturing is a novel process for producing high-quality pharmaceutical products [5]. When a galenic powder is dosed below that indicated on the package, the patient could receive an insufficient amount of the active ingredient, reducing the effectiveness of the treatment [6]. This can prolong the illness or require frequent dosage adjustments. If the weight of the galenic powder is greater than that indicated on the package, the patient could receive an excessive dose, causing serious adverse effects or toxicity, especially in drugs with a narrow therapeutic margin, where the difference between an effective dose and a toxic one is minimal [7]. Inaccuracy in dosing leads to variability in treatment outcomes, complicating the assessment of effectiveness and the ability of healthcare professionals to adjust treatment appropriately. Reliability in the administered dose is essential to ensure a consistent and predictable therapeutic response [8]. Inaccuracy in the weight of the packaged product puts patient safety at risk, especially with medications that require precise dosing [9], such as anticoagulants, cardiovascular drugs, or treatments for chronic diseases. Administering incorrect doses can lead to serious and even fatal complications [10]. Side effects resulting from inadequate doses can demotivate patients, reducing treatment adherence and increasing the risk of premature discontinuation. This could worsen the patient's condition and increase costs for their treatment and the healthcare system due to additional treatments and hospitalizations. Weight variations and inconsistencies during the packaging of galenic products have a significant impact on production costs, generating material waste, as incorrect doses must be discarded or reprocessed, increasing the use of additional raw materials and resources. Weight problems require additional controls and adjustments, increasing operating costs.

For the next 5 years, the global pharmaceutical market is expected to grow by 5% annually, requiring manufacturing companies to improve and maintain the quality of production and distribution of their pharmaceutical products [11].

Peru's economy is characterized by over 70% informal labor [12], which means that many domestic companies are not supervised by health quality regulatory bodies, making it increasingly difficult to identify or control unidentified or poorly controlled companies. In a globalized economy, where profit margins on the sale of pharmaceutical products are very tight, the use of electromechanical technology to control screw-type packaging machines is insufficient.

The research aims to apply a predictive model based on Artificial Neural Network (ANN) to optimize the accuracy of galenic powder packaging, for this purpose a mathematical algorithm will be trained with different ANNs and the results will be measured, choosing the most accurate and stable model.

ANNs work with large amounts of data, their ability to learn and update data allows them to predict errors and correct weight deviations during the packaging process in real time

In the food, pharmaceutical, cosmetics, and related industries, the accurate weight of the products offered is crucial to ensuring their quality and effectiveness. The main challenge in powder packaging is its precise measurement and dosage [13].

2 LITERATURE REVIEW

The proposed research is based on the application of mathematical algorithms that allow the improvement of processes, specifically, in our case, the packaging of sodium bicarbonate x 50 grams. Previous research has proposed methods that apply mathematical algorithms for preprocessing, balancing, data mining and selection, necessary for the development of more efficient predictive models, such as ANNs.

An essential aspect is model validation, which ensures that the predictions meet the accuracy requirements of the packaging MATLAB has been the most widely used tool in related work due to its ability to run and train predictive algorithms [14]. Variations of training algorithms available in MATLAB offer several options that allow improving the model fidelity when applied to galenical powder packaging. This research focuses on weight control, proposing an ANNs model that ensures accuracy and efficiency in galenic powder packaging.

2.1 Artificial Neural Networks (ANN)

ANNs are computer models that mimic the structure and function of the human brain. They are widely used in process prediction and classification, and are notable for their ability to identify trends from large volumes of data. Their biologically inspired design allows ANNs to adapt to different activities and environments, making them valuable tools for data management and promoting important advances in different fields of action where artificial intelligence and machine learning can be applied [15]. ANNs are trained to identify complex patterns and make accurate predictions. This makes them suitable tools for Quality Control and process automation in industry. The learning capacity of ANNs allows for significant improvements in industrial processes and optimization of their results, ensuring precision and efficiency [16].

2.2 Applications in the Pharmaceutical Industry

In the pharmaceutical industry, ensuring accurate packaging of galenic powders is critical to the correct administration of drugs to the patient. ANNs are used to anticipate weight errors and calibrate variables during dosing, which increases production efficiency and reduces waste of resources. This contribution of ANNs significantly improves process quality, ensuring stable and reliable results in drug production [17]. Recent research shows that ANNs stand out significantly in terms of accuracy and flexibility, making them the most advantageous option in various industrial applications compared to conventional methods. [18].

2.3 Predictive Models

Predictive models that use ANN, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), have proven to be highly effective in predicting process control variables. Their ability to handle different types of data highlights their great potential in various analysis and prediction uses [16]. ANNs have the ability to be trained using historical process data, allowing them to accurately predict the weight of galenic powders dosed into rigid containers and self-calibrate the variables that control this process in real time, guaranteeing consistent results. This characteristic makes them valuable tools to optimize industrial processes [15].

2.4 Challenges and Future Trends

Beyond their benefits, ANNs are hampered by the need to process large amounts of data for training and the difficulty in processing the results. These limitations can hinder their application and adaptation in certain environments. [1]. However, looking to the future, we need not only technically efficient models but also more understandable and easier-to-program models. Deep learning techniques are currently being integrated to increase the accuracy and stability of predictions. These advances project a significant improvement in the efficiency and applicability of ANNs [18].

Applying ANN to predict weight accuracy in powder packaging is a revolutionary industrial innovation. ANNs are trained and adapt to various process conditions, optimizing quality and efficiency. This demonstrates a significant technological advance in the industry, demonstrating its potential to optimize industrial processes and ensure stable and reliable results.

3 MATERIALS AND METHODS

The research conducted is of the applied type because it uses existing knowledge. Applied research considers all the norms, regulations, and statutes that govern social behavior, ensuring that its findings and methods are aligned with the legal and ethical context. [1].

As a sample, there is a weight record of 1,430 Sodium Bicarbonate bottles × 50 grams packaged with the METS-01 screw dosing machine from a Pharmaceutical Laboratory in Lima, taken randomly from the production line, at the discretion of the Quality Control Department, these correspond to 11 production batches on consecutive dates since January 2025, the variables that affect the operation of the machine during dosing were also considered.

The sample was collected using a precision digital scale that was initially certified by a competent laboratory specializing in its field. For weight control studies, the use of a precision balance will be necessary. [19].

The input variables involved in controlling the weight of the METS-01 are the rotational speed of the three-phase electric motor. This speed is adjusted by a frequency converter. This rotation drives the worm gear that transports the galenic powder and the electrical pulses per revolution read by the incremental encoder.

Figure 1 METS-01 screw packaging machine

The information obtained is analyzed, which will yield important insights and establish patterns, relationships, and trends for developing the prediction model.

To verify the current accuracy of the galenic powder packaging process, the Process Control Analysis (PCA) was carried out. Applying statistical control through variable control allows for greater control of the final product, guaranteeing its specifications and quality [20].

It was necessary to determine the distribution type of the sample data to clearly define the type of statistic to be used in the inferential statistics; therefore, the Student t-test was used. The effect of the model on the capacity indicators of the galenic powder packaging process was reviewed, and the Cp and Cpk in the control group and their optimization were identified after applying the model to the experimental group.

3.1 Type of Artificial Neural Network

It began with a regression analysis that examines the relationship between process variables, such as the rpm of the motor that drives the auger and the electrical counting pulses per motor revolution, and the weight history. The data was gathered, cleaned, and normalized to ensure consistency. A linear regression model was subsequently implemented as an initial approximation to understand the influence of the variables on the batch weight. This model provided us with a solid foundation for designing and training the neural network, using its results to define the appropriate structure, layers, and activation functions, adjusting the parameters to improve the model's accuracy. There are several types of ANN, each with its own characteristics, advantages, and disadvantages. To decide which type of neural network to use for the project, a mathematical algorithm was developed and trained with six different types of ANN. The results were verified, and it was decided that the research should be done with the RNN model because it offers greater advantages in accordance with our objective. The RNN is a type of ANNs. Once the neural network was implemented and the effect of the verified improvements was measured, the findings were compared with similar research to validate the proposed model. Using neural networks contributes to improved accuracy compared to the previous system [21].

4 ETHICAL CONSIDERATIONS

The owner of the MEPS-01 machine has given his or her consent, and data privacy and confidentiality will be respected. Ethical conduct in research must consider three general dimensions: the human, the political, and civil society [22].

5 RESULTS

5.1 Data Analysis and Visualization

The dataset to be analyzed by the model is created from the sample. The upper and lower limits, range, class number, class size, standard deviation, variance, mean, median, and mode are calculated. The histogram is created from this data.

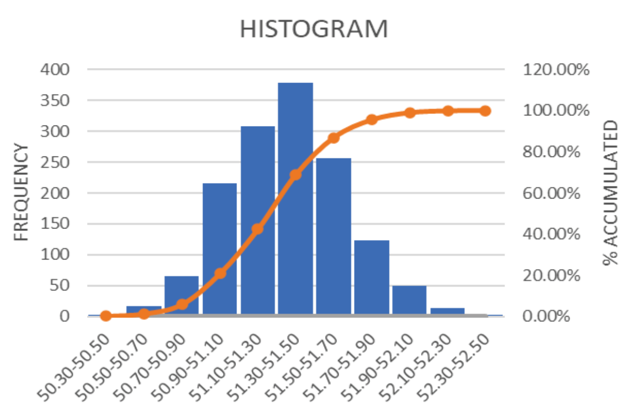


Figure 2 Visual representation of data

It is observed which frequency is almost symmetrical and bell-shaped (Fig. 2), suggesting that the data have a nearly normal distribution since the highest peak is in the center, specifically in frequency bar 379; as we move away to the left or right, the frequencies progressively decrease, indicating that there is less data at the extremes.

The frequencies do not show skew to the left or right, that is, they are not tilted, which reinforces the idea of a normal distribution.

To give a better visualization of the statistical diagram (Fig. 2) we will increase the interval to a smaller scale and propose an interval of 0.1, where the frequency and its distribution are found.

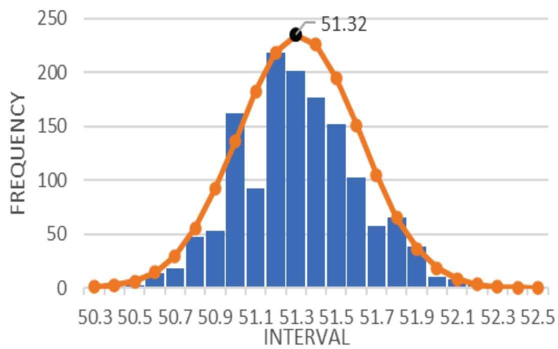


Figure 3 Data frequency distribution

In the Frequency Distribution (Fig. 3), it is verified that the distribution of the data is centered according to the mean; we can affirm that the sample weight data are within a normal distribution.

For the scatter plot, we group the data that were taken at the same time, 10 records per group, obtaining 143 groups, and the scatter diagram is developed.

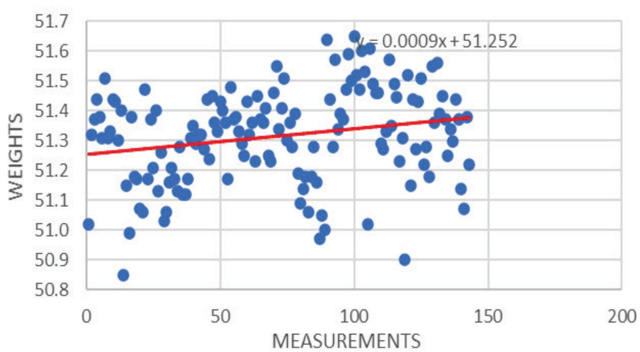


Figure 4 A wide dispersion is seen, that is, there are many outliers that do not follow the general trend line pattern

5.2 Analysis and Efficiency of the METS-01 Galenic Powder Packaging Machine

X-R charts are used for statistical monitoring of part quality control in industry. They allow for the detection of variability, consistency, control, and improvement of a production process. They are composed of the following elements:

- Upper control limit (*UCL*), or maximum tolerance.
- Lower control limit (*LCI*), or minimum tolerance.
- Control Line (*CL*), or average of the minimum and maximum tolerances.

Measurement variables, which in our research are the weights of the bottles packed with Sodium Bicarbonate × 50 grams in the Pharmaceutical Laboratory.

The Tab. 1 must be completed to calculate the control limits for means and ranges.

Table 1 Formulas to obtain the limits by means and ranges

Graphics for	*LCS*	*LC*	*LCI*
Averages	$\bar{\bar{x}} + A_2 \cdot \bar{R}$	$\bar{\bar{x}}$	$\bar{\bar{x}} - A_2 \cdot \bar{R}$
Ranks	$D_4 \cdot \bar{R}$	$\bar{R}$	$D_3 \cdot \bar{R}$

To perform the calculations, the values of A_2, D_3, and D_4 are required, which will be extracted from the XR Constants Table. n = 10 must be considered, then it is obtained.

Table 2 Summary of data obtained

Graphics for	*LCS*	*LC*	*LCI*
Averages	51.56	51.3	51.07
Ranks	1.417	0.8	0.178

Integrating the sample and Tab. 2, we have Fig. 5.

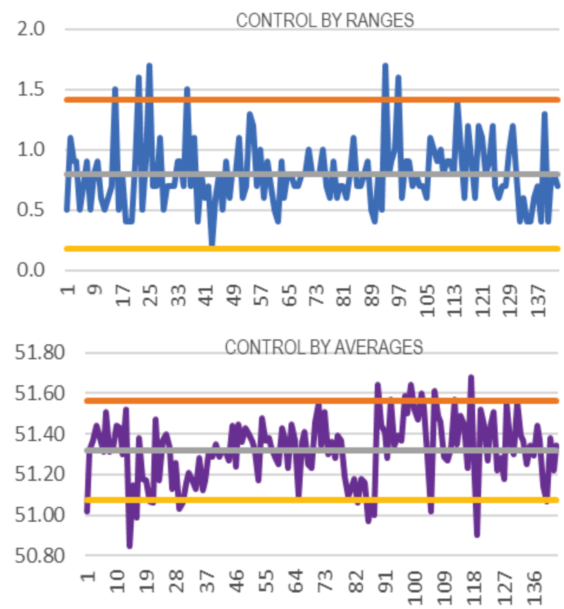


Figure 5 Graph of Averages and Ranges, the process is outperforming the LCS and LCI

Both graphs indicate that the process is statistically out of control. Because the operator's procedure was observed during packaging, we can confirm that the machine is out of control in terms of weight accuracy and consistency during the dosing of galenic powders into rigid packaging.

5.3 Analysis of Packaging Process Capacity

Packaging process capacity can be expressed as the packaging machine's ability to perform its intended purpose within the permitted tolerances. There are two main indicators: one C_p shows whether the data distribution can fit within the required production specifications, and the other C_{pk} shows whether the overall data average is located at the center of the limits.

The legal tolerable weight requirements for prepackaged products are specified, indicating predetermined nominal values [24]. From the Peruvian Metrological Standard [23], the tolerance for 50-gram presentations is –0.0 and +1 grams. In our case, the Control Line or production average LC = 51.3 grams, the Pharmaceutical Laboratory has seen fit to consider ±1 gram of tolerance for the analysis of the packaging process capacity. In this sense, USL = 52.3 grams and LSL = 50.3, with these limits, the control chart is drawn by specific averages.

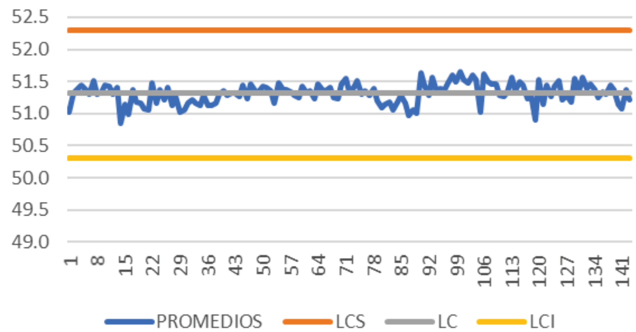


Figure 6 Specific means are considered as tolerances specifically given by the manufacturer

It can be seen that all means of the data subgroups are fully controlled, with slight alterations inherent to the process.

The process capability analysis C_p and the actual process capability index C_{pk} of the packaging will be carried out to determine if the process is capable of producing Sodium Bicarbonate × 50 grams according to the tolerance decided by the Pharmaceutical Laboratory in ±1 gram.

For C_p:

$$C_p = \frac{USL - LSL}{6 \cdot \sigma} = \frac{52.3 - 50.3}{6 \cdot 0.3031} = 1.0997 \quad (1)$$

For C_{pk}:

$$C_{pk} = \text{minimum}\left(\frac{USL - \overline{\overline{x}}}{3 \cdot \sigma}, \frac{\overline{\overline{x}} - LSL}{3 \cdot \sigma}\right) = \\ = \text{minimum}\left(\frac{52.3 - 51.3}{3 \cdot 0.3031}, \frac{51.3 - 50.3}{3 \cdot 0.3031}\right) = 1.0997 \quad (2)$$

$C_p > 1$, the process is moderately capable, and requires observation.

If $C_{pk} > 1$, the process is suitable for producing within the given specifications.

5.4 Analysis of Prediction Models

The mathematical algorithm was designed, and six Artificial Neural Network models were trained. Optimized Artificial Neural Network – TUNED_ANN, Deep Neural Network – DNN, Long Short-Term Memory Network – LSTM, One-Dimensional Convolutional Neural Network – CNN_1D, Multilayer Perceptron – MLP, Recurrent Neural Network – RNN.

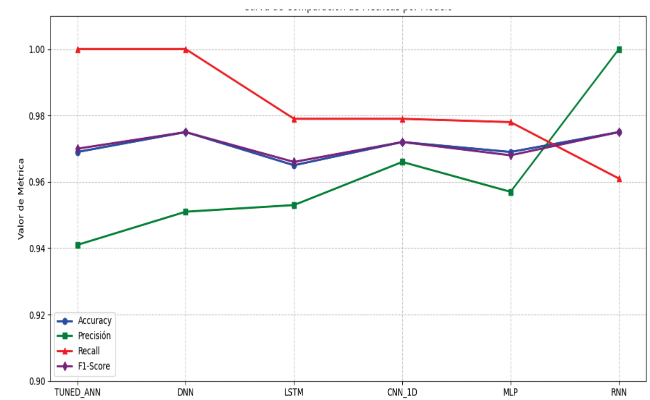


Figure 7 Metrics of the 6 models tested

The term added as LSTM means Long Short-Term Memory. The comparison of metrics by model allows us to compare between the reviewed models and conclude that the RNN model outperforms all the ANN models trained for the prediction of weight in the packaging of galenic powders in rigid packaging, therefore, it is stated that the RNN model is the best model to predict the accuracy of the weight of the packaging of Sodium Bicarbonate × 50 grams in rigid packaging with the METS-01 packaging machine.

5.5 The Indicators were obtained from the RNN Model

The Accuracy compares the overall performance of the models in terms of global precision. The RNN model obtains the highest Accuracy (0.975), which suggests high performance in correct predictions. The Precision shows how accurate each model's predictions are, avoiding false positives. The RNN model has a perfect precision (1.0000), ensuring that all its predictions are correct. The F1-score evaluates the balance between Precision and Recall. The RNN model has the best F1 Score (0.9822), which indicates that it correctly balances both metrics. Recall is the ability to effectively detect correct values. The RNN model has the best Recall (0.973), balancing both metrics.

We will work with an RNN with an LSTM architecture, which is trained to predict the binary classification of the weight dispensed by the METS-01 machine, based on the input variables. This classification is defined based on whether the predicted weight value is below or above the median, transforming a regression problem into a classification problem.

This RNN-LSTM architecture represents a systemic and integrated solution for addressing prediction problems in industrial processes, aligning with the fundamental principles of Systems Engineering, such as:

- Dynamic modeling of complex systems.
- Decision-making under uncertainty
- Performance optimization in multidimensional systems.

Interoperability between mathematical models, real data, and physical systems.

We evaluate whether our data belongs to a normal distribution. For this, we will work with the help of IBM SPSS STATISTICS Viewer software and Tab. 2. The

Kolmogorov-Smirnov normality test was performed, and a p-value of 0.2 was obtained. Since we have considered a significance level $a = 0.05$, it is necessary that $p > a$; therefore, it is accepted that "the data follows a normal distribution".

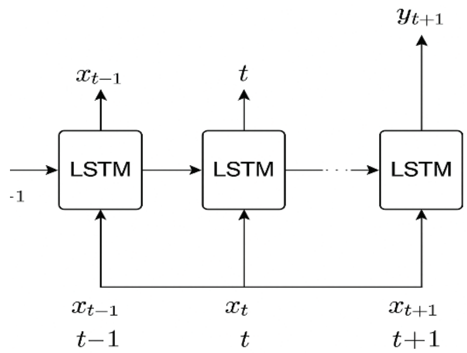


Figure 8 Sequential flow of information through time steps and the propagation of hidden states, which clearly illustrates the dynamic and adaptive nature of the model.

Groups of 5 data points were formed, obtaining 28 subgroups as the Control Group (CG), the data for USL and LSL are considered, and formulas 1 and 2 are applied to find Cp and Cpk for each subgroup. The defined RNN-LSTM model is applied to this control group and Cp and Cpk are obtained with values higher than 5, mathematically it is possible due to the high capacity of the model, but in industrial application this value is not realistic, therefore we will consider a statistically acceptable and technically viable value of C_p and $C_{pk} = 1.333$.

The groups and variables are reordered, and with the help of SPSS, $p < 0.001$ is obtained for the C_p Group of the CG and EG, and $p < 0.001$ for the C_{pk} Group of the CG and EG. Then it is true that: $p < 0.001$ a. Therefore, we can affirm that neural networks optimize the capacity indicators of the dosing process of galenic powders in rigid packaging.

From the Control Group, 30 samples were randomly selected as Initial Weight to perform the Student's t-test. The defined RNN-LSTM model was applied, and the Modeled Weight was obtained, with an average of 50.6 grams. Analyzed with SPSS software.

Table 3 From the Student t-test for sample correlation, $p < 0.001$

Sample correlations				
	N	Correlation	Significance	
			p of a factor	Two-factor p
Initial weight & model weight	30	–.762	< .001	< .001

Maintaining the same value of the significance level to $a = 0.05$, we have that $p < a$, therefore, it is stated that the prediction model based on neural networks determines the best accuracy of packaging galenic powders in rigid packaging.

Powder packaging between 50.2 and 50.7 grams is acceptable. We will consider precision as a dichotomous variable: 1 indicates weights that are within tolerance (precise) and 0 indicates weights that are outside the tolerance (imprecise). Binary language favors the application of prediction models such as RNN and the evaluation of performance metrics such as *Precision*, *Accuracy*, *F1-Score,* and *Recall*. The 1,430 data points were grouped into 286 groups and converted into binary language. The Student t-test for related samples is performed to determine if there is a statistically significant difference between the actual values and the predicted values, as well as the key result *p*-value (two-tailed sig.). With SPSS software, we have

Table 4 Paired sample test, $p < 0.083$ was obtained, then: $p > a$

Paired sample test				
	Standard deviation	Mean standard error	p of a factor	Two-factor p
Actual value & predicted value	0.102	0.006	< 0.042	< 0.083

It also performed simple linear regression.

Table 5 A linear correlation $R = 0.974$ indicates a positive and very strong correlation between predicted and actual values

Model summary					
Model	R	R^2	Adjusted R^2	Standard error of the estimate	Regression coefficient B
1	0.974	0.948	0.948	0.101	0.962

The coefficient of determination $R^2 = 0.948$, meaning that the model explains 94.8% of the variability in the accuracy of the actual weight. The adjusted $R^2 = 0.948$ confirms the stability of the model and that there is no overfitting or loss due to complexity. The standard error of the estimate indicates the average of the errors between the predicted and actual values. A low value, such as 0.101, suggests that the model is fairly accurate in its estimates.

Statistical indicators obtained confirm that the applied RNN-LSTM model effectively predicts the weight accuracy in the packaging of galenic powders, since, due to the three conclusive findings, it is stated that the RNN-LSTM model optimizes the accuracy of dosing galenic powders in rigid packaging.

6 DISCUSSION

Initially, data from the production line or control group demonstrate that the process is unable to meet the limits specified for the galenic powder packaging process, with an average C_p and C_{pk} of 0.295 and an average standard deviation of 0.13 grams. When the galenic powder packaging process capacity indicators show an indicator lower than 1, it indicates that the process is at considerable risk, which is a level below the permitted quality standards.

Levene's test shows a significant result ($F = 37.187$; $p < 0.001$), which confirms that the variability of the EG is much lower than that of CG. The research carried out confirms that by applying the RNN model to the control group, an experimental group is obtained with a better centered process and with a reduction in its dispersion, which directly affects the improvement of the process capacity indicators C_p and C_{pk} for the packaging of galenic powders. The results obtained in this research agree significantly with the findings described by [25] using the quality indicators C_p and C_{pk}, they also use RNN-LSTM to predict the process quality performance. In addition, the implementation of RNN allows it to reduce the standard deviation of the analysis process, that is, it has less variability, so the values of C_p and C_{pk} increase significantly. They report that the RNN model achieved process improvement through a marked reduction in the standard deviation, which increased the process capability indices, considering those obtained with models such as Random Forest or ARIMA. Similarly, in our research, RNN was applied to the control group, and an experimental group with a standard deviation of 0.025 grams was obtained. This considerable improvement reflects an increase in the process capability indicators, obtaining a C_p and C_{pk} of 1.333. This value shows that the process is stable and capable, according to standards in the galenic industry, and also indicates that the C_{pk} has a proportion of less than 0.01% of non-conforming products. Then (1.333/0.295 = 4.53), it is verified that the application of the model offers a 4.53-fold improvement in the capacity of the galenic powder packaging process. The Student t-test shows statistically significant results, $t = –50.254$ and $p < 0.001$.

The verification of the performance of the predictive model based on recurrent neural networks (RNNs) developed to design a prediction model for the accuracy of galenic powder packaging is based on the technical analysis of its performance indicators and their comparison with documented studies of a high methodological level. The model was designed to predict weight accuracy in the galenic powder packaging process, using sequential variables such as the rotation speed of the auger and the electrical pulses per revolution of the auger. In this context, key metrics such as *Accuracy*, *Precision*, *Recall* (sensitivity), and *F1-score* (harmonic mean of precision and sensitivity) were applied, obtaining values above 80% in all cases, demonstrating a high predictive capacity.

This performance is supported on a study [25], where a hybrid ensemble model is proposed that integrates multiple deep architectures to classify genetic mutations in cancer patients. In this work, the authors applied their model to the Kaggle clinical dataset "MSK-Personalized Medicine", obtaining high-level validation indicators: *Accuracy* = 80.6%, *Precision* = 81.6%, *Recall* = 80.6%, and *F1-score* = 83.1%. These results are lower than those obtained in our research "Predictive model based on neural networks to optimize the accuracy of dosing galenic powders in rigid containers", where the following were obtained: *Accuracy* = 97.5%, *Precision* = 100%, *Recall* = 97.3%, and *F1-score* = 97.5%. The findings reinforce the idea that when RNNs exceed 80% in their validation indicators, they are highly effective in modeling sequential data and can be used in an automated production line.

Furthermore, the work [26] establishes that predictive models with F1 and AUC metrics above 80% are suitable for implementation in industrial environments. The authors emphasize that the *F1-Score* provides a balanced metric, ideal for processes where false positives and false negatives have an economic impact, such as in weight control in the packaging of galenic powders. This assertion reinforces the reliability of the proposed model for plant application by avoiding losses due to overfilling or underfilling, which entails raw material losses, or deficient filling, also known as underfilling, which affects product quality.

Regarding statistical validation, the model underwent normality testing (Kolmogorov-Smirnov) and t-tests comparing means to evaluate the difference in performance between the control group with data from the production line and the experimental group obtained after applying the RNN model. A statistically significant difference was found ($p < 0.05$), indicating that the RNN model substantially improves process accuracy.

7 CONCLUSIONS

Regarding the variables that affect the capacity of the galenic powder packaging process, the C_p and C_{pk} indicators of the control group were optimized 4.53 times with the application of recurrent neural networks, using historical data as inputs.

From 1430 data, 143 groups were formed, three groups were omitted and 28 subgroups of 5 data each were formed, each of the 28 subgroups had their C_p and C_{pk} calculated for the control group and the experimental group, the data are integrated and the average C_p and C_{pk} for the CG = 1.333, likewise the average C_p and C_{pk} of the SG = 1.333. Then: (1.333 – 0.2954)/(0.2954) = 3.5125. Then, it is stated that the indicators of the capacity of the galenic powder packaging process C_p and C_{pk} have been optimized by 351.25% by the application of the RNN model.

To verify that the design of a prediction model based on neural networks determines the best accuracy of the packaging of galenic powders, 30 random data were taken as a control group to which the RNN model was applied and the experimental group was obtained, the averages were calculated for GC = 52.28 g and GE = 50.64 g, it is known that the tolerance allowed for packaging applying the RNN model is 50.2 to 50.7 grams, the average of the allowed tolerance being 50.45 grams, then the CG error: 51.28-50.45 = 0.83 grams and the EG *error*: 50.64 – 50.45 = 0.19 grams, from these data it is obtained that (0.83 – 0.19)/(0.83) = 0.7711, then it is stated that the RNN model improved by a 77.11% weight accuracy in packaging of Sodium Bicarbonate × 50 grams

To verify that a model based on neural networks can effectively predict the weight accuracy in the packaging of galenic powders, from the operating parameters of the system, it was necessary to experiment with the data taken from the production line (control group), the already trained RNN model was applied and results were obtained (experimental group) within the established tolerance of 50.2 to 50.7 grams, the metrics were observed, being the Accuracy 97.5% the Precision 100%, the Recall 97.3% and the F1-Score 97.5% these values are very favorable which ensure an effectiveness of the model over 97%.

Of the 1430 data taken from 11 production batches of Sodium Bicarbonate × 50 grams in rigid packaging, from the

Pharmaceutical Laboratory, the highest weight is 52.5 grams and the lowest is 50.3 grams, with an average of 51.3 grams of samples, applying the RNN model, the average is 50.6 grams, which allows us to affirm that there will be a saving of 0.7 grams per bottle of Sodium Bicarbonate × 50 grams, which is equivalent to 1.36% of raw material per bottle.

From the study carried out, it is stated that the proposed RNN model significantly improves weight accuracy in the packaging of galenic powders, considering historical variables.

8 REFERENCES

[1] Castro Maldonado, J. J., Gómez Macho, L. K., & Camargo Casallas, E. (2023). La investigación aplicada y el desarrollo experimental en el fortalecimiento de las competencias de la sociedad del siglo XXI. *Tecnura, 27*(75), 140–174. (in Spanish) https://doi.org/10.14483/22487638.19171

[2] Díaz, E. E., Díaz, C., Flores, L. C., & Heyser, S. (2009). Estudio de la Variabilidad de Proceso en el Área de Envasado de un Producto en Polvo. *Información tecnológica, 20*(6). (in Spanish) https://doi.org/10.4067/S0718-07642009000600013

[3] Gallo, L., Chanampa, L., Gonzales, A., Calcagno, A., Razuc, M., & Natalini, P. (2019). *Tecnología farmacéutica: Diseño y preparación de formulaciones*. Editorial de la Universidad Nacional del Sur. Ediuns. (in Spanish)

[4] United States Pharmacopeia. (2023). Perspectiva (Farmacopeica) de la USP sobre Fabricación Continua de Medicamentosa. *Pharmacopoeial Forum, 44*(6), 1–30. (in Spanish)

[5] Santamaría Alvarez, F. N., Kleinebudde, P., & Thies, J. (2024). Influence of refilling on dosing accuracy of loss-in-weight powder feeders in continuous manufacturing. *Powder Technology, 436*, 119501. https://doi.org/10.1016/j.powtec.2024.119501

[6] Mauro, L. (2024, February). Non-Sterile Clinical GALENIC LABORATORY: A Scientific Discipline between Laboratory Practice Clinical Pharmacy and Personalized Pharmacological Therapy. *Archives of Pharmacy & Pharmacology Research, 4*(1). https://doi.org/10.33552/APPR.2024.04.000579

[7] Luisetto, M. et al. (2024). Galenic Laboratory: State of the Art—A Scientific and Technological Discipline, Innovation and Management. *Open Access Journal of Pharmaceutical Research, 8*(2). https://doi.org/10.23880/oajpr-16000316

[8] Juárez-Hernández, J. E., & Carleton, B. C. (2022). Pediatric oral formulations: Why don't our kids have the medicines they need? *British Journal of Clinical Pharmacology, 88*(10), 4337–4348. https://doi.org/10.1111/bcp.15456

[9] Desideri, I., Martinelli, C., Ciuti, S., Uccello Barretta, G., & Balzano, F. (2022). Lopinavir/ritonavir, a new galenic oral formulation from commercial solid form, fine-tuned by nuclear magnetic resonance spectroscopy. *European Journal of Hospital Pharmacy, 29*(5), 259–263. https://doi.org/10.1136/ejhpharm-2020-002389

[10] da Silva, M. R. M., Dysars, L. P., dos Santos, E. P., & Ricci Júnior, E. (2020). Preparation of extemporaneous oral liquid in the hospital pharmacy. *Brazilian Journal of Pharmaceutical Sciences, 56*. https://doi.org/10.1590/s2175-97902019000418358

[11] Toktonalieva, N., & Toktonaliev, I. (2020). History and Background of the Implementation of Good Manufacturing Practice Standards in the Pharmaceutical Industry (Review). *Bulletin of Science and Practice, 6*(9), 182–191. https://doi.org/10.33619/2414-2948/58/17

[12] La Rosa Basurco, J. (2023, April). *El reto de formalizar*. Ministerio de Economía y Finanzas del Perú. (in Spanish)

[13] Techflow Pack. (2024, May 27). *Envasado eficiente de polvo; agilización de los procesos de envasado con máquinas avanzadas*. TECHFLOW PACKAGING SOLUTIONS ENGINEERING CORP. (in Spanish)

[14] Ishfaque, M., Dai, Q., ul Haq, N., Jadoon, K., Shahzad, S. M., & Janjuhah, H. T. (2022). Use of Recurrent Neural Network with Long Short-Term Memory for Seepage Prediction at Tarbela Dam, KP, Pakistan. *Energies, 15*(9), 3123. https://doi.org/10.3390/en15093123

[15] Varela-Arregoces, E., & Campells-Sanchez, E. (2011, June). Redes Neuronales Artificiales: una revisión del estado del arte, aplicaciones y tendencias futuras. *Universidad Simon Bolivar*, 18–27. (in Spanish)

[16] Perez Ortiz, J. A. (2002). *Modelos predictivos basados en redes neuronales recurrentes de tiempo discreto* [Doctoral thesis, Universidad de Alicante]. España. (in Spanish)

[17] George Reyes, C. E. (2019). Methodological strategy to develop the state of the art as a product of educative research. *Praxis Educativa, 23*(3), 1–14. (in Spanish) https://doi.org/10.19137/praxiseducativa-2019-230307

[18] Cruz Martinez, A. G. (2018). El estado del arte en la construcción de la tesis doctoral y la gestión de nuevos conocimientos. *Revista Caribeña de Ciencias Sociales*, Universidad Pedagógica Nacional, México. Retrieved October 26, 2024, from https://www.eumed.net/rev/caribe/2018/04/gestion-nuevos-conocimientos (in Spanish)

[19] Arcos, L., & Armas, A. (2019, June). Estabilidad en cuanto a color y peso, de resinas compuestas tipo flow tras contacto con bebidas gaseosa: estudio in vitro. *Revista Odontología Vital*, 59–64. (in Spanish) https://doi.org/10.59334/ROV.v1i30.138

[20] Atencia, C., Reyes, C., & Bautista, A. (2022, August). Aplicación de control estadístico de calidad en bloques bio-climáticos de arcilla mediante gráficos de control por variables. *Universidad Nacional Experimental de Táchira*, 202–2108. (in Spanish)

[21] Xavier, O. C., Pires, S. R., Marques, T. C., & Soares, A. da S. (2022, June). Identificação de evasão fiscal utilizando dados abertos e inteligência artificial. *Revista de Administração Pública, 56*(3), 426–440. (in Spanish) https://doi.org/10.1590/0034-761220210256

[22] Herrera Medina, N., Rivera Gutiérrez, S., & Espinoza-Navarro, O. (2022). Marco Ético y Jurídico en la Investigación Científica en Chile: Acreditación de Comités Ético Científicos. *International Journal of Morphology, 40*(4), 953–958. (in Spanish) https://doi.org/10.4067/S0717-95022022000400953

[23] INDECOPI. (2008). *Norma metrológica peruana* (pp. 1–27). Perú: Lima. (in Spanish)

[24] Pheng, T., Chuluunsaikhan, T., Ryu, G.-A., Kim, S.-H., Nasridinov, A., & Yoo, K.-H. (2022). Prediction of Process Quality Performance Using Statistical Analysis and Long Short-Term Memory. *Applied Sciences, 12*(2), 735. https://doi.org/10.3390/app12020735

[25] Aburass, S., Dorgham, O., & Al Shaqsi, J. (2024, December). A hybrid machine learning model for classifying gene mutations in cancer using LSTM, BiLSTM, CNN, GRU, and GloVe. *Systems and Soft Computing, 6*, 200110. https://doi.org/10.1016/j.sasc.2024.200110

[26] Esraa, R. A. (2024, September). Improving Industrial Quality Control by Machine Learning Techniques. *Journal La Multiapp, 5*(5), 692–711.

Authors' contacts:

Jaime Antonio Cancho Guisado, Mg.
(Corresponding author)
Universidad Nacional Federico Villarreal, Graduate School,
Jr. Prolongations Camaná N°1014, Lima 01, Perú
(511) 998945990,
jcancho@unfv.edu.pe

Ciro Rodriguez Rodriguez, Dr.
Universidad Nacional Mayor de San Marcos,
Lima, Perú
(511) 990427136
crodriguezro@unmsm.edu.pe

Ivan Carlo Petrlik Azabache, Dr.
Universidad Nacional Federico Villarreal,
Lima, Perú
(511) 992502130
ipetrlik@unfv.edu.pe

Milner David Liendo Arévalo, Mg.
Newman Escuela de Posgrado,
Tacna, Perú
(511) 841983001
Mdliendoa1051@gmail.com

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240506194710

Preliminary communication
Received: 2024-05-06, Accepted: 2024-09-11

Enhancement of Powder Mixed Electrical Discharge Machining Performance Using Nano SiO_2 Powder Additive

Dunya Adnan Ghulam*, Abbas Fadhil Ibrahim

Abstract: Nanopowder-mixed EDM was employed to machine metals that are challenging to machine by traditional processes. In this study, Inconel 718 was machined by adding nano-SiO_2 to a soybean oil dielectric with the use of magnetic field to enhance the process's functioning. The impact of machining factors (current, pulse on time, SiO_2 concentration, and magnetic field) on white layer thickness (*WLT*), heat-affected zone (*HAZ*), surface roughness (*Ra*), material removal rate (*MRR*), tool wear rate (*TWR*), and surface crack density (*SCD*) were studied. The findings demonstrated that adding nano-SiO_2 to the insulating fluid improved process's functioning, whereas the *WLT* and *HAZ* were enhanced by 37.68% and 42.16%, respectively. Also, *Ra*, *MRR*, and *TWR* were improved by 9.59%, 50.93%, and 66.44%, respectively, as compared to machining without powder addition. The SEM micrographs revealed that fewer cracks were produced on the workpiece's surface with 4 g/l of added SiO_2, and the *SCD* improved by 12.81%.

Keywords: Inconel 718; magnetic field assistance; NPMEDM; nano SiO_2; white layer thickness

1 INTRODUCTION

Shear force is employed in processes of traditional machining to form the machined parts to the required shape. When dealing with processed materials that possess elevated toughness and hardness levels, these processes become challenging to complete. This is because of the wear caused by the friction and the deformation process between the machined part and the used tool. Therefore, non-traditional processes were used, which involve thermal processes such as electrical discharge machining (EDM), which is one of the most widely used machining techniques to machine hard to machine materials by conventional processes [1]. In (EDM), the electrical energy transforms into thermal energy, which develops a plasma channel between the electrode and the workpiece, which causes the surface of the workpiece to vaporize after melting. EDM is effective for machining intricate shapes of any conductive material, regardless of its properties because it is contactless process [2-4]. A series of pulses are applied in the EDM process to both of the electrodes, i.e., the tool and the workpiece, which are separated by a very tiny gap called the machining gap, in which the plasma channel develops [5]. However, the application of EDM is limited by some aspects, like poor surface finish, craters, holes, voids, and cracks [6]. To overcome these limitations, a powder mixed electrical discharge machining (PMEDM) technique is used, in which powders are combined with the insulating fluid to modify the machined surface [7]. The added fine powders with suitable concentrations lower the breakdown strength of the dielectric fluid, modify the plasma channel, enlarge the gap distance, and helps the sparking to be uniformly distributed in the machining zone, resulting in an enhancement of machining performance [8]. Nickel based alloys are difficult to machine by traditional methods because of their high toughness and strength; therefore, the PMEDM process is an efficient way to machine them.

Many researchers have studied the machining of Inconel alloys using this process by adding different powders to various dielectric fluids. Inconel 718 machining and improving the performance of NPMEDM process was studied by adding nano alumina powder and using copper tungsten electrode. As compared to traditional electrical discharge machining, MRR was enhanced by 50% when using Nano alumina powder added to kerosene [9]. To enhance the performance of EDM process in purview surface integrity, *TWR*, and *MRR*, a 0.5 g/l of multi walled carbon nano tubes (MWCNT) was added to kerosene to machine Inconel 718 by [10]. Findings detected that MRR was improved by 44.95 %. *TWR* and *SR* decreased by 64.1 % and 14.1 % respectively as compared to traditional EDM. The reduction in recast layer thickness was 16.28% when using 0.5 g/l of MWCNT. In [11] authors studied graphene nanofluid added to deionized water used as a dielectric fluid. The impact of Ton, V, and I on *SR*, *TWR* and *MRR* were investigated. It was concluded that *SR*, *TWR*, *MRR* enhanced by 14 %, 2 %, and 20.1 % respectively. A. Ranjan et al. investigated Inconel 718 machining by adding four kinds of powders (aluminum, aluminum oxide, silicon carbide, as well as graphite) to EDM oil. The obtained results showed that when powders were added, the *MRR* improved by 5.93%, while when surfactant was added, the *MRR* improved by 13.56 %. Also, the SR was lowered by 43.9 % when adding powders and by 45.05 % when adding surfactant. The best SF was obtained when Gr powder was used due to its low density [12]. Influence of adding different concentrations of Nano SiC to EDM oil was examined for machining Inconel 718 superalloy by [13]. The authors concluded that adding Nano silicon carbide at 0.4 g/l enhanced the MRR by 163 %, while *SR* and *TWR* were eliminated approximately by 17 % and 24 %, respectively. Also, the WLT was significantly lowered after the addition of Nano silicon carbide. The effect of adding Nano alumina to deionized water on *TWR*, *SR*, as well as *MRR* was examined by [14]. It was concluded that *MRR* was higher by about 7 % as compared to conventional EDM. As well as the *TWR* and *SR* were eliminated by 6 % and 10 %, respectively. S. Kumar investigated a hybrid method of EDM used to operate Inconel

706, adding silicon carbide and boron carbide powders with magnetic field assistance. The results showed that the maximum material removal rate produced using the B_4C powder with the highest level of magnetic field intensity 0.66 T. The minimal surface roughness achieved was when using SiC powder without using a magnetic field [15]. In [16], authors studied the processes of Inconel 718 by introducing multi walled carbon nano-tubes to distilled water as an insulating fluid. It was discovered that the peak current, powder concentration, and pulse on time have significant effects on the responses; it was found that an ideal range of *MRR* and *TWR*, as well as less *SR*, could be achieved by combining Peak Current 17–20 A with Pulse on Time 270–350 µs and 1.50–2.25 g/l of multi walled carbon nano tubes concentration. Furthermore, other hard to machine materials were processed using different added powders; [17] studied the PMEDM process by adding molybdenum (Mo) powder to the insulating fluid to machine AISI H13 tool steel. The results indicated that the mechanical properties of the machined surface were improved and the fold hardness increased four times. V. T. Le studied the micro hardness and surface roughness of the SKD61 workpiece by adding tungsten carbide powder to the hydrocarbon dielectric fluid. Evaluation and analysis indicate that the addition of the powder features a beneficial influence on the characteristics of the workpiece at short pulse on time and low current. The surface roughness and micro hardness were enhanced by 53.25 % and 81.5 %, respectively, as compared to machining without powder addition [18]. G. Keskin et al. studied the effect of mixing nano graphite powder with kerosene to machine AA7075/ B_4C and AA7075/ B_4C + SiC composites. Current, pulse on time, and the reinforcement ratio were the machining parameters. Results showed that the optimum factors for material removal rate were found at 8 A of current, 150 µs of pulse on time, and 10 % of reinforcement ratio [19]. In [20] the authors tried to achieve the minimum thickness of the defect layer (recast layer) by adding graphite powder. Pulse on time, duty cycle, discharge current, and powder concentration were selected as input parameters to machine titanium alloy. The findings revealed that current was the most effective parameter on recast layer thickness, and the minimum measured thickness of it was at 1.5 A of applied current, 30 µs of applied pulse on time, 50 % of duty cycle, and 12 g/l of graphite powder concentration.

It was evident from the literature survey that there is a dearth of research on the subject of machining Inconel 718 with the use of SiO_2 nanoparticles in conjunction with the magnetic field assistance technique. Also, not much attention has been paid to studying the zone impacted by heat (*HAZ*) and the thickness of the recast layer (*WLT*) of the Inconel 718 machined surface. Therefore, within the current research, improvement of NPMEDM performance was investigated using a magnetic field assistance technique, and the effect of current, pulse on time, nano SiO_2 addition, as well as magnetic fields on process performance regarding the *WLT* and *HAZ*, in addition to the material removal rate (*MRR*), tool wear rate (*TWR*), surface roughness (*Ra*), and surface crack density (*SCD*), were evaluated.

2 EXPERIMENTAL WORK

Inconel 718 alloy was machined using (CHMER EDM) of the model CM 323+50N electrical discharge machine. 25 × 10 × 4 mm was the dimensions of the workpiece, and the chemical composition of Inconel 718 was as shown in Tab. 1. The workpiece's ultimate tensile strength and yield stress are 965 MPa and 550 MPa, respectively; its density is 8.912 g/cc; its thermal conductivity is 11.2 W/m·K; and the melting point of the workpiece is 1260–1335 °C.

Table 1 Chemical composition of Inconel 718

Element	Wt.%
C	0.03
Si	0.29
Mn	0.30
S	0.01
Cu	0.015
Cr	18.3
Ni	53
Mo	2.9
Co	0.8
Al	0.5
Ti	1.05
Nb	4.8

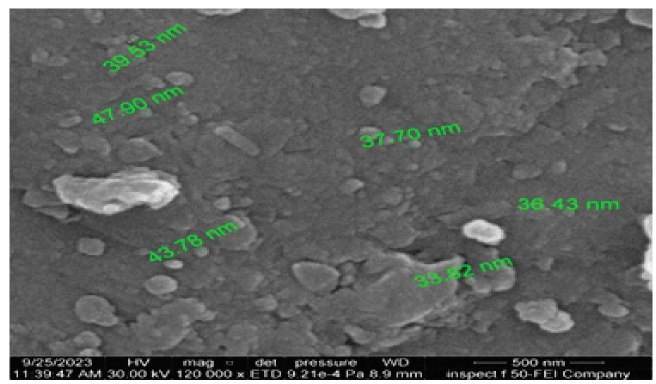


Figure 1 Nano-silicon oxide (SiO_2) particles

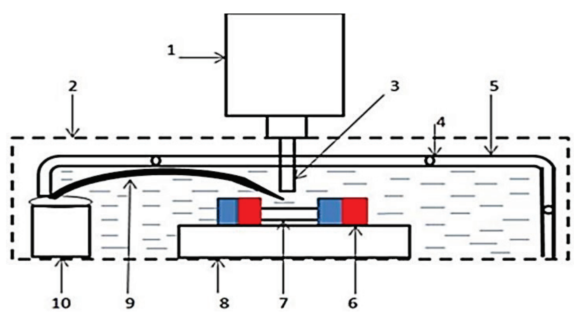


Figure 2 Scheme of MF-PMEDM: 1 – Servo system, 2 – Machining tank, 3 – Electrode, 4 – Hole, 5 – Pipe, 6 – Pair of permanent magnets, 7 – Workpiece, 8 - Vise, 9 – Flushing nozzie, 10 – Pump.

The tool electrode was made of copper; its dimensions were 15 × 10 × 70 mm. Nano-silicon oxide (SiO_2) was used as a powder additive (see Fig. 1). As an insulating fluid, soybean oil was employed in the 500 × 200 × 250 mm machining tank. Fig. 2 displays the schematic of the process, and it can be seen that the insulating fluid is flushed into the machining zone via the attached nozzle. A surfactant known as Tween60 was added to the dielectric to help keep the added nano SiO_2 suspended in the dielectric and not settle at the bottom of the tank. A magnetic field of (0.2 T) intensity was generated by a pair of permanent magnets; its intensity was examined using a digital gaussmeter device. The machining tank provided with a specific circulation system included a flushing nozzle, a pump, and a five millimeter diameter pipe that was attached within the processing tank's inside walls, this pipe was drilled with one millimeter diameter holes along its length in order to circulate the insulating fluid and avoid powder particles from collecting in the machining tank base (see Fig. 3).

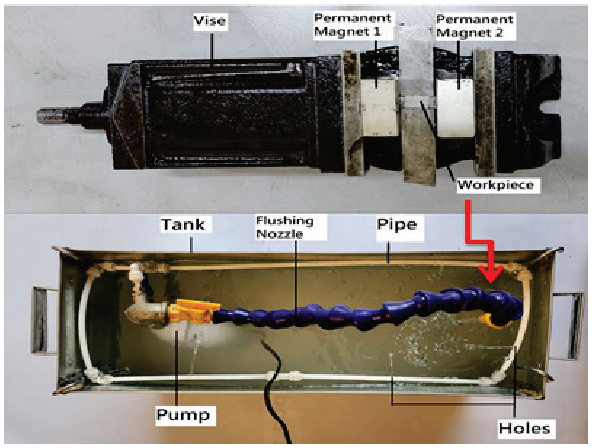


Figure 3 Machining tank

2.1 Selection of Process Parameters

Multi-level general full factorial design was used to design the experiments by Minitab 2022. The evaluated variable parameters were current (*I*), pulse on time (*Ton*), Nano SiO_2 concentration, magnetic field (*MF*) as shown in Tab. 2. The responses were white layer thickness (*WLT*), heat affected zone (*HAZ*), surface roughness (*Ra*), material removal rate (*MRR*), tool wear rate (*TWR*), and surface crack density (*SCD*). The fixed parameters were listed in Tab. 3.

Table 2 Input variables parameters

Parameter	Level 1	Level 2	Level 3
I	8 A	16 A	/
Ton	100 µs	200 µs	/
Con.	0 g/l	2 g/l	4 g/l
MF	0 T	0.2 T	/

Table 3 Fixed parameters

Parameter	Values
Pulse off time (Toff)	75 µs
Powder	SiO_2
Polarity	Straight
Gap voltage	240 V
Surfactant type	Tween60
Surfactant con.	1 ml/l
Depth of cut	1 mm

2.2 Evaluation of Experimental Outputs

An Italian OPTICA MET serial optical microscope (with high resolution optics, precise X-Y movement controls, and 10× to 100× magnification) was employed to measure the thicknesses of the recast layer and the zone impacted by heat.

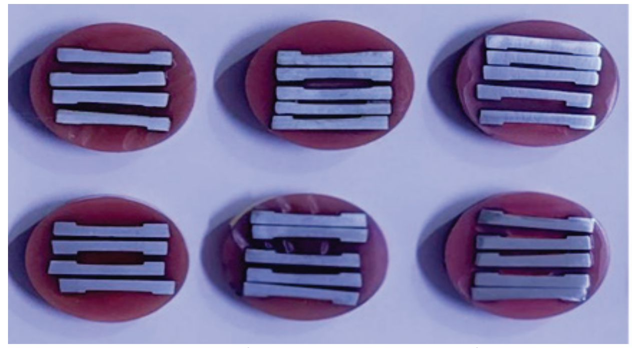


Figure 4 Prepared workpieces for measuring the thickness of white layer and heat affected zone

Prior to determining the thickness, the machined workpieces were prepared as shown in Fig. 4. The workpieces were set up in a suitable mold for grinding and polishing processes; emery paper with grit sizes of 100, 500, 1000, 1200, and 2000 was used for grinding; and polishing paper with alumina powder was used for the polishing process, followed by the etching process; then the average thicknesses of *WL* and *HAZ* were computed using the 400× magnification of the optical microscope.

Average of three readings for surface roughness was evaluated for all specimens by the Maher Fedral pocket Surf-PS1 device (the cutoff l_c was 0.25 mm automatic, the traversing length L_t was 1.75 mm automatic, and the number *n* of sampling lengths was 5).

Eq. (1) was used to calculate *MRR*, which is the specimen's weight difference before and after machining divided by the machining time and the density of Inconel 718 (0.00822 g/mm^3).

$$MRR = \frac{(W_a - W_b)}{\rho \cdot t}. \tag{1}$$

Where: W_b - the workpiece's weight in grams prior to machining; W_a - the workpiece's weight in grams following machining; ρ - the density of the Inconel 718 (0.00822 g/mm^3); t - machining time (min).

Electrode wear rate (*EWR*) which is the electrode weight that differs between electrode tool before and after machining relative to the machining time and copper density (0.008906 g/mm^3), it can be calculated as in Eq. (2).

$$EWR = \frac{(Ewt._b - Ewt._a)}{\rho \cdot t}. \tag{2}$$

Where: $Ewt._b$ - the weight of the electrode before machining (g); $Ewt._a$ - the weight of the electrode after machining (g); ρ - the density of the copper (0.008906 g/mm^3); t - machining time (min).

SCD was computed by dividing the average microcracks lengths by the micrograph's area that was captured by Axia Chemi SEM thermo scientific equipment (with thermionic tungsten filament source, automated sample navigation, a backscatter detector, and live quantitative elemental mapping), as shown in Eq. (3).

$$SCD = \frac{C_L}{A}. \tag{3}$$

Where: C_L - crack's length μm (average); A - micrograph's area (μm^2).

3 RESULTS AND DISCUSSION

3.1 Influencing Factors on White Layer Thickness (*WLT*) and Heat Affected Zone (*HAZ*)

The hardness of the machined surface is affected by *WLT*; therefore, it is essential to evaluate the effect of machining parameters on it. The *WL* develops as a consequence of the re-solidification of a portion of the melted material in the machining gap as it cools with dielectric fluid through the flushing process [21]. Since the EDM process is a thermal process, the extremely high temperature that is generated during the process alters the subsurface region structure down to a certain depth, termed the heat affected zone (*HAZ*) [22].

Adding SiO_2 nanoparticles to soybean oil result in reducing *WLT* due to the uniform distribution of energy on the machined surface; and the *WLT* was lowered due to the existence of nano SiO_2 in the machining zone, which in turn made the flushing technique more effective by enlarging the distance between the tool and the specimen and eliminating the re-solidification of debris on the machined surface. Applying greater values of pulse on time and current causes more metal to melt due to the plasma channel's raised discharge energy; therefore, more debris re-solidified and thicker *WL* produced. *MF* caused a slight rise in *WLT* when applied with the highest current value, i.e., 16 A, as shown in Fig. 5.

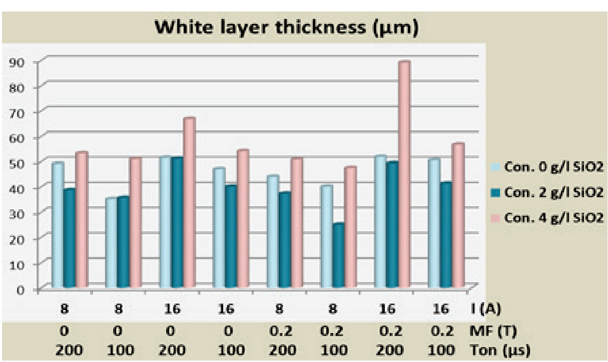


Figure 5 Effect of machining parameters on WLT of SiO_2 addition

Experimental results showed that the minimum measured *WLT* was 24.99 μm, when applying the machining conditions of run 14 (see Tab. 4). Microscopically pictures of machined specimens shown in Fig. 6. According to Tab. 5, which represents the ANOVA for *WLT*; it is noticeable that the model is statistically significant, and the more efficient parameter was current; next comes SiO_2 nanoparticles concentration and *Ton*; the effect of *MF* was insignificant.

Table 4 Results of measured *WLT*, *HAZ*, *MRR*, and *TWR*

I, A	*Ton*, μs	*Con.*, g/l	*MF*, T	*WLT*, μm	*HAZ*, μm	*Ra*, μm	*MRR*, m^3/min	*TWR*, mm^3/min
16	200	0	0.2	49.12	93.34	4.457	9.033	0.0298
8	200	2	0.2	35.2	94.557	4.413	3.619	0.0117
16	100	4	0.2	51.51	111.8	4.737	9.745	0.1684
8	100	4	0	47.02	103.02	4.304	3.266	0.0150
8	200	4	0	43.92	137.5	4.754	3.789	0.0241
8	200	0	0	40.1	92.49	4.569	3.766	0.1276
8	100	0	0	51.84	115.2	5.297	3.018	0.0983
16	100	0	0.2	50.5	73.33	4.828	8.839	0.4685
16	100	2	0.2	38.66	108	5.594	9.582	0.1567
8	100	4	0.2	35.66	74.99	5.32	3.043	0.0172
8	200	0	0.2	51.12	74.02	5.945	3.282	0.1226
16	200	4	0	40.1	132	5.852	12.11	0.1277
8	200	2	0	37.25	154.9	4.843	3.420	0.0159
8	100	2	0	24.99	53.5	4.634	3.122	0.0135
16	200	0	0	49.33	103.25	5.246	11.04	0.2824
16	200	2	0.2	41.32	107.49	5.137	10.50	0.0592
16	100	2	0	53.33	177.78	4.447	9.789	0.1402
8	200	4	0.2	50.98	86.66	4.23	3.602	0.0189
8	100	0	0.2	66.92	276.15	4.513	3.030	0.1060
16	200	4	0.2	54.17	138.35	4.347	13.63	0.0989
16	200	2	0	50.83	179.17	4.416	12.08	0.0996
8	100	2	0.2	47.49	207.83	4.469	3.835	0.0182
16	100	4	0	89.16	153.33	4.789	10.58	0.1321
16	100	0	0	56.66	140.57	4.63	10.40	0.4205

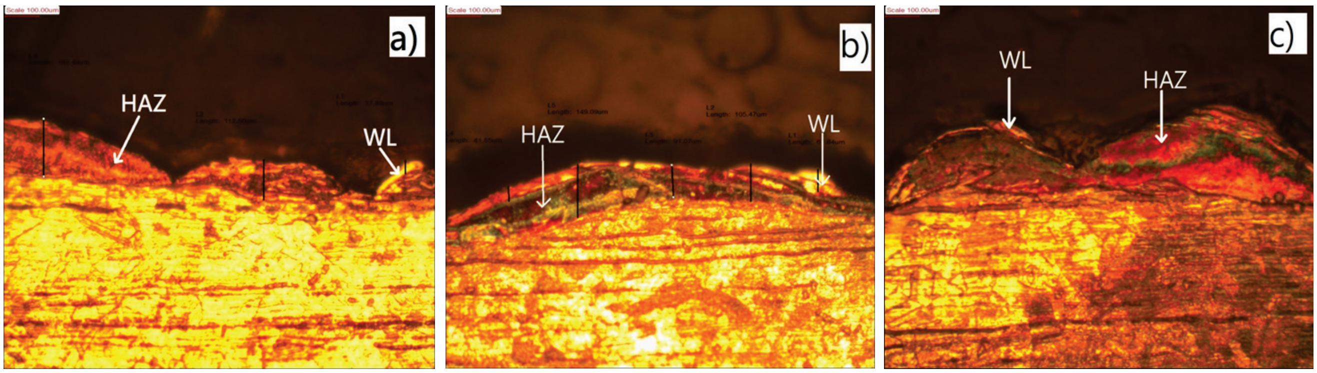
Table 5 ANOVA for WLT of SiO_2 nanoparticles addition

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	4	2813.2	703.29	19.71	0.000
Linear	4	2813.2	703.29	19.71	0.000
I	1	841.5	841.55	23.58	0.000
Ton	1	493.3	493.35	13.83	0.001
Con.	2	1478.3	739.13	20.71	0.000
Error	19	678.0	35.68		
Total	23	3491.1			

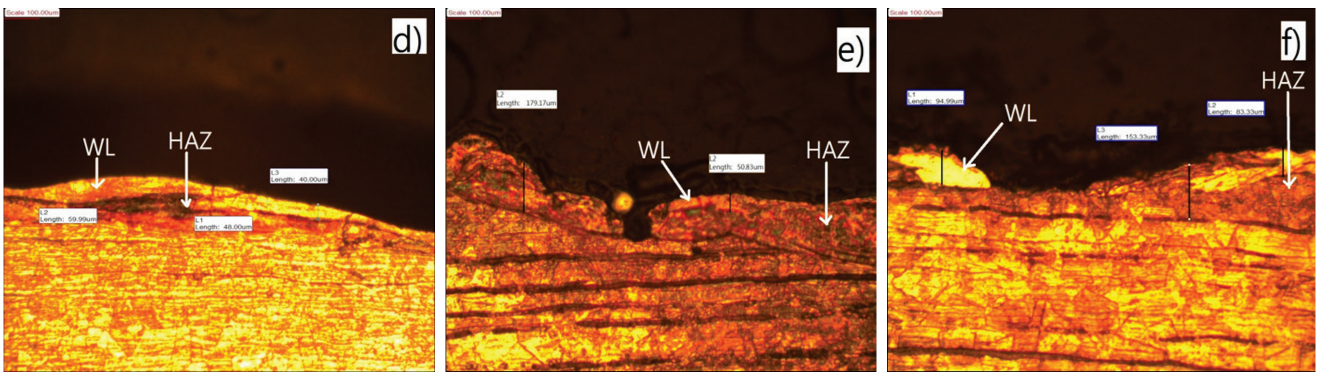
3.2 Influencing Factors on Surface Roughness (*Ra*)

In NPMEDM, the discharge gap widens because of the existence of the added nanoparticles in the insulating medium, which disperse the discharge over the machined workpiece. In addition, the presence of these powder particles contributes to lowering the energy of the discharge, which enhances the surface quality by enabling more effective and regulated localized heat delivery [23]. Better surface finish was obtained when adding nano particles of SiO_2 to soybean oil due to the improvement in the flushing mode as a result of nanoparticles existence in the dielectric fluid which enlarge the machining gap distance. Experimental results observed that adding 2 g/l of SiO_2 causes a rise in Ra in comparison with machining without adding SiO_2 nanoparticles; while adding it with 4 g/l powder concentration led to improve the *Ra* and reduce it to values less than the values of machining without powder addition because the process became more stable, as shown in Fig. 7.

Moreover, increasing in applied current and pulse on time cause increasing in surface roughness. The minimum obtained *Ra* was 4.23 µm when applying current with 8 A, *Ton* 100 µs, 4g/l SiO_2 concentration, and without employing MF assistance, as shown in Tab. 4. The maximum measured *Ra* was 5.9447 µm when 2 g/l of SiO_2 was added, at 16 A applied current, 200 µs applied *Ton*, and without using MF, this measured value reduced to 4.5127 µm when adding the nanopowder with 4 g/l and fixing other parameters. The specimens that were machined when nano SiO_2 was added at 2 g/l and without applying MF showed a large increment in surface roughness as compared to machining without powder addition, as shown in the first four groups of columns compared to the second four groups in Fig. 7. While the surface roughness reduced slightly but was still higher than machining without powder addition when adding nano SiO_2 at 2 g/l and applying 0.2 T of magnetic field due to the effect of the magnetic field on the movement of powder particles, which contributes to regulating their movement.



a) *I* = 8 A, *Ton* = 200 µs, *Con.* = 0 g/l, *MF* = 0.2 T b) *I* = 16 A, *Ton* = 200 µs, *Con.* = 0 g/l, *MF* = 0.2 T c) *I* = 8 A, *Ton* = 200 µs, *Con.* = 2 g/l, *MF* = 0.2 T



d) *I* = 16 A, *Ton* = 200 µs, *Con.* = 2 g/l, *MF* = 0.2 T e) *I* = 8 A, *Ton* = 200 µs, *Con.* = 4 g/l, *MF* = 0.2 T f) *I* = 16 A, *Ton* = 200 µs, *Con.* = 4 g/l, *MF* = 0.2 T

Figure 6 Microscopically pictures of WLT and HAZ

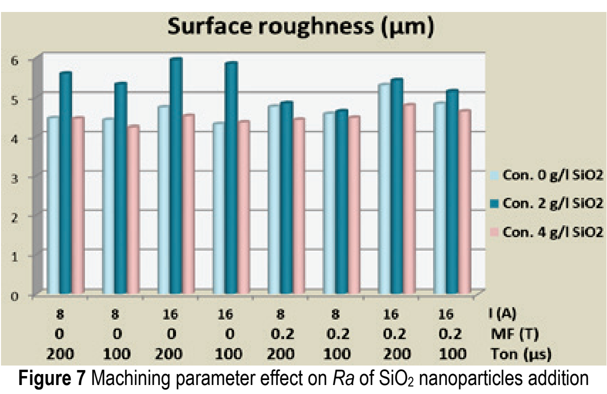


Figure 7 Machining parameter effect on *Ra* of SiO_2 nanoparticles addition

Table 6 ANOVA for Ra of SiO_2 nanoparticles addition

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	3	3.5871	1.19570	12.57	0.000
Linear	3	3.5871	1.19570	12.57	0.000
I	1	0.4243	0.42429	4.46	0.047
Con.	2	3.1628	1.58141	16.63	0.000
Error	20	1.9021	0.09511		
Total	23	5.4892			

According to ANOVA table it appears that the concentration was the most effective parameter of *Ra* followed by current effect, and the effect of other parameters were limited (see Tab. 6).

3.3 Influencing Factors on Material Removal Rate (*MRR*)

MRR is a remarkable technical indicator in the production process of products. Production efficiency in powder mixed electrical discharge machining is dependent on the quantity of material removal per unit time. Higher *MRR*, i.e., higher production efficiency [24].

When current reaches its maximum value, the current density rises and produces a lot of thermal energy in the machining gap, which raises the *MRR*. Also, an increase in pulse on time can be attributed to the increased amount of heat generated in the machining zone, which causes the *MRR* to increase these outcomes were found to be in agreement with the study of [25]. Current is the most effective parameter on *MRR*; it is directly correlated to *MRR*, as was confirmed in the experimental results of this study. When current increases from 8A to 16A, the discharge energy increases, which leads to the melting and evaporation of the machined material, as shown in Fig. 8. The addition of nano SiO_2 to soybean oil dielectric improved *MRR*; the presence of nano particles raised the dielectric's thermal conductivity; and the frequency of the sparks increased consequently enhancing the *MRR*. When the concentration of nano SiO_2 increases from 2 g/l to 4 g/l, *MRR* increases. The minimum measured *MRR* was 3.0184 mm^3/min when applying the machining conditions without adding nanoparticles, at 8 A of applied current, 100 µs of pulse on time, and without using a magnetic field, while when adding 2 g/l of nano SiO_2, the *MRR* increased to 3.1218 mm^3/min, and when adding 4 g/l of SiO_2, it was increased to 3.2665 mm^3/min, after fixing other process parameters. The maximum obtained *MRR* was 13.6326 mm^3/min; at 16A applied current, 200 µs applied pulse on time, 4 g/l nano SiO_2 concentration, and 0.2T of MF intensity, as listed in Table 4.

From the analysis of variance (ANOVA) for *MRR* listed in Tab. 7, it can be seen that the model is statistically significant, whereas *P* values are less than 0.05. Current was the most effective parameter on *MRR*.

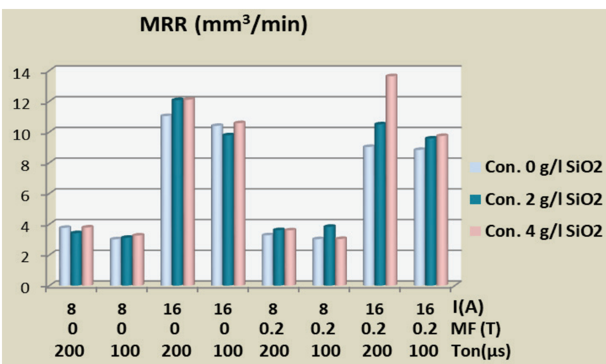


Figure 8 Effect of input parameters on *MRR*

Table 7 Analysis of variance for *MRR*

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	7	325.756	46.537	80.77	0.000
Linear	4	321.028	80.257	139.29	0.000
I	1	312.016	312.016	541.52	0.000
Ton	1	5.627	5.627	9.77	0.007
Con.	2	3.384	1.692	2.94	0.082
2-Way Interactions	3	4.728	1.576	2.74	0.078
I×Ton	1	2.217	2.217	3.85	0.067
I×Con.	2	2.511	1.256	2.18	0.146
Error	16	9.219	0.576		
Total	23	334.975			

3.4 Influencing Factors on Tool Wear Rate (*TWR*)

TWR represents the amount of eroded material from the used tool electrode. It can be calculated by dividing the lost weight of the electrode before and after machining by the machining time multiplied by the density of the electrode material. The accuracy of the machine part is largely affected by *TWR* because the machined part is an exact copy of the electrode profile. *TWR* is affected by tool shape, tool material, melting point, current, and pulse on time [26].

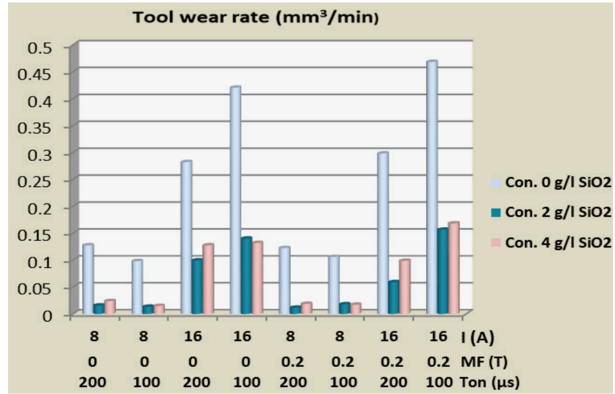


Figure 9 Effect of input parameters on TWR

As compared to the lowest obtained value for measured *TWR* in [27], which is 0.244 mm^3/min at 40 A of applied current and 400 µs of applied pulse on time to machine Inconel 718, the present study achieved lower values for measured *TWR*. The effect of machining parameters on *TWR* is shown in Fig. 9. Whereas increasing current from 8 A to 16 A caused an increase in *TWR* because of the greater discharging that generated more heat in the machining gab, which in turn led the machined metal to melt and evaporate. When *Ton* increased from 100 µs to 200 µs, it led to a slight reduction in *TWR*. This could be explained by the longer pulse on time leading to more metal to melt from the workpiece and more molten material to solidify on the surface of the electrode; thus, the MRR rises and the *TWR* reduces. Nanoparticles of SiO_2 added to soybean oil dielectric have greatly improved *TWR* through stabilizing the process by enlarging the machining gab distance between the tool and the specimen and enhancing the flashing mode, thus reducing *TWR*. On the other hand, MF has no impact on the *TWR*.

Table 8 Analysis of variance for *TWR*

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	5	0.25884	0.051768	11.39	0.000
Linear	4	0.22935	0.057338	12.62	0.000
I	1	0.10599	0.105991	23.33	0.000
Ton	1	0.02260	0.022596	4.97	0.039
Con.	2	0.10077	0.050383	11.09	0.001
2-Way Interactions	1	0.02949	0.029489	6.49	0.020
I×Ton	1	0.02949	0.029489	6.49	0.020
Error	18	0.08178	0.004543		
Total	23	0.34062			

The minimum value for *TWR* obtained was 0.01174 mm^3/min at 8 A applied current, 200 µs applied *Ton*, 2 g/l SiO_2 nanoparticles, and 0.2 T MF intensity. The maximum measured *TWR* was 0.4685 mm^3/min without powder addition at 16 A applied current, 100 µs applied *Ton*, and 0.2 T MF, while this measured value reduced to 0.1684 mm^3/min when adding 4 g/l of nano SiO_2 and decreased to 0.156676 mm^3/min when adding 2 g/l SiO_2 nanoparticles and fixing other parameters that demonstrate the effect of powder addition to the dielectric (see Tab. 4). ANOVA of the experimental results demonstrated that current has the greatest impact on *TWR*, thereafter powder concentration and *Ton*. The impact of MF was very insignificant on *TWR* compared to other parameters. Tab. 8 represents the ANOVA for *TWR*; the model is statistically significant whereas P-values are less than 0.05.

3.5 Surface Crack Density (*SCD*)

In NPMEDM, temperature rises beyond the material's melting point to melt and evaporate the machined material; simultaneously, a rapid cooling happens, resulting in thermal stress that is higher than the machined material's fracture strength and leads to crack formation. Cracking is among the more important surface flaws that reduce the corrosion and fatigue resistance of the material [28]. Due to increases in discharge energy, microcracks generates at greater current and pulse on time applied values, which leads to an increase in thermal tendency within the solidified layer and generates more residual stresses, thus increasing the *SCD* [29]. Powder adding to the dielectric technique was used to reduce the *SCD*. When nano SiO_2 was added, a greater number of sparks were produced in the gap, the energy of the discharge was reduced, and the plasma channel became larger, resulting in fewer cracks. The machined surface was seen to have complicated looks through the SEM micrograph, which included spherical particles, craters, voids, and cracks. As shown in Fig. 10a, the surface machined by adding nano SiO_2 has fewer cracks as compared to Fig. 10b, which was processed with no adding nano SiO_2 at the same magnification. To compute the amount of cracking, the *SCD* measures the average cracks' length per unit area. *SCD* for specimen machined at 16 A applied current, 200 µs applied *Ton*, 0.2 T *MF* intensity, and adding 4 g/l nano SiO_2 was 0.00313 $\mu m/\mu m^2$, and the *SCD* of specimen machined with the same applied conditions but without the addition of nano SiO_2 was 0.00359 $\mu m/\mu m^2$.

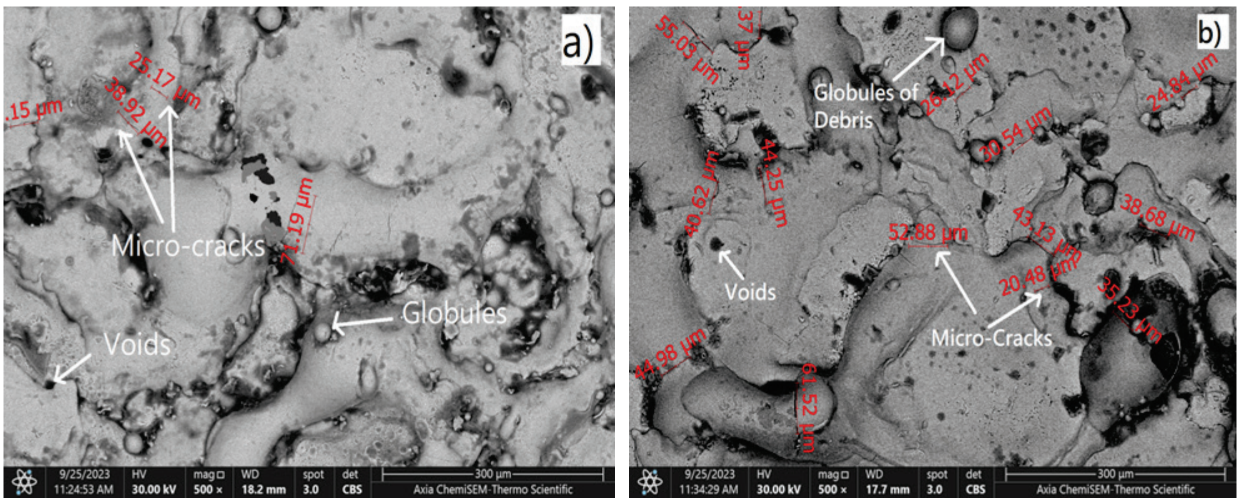


Figure 10 Cracks of machined surface with *I* = 16 A, *Ton* = 200µs, *MF* = 0.2 T, and (a) 4 g/l of nano SiO_2, (b) without adding powder.

4 CONCLUSIONS

For the purpose of enhancing its performance, the NPMEDM was examined in the current study. Nanoparticles of SiO_2 were introduced to soybean oil, a fluid employed as insulation. Practical experiments were concluded to evaluate the influence of controlling factors, in the context of peak current, pulse on time, SiO_2 nanoparticle concentration, and magnetic field, on NPMEDM performance with respect to *WLT*, *HAZ*, *Ra*, *MRR*, *TWR*, and *SCD*.

The outcomes of the experiments demonstrated that the addition of SiO_2 nanoparticles to the insulating fluid led to improved process's performance with regard of all the outputs (*WLT*, *HAZ*, *Ra*, *MRR*, *TWR*, and *SCD*). The minimum *WLT* and *HAZ* were 24.99 and 53.5 µm, respectively, when applying the machining conditions of 8A of applied current, 100 µs of applied *Ton*, and 2 g/l of SiO_2 nanoparticles without using *MF*. The lowest amount attained of *Ra* was 4.23 µm when applying 8 A of current, 100 µs of *Ton*, 4 g/l of SiO_2 concentration, and without employing *MF*. The maximum obtained *MRR* was 13.6326 mm^3/min at 16 A applied current, 200 µs applied pulse on time, 4 g/l nano SiO_2 concentration, and applying 0.2 T of *MF* intensity. The minimum value for *TWR* obtained was 0.01174 mm^3/min at 8A applied current, 200 µs applied *Ton*, 2 g/l SiO_2 nanoparticles concentration, and 0.2 T *MF* intensity. Lower *SCD* was obtained when adding 4 g/l of SiO_2 nanoparticles to the used insulating fluid; it was 0.00313 $\mu m/\mu m^2$, while the *SCD* was 0.00359 $\mu m/\mu m^2$ when machining without adding SiO_2 nanoparticles to the insulating fluid. Run #14 has the optimal combination of applied process parameters (current of 8 A applied value, pulse on time of 100 µs, nano SiO_2 of 2 g/l, and 0.2 T of magnetic field intensity).

NPMEDM is a crucial step in the machining of superalloys, though its mechanisms are still poorly understood, and this study aims to improve the process performance. Subsequent research efforts could delve deeper into the underlying mechanism. To fully understand how the type of the nanoparticle, its shape, and its size affect the machined part's corrosion resistance, fatigue, and surface integrity, more research is required. Further, investigations might be looking into optimization of nanoparticles parameters to improve the effectiveness and reliability of the NPMEDM. Future research can yield insightful information by addressing these factors, and this could help in the development of more efficient machining for advanced materials.

5 REFERENCES

[1] Mookam, N., Sunasuan, P., Madsa, T., Muangnoy, P. & Chuvaree, S. (2021). Effects of Graphite and Boron Carbide Powders Mixed into Dielectric Fluid on Electrical Discharge Machining of SKD 11 Tool Steel. *Arab. J. Sci. Eng.*, 46(3), 2553-2563. https://doi.org/10.1007/s13369-020-05156-4

[2] Payal, H., Bharti, P. S., Maheshwari, S. & Agarwal, D. (2020). Machining characteristics and parametric optimisation of inconel 825 during electric discharge machining. *Teh. Vjesn., 27*(3), 761-772. https://doi.org/10.17559/TV-20190214135509

[3] Ibrahim, A. F. (2022). Evaluation of Surface Roughness And Material Removal Rate in Electrical Discharge Machining of Al-Alloy With 10% SiC. *IIUM Eng. J., 23*(1), 349-357. https://doi.org/10.31436/IIUMEJ.V23I1.2114

[4] Shirguppikar, S. et al. (2021). Assessing the effects of uncoated and coated electrode on response variables in electrical discharge machining for ti-6al-4v titanium alloy. *Tribol. Ind., 43*(4), 524-534. https://doi.org/10.24874/ti.1020.12.20.03

[5] Singh, R. K., Tiwari, S. K., Srivastava, S. C. & Kumar, B. (2023). Hybrid Taguchi-GRA-CRITIC Optimization Method for Multi-Response Optimization of Micro-EDM Drilling Process Parameters. *Teh. Vjesn., 30*(3), 804-814. https://doi.org/10.17559/TV-20220601114015

[6] Nguyen, T. T. & Nguyen, K. D. (2021). Multi-attribute optimization of the wedm process for surface characteristics and productivity. *Teh. Vjesn., 28*(2), 473-480. https://doi.org/10.17559/TV-20190728094458

[7] Gul, I. A., Abdul-Rani, A. M., Al-Amin, M. & Garba, E. (2023). Elucidating Powder-Mixed Electric Discharge Machining Process, Applicability, Trends and Futuristic Perspectives. *Machines, 11*(3). https://doi.org/10.3390/machines11030381

[8] Kavade, M. V., Mohite, S. S. & Unaune, D. R. (2019). Application of metal powder to improve metal removal rate in Electric Discharge Machining. *Mater. Today Proc*., 16, 398-404. https://doi.org/10.1016/j.matpr.2019.05.107

[9] Ahmad, S. et al. (2017). Powder Mixed Dielectric in Electrical Discharge Machining of Inconel 718. *J. Mech. Eng., SI 4*(3), 110-127.

[10] Jadam, T., Sahu, S. K., Datta, S. & Masanta, M. (2019). EDM performance of Inconel 718 superalloy: application of multi-walled carbon nanotube (MWCNT) added dielectric media. *J. Brazilian Soc. Mech. Sci. Eng., 41*(8). https://doi.org/10.1007/s40430-019-1813-9

[11] Paswan, K., Pramanik, A. & Chattopadhyaya, S. (2020). Machining performance of Inconel 718 using graphene nanofluid in EDM. *Mater. Manuf. Process., 35*(1), 33-42. https://doi.org/10.1080/10426914.2020.1711924

[12] Ranjan, A., Chakraborty, S., Kumar, D. & Bose, D. (2020). An Investigation on Surfactant Added PMWEDM of Inconel 718. *Int. J. Automot. Mech. Eng., 17*(3), 8140-8149. https://doi.org/10.15282/ijame.17.3.2020.07.0611

[13] Elumalai, B., Gowri, S., Hariharan, P. & Arun Pillai, K. V. (2022). Experimental investigations on μeD milling of inconel 718 with nano SiC abrasive mixed dielectric. *Mater. Res., 25*. https://doi.org/10.1590/1980-5373-MR-2021-0468

[14] R. K. Patel and M. K. Pradhan, (2023). Machining of nickel-based super alloy inconel 718 using alumina nanofluid in powder mixed electric discharge machining. *Mater. Res. Express, 10*(3). https://doi.org/10.1088/2053-1591/acbae8

[15] Kumar, S., Goud, M. & Suri, N. M. (2021). An Investigation of Magnetic-field-assisted EDM by Silicon and Boron Based Dielectric of Inconel 706. *Silicon, 13*, 4747-4755. https://doi.org/10.1007/s12633-020-00776-9

[16] Khatun, S., Amin, A. K. M. N., Mashuk, M. R., Tuli, N. T., Jerin, I., Bashar, M. S. & Rumi Md, J. U. (2024). Parameter optimization and surface integrity aspects in MWCNT-based nano-PMEDM process of Inconel 718. *Res. Eng. Struct. Mater., 10*(1), 271-304. https://doi.org/10.17515/resm2023.18ma0530rs

[17] Amorim, F. L., Dalcin, V. A., Soares, P. & Mendes, L. A. (2017). Surface modification of tool steel by electrical discharge machining with molybdenum powder mixed in dielectric fluid. *Int. J. Adv. Manuf. Technol.*, *91*(1-4), 341-350. https://doi.org/10.1007/s00170-016-9678-x

[18] Le, V. T. (2021). The role of electrical parameters in adding powder influences the surface properties of SKD61 steel in EDM process. *J. Brazilian Soc. Mech. Sci. Eng., 43*(3), 1-19. https://doi.org/10.1007/s40430-021-02844-6

[19] Keskin, G., Salunkhe, S., Küçüktürk, G., Pul, M., Gürün, H. & Baydaroğlu, V. (2023). Optimization of PMEDM process parameters for B4C and B4C+SiC reinforced AA7075 composites. *J. Eng. Res.* https://doi.org/10.1016/j.jer.2023.09.012

[20] Rodic, D., Gostimirovic, M., Sekulic, M., Savkovic, B. & Aleksic, A. (2023). Study and Optimization Defect Layer in Powder Mixed Electrical Discharge Machining of Titanium Alloy. *Processes, 11*(4). https://doi.org/10.3390/pr11041289

[21] Muthuramalingam, T. & Phan, N. H. (2021). Experimental Investigation of White Layer Formation on Machining Silicon Steel in PMEDM Process. *Silicon, 13*, 2257-2263. https://doi.org/10.1007/s12633-020-00740-7

[22] Talla, G., Gangopadhyay, S. & Biswas, C. K. (2023). Influence of powder-mixed EDM on surface morphology and metallurgical alterations of Inconel 625. *Aust. J. Mech. Eng., 21*(5), 1533-1546. https://doi.org/10.1080/14484846.2021.2022580

[23] Paswan, K. et al. (2023). An analysis of microstructural morphology, surface topography, surface integrity, recast layer, and machining performance of graphene nanosheets on Inconel 718 superalloy: Investigating the impact on EDM characteristics, surface characterizations, and opt. *J. Mater. Res. Technol., 27*, 7138-7158. https://doi.org/10.1016/j.jmrt.2023.11.080

[24] Ming, W. et al. (2022). Research on EDM Performance of Renewable Dielectrics under Different Electrodes for Machining SKD11. *Crystals, 12*(2). https://doi.org/10.3390/cryst12020291

[25] Khadar, B. S. Jagannadha, R. M. V. & Murahari, K. (2021). Multi-Objective Optimization of Process Parameters for Powder Mixed Electrical Discharge Machining of Inconel X-750 Alloy Using Taguchi-Topsis Approach. *Strojnícky časopis - Journal of Mechanical Engineering, 71*(1), 1-18. https://doi.org/10.2478/scjme-2021-0001

[26] Gerçekcioğlu, E. & Albaşkara, M. (2023). Multi-Response Optimization of Electrical Discharge Machining of 17-4 Ph Ss Using Taguchi-Based Grey Relational Analysis. *Arch. Metall. Mater., 68*(3), 861-868. https://doi.org/10.24425/amm.2023.145448

[27] Lajis, M. A. & Ahmad, S. (2015). Machinability performance of powder mixed dielectric in electrical discharge machining (EDM) of inconel 718 with copper electrode. *Int. J. Mech. Mechatronics Eng., 15*(4), 37-52.

[28] Reddy, V. V., Kumar, A., Valli, P. M. et al. (2015). Influence of surfactant and graphite powder concentration on electrical discharge machining of PH17-4 stainless steel. *J Braz. Soc. Mech. Sci. Eng., 37*, 641-655. https://doi.org/10.1007/s40430-014-0193-4

[29] Srivastava, S., Vishnoi, M., Gangadhar, M. T. & Kukshal, V. (2023). An insight on Powder Mixed Electric Discharge

Machining: A state of the art review. *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf., 237*(5), 657-690.
https://doi.org/10.1177/09544054221111896

Authors' contacts:

Dunya Adnan Ghulam
(Corresponding author)
Production Engineering and Metallurgy Dept.,
University of Technology,
Alsina'a Street, 10066 Baghdad, Iraq
dunia.a.ghulam@gmail.com

Abbas Fadhil Ibrahim, Professor
Production Engineering and Metallurgy Dept.,
University of Technology,
Alsina'a Street, 10066 Baghdad, Iraq
abbas.f.ibrahim@uotechnology.edu.iq

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240525152340

Preliminary communication
Received: 2024-05-25, Accepted: 2024-05-27

Innovative Culture: A Predictor of Digital Transformation in Technical Manufacturing Companies

Marin Grčić, Ljerka Luić*

Abstract: The innovative culture and the implementation of new digital technologies in the overall operation of enterprises are predictors of their digital transformation process. This paper, based on a study conducted on a convenient sample of 42 technical manufacturing companies in Croatia, examines their awareness and determines their readiness for digital transformation, with particular emphasis on the manufacturing process itself. Accepting the results of related international studies, according to which the Republic of Croatia is placed in the lower part of the readiness scale for business operations in the digital age, the hypothesis was set that Croatian enterprises are insufficiently familiar with the benefits of digital transformation achieved through the development of a digital innovative culture. Based on the analysis of the results, the reasons for the effective semi/unpreparedness of technical manufacturing companies to enter the process of digital transformation are determined, inductive conclusions are drawn, and suggestions for ways to encourage innovative culture are given.

Keywords: awareness; digital transformation; innovative culture; technical manufacturing company

1 INTRODUCTION

The exceptionally rapid progress and development of digital technologies require employees to adapt to changes quickly, learn new skills, and work in teams, as jobs become more specialized, making it challenging for one person to perform and be familiar with all work segments (or at least most of them). Furthermore, the magnitude of changes we witness daily has led to various work concepts and hours, the disappearance of certain jobs and the emergence of new ones. Computer systems have become the backbone of our lives, both private and business. It is no surprise, therefore, that companies wishing to survive and advance on the market invest significant human, material and financial resources in developing their operations and IT solutions. This trend has already been well established and is constantly increasing in the developed countries. In Croatia, however, signs of movement in this direction are observed, especially among some more technologically aware, younger and innovative business environments, and yet there is an impression that an insufficient number of companies understand the importance of changing business and production methods. It is possible that the reasons for this also lie in the socio-economic context in which companies operate, at both local and national levels, and the administrative burden that hinders or suffocates the possibility of change. The research described in this paper seeks to argumentatively determine the attitudes of Croatian technical manufacturing companies, specifically those engaged in production within the field of mechanical engineering (domestic or foreign-owned), regarding digital transformation, i.e., the need to reshape the existing business-production models. [1]

The paper analyzes the awareness of Croatian technical manufacturing companies regarding the application of digital technologies in business, i.e., their readiness for digital transformation, with special emphasis on the digital transformation of the manufacturing process. In this context, it is tentatively possible to estimate how the results presented in this paper will contribute, in an applicative sense, to a better understanding of the Croatian economy in the domain of the manufacturing industry regarding readiness to face new challenges brought by the modern era marked by the rapid development of high technologies.

2 DIGITAL TRANSFORMATION, INNOVATIVE CULTURE

At the beginning of the paper, it is important to define the concepts of digital transformation and innovative culture, which are often used as synonyms in the context of introducing digital technologies and solutions into modern business in business practice, as well as in professional and scientific publications. For this reason, theoretically grounded explanations are provided below, while their practical application is explained in the introductory part of the survey questionnaire.

2.1 Digital Transformation

According to Grčić [1], digitization is basically the conversion of analog into digital information, which can be stored, processed, and transmitted further using computers, without substantial changes in the business process itself. It is, therefore, a simple conversion of data into a digital format, a conversion that does not imply the digital transformation of business processes. This aspect is defined by the term digitalization, which the same author describes as a process that means "...enabling, improving, and/or transforming business activities, roles, models, activities using digital technologies. It mainly refers to management systems and their implementation through digital data and processes, i.e., the use of digital technologies to change business models and create new revenues and values." [1]

The term digital transformation denotes the process within which business organizations, in the context of this paper, technical manufacturing companies, completely change their business processes and the way of doing business. Thus, according to Von Leipzig et al. [2], digital business transformation is currently one of the most frequently used terms in business publications, on business portals, and in business meetings. According to Tekic and Koroteev [3], there is probably no successful manager who does not know what digital transformation means for the operation of their organization. According to Spremić [4], digital transformation refers to the intensive application of digital technology and resources to convert these resources into new revenues, business models, and ways of doing business. Transformation occurs when a company decides to fundamentally change its business processes, strategies, activities, hierarchical and organizational structure in a relatively short period of time, all in order to better connect these processes and strategies and ultimately provide a better competitive advantage for the company on the market. Although Schallmo et al. [5] present arguments from which it is evident that the term digital transformation is still not clearly defined, for the purposes of this paper, the definition of the authors Verhoef et al. has been accepted, according to which "Digital transformation is defined as a change in the way a company uses digital technologies to develop a new digital business model that helps create and provide greater value for the company" [6]. Furthermore, the term digital transformation can be described as "a thorough change in the organization and the traditional way of doing business using digital technologies and applying new business models with the aim of improving the performance of the organization and faster adaptation in an environment that is constantly and rapidly changing" [7]. According to the European Commission, "digital transformation today is no longer a matter of choice - it is inevitable, necessary, and unavoidable" [8], and implies a complete transformation of business, but also attitudes, i.e. business culture, and the need for continuous learning and improvement of both employees and persons in leading positions. It also implies changes in relationships with (potential) customers, advertising methods, and the production of products themselves. In essence, "it encompasses much more than just technologies and relates to radical changes in the way of thinking and perceiving the business of the organization" [9].

Many business entities, their directors, and managers are still not fully or sufficiently familiar with the comprehensiveness of the term digital transformation and "equate it with the digitalization of existing business models and/or analog processes and solutions of information and communication technology" and are often not aware of the necessity of the active role of "the management structure of the company for the implementation of the transformation at all levels" [10]. If companies understand the strategic role of new digital technologies "they can create new values, improve user experiences, simplify operations or create completely new business models and sources of income. But more than the technology itself, the success of digital transformation is based on people" [11]. Namely, "educating employees on digital topics is measure no. 1 when it comes to planning digital transformation" [9]. In addition to this, many other measures are available to companies in terms of improving their digital competitiveness, such as: "prioritizing digital transformation"; "creating a digital transformation strategy, i.e. a vision of the desired future"; "engaging experts"; "creating a digital culture"; "implementing an appropriate digital transformation management structure"; "quality human resource management"; "developing collaborations and partnerships"; "greater agility and better preparedness for the unexpected" [9].

2.2 Digital Transformation of Production

Today, almost all initiatives aimed at digital transformation are built on technologies and solutions of the so-called Third Platform [12], which includes mobile technology, cloud data storage, big data set analysis, and the use of social media. More recently, the term Industry 4.0 is increasingly used, which implies the introduction of digital technologies into (almost) all production processes and is based on "automated technology networked through sensors and communication elements [...] thus connecting the real and virtual world in the form of the so-called cyber-physical systems, such as autonomous robots" [13]. In such industrial systems, machines are not viewed as independent units, but are connected "into a kind of community that is in mutual interaction and cooperation, autonomously and "intelligently" [13]. They imply the use of technologies "that lead to the automation of certain processes in production and/or provision of services" [13]. The most important among these technologies are the following: artificial intelligence, "big data", robots, simulations, horizontal and vertical integrations of systems, the Internet of Things, cybersecurity, cloud technology, three-dimensional printing, augmented reality [13]. The use of digital technologies and the networking of different sectors enable technical manufacturing companies to obtain information about the production process more efficiently and optimize it due to the availability of external and internal data, which ultimately gives them a competitive advantage on the market. Furthermore, they enable the development of individualized production based on customer requirements and better connectivity with the latter.

"Industry 4.0 has been in place for the last 10 years, benefiting the industry but also highlighting its shortcomings. Now, the time for Industry 5.0 has arrived. Smart factories have increased business productivity; however, Industry 4.0 has its limitations. [...] Industry 5.0 is changing the paradigm and offers a solution by reducing the emphasis on technology and focusing on collaboration between humans and machines. [...] In modern business, with advanced technological developments, Industry 5.0 is necessary for gaining competitive advantages as well as fostering economic growth for factories." [14]

2.3 Digital Transformation Strategies

Digital transformation strategies in each company take on a different perspective and are aimed at achieving different goals. In any case, it is necessary to "predict the costs of the process as concretely as possible and set measurable performance indicators for implementation in the short and long term" [9]. If the company's digital strategy is designed in a quality way, based on detailed analyses of all business elements and indicators of positive and negative characteristics of the business and taking into account existing trends on global markets, it will ensure "a transparent and organized system within a successful [sic] company or organization. Such a system includes the management of technologies, a structured process of collecting ideas, selecting adequate technologies, preparing a concept, commercialization, and monitoring performance on the market or within the company" [9]. In order for the digital strategy to ultimately be successful, "a clear, concrete, and measurable implementation plan is needed that will include a series of measures, steps, and activities with success indicators and a time frame, but also tools that will be applied" [9]. Certainly, digital transformation is a journey, not a destination. Even when a company reaches a level of digital maturity, continuous investment, innovation and quick response to market changes are required. Continuous investment allows new generations to be agile from the very beginning [15]. This means that they know how to adapt quickly and flexibly to changes, unforeseen elements and situations, and the needs of their customers and users.

2.4 Preconditions for Digital Transformation

According to Gurbaxani and Dunkle [16] and Warner and Wager [17], successful digital transformation of business requires a change in organizational thinking and corporate culture, which is considered one of the key dimensions of the digital transformation process. Furthermore, Meske and Junglas [18] note that although some managers believe that transformation always results in the cannibalization of existing products and services, this is not the case. Digital transformation provides companies and their employees with a long-term perspective of sustainable business and is necessary for achieving and maintaining the competitiveness of companies. Organizing and implementing good business transformation is not the responsibility of employees, but it provides them with the opportunity to better focus on business priorities, be ready to respond to environmental changes (agility), and better long-term career development opportunities, knowledge, and skills. According to Haffke et al. [19], for successful digital transformation, it is important to have someone for whom it is the primary job. For this reason, more and more companies are adding functions of digital transformation directors or board members for managing digital transformation to the management structure, who should be at the highest level of management, and who will take responsibility for implementing digital transformation. Thus, Burilović [10] notes that when planning a digital transformation strategy, it is important to know that in most cases it will not bring quick and short-term results, and it is important to define the long-term interests of the company. It is also very important to predict the costs from the very beginning. Of course, it is impossible to fully predict the costs associated with changing the business culture or organizational structure, but with good planning, unnecessary resistance in project implementation can be avoided. Furthermore, at the beginning of implementation, it is necessary to set measurable performance indicators. Only in this way can the results of digital transformation be evaluated after implementation. It is important to accept the fact that investing in digital transformation is not a cost, but an investment that will specifically pay off in the future.

From the analysis of recent sources, it is evident that most authors who have researched the success of digital transformation implementation in companies highlight the following key elements: change in the organizational way of thinking and corporate culture, personal commitment and responsibility of the most influential stakeholder in the organization, commitment of top management to digital transformation, defining and updating the strategy, developing a business plan that should contain components of digital transformation, the function of the director of digital transformation who will take responsibility for implementing digital transformation, defining the long-term interests of the company, predicting costs from the very beginning, setting measurable performance indicators, and educating employees.

2.5 Advantages and Barriers to Digital Transformation

The advantages of digital transformation can be seen through the following elements: the possibility of storing and quickly accessing a large amount of data and its more efficient linking into sets based on certain characteristics or parameters, for the purpose of planning the production and business, as well as developing short-term and long-term strategies; connectivity of production and administration and all other sectors and departments; networking of headquarters, all production plants, departments, and branches, which enables better communication among employees and synchronization of daily activities; higher quality and more direct communication with clients and the possibility of creating personalized product and service offers; networking of all processes that enables the creation of a more efficient monitoring and control system, as well as timely elimination of defects, breakdowns, problems in stock supplies, etc.; acceleration of production; reduction of material costs; better management of working hours; higher quality of making products and service offers. [1]

The main obstacle to digital transformation is the insufficient number of experts in companies who would be able to technically design, organize and implement digital transformation in a quality way, which significantly hinders, slows down or completely postpones the process itself. For this reason, companies must employ a new person or more of them or, alternatively, hire external consultants, which requires additional costs, which are already high due to technical needs, creating a new or more suitable infrastructure, purchasing new IT tools, training employees. Furthermore, due to the high concentration and huge amount of information, and their constant "movement" in the digital world, it is necessary to further strengthen security systems and ensure the impregnability of virtual storage spaces. In Croatia, according to research [7, 9], the biggest obstacles to the implementation of digital transformation are: "financial reasons (costs)"; "too much focus on existing business priorities" leading to the postponement of "all new strategic directions"; "lack of time for digital transformation implementation", i.e. lack of time primarily for training and adequate preparation of employees for changes that inevitably come with digital transformation, changes that affect the organizational and innovative culture of companies.

2.6 Innovative Culture

Knowledge and skills necessary for effective job performance in the modern environment are changing and developing at an unprecedented speed. Accepting and encouraging innovation, with special emphasis on curiosity, agility, and experimentation [20], is key to developing competencies necessary for functioning in such an environment. Innovation also requires not only new ideas, but also employees and management who possess the knowledge and skills to recognize and implement innovation [21]. Given that innovativeness is recognized as a "driver of competitive advantage, companies go beyond their own borders to find and generate new knowledge" [22]. In this regard, creativity contributes to innovation through creative breakthroughs and strategic connections of seemingly unrelated elements [23]. One of the drivers and sources of innovativeness is certainly digitalization, considering that digital technologies enable greater collaboration between science and innovation, internationalization, and openness to the public [24].

As Tohidi and Jabbari pointed out, "innovation is a process that begins with the introduction of an idea that will become a new function and this makes it distinguished from creation" [25], while digital and information and communication technologies provide new technical and business opportunities that significantly change the structure of traditional technical and business systems [26] creating preconditions and the necessary infrastructure for managing large amounts of data, which are converted into specific knowledge in the process of digital transformation, into innovative ideas which are the first step in creating and establishing an innovative culture of companies.

3 MATERIALS AND METHODES

Awareness of the importance of implementing a comprehensive digital transformation of business, with an emphasis on the production process, and the readiness of companies for the same, was investigated using the survey method. For this purpose, a survey questionnaire was created that contained 35 research questions for whose conceptual framework the source was used [27], 5 general questions about the company, and 4 general questions about the respondent (1. Your current position in the company is; 2. Your level of education is; 3. The email address where you would like to receive the research results; 4. Your comments and suggestions related to the research topic). The questionnaire in digital form was sent to publicly available e-mail addresses of 200 technical manufacturing companies based in the Republic of Croatia. The following sources were used to determine the research sample: the business search engine of the website www.tvrtke.hr, court and crafts registers, and data from the Croatian Chamber of Economy. Furthermore, publicly available materials on similar research were studied with the aim of more precisely formulating the research hypothesis, which is as follows,

H1: The majority of technical manufacturing companies in Croatia are not sufficiently aware of the impact of digital transformation of production on creating an innovative culture and the overall business success of companies.

42 companies (21% of the total number of 200 technical manufacturing companies in Croatia) responded to the questionnaire. The results of the research are presented below, grouped according to subtopics. Namely, after the introductory questions, companies were asked to select the stage in which they are in terms of implementing digital transformation (1. They do not plan to implement it, 2. They plan to implement it, 3. They are in the implementation phase, 4. They have implemented digital transformation). Depending on the selected stage, the questionnaire was directed to questions specific to the selected (chosen) stage of digital transformation implementation.

To summarize the information contained in the collected data and to determine the main facts about the researched issue, to analyse the obtained results, their graphical and tabular presentation, to determine the distribution of results, and to compare and position individual results within the group, descriptive statistics were used. The methods, both numerical and graphical, are presented in the following chapter.

4 RESULTS

The results of the research, processed using the method of descriptive statistics, are presented below.

4.1 Company Data

Companies that completed the questionnaire are mostly engaged exclusively in production or, alternatively, in production and sales, as shown in Fig. 1, and the majority of them have between 10 and 50 employees, as evident from Fig. 2.

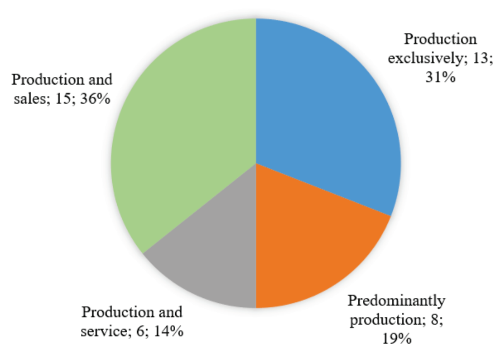


Figure 1 Company activity [1]

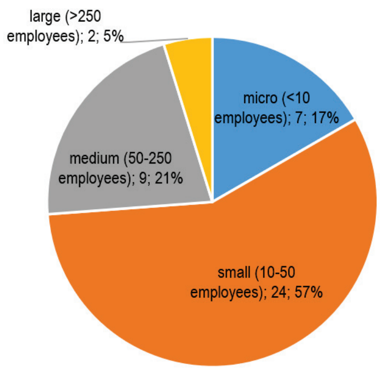


Figure 2 Company size [1]

Furthermore, in terms of ownership type, a large majority of responses came from private companies with domestic ownership, as evident from Fig. 3, all of which fall into the category of small companies. Only two companies have more than 250 employees: one of them is from the public sector, and the other is private with foreign ownership. They operate exclusively (the first) or predominantly (the second) on the foreign market.

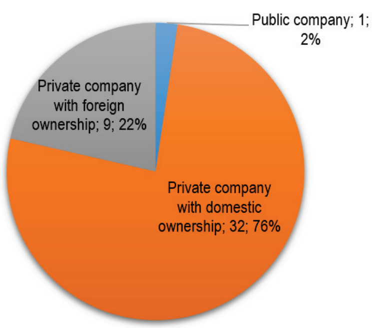


Figure 3 Type of ownership [1]

Companies participating in the research generally operate mostly on the domestic market, but about a quarter of them mainly do business with foreign countries, and 17% operate exclusively on the foreign market, as evident from Fig. 4. No company operates exclusively on the Croatian market. In line with expectations, all companies that operate exclusively (3 responses) or predominantly (6 responses) on the foreign market are in foreign ownership. One company (in domestic private ownership, engaged exclusively in production) responded that it operates both on the domestic and foreign markets.

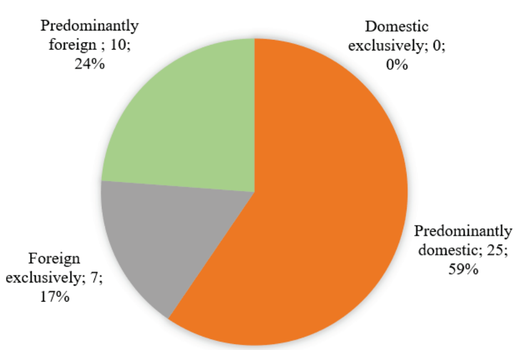
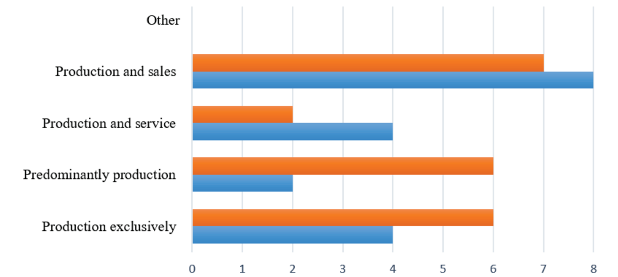


Figure 4 Market of operation [1]

All companies which stated that they mainly operate on the domestic market are private companies with domestic ownership, while those that operate only in the foreign market are exclusively engaged in production.

4.2 Analysis of the Status of Digital Transformation Implementation

The comparison of company activities with the stage in which the companies are in terms of digital transformation, considering those companies that plan to implement it or are in the implementation phase, is shown in Fig. 5.

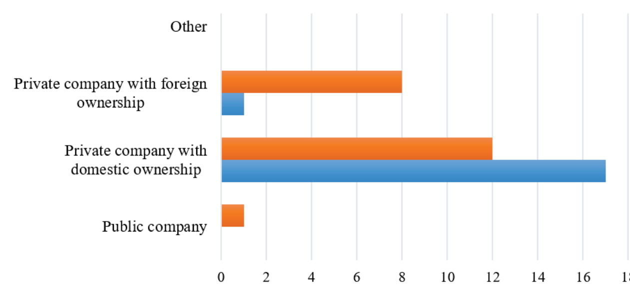


	Production exclusively	Predominantly production	Production and service	Production and sales	Other
Implementing digital transformation stage	6	6	2	7	0
Planning to implement digital transformation	4	2	4	8	0

Figure 5 Comparison of activities with the implementation stage of digital transformation [1]

No particular differences are observed among the types of activities of companies which are in the implementation stage of digital transformation, while among those that only plan to implement it, prevail companies which are engaged in production and sales.

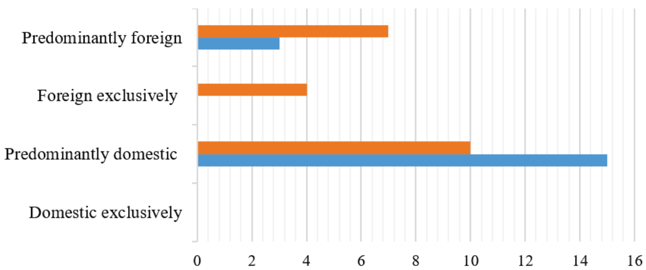
If, on the other hand, we compare the ownership type with the stage of digital transformation in which the companies are, we observe that the majority of companies that implement it or plan to implement it are privately owned with domestic ownership, as evident from Fig. 6.



	Public company	Private company with domestic ownership	Private company with foreign ownership	Other
Implementing digital transformation stage	1	12	8	0
Planning to implement digital transformation	0	17	1	0

Figure 6 Comparison of ownership type and digital transformation implementation stage [1]

This data can be interpreted as an increasing awareness of private entrepreneurs in Croatia regarding innovations recommended by the modern era, and in some segments, even required both on the production and general business level. However, a greater willingness to change is noticed in companies with foreign ownership, which may also depend on the "traditional" business culture of Croatia, which is quite oriented towards retaining the existing models and expressing less willingness to take risks and step out of the "comfort zone".

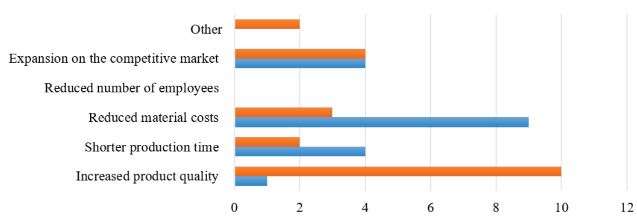


	Domestic exclusively	Predominantly domestic	Foreign exclusively	Predominantly foreign
Digital transformation implementation stage	0	10	4	7
Planning to implement digital transformation	0	15	0	3

Figure 7 Comparison of market type and digital transformation implementation stage [1]

Compared with the type of market in which they operate, shown in Fig. 7, it is possible to observe that the majority of companies operate mostly on the domestic market, which particularly applies to companies which only plan to implement digital transformation.

The comparison of expectations from the implementation of digital transformation depending on the stage of digital transformation implementation is shown in Fig. 8.



	Increased product quality	Shorter production time	Reduced material costs	Reduced number of employees	Expansion on the competitive market	Other
Digital transformation implementation stage	10	2	3	0	4	2
Planning to implement digital transformation	1	4	9	0	4	0

Figure 8 Comparison of expectations from implementation and the implementation stage of digital [1]

It is evident in Fig. 8 that companies that have initiated the digital transformation process expect an increase in product quality after its implementation. Companies that plan digital transformation primarily expect a reduction in material costs.

4.3 Priorities and Effects of Digital Transformation of Production

From Fig. 9, it is evident that as a priority of digital transformation of production, respondents primarily place product manufacturing (62% of them), followed by technology (24% of them), while product design (7% of them) and material procurement (7% of them) are at the end of priorities (according to 14% of respondents).

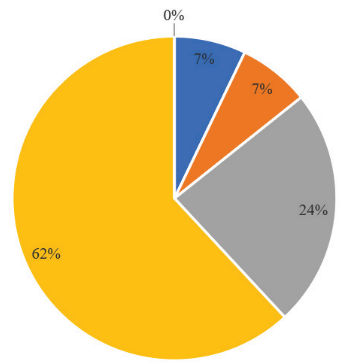


Figure 9 Priorities in the process of digital transformation of production [1]

As a result of the implementation of the digital transformation process of production, the majority of respondents expect primarily higher product quality, expansion in the competitive market and lower material costs, as evident from Fig. 10.

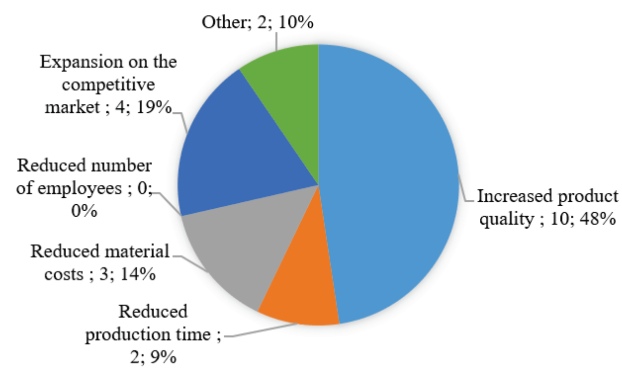


Figure 10 Expected results of digital transformation of the production process [1]

In the category "Other", the following responses were obtained: "it is difficult to single out one benefit", "better insight and monitoring of efficiency and profitability". It can be interpreted as positive data that no company stated that it expects a reduction in the number of employees, which is often mentioned as a potential disadvantage in the implementation of digital transformation of business.

5 DISCUSSION

The subject research, conducted on a sample of 21% of technical manufacturing companies in Croatia, yielded results whose analysis makes it possible to determine the degree of awareness of manufacturing companies from the technical sector. Based on the analysis of the results, it is evident that technical manufacturing companies, especially those with foreign ownership, are significantly aware of the role of digital transformation and the radical transformation of business it brings about. However, still an insufficient number of them (21 out of 42 companies, 50% of the companies that participated in the research) dared to take this step and fully implement digital transformation. In terms of the production process itself, the majority of companies believe that product manufacturing should be the priority in digital transformation, followed by technology development. These are areas that require the greatest additional investments, in term of both material and time. These responses are consistent with expectations because this sector is most susceptible to human errors, so it is important for the company to create a system in which product defects and rejects are minimized, and the consumption of materials and energy is optimized. Since technology is still prone to errors and malfunctions, it is important that people who will manage and/or supervise machines, robots, and software programs used in the production itself are adequately and well-trained and ready to recognize potential problems and react to them in a timely manner. The entire transformation process should be overseen and led by the company's management structures to ensure that digital transformation is carried out in a holistic manner, taking into account all potential business goals and executive functions. "Croatia should accelerate its efforts in the area of digitalisation of businesses. In particular, it should raise awareness about the benefits of business digitalisation, provide public support for workshops and trainings, increase participation in existing (funding) schemes, especially among SMEs. Croatia should intensify its efforts to support the development and deployment of trustworthy, secure, sovereign advanced technologies and solutions, especially for AI, cloud, big data, including through the availability of legal and technical support and procurement procedures." [27]

6 CONCLUSION

Based on all of the above, it can be concluded that technical manufacturing companies in Croatia are predominantly aware of the importance of investing in digital technologies and the need for implementing digital transformation of the entire business, especially the production process. Namely, a larger part of the surveyed companies is in the stage of implementing or planning the implementation of the same. Therefore, the obtained results only partially support the initial research hypothesis, which claimed that the majority of technical manufacturing companies in Croatia are not sufficiently aware of the impact of digital transformation of production on creating an innovative culture and the overall business success of companies. It is possible that this is the case because technical manufacturing companies must "fight" more with strong global competition that regularly follows and adopts trends in business and production and continuously improves its work performance. Since these companies are aware of the need for digital transformation, they are on the right path to achieving it. The time period in which they will succeed and the results they will achieve largely depend on the availability of experts necessary for the implementation, financial resources, as well as national strategies, infrastructures, potential administrative limitations, and willingness to invest in entrepreneurship.

Furthermore, the research showed that technical manufacturing companies in Croatia are in principle ready for digital transformation, but it can be assumed that they lack external support and a thorough development of the strategy, and possibly material resources, for relatively rapid implementation. Therefore, it would be useful, through further research, to determine ways to effectively implement digital transformation of production. Furthermore, in order to increase the number of quality workforce and experts in the field of digital technologies applied to production and the mechanical engineering sector in general (without which it is not possible to fully successfully implement digital transformation of production), it would be necessary to modernize the education system at all levels, especially higher education, and bring it closer to the real sector and the labour market, i.e., systematically work on developing practical digital competencies of students and their active involvement in the processes of innovative culture.

Acknowledgements

Authors express their gratitude to University North in supporting the publication of this article through scientific research project "d-Learning: Digital Competences for Innovation Culture" funds.

7 REFERENCES

[1] Grčić, M. (2021). Digital transformation of production-technical companies business. *Master's thesis*. Karlovac: Karlovac University of Applied Sciences. https://urn.nsk.hr/urn:nbn:hr:128:537009

[2] von Leipzig, T., Gamp, M., Manz, D., Schöttle K., Ohlhausen, P., Oosthuizen, G., Palm, D. & von Leipzig, K. (2017). Initialising Customer-orientated Digital Transformation in Enterprises. *Procedia Manufacturing, 8*, 517-524. https://doi.org/10.1016/j.promfg.2017.02.066

[3] Tekic, Z. & Koroteev, D. (2019). From disruptively digital to proudly analog: A holistic typology of digital transformation strategies. *Business Horizons, 62*(6), 683-693. https://doi.org/10.1016/j.bushor.2019.07.002

[4] Spremić, M. (2017). *Digitalna transformacija poslovanja*. Zagreb: Faculty of Economics & Business, University of Zagreb. (in Croatian)

[5] Schallmo, D., Williams, C. A. & Boardman, L. (2017). Digital transformation of business models — best practice, enablers, and roadmap. *International Journal of Innovation Management, 21*(8), 1-17. https://doi.org/10.1142/S136391961740014X

[6] Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Qi Dong, J., Fabian, N. & Haenlein, M. (2021). Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research, 122*, 889-901. https://doi.org/10.1016/j.jbusres.2019.09.022

[7] Apsolon. (2019). *Digitalna transformacija u Hrvatskoj 2019.* https://apsolon.com/publikacije/digitalna-transformacija-u-hrvatskoj-2019/ (Accessed on March 20, 2024) (in Croatian)

[8] European Commission Representation in Croatia. (2019). *What is the digital transformation and what changes do we expect?* https://croatia.representation.ec.europa.eu/news/sto-je-zapravo-digitalna-transformacija-i-kakve-nas-promjene-

ocekuju-2019-07-26_hr?prefLang=en&etrans=en (Accessed on March 24, 2024)
[9] Apsolon. (2020). *Digitalna transformacija u Hrvatskoj 2020.* https://apsolon.com/publikacije/digitalna-transformacija-u-hrvatskoj-2020/ (Accessed on March 20, 2024) (in Croatian)
[10] Burilović, L. (2020.). Digital transformation of retail business. *Poslovna izvrsnost, 14 (2)*, 197-221. https://doi.org/10.22598/pi-be/2020.14.2.197
[11] Mehulić, D. (2020). *Pandemija ubrzava tempo digitalne transformacije, nitko nije imun*. https://www.iztzg.hr/UserFiles/file/novosti/2020/COVID-19%20radovi/Mehuli%C4%87-D_2020.pdf (Accessed on March 20, 2024) (in Croatian)
[12] Auriga. (2016). *Digital Transformation: History, Present, and Future Trends*. https://auriga.com/blog/2016/digital-transformation-history-present-and-future-trends/ (Accessed on March 14, 2024)
[13] Hrbić, R. & Grebenar, T. (2021). *Procjena spremnosti hrvatskih poduzeća na uvođenje tehnologija I4.0*. Zagreb: Croatian National Bank. https://www.hnb.hr/documents/20182/3776564/i-062.pdf/5720fb3d-20b6-e592-bb23-20cd61c19680 (in Croatian)
[14] Adel, M. (2022). Future of industry 5.0 in society: human-centric solutions, challenges and prospective research areas. *Journal of Cloud Computing, 11*, 40. https://doi.org/10.1186/s13677-022-00314-5
[15] Evans, N. D. (2017). Assessing your organization's digital transformation maturity. *CIO*. https://www.cio.com/article/230462/assessing-your-organization-s-digital-transformation-maturity-2.html
[16] Gurbaxani, V. & Dunkle, D. (2019). Gearing Up For Successful Digital Transformation. *MIS Quarterly Executive, 18*(3), 209-220. https://doi.org/10.17705/2msqe.00017
[17] Warner, K. S. R. & Wager, M. (2019). Building dynamic capabilities for digital transformation: An ongoing process of strategic renewal. *Long Range Planning, 52*(3), 326-349. https://doi.org/10.1016/j.lrp.2018.12.001
[18] Meske, C. & Junglas, I., (2020). Investigating the elicitation of employees' support towards digital workplace transformation. *Behaviour & Information Technology, 40*(11), 1120-1136. https://doi.org/10.1080/0144929X.2020.1742382
[19] Haffke, I., Kalgovas, B. & Benlian, A. (2016). The Role of the CIO and the CDO in an Organization's Digital Transformation. Ågerfalk, P. J., Levina, N. & Kien, S. S. (eds.), *Proceedings of the International Conference on Information Systems - Digital Innovation at the Crossroads (ICIS 2016)*, *1*(7), 1-20. Dublin: Ireland. https://core.ac.uk/download/pdf/301370191.pdf
[20] Baričević, M. & Luić, L. (2023). From Active Learning to Innovative Thinking: The Influence of Learning the Design Thinking Process among Students. *Education Sciences, 13*(5), 455, 1-19. https://doi.org/10.3390/educsci13050455
[21] Luić, L. & Glumac, D. (2009). The role of ICT technology in the knowledge society. Milovanović, B. D. (ed.) *Proceedings of the 9th International Conference on Telecommunication in Modern Satellite, Cable, and Broadcasting Services (TELSKS 2009), 1,* 310-313. Niš, Serbia. https://doi.org/10.1109/TELSKS.2009.5339515
[22] Silva, T. F. P. & Marques, J. P. C. (2020). Human-Centered Design for Collaborative Innovation in Knowledge-based Economies. *Technology Innovation Management Review, 10*(9), 5-15. https://doi.org/10.22215/timreview/1385
[23] Taylor, S. P. (2017). What Is Innovation? A Study of the Definitions, Academic Models and Applicability of Innovation to an Example of Social Housing in England. *Open Journal of Social Sciences, 5*(11), 128-146. https://doi.org/10.4236/jss.2017.511010
[24] Babić, N., Luić, L. & Rončević, A. (2022). Impact assessment of digital skills on D&R investment. Luic, L., Martincevic, I. & Sesar, V. (eds.) *Proceedings of the 83rd International Scientific Conference on Economic and Social Development - "Green Marketing"*, 86-96. Varaždin: University North. https://www.esd-conference.com/upload/book_of_proceedings/Book_of_Proceedings_esdVarazdin2022_Online.pdf
[25] Tohidi, H. & Jabbari, M. M. (2012). The important of Innovation and its Crucial Role in Growth, Survival and Success of Organizations. *Procedia Technology, 1*, 535-538. https://doi.org/10.1016/j.protcy.2012.02.116
[26] Savić, G. & Luić, L. (2016). Business intelligence in managing of technical-information system. Sawik, T. (ed.) *Proceedings of the 13th International Conference on Industrial Logistics*, 231-238. Krakow: AGH University of Science and Technology. http://www.icil.zarz.agh.edu.pl/images/papers/Savic_Luic.pdf
[27] European Commission. (2023). *Country reports - Digital Decade report 2023. Digital Decade Country Report 2023. Croatia.* https://digital-strategy.ec.europa.eu/en/library/country-reports-digital-decade-report-2023 (Accessed on March 24, 2024)

Authors' contacts:

Marin Grčić, mag. ing. mech.
Karlovac University of Applied Sciences,
Trg J. J. Strossmayera 9, 47000 Karlovac, Croatia
E-mail: grcicmarin92@gmail.com

Ljerka Luić, Full Professor
(Corresponding author)
University North,
Jurja Križanića 31b, 42000 Varaždin, Croatia
E-mail: ljluic@unin.hr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240603194203

Preliminary communication
Received: 2024-06-03, Accepted: 2024-08-01

Comparative Insights on Vehicles' Deceleration Measurements

Jasmin Šehović*, Mirsad Trobradović

Abstract: This paper investigates the possibilities of measuring vehicle deceleration using modern devices based on GPS (Global Positioning System) and MEMS (Micro Electro Mechanical System) technologies, with the aim of replacing traditional inertial devices. The experiments were carried out on a passenger vehicle and a construction machine, using three different measuring devices: an inertial decelerometer, a GPS device (Racelogic DriftBox) and a MEMS accelerometer. The measurement results show good mutual agreement between recorded results from devices used for this paper, with an emphasis on the precision and ease of use of modern devices. The conclusion emphasizes the justification of using GPS and MEMS devices in testing the dynamic characteristics of vehicles due to their high frequency of data acquisition, simple processing, absence of the need for calibration, and greater practicality. This research suggests that such devices will be increasingly used in the future to analyse vehicle dynamics.

Keywords: construction machine; deceleration; GPS; MEMS; vehicle

1 INTRODUCTION
1.1 Background

Deceleration is an important dynamic characteristic of a vehicle that provides information about its ability to change speed in a certain time interval. It is often used as a measure of vehicle performance and safety when stopping the vehicle.

When braking is considered, deceleration is the rate at which a vehicle can reduce its speed after applying the brakes. Effective deceleration is key to safe driving, as it allows the driver to control the vehicle and adjust speed to changing road conditions. Investigation of various dynamic characteristics of the vehicle, among which is the deceleration, was most often performed with devices that are demanding and complex to handle. In recent times, devices based on GPS (Global Positioning System) and MEMS (Micro Electro Mechanical System) technologies have been increasingly introduced to test the dynamic characteristics of vehicles.

1.2 Objectives

The aim of this paper is to investigate the possibility of measuring deceleration using modern devices based on GPS and MEMS technology, which should replace inertial devices. To justify the use of these devices on all types of vehicles, the measurement was carried out on a passenger vehicle and a construction machine to compare the matching results between individual devices. The obtained results of this research show the justification of using devices based on GPS and MEMS technologies for measuring the dynamic characteristics of vehicles, both passenger and heavy construction machines.

The use of devices based on GPS and MEMS (Micro-Electro-Mechanical Systems) technologies for measuring the dynamic characteristics of vehicles brings significant advantages. GPS technology enables highly precise real-time tracking of vehicle positions, which is crucial for analysing vehicle movement and routes. When GPS is combined with MEMS sensors, which provide detailed information on acceleration, velocity, and other dynamic parameters, a comprehensive and accurate information of the vehicle's dynamic behaviour can be obtained. These sensors are compact, energy-efficient, and relatively inexpensive, making them ideal for widespread application in the automotive industry. By using these technologies, researchers can better understand vehicle performance, optimize design, and improve safety features, thereby contributing to the enhancement of efficiency and safety in traffic.

1.3 Literature Review

Apart from GPS and MEMS devices, other types of devices are also used, as shown in [1]. A wireless accelerometer mounted on the inner rim of the wheel was used here. It has been shown that this method can be used to monitor the condition of the screw joints that hold the wheel to the hub. In the paper [2], a triaxial accelerometer based on MEMS technology was used to analyse the obstacles on the road. Along with a smartphone camera and a model based on neural networks, a new model was created to analyse these phenomena. In a study [3], the application of MEMS accelerometers for evaluating the dynamics of cargo securing on road vehicles is investigated. Safety is a crucial aspect of successful cargo transport, particularly in road traffic, where the dynamics of a vehicle during routine events such as braking, steering, and evasive manoeuvres vary at different locations within the vehicle. The authors developed a smartphone-based system utilizing a dual acquisition method capable of capturing images of road surface anomalies and measuring the vehicle's acceleration upon their detection. The study highlights the importance of MEMS accelerometers in monitoring accelerations and related forces acting on cargo during transport. The use of MEMS accelerometers provided reliable data for assessing the impact of dynamic events on cargo, enabling adjustments to cargo packaging, securing methods, and even modifications to transport routes. The results emphasize that MEMS devices play a crucial role in enhancing transport safety by providing valuable insights into the dynamics of securing the cargo on road vehicles. The research presented in the paper [4] focuses on the accuracy assessment of three selected devices used for distance measurement in dynamic vehicle tests: accelerometer (XL meter), a global position measurement device (VBOX), and a constructed 5th wheel providing precise measurements of distance, speed, and acceleration. The experiments involved attaching all three devices to the vehicle during various driving conditions. Results indicated that GPS-based devices, such as VBOX, outperformed accelerometers, particularly in locations with a clear view of the sky. Another application of GPS based measuring accelerometers can be seen in [5]. Here the authors investigated the influence of various suspension configurations regarding comfort in N2 type vehicles. Vehicles safety in traffic with regards to proper functioning is ensured through periodic technical inspections. As vehicles are increasingly complex for traditional inspection, sophisticated test equipment must be used. An example of the analysis of the periodical technical inspection of vehicles in Slovakia is presented in [6]. Here, it was pointed out that the previously used inertial decelerometers are being replaced with modern triaxial accelerometers. These accelerometers are able to perform remote reading of the vehicle braking parameters, i.e. deceleration. Various examples of recent investigation into possibility of using triaxial accelerometers for the vehicle dynamics investigations can be seen as an example in [7, 8]. Research [9] shows that sensors have indeed become a reality in modern construction machines. This paper addresses the challenges of modelling construction machine dynamics, while also considering their low velocities and unique operating conditions. The proposed vehicle model, tailored for construction machines, is validated using experimental data obtained from a series of tests. The dynamic model includes considerations for vehicle and wheel dynamics, weight transfer, and the transmission system. It effectively models wheel slip behaviour, especially crucial for low-speed systems typical in construction machinery. Experimental validation, conducted using a range of sensors, including triaxial accelerometers, demonstrates a close match between the system model and collected data. Additional examples of the use of sensors to analyse construction machines are presented in [10, 11]. Paper [12] investigates the application of sensor-based technology to improve safety management on construction sites. Given the complexity of construction sites and the increased amount of information, traditional safety management faces challenges. Sensor based technology provides efficient ways to collect, identify and process information, thus enabling the creation of a new generation of methods to improve safety management in the construction sector. The paper reviews various sensor-based technologies that have been applied to safety management on construction sites, including location-based technologies, visual recognition, and wireless sensor networks. Among others, it emphasizes the importance of triaxial accelerometers as essential elements for future analysis of the processes that construction machines are subjected to.

Previous research indicates the presence of triaxial accelerometers as sensors to test the dynamic characteristics of vehicles, along with all the other phenomena related to movement and performance of work using vehicles. It has been proven several times that triaxial accelerometers based on MEMS technology are relevant devices for measuring vibrations on internal combustion engines and vehicles or for road profile investigations [13-15].

1.4 Structure of the Paper

The paper is composed of four chapters. In the first chapter, an introduction to the topic of the paper is given, the relevant literature is reviewed, and the aim of the paper is set. The second chapter presents the measurement method and the devices that were used for measurement. The third chapter presents the measurement results and a comparison between individual types of devices for both tested vehicles. The fourth chapter is the conclusion of the paper, where the research is summarized and an assessment of the possibility of using these devices is given. At the end, references are listed.

2 MEASUREMENT METHOD AND DEVICES

2.1 Overall Approach

To compare the deceleration measurements on two different types of vehicles, an experimental deceleration measurement was performed on one passenger vehicle and one construction machine.

To measure the vehicle's deceleration, it is necessary to provide an area where it can be done safely, while respecting the safety of other road users. The length of the test track should be at least 75 m for a passenger vehicle and 50 m for a construction machine. The minimum velocity at the moment of braking start should be 40 km/h for a passenger vehicle. For a construction machine, a velocity of at least 40% of its maximum velocity should be achieved. When the desired velocity is reached, it is necessary to press the service brake pedal and maintain the pressure evenly until the vehicle comes to a complete stop. Deceleration is recorded using the selected deceleration measuring device. The tested vehicles are shown below.

2.2 Passenger Vehicle

The Škoda Kamiq passenger vehicle, shown in Fig. 1, was used to measure deceleration. The tested vehicle is powered by a turbocharged otto engine, with a volume of 1498 cm^3, maximum power 110 kW which is achieved from 5000 to 6000 min^{-1} and a torque of 250 Nm which is achieved from 1500 to 3500 min^{-1}. The engine is paired with a seven-speed DSG automatic transmission. The vehicle curb weight is 1279 kg, and the maximum speed is 215 km/h. The braking system is a double circuit hydraulic system with self-ventilating discs on all wheels. The vehicle is equipped with modern active safety systems (ABS, ESP, ASR, etc.).

Figure 1 Passenger vehicle – Škoda Kamiq

2.3 Construction Machine

The second vehicle that was tested, shown in Fig. 2, is a Caterpillar 444 combined construction machine powered by a diesel engine with a volume of 3600 cm^3 and a power of 82 kW. The vehicle curb mass is 9500 kg, and the maximum speed is 40 km/h.

Figure 2 Construction Machine – Caterpillar 444

2.4 Measurement Equipment and Setup

As a part of this experiment, three devices were used to measure deceleration. The first device is of the inertial type, which until now has been most often used for these purposes. The device with its characteristics is shown in Fig. 3. The other two devices are based on modern technologies of GPS (Global Positioning System) and MEMS (Micro Electro Mechanical System) accelerometers. These two devices with their characteristics are shown in Figs. 4 and 5.

Deceleration measurement range	0 – 10 m/s^2
Force on the pedal range	0 – 1000 N
Pressure range	0 – 20 bar
Measurement accuracy	± 1%
Power supply	230 V/15V AC
Battery	6 V / 1.2 Ah
Weight	1.1 kg
Dimensions	120 x 65 x 245 mm

Figure 3 Inertial decelerometer [16]

The goal of the measurements is to investigate:

- The possibility of using devices based on MEMS and GPS technology instead of inertial devices for measuring vehicle deceleration and
- Flexibility of measuring devices on different types of vehicles.

Measurement range	0 – 4g
Measurement resolution	0,01g
Measurement accuracy	± 1%
GPS accuracy	± 5 m
GPS frequency	10 Hz
Memory type	SD card
Battery	6 – 28 V, 100 mA
Mass	0.266 kg
Dimensions LxWxH	113x63x96 mm

Figure 4 DriftBox GPS measurement device [17]

The set goal is important because of the measurement of the efficiency of the braking system in all types of vehicles, which cannot be tested on the inspection lines of the vehicle technical inspection stations. The most important value that needs to be determined is the maximum deceleration and matching of the character of the results obtained by measurements with these devices. The results of measuring deceleration with devices based on GPS and MEMS technology are compared to the results obtained by measuring on an inertial device.

Measurable range of acceleration	± 100G
Frequency of data acquisition	100 ÷ 20000 Hz
Possible data records	up to 4 10^9 points
Weight	65 g
Dimensions	76.2 x 29.8 x 15 mm
Temperature work range	- 40 ÷ 80 °C
Type of PC connection	USB
Battery lifetime (with one charge)	5.5 ÷ 15.5 h
Measurement accuracy	± 5%

Figure 5 MEMS accelerometer [18]

The measurement setup in a passenger vehicle, for the purpose of measuring deceleration, is depicted in Fig. 6.

Figure 6 Measurement setup in a passenger vehicle

The measurement setup in the construction machine had to be adapted to the specific construction of the machine cabin. Fig. 7 shows measurement setup of these devices in the cabin of the construction machine.

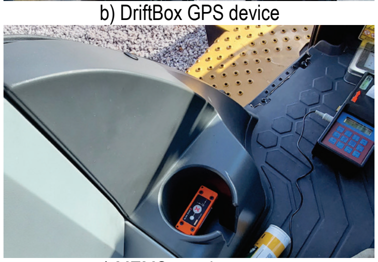
The decelerations measured by the mentioned devices were collected and processed using appropriate software programs developed by the manufacturers of these devices. The results are presented on common diagrams, where along with the display of the deceleration, the change in the velocity of the vehicle and the distance travelled during the braking are also given. The data recording frequency of the Racelogic DriftBox device was 10 Hz, and the MEMS accelerometer was 150 Hz.

3 RESULTS AND DISCUSSION

Several measurements were performed on both vehicles. The measurements were performed on the test track, which is isolated from the rest of the traffic, to ensure safety when conducting the experiments. After the setup of the measuring devices from Figs. 6 and 7, several test measurements were conducted to check the accuracy of the measuring devices. The device in Fig. 3 is an inertial device and printed data on the measured deceleration were obtained on it, which were then processed in digital form.

a) Inertial decelerometer

b) DriftBox GPS device



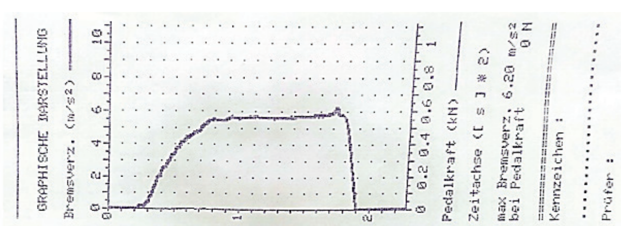
c) MEMS accelerometer

Figure 7 Measurement setup in a construction machine

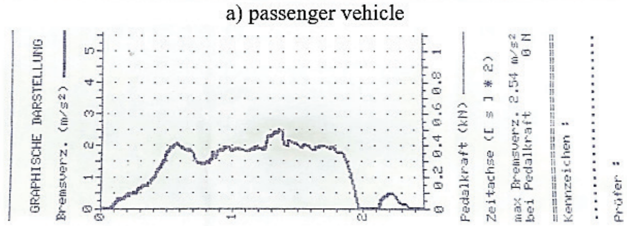
3.1 Measurement Results

The first measured set of decelerations of the passenger vehicle was measured from the initial velocity of 55 km/h, and the of the construction machine from 20 km/h. Fig. 8 depicts the results of deceleration measurements, for both vehicles, obtained in printed form from the inertial decelerometer shown in Fig. 3.

In Fig. 8, the time axis is given with a factor of 2. Therefore, for the exact time axis, the time from the axis on the diagram should be multiplied by 2.



a) passenger vehicle



b) construction machine

Figure 8 Results obtained from inertial decelerometer

It has already been pointed out that along with the measurement of deceleration by means of an inertial device, the measurement was also performed with two other devices, shown in Figs. 4 and 5. Fig. 9 depicts the results of measuring the deceleration of a passenger vehicle with the devices, while the results for the construction machine are depicted in Fig. 10.

In addition to deceleration, the Racelogic DriftBox device can record other kinematic parameters when the vehicle is moving, so that in addition to the results of the deceleration, the change in speed and the distance travelled while the vehicle is being stopped are also obtained.

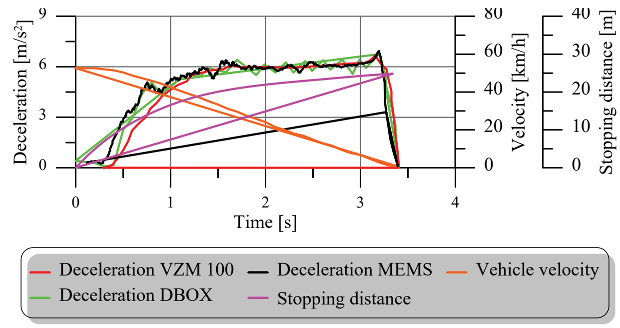


Figure 9 Deceleration results for passenger vehicle, first set

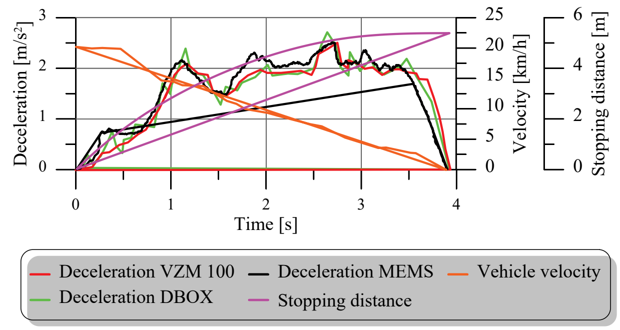


Figure 10 Deceleration results for construction machine, first set

3.2 Evaluation

Based on the obtained results of vehicles' deceleration measurements depicted in Figs. 9 and 10, it can be concluded that all three measuring devices for both vehicles match well when it comes to the nature or character of the measurement results. It is clearly seen that devices based on GPS and MEMS technology work at higher data acquisition frequencies than the inertial decelerometer. That is why they can be used to measure data that are of dynamic nature. This is especially depicted in construction machines. Oscillation of the recorded results with the construction machine (Figure 10) is a consequence of the rigid construction of the machine, its mass and the performance of the suspension system.

Another important parameter obtained by measurements are the values of the maximum decelerations of the given vehicles. Tab. 1 shows the results of the measured maximum deceleration for all three measuring devices, for the first set of measurements from Figs. 9 and 10. In addition to these values, the percentage deviations of the deceleration measured with MEMS and GPS devices in relation to the deceleration measured with an inertial device, as a reference measuring device, were also calculated.

Table 1 Measurement results – first set

Measurement device	Passenger vehicle			Construction machine		
	v = 55 km/h			v =20 km/h		
	a (m/s^2)	r (%)		a (m/s^2)	r (%)	
DriftBOX GPS	6.25	0.8	-	2.60	2.3	-
Inertial device	6.20		0.95	2.54		0
MEMS acc.	6.26	-		2.54	-	

In Tab. 1, a is the measured maximum deceleration of the vehicle, while r is the deviation between the measured acceleration values for the two devices (DriftBOX – Inertial, Inertial – MEMS) and is calculated as the quotient of the difference between the measured deceleration values from the two devices ($a_{\max}$, $a_{\min}$) and the larger value of those two ($a_{\max}$) multiplied with 100:

$$r = \frac{a_{\max} - a_{\min}}{a_{\max}} \cdot 100\ (\%). \tag{1}$$

In addition to the previously discussed deceleration measurements on both vehicles for the specified conditions, two more measurements were performed for different velocities of the vehicles at the moment of braking process start. Tab. 2 shows the measurement results for the other two velocities for both vehicles.

Table 2 Measurement results – second and third set

Measurement device	Passenger vehicle			Construction machine		
	v = 45 km/h			v =25 km/h		
	a (m/s^2)	r (%)		a (m/s^2)	r (%)	
DriftBOX GPS	6.99	0	-	3.27	1.2	-
Inertial device	6.99		0	3.31		0.6
MEMS acc.	6.99	-		3.29	-	
Measurement device	Passenger vehicle			Construction machine		
	v = 65 km/h			v =35 km/h		
	a (m/s^2)	r (%)		a (m/s^2)	r (%)	
DriftBOX GPS	7.54	0.53	-	3.92	2	-
Inertial device	7.5		1.31	4.00		0.25
MEMS acc.	7.6	-		3.99	-	

3.3 Interpretation

If the results of both vehicles are summed up, there is clear justification in using modern devices based on MEMS or GPS technology for several reasons:

- the measurement results match, both in the character of the signal and maximum deceleration values,
- the deviations of the measured maximum deceleration values are within the limits prescribed by the device manufacturer, i.e. according to the device accuracy,
- data acquisition and processing are more simplified than with inertial devices,
- there is no need for their laboratory calibration as in the case of an inertial device,
- by working with a higher frequency of data acquisition, they are suitable for sensitive measurements in a range of dynamic loads,
- they are easier to handle and transport.

It is important to note that, since the displayed devices are properly calibrated, and their accuracy is clearly declared by the manufacturer, the measurement error between the values of individual decelerations is minimal. This is clearly seen from Figs. 9 and 10, where the measured deceleration results match very well in character and obtained values.

4 CONCLUSIONS

The aim of this paper was to investigate whether measuring devices based on GPS and MEMS technology are suitable for use when researching the dynamic characteristics of construction machines travel, with emphasis on deceleration. To investigate this, a deceleration measurement experiment of a construction machine was carried out. Deceleration was measured with a traditional Maha VZM 100 inertial decelerometer, as well as with modern Racelogic DriftBOX measuring devices and a triaxial accelerometer manufactured by Mide Technologies. The other two devices are based on GPS and MEMS technologies respectively. The recorded results of the deceleration of the construction machine showed good matching of the three devices, and a very small deviations in measured maximal decelerations for each vehicle at different stopping velocities. This concludes that modern devices can be successfully used to determine the dynamic characteristics of construction machines travel. This is important when analysing the performance of these vehicles, their load and predicting operating modes. To prove that these devices are universal when it comes to testing the dynamic characteristics of vehicles, in addition to testing a construction machine, measurements were also performed on a passenger vehicle. The measurement methodology is the same for both vehicles. Also, as for the construction machine, the obtained results showed that the measuring devices match well. Although research dealing with the same topic is very scarce, the ones available showed good results in the use of GPS or MEMS devices for determining the dynamic characteristics of vehicles [3, 19, 20]. This shows that the experiments and research results in this paper justified the use of GPS and MEMS devices. From all the above and the results presented, it is reasonable to expect that devices based on GPS and MEMS technology will be increasingly represented in the testing of dynamic characteristics of all types of vehicles in the future. With the very development of the ADAS Systems (Advanced Driver Assistance System) and further integration into vehicles, it can be expected that devices based on GPS and MEMS technologies will be a standard part of the vehicle's equipment for monitoring the dynamic characteristics of the vehicle. These measuring devices can be used in automated vehicles for logistics and transport of goods, so that the dynamics of the movement of such vehicles can be monitored [19]. Further research will be focused on these systems. Although the results obtained by this research are very good, current GPS and MEMS devices still have certain limitations. Some limitations are the strength of the GPS signal, the need to use special software to process and analyse the results, the sensitivity of the device and the expertise required to operate them.

5 REFERENCES

[1] Borecki, M., Rychlik, A., Olejnik, A., Prus, P., Szmidt, J. & Korwin-Pawlowski, M. L. (2020). Application of Wireless Accelerometer Mounted on Wheel Rim for Parked Car Monitoring. *Sensors, 20*(21), 6088. https://doi.org/10.3390/s20216088

[2] Lee, T., Chun, C. & Ryu, S. (2021). Detection of Road-Surface Anomalies Using a Smartphone Camera and Accelerometer. *Sensors, 21*(561). https://doi.org/10.3390/s21020561

[3] Gnap, J., Jagelčák, J., Marienka, P., Frančák, M. & Kostrzewski, M. (2021). Application of MEMS Sensors for Evaluation of the Dynamics for Cargo Securing on Road Vehicles. *Sensors, 21*(2881). https://doi.org/10.3390/s21082881

[4] Kuchar, P., Janoško, I., Holubek, M., Čedik, J. & Pexa, M. (2022). The Accuracy Assessment of Devices Used for Distance Measuring in Dynamic Vehicle Tests. *Acta Technologica Agriculturae, 3*/2022, 150-156. Nitra, Slovaca Universitas Agriculturae Nitriae. https://doi.org/10.2478/ata-2022-0023

[5] Mitroi, M. F. & Chiru A. (2020). Determinations regarding the influence on movement and comfort of different elastic suspension structures in N2 type vehicles. *IOP Conf. Series: Materials Science and Engineering 997*(012049). https://doi.org/10.1088/1757-899X/997/1/012049

[6] Tapak, P., Kocur, M., Rabek, M., Matej, J. (2023). Periodical Vehicle Inspections with Smart Technology. *Appl. Sci., 13*(7241). https://doi.org/10.3390/app13127241

[7] Galvagno, E., Mauro, S., Pastorelli, S. & Tota, A. (2020). A Smart Measuring System for Vehicle Dynamics Testing. *SAE Technical Paper* 2020-01-1066. https://doi.org/10.4271/2020-01-1066

[8] Widner, A., Tihanyi, V. & Tettamanti, T. (2022). Framework for Vehicle Dynamics Model Validation. *IEEE Access, 10*, 35422-35436. https://doi.org/10.1109/ACCESS.2022.3157904

[9] Alexander, A. & Vacca, A. (2017). Longitudinal vehicle dynamics model for construction machines with experimental validation. *International Journal of Automotive and Mechanical Engineering, 14*(4), 4616-4633. https://doi.org/10.15282/ijame.14.4.2017.3.0364

[10] Jiang, Y. & He, X. 82020). Overview of Applications of the Sensor Technologies for Construction Machinery. *IEEE Access 8*, 110324-110335. https://doi.org/10.1109/ACCESS.2020.3001968

[11] Casoli, P., Scolari, F., Minav, T. & Rundo, M. (2020). Comparative Energy Analysis of a Load Sensing System and a Zonal Hydraulics for a 9-Tonne Excavator. *Actuators, 9*(39). https://doi.org/10.3390/act9020039

[12] Zhang, M., Cao, T. & Zhao, X. (2017). Applying Sensor-Based Technology to Improve Construction Safety Management. *Sensors, 17*(8). https://doi.org/10.3390/s17081841

[13] Šehović, J. (2023). Contribution to the Diagnostics of the Internal Combustion Engine Timing Mechanism Chain. In: Ademović, N., Kevrić, J. & Akšamija, Z. (Eds) *Advanced Technologies, Systems, and Applications VIII. IAT 2023. Lecture Notes in Networks and Systems, 644*. Springer, Cham. https://doi.org/10.1007/978-3-031-43056-5_35

[14] Šehović, J. (2022). Application of MEMS Accelerometers in Measuring Vertical Oscillations in Motor Vehicles. In: Ademović, N., Mujčić, E., Akšamija, Z., Kevrić, J., Avdaković, S. & Volić, I. (eds) *Advanced Technologies, Systems, and Applications VI. IAT 2021. Lecture Notes in Networks and Systems, 316*. Springer, Cham. https://doi.org/10.1007/978-3-030-90055-7_24

[15] Shi, B., Shen, S., Liu, L. & Wang, X. (2021). Estimation of Vehicle Speed from Pavement Stress Responses Using Wireless Sensors. *Journal of Transportation Engineering, Part B: Pavements 147*(3). https://doi.org/10.1061/JPEODX.0000288

[16] VZM 100 Decelerometer, Version 4 of the Operating Manual dated 20. (2000). D1 2002BA1-GB04, MAHA GMBH & CO. KG.

[17] Racelogic. (2009). DriftBox Manual.

[18] Retrieved from https://www.mouser.com (Accessed 08.02. 2024)

[19] Lerher, T. & Bencak P. (2022). Advanced Technologies in Logistics Engineering: Automated Storage Systems with Shuttles integrated with Hoisted Carriage. *Tehnički glasnik, 16*(3), 336-342. https://doi.org/10.31803/tg-20220509104609

Authors' contacts:

Jasmin Šehović, Assistant Professor PhD
(Corresponding author)
University of Sarajevo, Faculty of Mechanical Engineering,
Vilsonovo šetalište 9, 71000 Sarajevo, Bosnia and Herzegovina
sehovic@mef.unsa.ba

Mirsad Trobradović, Assistant Professor PhD
University of Sarajevo, Faculty of Mechanical Engineering,
Vilsonovo šetalište 9, 71000 Sarajevo, Bosnia and Herzegovina
trobradovic@mef.unsa.ba

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240517182356

Subject review
Received: 2024-05-17, Accepted: 2024-05-22

Dynamic Model of Spur Gear with Friction and Crack in Tooth Root

Aleš Belšak*, Matej Ozebek, Mario Hirz

Abstract: A mathematical model developed to support investigation of pairs of gears is presented. It includes the effects of actuation, the loads, the changing tooth stiffness, and the changing friction between tooth flanks during engagement. Friction between the tooth flanks generates an alternating non-symmetrical periodic function and the behaviour during the contact. The system actuation is modelled for different sources and the tooth stiffness. The mathematical model of the mechanical system of gear has the form of a system of two non-homogeneous nonlinear second-order differential equations.

Keywords: dynamic analysis; gear fatigue crack; gears; surface friction

1 INTRODUCTION

A number of measurements and numerical analyses is necessary for the diagnostics of mechanical systems prior to developing advanced algorithms for predicting damages and the remaining life-time of the system. This applies exemplarily for gear sets. Here, it is required to use mathematical models for analyses and optimization, which depend to a great extent on the quality of the applied diagnostics and the perception of the mechanical system.

This paper deals with a mathematical model that is developed to support investigation of pairs of gears. The accuracy of the model depends on the modelling requirements and the quality of observation, and available mathematical tools and software play an important role when developing relevant models.

In the past, various mathematical models have been developed by a number researchers, primarily focusing on gear dynamics, considering a multitude of different influencing quantities.

Exemplarily, a nonlinear time-varying dynamic model of a spur gear pair is studied by [1] for the prediction of modulation sidebands, which are caused by different modulation internal excitation. They pay attention to backlash, modulation time-varying mesh stiffness, and modulation transmission error.

The impact of dynamic backlash and rotational speed on the six-degrees-of-freedom model of the gear system with the time-varying meshing stiffness is studied by [2]. Here, it is possible to determine the relationship between dynamic backlash and centre distance.

A dynamic model of three shafts and two pairs of gears in mesh, with 26 degrees of freedom, with the effects of variable tooth stiffness, pitch and profile errors, friction and a localized tooth crack on one of the gears is presented by [3].

A two-stage planetary gearbox is presented by [4]. They establish a system coupling torsional dynamical model, which takes into account the time-varying mesh stiffness, friction forces and interstage coupling factors. The friction and lubrication states are classified in order to analyse the calculation of friction coefficients under various conditions.

The dynamic performance of the gear is dealt with by [5]. In the dynamic model, the profile deviation between involute gear and microsegment gear is considered as a displacement excitation.

An analytical model for helical gears is proposed by [6]; their model characterizes the contact plane dynamics and captures, owing to the sliding friction, the velocity reversal at the pitch line.

A dynamic wear prediction methodology is described by [7] to research the coupling effects between surface wear and dynamics of spur gear systems. Here, a quasi-static wear model and a translational-rotational-coupled nonlinear dynamic model are combined.

On the basis of the Coulomb model, [8] deal with an analytical analysis of tooth friction excitations in errorless spur and helical gears.

In their paper, [9] deal with a new gear surface roughness induced noise source model, with the sliding contacts between meshing gear teeth being considered. They use a linear time-varying model of a spur gear pair (with sliding friction) for the calculation of the instantaneous sliding velocity between pinion and gear teeth.

The dynamics of the gear pair is connected with sliding friction and a crack in the tooth root is associated with different loading level by [10].

As an enhancement of the topics treated in the different cited works, in our work we address the impact of wear upon the vibrations of gears. For this purpose, a mechanical model with reduced masses is developed. In this model, the function of friction is defined by taking into consideration the knowledge of the engagement conditions and theories of friction in the Hertz contact. Besides, function of stiffness changes along the line-of action at the gear tooth surface and an error function in the form of a fatigue crack in the tooth, by which the stiffness is reduced, are added to the model..

2 GEAR PAIR MODEL

The engaged pinion and gear can be substituted by a variety of alternative mechanical models (oscillating models), thus allowing observations of responses by the dynamic system. Different approaches and forms of alternative mechanical models can be found in literature [11-14].

When it comes to modelling gears, an approach based on reduced masses and displacements is often used. The reduced masses are calculated on the basis of the moment of inertia of the gear and the rotations are transformed into translations. The gear pair is modelled as a system of non-homogeneous, non-linear second-order differential equations – a system of equations of motion. Different forms of excitation, damage to the tooth and the mechanism of friction at the contact point between the tooth flanks in the system's pinion and gear geometry are represented in the selected model. The stiffness of the tooth during engagement and the engagement conditions at a rotational frequency are changed to excite the model internally. The first model involves linear contact stiffness of tooth flanks, whereas the improved model involves non-linear stiffness and it depends on the point of contact of the tooth flanks. The connection between the discs is modelled as a shock absorber with stiffness K and damping C that transmit the drive torque in the form of tangential force from pinion to gear. Both masses are rotated by the drive torque within the selected time-varying form. The model of pinion and gear is divided into two free masses with all the external influences, Fig. 1. The translation x of the spring and damper and the rotation θ of the gear are included in the differential equations. The equilibrium Euler's dynamic equation is defined for each gear. Two non-homogeneous second-order differential equations form the system of equations of motion for the gear pair. Drive torque M_{p1} and the output torque M_{p2} provide excitation. Below is the system of equations of motion [11]:

$$\begin{aligned} J_1\ddot{\theta}_1 + M_{tr1} + C(\dot{x}_1 + \dot{x}_2)R_i + K(x_1 + x_2)R_i = M_{p1} \\ J_2\ddot{\theta}_2 + M_{tr2} + C(\dot{x}_1 + \dot{x}_2)R_i + K(x_1 + x_2)R_i = M_{p2} \end{aligned} \quad (1)$$

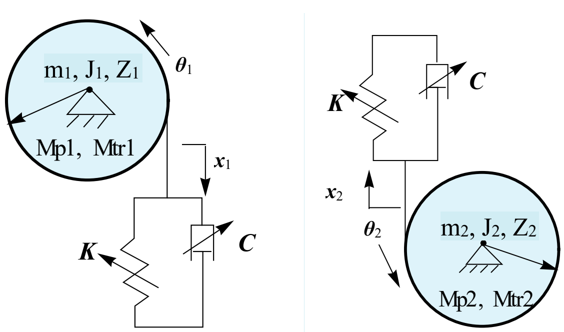


Figure 1 The model of gears

It is possible to divide the model of the pinion and gear set, Fig. 2, into two single masses, each of them as a body with rotational inertia J_i and with all external influences, depending on the translational displacements $x_1=\partial\theta_1 R_1$ and $x_2=\partial\theta_2 R_2$, damping C and the stiffnesses of springs K, acting on the radius R_i. In polar coordinates, the second derivative of the rotation is the angular acceleration of the rotational body. M_{tr} refers to the friction in the bearing.

For each body with external forces and torques acting on it, it is possible to define Equilibrium Euler dynamic Eq. (2). Considering small rotations in the same rotational directions as the two bodies, Fig. 2, it is possible to write as follows [10]

$$\begin{aligned} J_1\ddot{\theta}_1 + M_{tr1} + C(\partial\dot{\theta}_1 R_1 - \partial\dot{\theta}_2 R_2)R_1 + K(\partial\theta_1 R_1 \pm \partial\theta_2 R_2)R_1 = M_{p1} \\ J_2\ddot{\theta}_2 + M_{tr2} + C(\partial\dot{\theta}_2 R_2 - \partial\dot{\theta}_1 R_1)R_2 + K(\partial\theta_2 R_2 \mp \partial\theta_1 R_1)R_2 = M_{p2} \end{aligned} \quad (2)$$

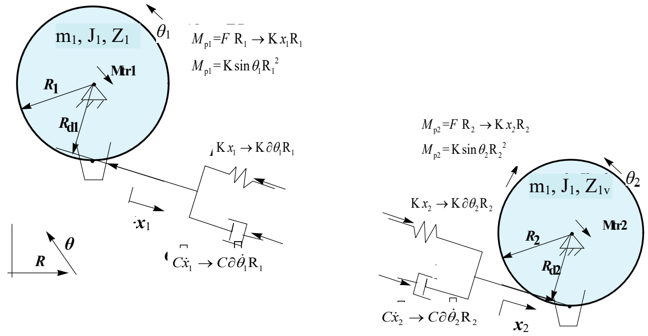


Figure 2 Forces and torques acting on gears

The gear pair is modelled precisely when the following factors are taken into consideration: rotation, the time-dependent changes of the drive torque M_{p1}, the output torque M_{p2}, damping C and the stiffness K of the tooth pair, as a real gear pair. If the tooth pair's stiffness, area pressure, Hertz contact relationship and the friction model during the contact are taken into consideration, the level of the model is improved.

$$\begin{aligned} J_1\ddot{\theta}_1 = M_{p1} - M_{tr1} \pm F_{trz1}R_{trz1} - C\partial\dot{\theta}_1 R_1 R_1 + C\partial\dot{\theta}_2 R_2 R_1 - K\partial\theta_1 R_1 R_1 + K\partial\theta_2 R_2 R_1 \\ J_2\ddot{\theta}_2 = M_{p2} - M_{tr2} \pm F_{trz2}R_{trz2} - C\partial\dot{\theta}_1 R_1 R_2 + C\partial\dot{\theta}_2 R_2 R_2 - K\partial\theta_1 R_1 R_2 + K\partial\theta_2 R_2 R_2 \end{aligned} \quad (3)$$

The above system of equations of motion (Eq. 3) is edited so that excitations of the external torques M_{p1} and M_{p2} appear on the right, whereas all the other variables are on the left. In the model, K stands for the replacement stiffness coefficient and C for the total damping of the gear pair.

$$\begin{aligned} J_1\ddot{\theta}_1 = M_{p1} - M_{tr1} \pm F_{trz1}R_{trz1} - C\partial\dot{\theta}_1 R_1 R_1 + C\partial\dot{\theta}_2 R_2 R_1 - K\partial\theta_1 R_1 R_1 + K\partial\theta_2 R_2 R_1 \\ J_2\ddot{\theta}_2 = M_{p2} - M_{tr2} \pm F_{trz2}R_{trz2} - C\partial\dot{\theta}_1 R_1 R_2 + C\partial\dot{\theta}_2 R_2 R_2 - K\partial\theta_1 R_1 R_2 + K\partial\theta_2 R_2 R_2 \end{aligned} \quad (4)$$

For solving the system of equations of motion (Eq. 4), direct numerical integration is used. The torque transfer between the engaged teeth involves the force from tooth to tooth, which includes friction.

The friction force direction is tangential to the flanks of the tooth at the current point of engagement. On the basis of the engagement conditions it is clear that the tooth force is oriented in the direction of angle α with respect to the engagement curve. The force acting on the tooth is not constant anymore. It, however, varies in mode (single, double) and the position and direction on the tooth flank and the radius of the associated engagement point. If the position of associated tooth engagement point changes, the tooth stiffness K changes in accordance with the position of the engagement point. Simultaneously, the friction, representing the damping, becomes an additional exciting force of the mechanical system.

Normally, pinion and gear work have no backlash, and the bearings are not completely rigid. This forms additional sources of excitation. Inertia moments for both pinion and gear are necessary for the equations of motion of the developed system. The following data is considered when calculating the inertia moments: material density ρ, modulus m, the number of teeth Z and width b and the coefficient of profile shifting x [10].

$$J_i = \left(\frac{\pi Z_i^4}{32} + \frac{B_0 x_i}{10^3}\right)\left(\frac{m_t}{10^3}\right)^4 \cdot \rho \cdot b \tag{5}$$

$$B_0 = -2818,2 + 267,74 Z_i - 837.38 Z_i^2 + 0.8443 Z_i^3 \tag{6}$$

3 STIFFNESS OF THE TEETH PAIR

On the pitch diameter, the contact gear force can be split into tangential and normal components. In the mechanical model, the actual force acts on the spring K and the damper C on the pitch radius.

The position of the engagement point changes on the tooth flank resulting in changes of the combined tooth stiffness with time and position. In the substituted model, the spring K represents the stiffness of the teeth pair depending on the elasticity of the tooth flank and the local elasticity of the contact surface. Four serially connected springs are applied to model the serial set-up of the tooth pair during the engagement. The stiffness of this spring system is calculated as a harmonic mean.

$$\frac{1}{k_{\text{povp.}}} = \frac{1}{k_1} + \frac{1}{k_2} + \frac{1}{k_{1\text{hz}}} + \frac{1}{k_{2\text{hz}}} \rightarrow k_{\text{povp.}} = \frac{k_1 k_2 k_{1\text{hz}} k_{2\text{hz}}}{\left(k_1 + k_2 + k_{1\text{hz}} + k_{2\text{hz}}\right) n} \tag{7}$$

The stiffness of the tooth at any radius between the root and the top land being k_1 and k_2 (n refers to the number of teeth engaged), whereas k_{hz} stands for the surface stiffness. Thus, the Hertz surface stiffness, which is given by [14], represents a part of the overall stiffness. The stiffness of a pair of contact surfaces is assumably constant during the entire period of engagement. The modulus E_y and the Poisson ratio ν as well as the shape of the contact define the surface stiffness h_{hz}.

$$k_{\text{hz}} = \frac{\pi \cdot E_y}{4(1-\nu^2)} \tag{8}$$

The material, loads and geometries of the surfaces in contact additionally influence the shape of the contact.

4 FRICTION FORCE AND TORQUE

The transferred torque acts on the tooth with a pulsating rectangular variable wave force during the operation. The drive torque is represented by the dynamic forces.

$$M_{\text{p1}} = (F_{\text{dinA}} + F_{\text{dinB}}) r_{\text{ON2}} \tag{9}$$

The friction in the system includes the friction in the bearings of the shaft and the friction at the Hertz contact point produced between the tooth flanks. The mechanisms of friction and relative gliding, and the lubrication within the dynamic model of the gear pair are taken into consideration.

In Eq. (10) the average sliding velocity is calculated at any point of engagement Y on the radius r_Y. In the equation, the rotation speed v at the pitch radius, the base radius R_b, and the pressure angle Φ are considered.

$$v_{\text{sr-dsr}} = \frac{v_{\text{obodna}}}{2}\left(\frac{1}{R_1} + \frac{1}{R_2}\right) R_{\text{b1}} \beta_{\text{r1}} \tag{10}$$

A function for describing the friction μ_A at a single point of engagement was developed by [12]:

$$\mu_A = \frac{v_{2A} - v_{1A}}{F_A t_A} \int_{-a_0}^{+a_0} \eta_A \, dx \tag{11}$$

F_A stands for the dynamic force transmission, t_A for the thickness of the oil film, v for the relative sliding velocity, a_0 for the radius of the Hertz contact and η_A for the dynamic viscosity of the lubricant.

We took the following factors into consideration when developing the dynamic model of gear and pinion. The change of the tooth force, sliding with direction of changing, rolling friction between the tooth flanks and in the bearings, and an unsatisfactory driving torque.

The model is primarily stimulated by the engagement and with the altering tooth stiffness; however, several facts need to be observed, otherwise a problem related to the determination of the damping occurs, which must be set up experimentally by observation. Friction forces for double (F_{trA}) and single (F_{trC}) engagements are:

$$F_{\text{trA}} = \mu F_{\text{nA}}, \quad F_{\text{trC}} = \mu F_{\text{nC}} \tag{12}$$

The torque of sliding friction can be as follows:

$$M_{\text{trY}} = F_{\text{trY}} r_{\text{N1Y}} \tag{13}$$

The friction torque distribution for typical engagement of a gear pair is shown in Fig. 3.

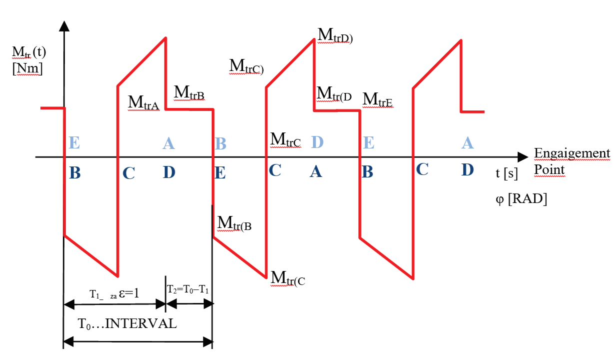


Figure 3 Friction torque at the intervals of engagement of tooth pair

A repetitive engagement behaviour for the gear calculated with Wolfram Mathematica is presented in Fig. 4. The frequency of engagement is taken into consideration and friction torque is calculated with a constant friction coefficient of 0.5 at constant torque transfer of 20 Nm. [10]

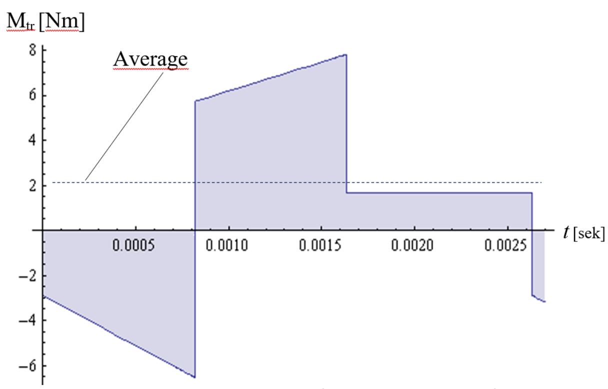


Figure 4 Friction torque distribution for one engagement of tooth pair

5 SOLVING THE MATHEMATICAL MODEL

A system of two second-order, non-linear, non-homogeneous differential equations forms equations of motion for the extended model of pinion and gear. The system of differential equations is prepared for solving, using numerical integration with software Wolfram Mathematica. By inserting numerical values for real pinion and gear from the data in Tab. 1 and the definition of the initial conditions.

Table 1 Data of the pinion and gear [10]

Parameters of the gear:	
$m_n = 4$	module
$z_1 = 19$	number of teeth for the pinion
$z_2 = 34$	number of teeth for the gear
$i = 1.789$	gear ratio
$n_1 = 20\ s^{-1}$	revolutions for the pinion shaft
$M_{p1}(t) = 20$ Nm	drive torque
$M_{p2}(t) = i \cdot M_{p1}(t)$	output torque
$b_1 = b_2 = 0{,}02$ m	tooth thickness
$r_1 = 0.03$ m	bearing radius
$l_1 = 0.02$ m	width of the bearing
$\alpha_n = 20°$	pressure angle
$\beta = 0°$	helix angle
$\varepsilon = 1.61$	overall degree of engagement
$a = 0.106$ m	wheel base
$c = 0.08$	damping coefficient (the tooth flanks)
$k = 462.1106$ N/m	average tooth pair stiffness
$\rho = 7800\ kg/m^3$	material density
$Ey = 210000$ MPa	Yield point of the tooth material
$\nu = 0.3$	Poisson's ratio
$\sigma_{el} = 500$ MPa	allowed stress for the tooth material
$J_1 = 0.00051095054\ kgm^2$	inertial moment of the cylinder – pinion
$J_2 = 0.00523938275\ kgm^2$	inertial moment of the cylinder – gear
$m_1 = 0.708$ kg	mass of the pinion
$m_2 = 2.266$ kg	mass of the gear

In our model, the influence of a crack in a tooth root is used as a reduction of stiffness of one gear tooth of the pinion.

We acquired responses of gear rotation from the numerical integration of differential equations as presented in Fig. 5. The response of the convergent transition responding at each change and a semi-stationary form of repetition are presented in the diagram. An increase in rotation and change of response indicate the single change in the stiffness of the gear tooth with a crack in tooth root. The second square wave shows the response of a tooth pair with a crack in a tooth root with tooth stiffness reduced by 35 %, all other responses are from the tooth pair without cracks in tooth root.

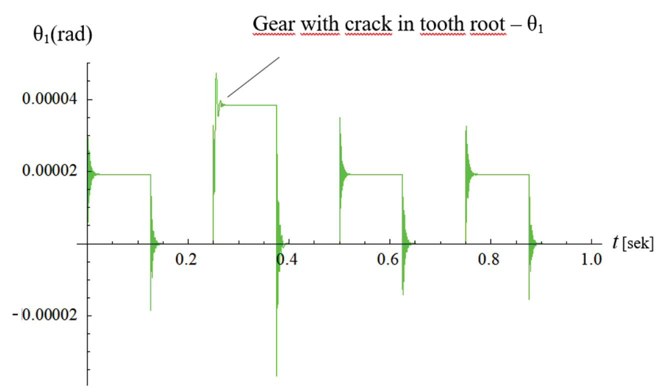


Figure 5 Detailed response of gear rotations of the considered model

Due to damping or energy dissipation of a non-conservative model, the amplitude is reduced. In accordance with the physical properties, any decreasing stiffness of the tooth causes an increase of the amplitude, whereas the oscillation frequency does not change.

In Fig. 6, Fig. 7 and Fig. 8, the tooth gear rotation under the same operating parameters and different lengths of fatigue cracks is presented, reducing tooth stiffness by 10 %, followed by 20 % and finally by 35 %.

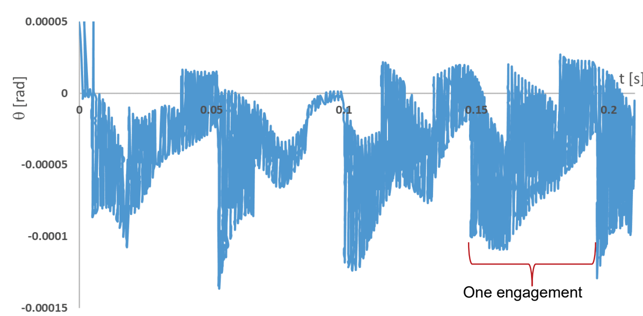


Figure 6 Tooth gear rotation with a 10 % reduction of tooth stiffness (M_{p1} = 130 Nm)

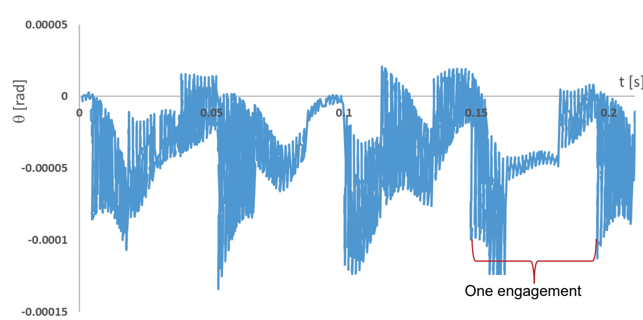


Figure 7 Tooth gear rotation with a 20 % reduction of tooth stiffness (M_{p1} = 130 Nm)

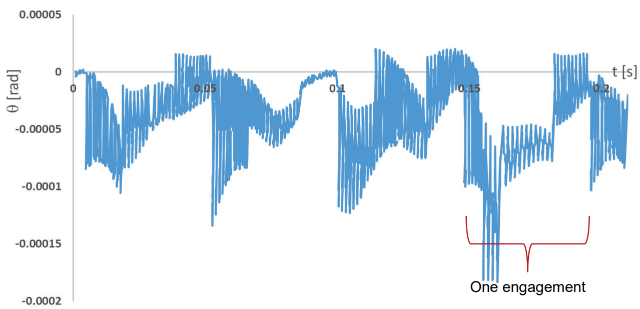


Figure 8 Tooth gear rotation with a 35 % reduction of tooth stiffness (M_{p1} = 130 Nm)

In Fig. 6 to Fig. 8 it is shown how a damaged tooth influences the tooth gear rotation – with the same parameters, being the drive torque, friction torque and variable stiffness, which represent additional excitation mechanisms applied in the mathematical model. If we observe the part of the diagram presenting the engagement of the gear pair with a crack in the tooth root, it is evident that in case of an increasing crack, which reduces tooth stiffness, gear rotation amplitude is redistributed in time domain. With reduction of tooth stiffness amplitude, the tooth gear rotation is increased.

6 CONCLUSION

The paper introduces the development of a mathematical model of a spur gear. The model includes a mathematically described variable excitation resulting from a friction model at the contact point of the tooth flanks. The shapes of the excitations make it possible to compare the model and the responses of the real gear drives. The excitation with input and output torque, variable stiffness of the gears, damage of a pinion with a fatigue crack in the tooth root, variable friction and damping are used in the replacement mechanical model of the gear pair. The nonlinear differential systems of equations are solved by numerical integration. As a result, the fluctuating nature of friction between the teeth flanks is confirmed by the developed and introduced excitation function of friction. The obtained results of the model represent the basis for comparing the results acquired with measurements, where it is difficult to locate a fault on the gear due to various influences. At the same time, the introduced model enables a more efficient search for suitable methods for noise reduction or signal filtering of measured signals in the course of gear investigation and optimization.

7 REFERENCES

[1] Wang, Z., Zhang, L., Luo, Y. & Chen, C. (2017). Dynamic Model of Spur Gear Pair with Modulation Internal Excitation. *International Journal of Rotating Machinery, 2017*, 1-8. https://doi.org/10.1155/2017/1264904

[2] Liu, Jie, Liu, S., Zhao, W. & Zhang, L. (2019). Dynamic characteristics of Spur Gear Pair with Dynamic Center distance and backlash. *International Journal of Rotating Machinery, 2019*(11), 1-9. https://doi.org/10.1155/2019/2040637

[3] Jia, S., Howard, I. & Wang, J. (2003). The dynamic modeling of multiple pairs of spur gears in mesh, including friction and geometrical errors. *International Journal of Rotating Machinery, 9*(6), 437-442. https://doi.org/10.1155/s1023621x03000423

[4] Xiao, Z., Chen, F. & Zhang, K. (2021). Analysis of dynamic characteristics of the multistage planetary gear transmission system with Friction Force. *Shock and Vibration*, *2021*, 1-10. https://doi.org/10.1155/2021/8812640

[5] Xiong, Y., Huang, K., Wang, T., Chen, Q. & Xu R. (2016). Dynamic Modelling and Analysis of the Microsegment Gear, *Shock and Vibration, 2016*, 1-13. https://doi.org/10.1155/2016/9691647

[6] He, S., Gunda, R. & Singh, R. (2007). Inclusion of Sliding Friction in Contact Dynamics Model for Helical Gears. *Journal of Mechanical Design, 129*(1), 48-57. https://doi.org/10.1115/1.2359474

[7] Yang, Y. & Zhang, J. (2016). Investigation on coupling effects between surface wear and dynamics in a spur gear system. *Tribology International, 101*, 383-394. https://doi.org/10.1016/j.triboint.2016.05.006

[8] Velex, P. & Sainsot, P. (2002). An analytical study of tooth friction excitations in errorless spur and helical gears. *Mechanism and Machine Theory, 37*(7), 641-658. https://doi.org/10.1016/S0094-114X(02)00015-0

[9] Kim, S. & Singh, R. (2007). Gear surface roughness induced noise prediction based on a linear time-varying model with sliding friction. *Journal of Vibration and Control, 13*(7), 1045-1063. https://doi.org/10.1177/1077546307078829

[10] Ozebek, M., Flašker, J. & Belšak, A. (2009). Diagnostika zobniške dvojice z uporabo mehanskega modela. *Magistrsko delo.* M. Ozebek, Maribor. (in Slovenian)

[11] Parker, R. G., Vijayakra, S. M. & Imajo, T. (2000). Non–linear dynamic response of a spur gear pair modelling and experimental comparisons, *Journal of sound and Vibration*, *237*(3), 435-455. https://doi.org/10.1006/jsvi.2000.3067

[12] Dowson, D. & Higginson, G. R. (1960). The effect of material properties on the lubrication of elastic rollers. *Journal of Mechanical Engineering Science, 2*(3), 188-194. https://doi.org/10.1243/JMES_JOUR_1960_002_028_02

[13] Kuang, J. H. & Lin, A. D. (2001). The effect of tooth wear on the vibration spectrum of a spur gear pair. *Journal of vibrations and acoustics, 123*(3), 311-317. https://doi.org/10.1115/1.1379371

[13] Kuang, J. H. & Lin, A. D. (2001). The effect of tooth wear on the vibration spectrum of a spur gear pair. *Journal of vibrations and acoustics, 123*(3), 311-317. https://doi.org/10.1115/1.1379371

[14] Yang, D. C. H. & Sun, Z. S. (1985). A rotary model for spur gear dynamics. *Journal of Mechanisms, Transmissions, and Automation in Design, 107*(4), 529-535. https://doi.org/10.1115/1.3260759

Authors' contacts:

Aleš Belšak, Assist. Prof. Dr.
(Corresponding author)
University of Maribor, Faculty of Mechanical Engineering,
Smetanova ulica 17, 2000 Maribor, Slovenia
+386 2 220 7710, ales.belsak@um.si

Matej Ozebek, MSc
Higher Vocational School,
Pot na Lavo 22, 3000 Celje, Slovenia
+386 3 428 5853, matej.ozebek@sc-celje.si

Mario Hirz, Assoc. Prof. Dipl.-Ing. Dr. techn.
University of Maribor, Faculty of Mechanical Engineering,
Inffeldgasse 11/II, 8010 Graz, Austria
+43 316 873 35220

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240518145453

Subject review
Received: 2024-05-18, Accepted: 2024-05-22

Teaching Predictive Maintenance using Industrial AI Tools

Ihor Savchenko*, Herbert Fleck, Peter Novotny, Helmut Ropin

Abstract: A new concept of an educational tool using predictive maintenance techniques is proposed. The tool is a part of a larger 'learning and research factory' that aims to enhance the current educational processes for industrial engineering students. The factory is organized as close as possible to the real business situation, so the students "learn by doing", gaining both theoretical knowledge and practical skills. Two use cases presented by the authors intended to effectively teach the principles of predictive maintenance. Basic elements of data science, machine learning and statistical analysis are used to prognose possible anomalies in the production process and react actively before a harmful event occurs. The authors outline ways for further development of the tool, including using it for other educational purposes.

Keywords: AI; anomalies; digital twin; machine learning; predictive maintenance

1 THEORETICAL OVERVIEW

1.1 AI in Education

Over the last decade, the artificial intelligence (AI) industry experiences an unprecedented boom. There is a multitude of various definitions of the AI term, such as "a branch of computer science devoted to developing data processing systems that perform functions normally associated with human intelligence, such as reasoning, learning, and self-improvement" [1], "a framework deployed with the objective of building intelligent systems that can creatively solve a given problem" [2] or "a paradigm that endows machines with intelligence, aiming to teach them how to work, react, and learn like humans" [3]. Most of these definitions clearly present the core of the idea: a machine able to "think like a human".

Rapid development of AI-related techniques and tools in almost all traditional areas of human life allowed many scientists and specialists to speak about the "AI revolution" starting as early as in 2017 [4-6]. The education is not an exception. Being a key element of a well-functioning society, the education as a process of passing and enhancing knowledge and skills through generations has been already deeply influenced by AI, with new ideas and theories appearing every day. In the first months of 2024 only, the concepts of using AI in healthcare education [7], language learning [8], social studies [9] and mathematics education [10] have been proposed and reviewed in multiple papers. The effectiveness of AI tools has been shown both on school [11] and college/university levels [12], as well as in different forms of adult (continuing) education [13]. Most researchers agree that correctly chosen AI tools greatly enhance the traditional learning process and will become an irreplaceable element of general education in the nearest future. Still, the reckless use of AI raises many ethical concerns [14], especially in areas like education, which has for thousands of years been an example of "human – human" interaction.

1.2 Predictive Maintenance

The concept of maintenance can be as old as the civilization itself, but after the onset of the Industrial revolution it gained a truly scientific meaning as a set of processes to avoid machine failures and to improve a machine's health condition [15].

With further industrial development the theories of maintenance became more complex, until the new digital revolution of the 21st century introduced the so-called "Industry 4.0" and specific maintenance techniques related to it. Achouch et al. defined predictive maintenance as "an approach that consists of improving the performance and efficiency of the manufacturing process by… offering the possibility of interventions through the prediction of failures" [16]. They regarded this approach as the most effective and developed compared with traditional ways like reactive (post-failure) and preventive (scheduled) maintenance. Cakir et al., by applying machine learning algorithms to predictive maintenance, postulated that this approach provides the longest life and highest reliability of equipment, as well as the most environmentally sound and cost-effective solutions [17]. In the Fig. 1, a standard predictive maintenance workflow is shown.

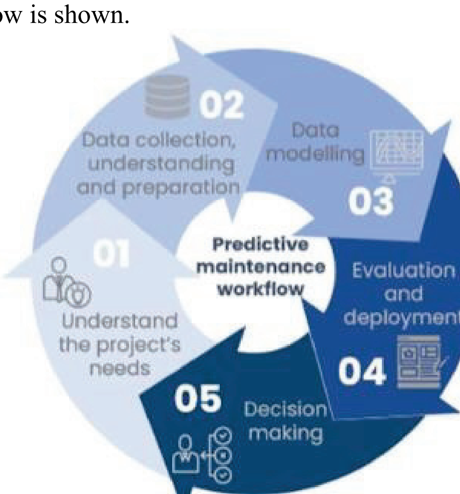


Figure 1 Predictive maintenance workflow [16]

Unlike the older approaches, predictive maintenance uses proactive methods to reduce cost and increase machine uptime. Predictive maintenance aims to foresee when a component or system will no longer fulfill its function. A key indicator in such a prediction is a remaining useful life (RUL), showing the amount of time during which the equipment is supposed to work normally before the need for repair/replacement occurs. Various AI techniques can be helpful in making such forecasts. In the systematic literature review provided by Van Dinter at al., the importance of digital twins (precise digital replicas of physical equipment) in predictive maintenance models was summarized and underlined. [15]

In the same paper, most popular statistical methods applied in predictive maintenance models were listed:

- probability distribution functions;
- Kalman and particle filters;
- Monte Carlo methods;
- Principal Component Analysis (PCA).

At the same time, several authors pointed out the challenges in applying predictive maintenance approaches as well as downsides of the model itself [18, 19]. The main issues can be divided into four categories:

- financial and organizational;
- data source;
- repair activities;
- deployment limits.

It has been specifically noted that the integration, support and updating of predictive maintenance techniques can be too expensive and not economically efficient for smaller industries and areas of work [20].

2 EDUCATIONAL TOOL

In this paper, the authors, representing a team of researchers affiliated with FH Joanneum (Austria) and FESTO (Germany), present a specific case: using a system of AI tools for teaching the basics of predictive maintenance for business engineering students on a bachelor/master level. The AI system consists of three main elements, integrated into a wider data generation and analysis system. The system architecture is presented in Fig. 2, while the three main elements are described below.

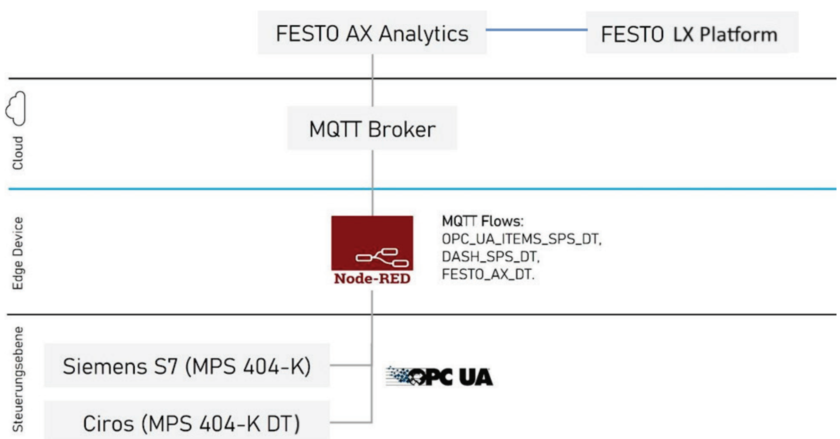


Figure 2 System architecture

2.1 Working Station

The working station is a simplified model of a real-life production equipment and a data generator at the same time. The station exists in two forms: a physical one (FESTO modular production system, or MPS) and its digital twin, created with CIROS virtual reality simulation tool. MPS consists of four consecutive stations simulating typical operations of the Conveyor Belt Production Process: distribution of source materials, measuring, assembling (joining two materials) and sorting the production by quality. This is an intuitively understandable, but hardly transportable learning appliance, driven electrically and by compressed air. On the other hand, CIROS simulation tool requires almost no maintenance and allows to process large productional quantities in short time without any manual effort. The digital twin in question was developed and implemented during the COVID-19 pandemic and continues to be used because of its effectiveness in distance learning. Both physical and virtual stations regularly produce the streams of data related to the different aspects of the micro factory in use. These data (e.g. speed of conveyor belt or time required to eject an element out of a cylinder) are recorded and stored for comparison and later statistical analysis. More details on the physical station and the CIROS digital twin are provided in the previous work of the authors. [21]

The working station used in the workshop is a part of a larger MPS 404-K learning factory developed by FESTO and designed to address economic topics in a production simulation. It provides a comprehensive understanding of an industrial production process, starting from the sensor technology through to the ERP system.

With MPS, the students have the opportunity to learn key performance indicators such as availability and efficiency and get a realistic impression of industrial processes.

The data generated by MPS are captured by its sensors and transferred to the middleware level by OPC UA specification. Then the data are processed and filtered by the Node-RED middleware. When only the relevant data are left, they are transferred further using MQTT – a standard protocol for communication between "internet of things" devices and other computers. OPC UA and MQTT brokers play crucial roles in transferring data between the learning factory and the FESTO AX software. FESTO AX analyses the obtained data and displays results in a dashboard.

2.2 FESTO AX

FESTO Automation Experience (AX) is an industrial AI- and machine learning (ML)-based solution designed to gain valuable insights from the system data. By analyzing live data in real time, FESTO AX enables quick identification of anomalies that can lead to reduction in productivity, quality problems or high-energy costs.

For the purpose of predictive maintenance, this means that through real-time analysis of data FESTO AX can early identify possible anomalies. This enables proactive (i.e. predictive) maintenance instead of reactive one.

FESTO AX applies ML to the collected data to recognize specific patterns and relations. Specifically, it uses k-means clustering, an unsupervised, easy-to-perform ML algorithm, to group the generated data into clusters based on their characteristics. K-means clustering is often used for anomalies detection and similar tasks. Here are the basic steps of this algorithm:

- Define the number of clusters K.
- Position the starting cluster centroids. This can be done by arbitrarily separating the data points into K clusters, then computing their centroids.
- Iterate over all data points and calculate the distances to the centroids of all clusters. In our example, the most common Euclidean distance was used. Then each data point was assigned to the cluster with the nearest centroid.
- Recalculate the centroids of new clusters.
- Repeat step 3 until the centroids remain stable.

For the workshop's purposes, a simplified clustering model was chosen with three expected clusters:

- "no anomalies"
- "conveyor belt attrition"
- "membrane tear in valve".

In addition, there were only two data features:

- "conveyor belt movement time"
- "suction cup active time".

Before the k-means procedure, features' data were normalized.

After the k-means calculation, the schematical results can be displayed as in the Fig. 3. In this case, three definitive clusters were formed based on two types of anomalies. In both cases, anomalies were caused by imperfect equipment parts (conveyor belt and valve membrane) and, as a result, slower operational times of these parts. Data points related to the normal settings are marked with blue, to the damaged conveyor belt – with red and to the damaged valve – with orange. Average distance between clusters was in the range of 30-40 ms for the belt and 40-50 ms for the valve.

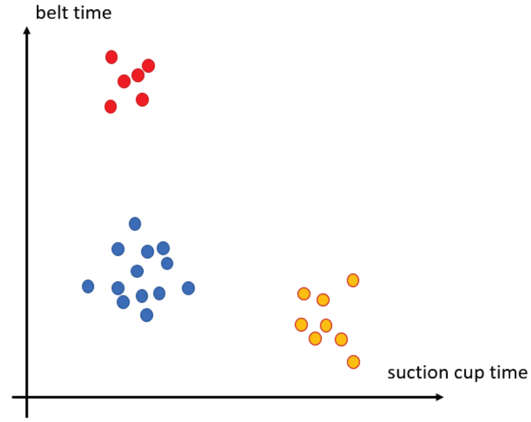


Figure 3 k-means algorithm results

The groups situated too far from the predefined normal criteria are marked as potentially dangerous (causing anomalies). For example, the system still functions with a slower conveyor belt or a damaged vacuum membrane, but without necessary measures, these states can result in a station crash. By identifying problems early and solving them users can reduce unplanned downtime and extend the service life of equipment. This leads to higher availability and efficiency indicators.

For anomaly detection and classification FESTO AX uses a practice-oriented "Human in the Loop" approach. This means that data are recorded, models are trained and anomalies are detected in real time, during the production process. The user helps the model to "learn" by confirming the fact of anomaly and classifying the anomalies. When there is a match, the anomaly is detected and the corresponding safety measure is proposed.

The results of system health monitoring in the form of "health scores" are displayed in FESTO AX user interface. It consists of following screens:

- Dashboards containing various widgets for data visualization, e. g. in a live chart form;
- Assets containing information about various machines and its components, e. g. conveyor belt or vacuum suction cup;
- Infrastructure providing connection between assets and dashboards;
- Data for different quantitative values collected by system sensors, e. g. duration of a suction cup movement;
- Models for various ML models based on collected data;
- Analytics where the user can choose an analysis object, define its features and the aim of analysis, assign the data, choose the model type and the visualization parameters.

2.3 FESTO LX

FESTO Learning Experience (LX) is a digital learning platform from FESTO that initially enabled companies and employees to expand their knowledge in the fields of automation technology, electrical engineering and pneumatics. The platform can be expanded to other topics, providing a wide range of online courses for beginners as well as for advanced learners. FESTO LX allows each user to study at their own preferred pace.

Main elements of the platform are learning paths and nuggets. A learning path is an interactive didactic module with a clear "begin to end" structure, helping to establish long-term knowledge and skills by practical problem solving. In this path, students learn theoretical basics, after that they gain skills and apply them in practice step by step. A learning path can contain different learning methods and media, such as online courses, e-learning modules, interactive simulations, videos, audio and text files as well as practical exercises. It can convey both theoretical and practical knowledge and help to improve or expand learners' skills in specific areas.

Nuggets can be defined as compact learning units in FESTO LX, each of which is dedicated to a specific topic and usually only last a few minutes. They are designed to teach a specific concept or skill in a short time. Nuggets allows for more personalization and flexibility in the learning process. It is important to note that nuggets are not simple "building blocks" of a specific learning path. They can be used independently, or be included in different paths at the same time.

Main types of nuggets are Task, Info, and Question-nuggets.

- Task-nuggets consist of series of tasks to perform using physical or digital equipment. They are usually accompanied by instructions, motivation explanations and list of helpful materials.
- Info-nuggets provide general theoretical information on the subject learned in the module.
- Question-nuggets serve to assess the users' skills and knowledge after performing the tasks. The proposed questions can be of various types, e.g. single-choice and multiple-choice tests, fill-the-gap and match-pair exercises, etc.

The learning path in question, named "FESTO AX and Predictive maintenance at the MPS station", was divided into three modules. In the first module, the theoretical foundations of the learning path were laid down. This is followed in the second module by practical implementation of these fundamentals with examples from FESTO AX, such as the creating analysis objects, models and dashboards. The last module presents use cases in which the content learned can be applied again on a physical station.

An example of a task nugget page, instructing the user (in German) to find pre-defined system features, is shown in Fig. 4.

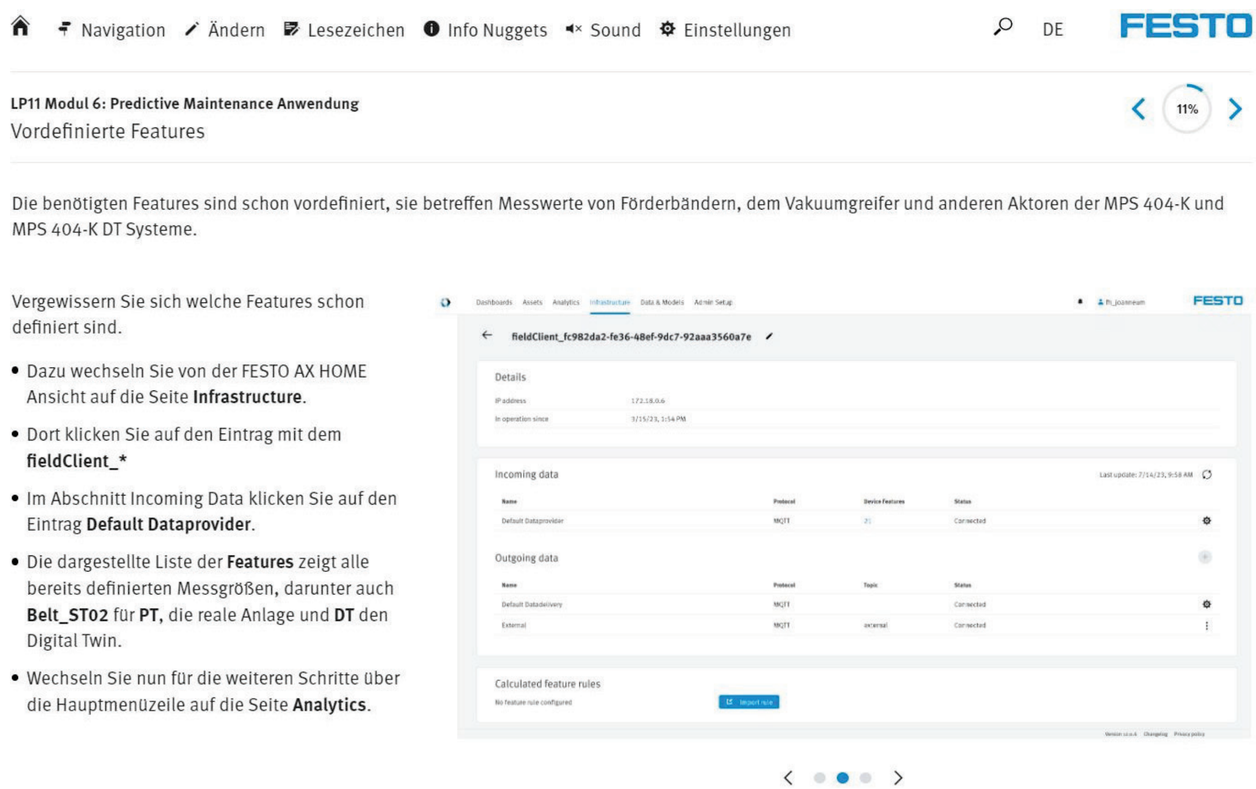


Figure 4 FESTO LX page outlook (task nugget)

2.4 Use Cases

Within the frame of the authors' workshop, two use cases for business engineering students were prepared.

First, one related to the presumed attrition of conveyor belt elements of the station 2 "Measuring". The effect of attrition was simulated by a potentiometer with a control knob installed at the front panel of the station. By turning the knob, users can regulate voltage at the electrical motor driving the conveyor belt, thus influencing the belt's speed.

In the course of the workshop, students first created a new analysis object with a feature related to the speed of the conveyor belt. Then, a model was created and configured using previously collected speed data for training. The model was assigned to the analysis object in the Analytics screen of AX. At the end of preparation, a dashboard was created with a gauge widget for displaying the health score of the analyzed object. and a chart widget displaying the data in real time. After that, three tests were conducted: without changing the motor voltage, with changing it to 70% of an optimum (simulating the attrition) and back with full voltage (simulating the situation after repairs). The health score was over 90% after the first and the third tests, while the slowing down of the conveyor belt resulted in the health score between 40% and 60%. The dashboard displaying the optimal situation is presented in Fig. 5.

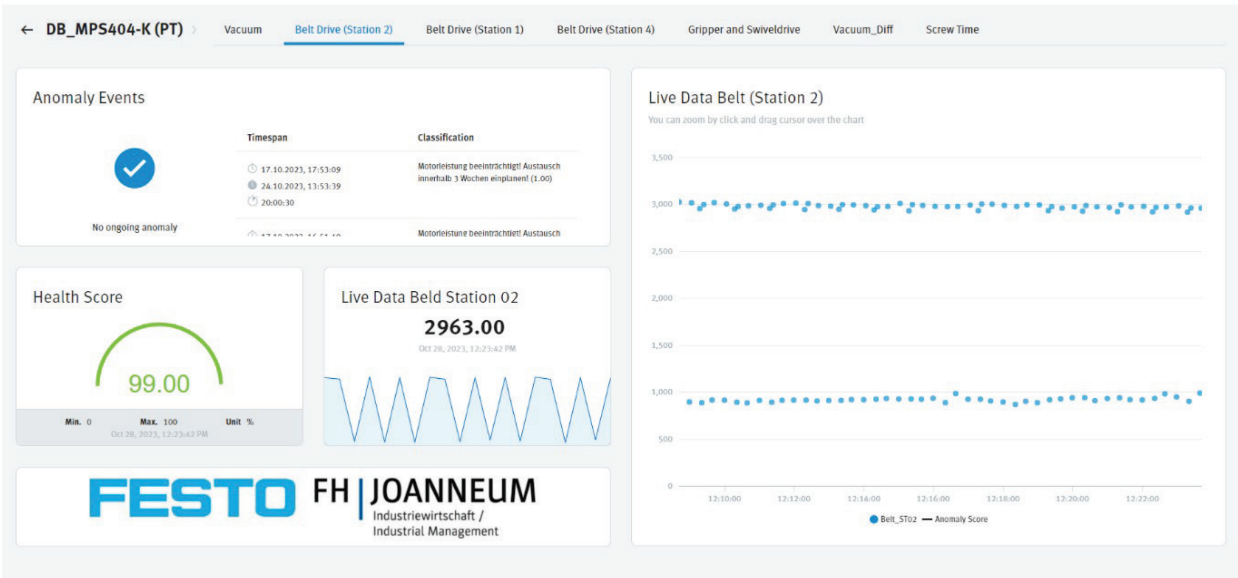


Figure 5 FESTO AX dashboard (use case 1 "Conveyor belt", optimal state)

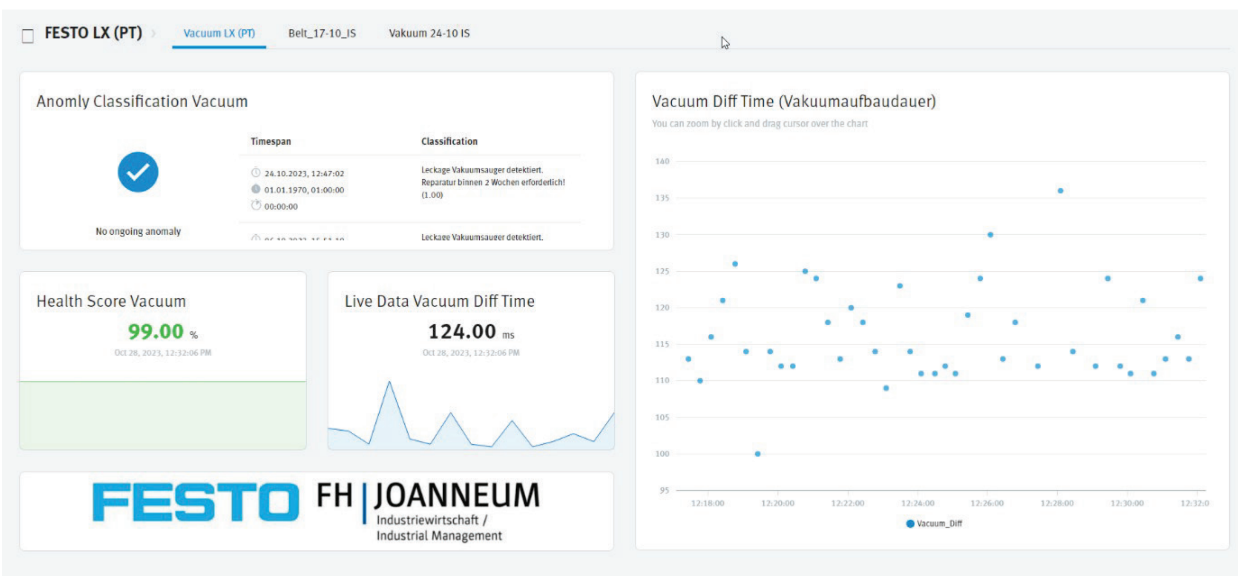


Figure 6 FESTO AX dashboard (use case 2 "Vacuum suction cup", optimal state)

In the second use case, the possible effects of vacuum suction cup attrition at the station 3 "Joining" were studied. The students prepared the AX tools in a similar procedure, this time choosing the suction cup as the feature of the analysis object. The problems of the suction cup were simulated by installing a throttle valve and a valve bypassing the throttle valve in line to the tube providing compressed air for the suction cup. By closing and then opening the bypassing valve, the students observed differences in the health score – first falling below 50%, then returning to normal values. The dashboard displaying the optimal situation is presented in Fig. 6.

Combined with theoretical information and control questions provided in corresponding nuggets, the students demonstrated high level of mastery of the predictive maintenance basics.

3 CONCLUSION

In the course of the proposed workshop, a concept for using a three-faceted industrial learning tool was described and tested. By combining a physical production station / learning factory, an AI interactive solution (AX) and a digital learning platform (LX), the high level of student involvement and skills transfer was reached.

To conduct the experiment, the physical station was enhanced by adding two components allowing for better simulation of anomalies and subsequent learning of predictive maintenance methods: a speed regulator of the conveyor belt and an air throttle valve for vacuum at the joining station. Both elements are intuitively understandable and easy to use. Later they were programmed to the digital twin model as well.

Based on the workshop, a learning path dedicated to predictive maintenance was created. The path, contains both theoretical part and a set of practical tasks regarding two use cases (slow conveyor belt and lack of air pressure), with corresponding questions for students. Both use cases are adjusted for a physical twin as well as for the digital one, allowing to combine the educational approaches. The workshop results further expanded the findings of authors presented in the 2021-2023 papers on business process modeling and learning factory simulation.

Main parameters like anomaly detection limits and model smoothness are set and justified, the results are visualized with dashboards and Excel tables.

4 FUTURE USE

By its nature, the learning concept described in this paper can be used not exclusively for teaching predictive analytics, but for other, broader educational purposes, for example:

- machine learning basics, especially k-means algorithm
- Node-RED coding
- Data visualization in Grafana or Snowflake
- data transfer by OPC-UA and MQTT
- Technical skills like computer-added design, rapid prototyping and managing electrical and pneumatical components using programmable controllers.

Possible future steps in using FESTO AX as well as positive experience based on the workshop are depicted in Fig. 7.

Regarding predictive maintenance itself, authors plan to use additional AI platforms like DataRobot to share ideas and concepts with other specialists in the field. Any dataset stored on AI platforms can be shared with the scientific community, furthering the authors' ideas. After more workshops and tests at FH Joanneum, the authors plan to cooperate with other technical schools and universities, in Austria as well as abroad.

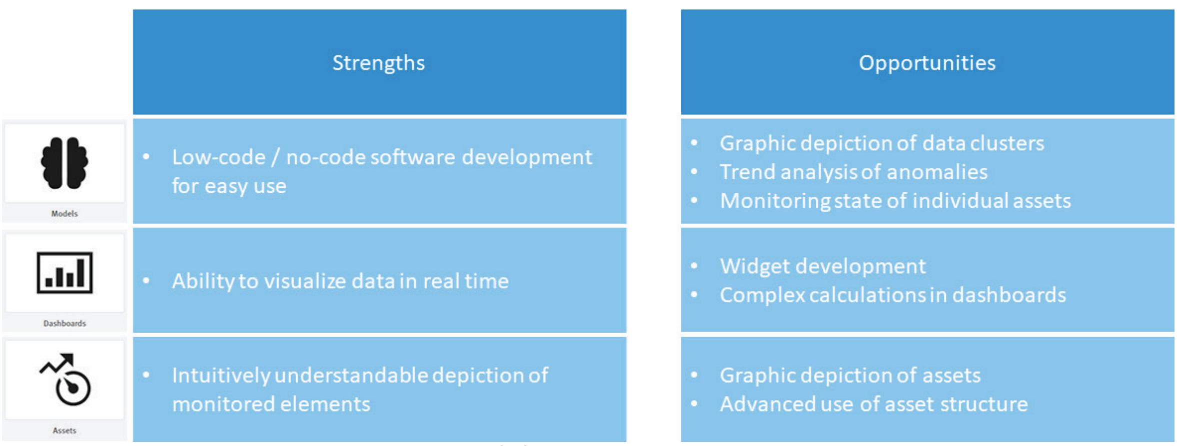


Figure 7 FESTO AX strengths and opportunities

5 REFERENCES

[1] Mason, J., Peoples, B. E. & Lee, J. (2020). Questioning the scope of AI standardization in learning, education, and training. *Journal of ICT Standardisation*, 107-122. https://doi.org/10.13052/jicts2245-800x.822

[2] Sundar, L. K. S., Muzik, O., Buvat, I., Bidaut, L. & Beyer, T. (2021). Potentials and caveats of AI in hybrid imaging. *Methods, 188*, 4-19. https://doi.org/10.1016/j.ymeth.2020.10.004

[3] Zhang, C., Patras, P. & Haddadi, H. (2019). Deep learning in mobile and wireless Networking: a survey. *IEEE Communications Surveys and Tutorials, 21*(3), 2224-2287. https://doi.org/10.1109/comst.2019.2904897

[4] Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures, 90*, 46-60. https://doi.org/10.1016/j.futures.2017.03.006

[5] Harari, Y. N. (2017). Reboot for the AI revolution. *Nature, 550*(7676), 324-327. https://doi.org/10.1038/550324a

[6] Hallak, J. & Azar, D. T. (2020). The AI revolution and how to prepare for it. *Translational Vision Science & Technology, 9*(2), 16. https://doi.org/10.1167/tvst.9.2.16

[7] Naqvi, W. M., Sundus, H., Mishra, G. V., Muthukrishnan, R. & Kandakurti, P. K. (2024). AI in Medical Education Curriculum: The Future of Healthcare Learning. *European Journal of Therapeutics*. https://doi.org/10.58600/eurjther1995

[8] Gruzdeva, M. L., Natalia, F., Smirnova, Z. V., Tsymbalov, S. D. & Garin, A. P. (2024). The use of artificial intelligence in teaching foreign languages. *Advances in science, technology & innovation*, 261-265. https://doi.org/10.1007/978-3-031- 49711-7_44

[9] Yetisensoy, O. & Karaduman, H. (2024). The effect of AI-powered chatbots in social studies education. *Education and*

Information Technologies. https://doi.org/10.1007/s10639-024-12485-6

[10] Fu, A. C. (2024). Investigation of recent advances related to AI in mathematics education. *Applied and Computational Engineering, 37*(1), 86-89. https://doi.org/10.54254/2755- 2721/37/20230476

[11] Cu, B. H. & Fujimoto, T. (2023). AI Education for Middle/ High School Level: A Proposal of a System that Supports Teachers to Design Their AI Lessons. Lecture notes in networks and systems, 12-20. https://doi.org/10.1007/978-3-031-27470-1_2

[12] Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education, 20*(1). https://doi.org/10.1186/s41239-023-00408-3

[13] Digel, S., Krause, T. & Biel, C. (2023). Enabling individualized and adaptive learning – the value of an AI-Based recommender system for users of adult and continuing education platforms. *Communications in computer and information science*, 797-803. https://doi.org/10.1007/978-3-031-36336-8_121

[14] Paschal, M. J. & Melly, I. K. (2023). Ethical guidelines on the use of AI in education. *Advances in educational technologies and instructional design book series*, 230-245. https://doi.org/10.4018/979-8-3693-0205-7.ch013

[15] Van Dinter, R., Tekinerdogan, B. & Çatal, Ç. (2022). Predictive maintenance using digital twins: A systematic literature review. *Information & Software Technology, 151*, 107008. https://doi.org/10.1016/j.infsof.2022.107008

[16] Achouch, M., Dimitrova, M., Ziane, K., Karganroudi, S. S., Dhouib, R., Ibrahim, H. & Adda, M. (2022). On Predictive Maintenance in Industry 4.0: Overview, models, and challenges. *Applied Sciences, 12*(16), 8081. https://doi.org/10.3390/app12168081

[17] Çakir, M., Güvenç, M. A. & Mistikoğlu, S. (2021). The experimental application of popular machine learning algorithms on predictive maintenance and the design of IIoT based condition monitoring system. *Computers & Industrial Engineering, 151*, 106948. https://doi.org/10.1016/j.cie.2020.106948

[18] Nunes, P., Santos, J. & Ribeiro, S. (2023). Challenges in predictive maintenance – A review. *CIRP Journal of Manufacturing Science and Technology, 40*, 53-67. https://doi.org/10.1016/j.cirpj.2022.11.004

[19] Daoudi, N., Zaki, S. & Mohamed, A. (2023). Machine Learning based Predictive Maintenance: review, challenges and workflow. *Lecture notes in networks and systems*, 71-88. https://doi.org/10.1007/978-3-031-43524-9_6

[20] Florian, E., Sgarbossa, F. & Zennaro, I. (2021). Machine learning-based predictive maintenance: A cost-oriented model for implementation. *International Journal of Production Economics, 236*, 108114. https://doi.org/10.1016/j.ijpe.2021.108114

[21] Savchenko, I., Ropin, H., Novotny, P. & Fleck, H. (2023). Data Science for business Engineers: Using a digital twin learning factory and cloud solutions for education. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4479483

Authors' contacts:

Ihor Savchenko, BSc
(Corresponding author)
FH Joanneum,
Werk-VI-Straße 46, 8605 Kapfenberg, Austria
+4367762527133, igor.savchenko@tavr.at

Herbert Fleck, DI
HITEL GmbH,
Redtenbachergasse 25/1/12, 1160 Vienna, Austria
herbert.fleck@hitel.at

Peter Novotny
FH Joanneum,
Werk-VI-Straße 46, 8605 Kapfenberg, Austria
peter.novotny@tavr.at

Helmut Ropin, DI
FH Joanneum,
Werk-VI-Straße 46, 8605 Kapfenberg, Austria
helmut.ropin@fh-joanneum.at

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240518201700

Subject review
Received: 2024-05-18, Accepted: 2024-05-22

Improving Manufacturing Processes through Artificial Intelligence - Example of Printed Circuit Board Manufacturing

Jiri Tupa*, Andrea Benesova, Frantisek Steiner, Tomas Rericha

Abstract: The advent of Artificial Intelligence (AI) in manufacturing has heralded a new era of industrial revolution, characterised by unprecedented efficiency, productivity, and innovation. This critical review delves into the application of AI technologies in the manufacturing sector, scrutinising their impact on process enhancement and addressing the spectrum of opportunities and challenges they present. By thoroughly analysing recent studies, industry reports, and case examples, this paper outlines the transformative potential of AI in various manufacturing domains, including predictive maintenance, supply chain optimisation, quality control, and intelligent manufacturing. However, the paper does not shy away from discussing the critical challenges facing the deployment of AI in manufacturing. These include technical limitations, data privacy and security concerns, the need for substantial investment, and the socio-economic implications of workforce displacement and skill gaps. Concluding with a forward-looking perspective, the review suggests practical strategies for overcoming these hurdles, such as fostering public-private partnerships, investing in AI literacy and training, and adopting ethical guidelines for AI use.

Keywords: artificial intelligence; lean management; predictive maintenance; printed circuit board; process improvement; quality management

1 INTRODUCTION

Artificial Intelligence (AI) represents a significant technological opportunity in manufacturing, offering a wide range of applications that lead to increased efficiency, productivity, and innovation. By utilising AI in manufacturing, businesses can automate complex processes, optimise production lines, reduce manufacturing costs, and minimise the duration of manufacturing processes. AI, employing intelligent algorithms, allows machines to learn from data, predict production errors, conduct predictive maintenance, and enhance the overall quality of products. AI also enables the personalisation of products based on individual customer preferences and the flexible response to changing market conditions. With these capabilities, AI opens the door to significant innovations in the manufacturing sector, supports sustainability, and improves the competitiveness of companies in the global market.

On the other hand, applying AI in manufacturing can bring certain risks that may negatively impact manufacturing processes. Therefore, this article aims to critically summarise the advantages and disadvantages of using AI to improve manufacturing processes. Specifically, attention will be focused on these areas:

- Lean manufacturing
- Predictive maintenance
- Internal logistics and SCM
- Production planning
- Quality management.

Based on the previous introduction, the paper tries to answer this research question:

"How can artificial intelligence technologies be effectively integrated into existing manufacturing processes to enhance efficiency, reduce costs, and improve product quality while also addressing the challenges of implementation and workforce adaptation?"

This research question is designed to explore the potential benefits of AI in manufacturing (such as efficiency and quality improvements) and the practical aspects of integrating these technologies into current systems. It considers the economic (cost reduction) and human factors (workforce adaptation), ensuring a comprehensive review of AI's impact on manufacturing. This broad yet focused approach encourages an exploration of multiple dimensions - technological, economic, and human resource management - making it suitable for a critical review.

2 RESEARCH METHODOLOGY

The research is based on a combination of methods research framework, integrating both qualitative and quantitative analyses to critically review the role of Artificial Intelligence (AI) in manufacturing processes. This approach allows for a holistic understanding of AI's transformative potential in manufacturing, encompassing efficiency, productivity, innovation, and the challenges therein.

Data Collection: Data collection involved a comprehensive literature review, focusing on recent studies, industry reports, and case examples. The selection criteria emphasised relevance to AI applications in manufacturing, including predictive maintenance, supply chain optimisation, quality control, and intelligent manufacturing solutions. This approach ensured a robust foundation of evidence for analysis.

Analysis Method: The study utilised a thematic analysis to identify common themes and trends within the literature, facilitating an in-depth exploration of AI's contributions and challenges in manufacturing.

Pre-implementation Study Approach: This kind of study is essential in ensuring the successful adoption and integration of new technologies, methods, or processes within an organization. This study helps identify potential challenges, estimate benefits, and refine implementation strategies before full-scale deployment. A practical implementation of this study was conducted through Printed Circuit Board (PCB) manufacturing and assembly. This case study provided concrete example of AI's potential improvements in process efficiency, predictive maintenance, and quality management within a specific manufacturing domain.

SWOT Analysis: The research methodology incorporated a SWOT analysis as part of results pre-implementation study. The aim is too comprehensively critically evaluate the strengths, weaknesses, opportunities, and threats associated with AI integration in manufacturing. This analysis facilitated a balanced view of AI's potential benefits and the challenges needing strategic consideration.

Conclusions and Recommendations: Findings from the literature review, case study, and analyses conducted, the study concludes with actionable recommendations for industry practitioners.

This methodology establishes a comprehensive framework for exploring AI's role in improving manufacturing processes, providing a robust basis for the paper's critical review.

3 LITERATURE REVIEW

Lean management and AI are two important concepts in the manufacturing industry nowadays. Lean manufacturing reduces waste and improves productivity [1]. It has been shown to increase plant efficiency and decrease processing times in major manufacturing operations [2]. On the other hand, AI has the potential to improve manufacturing efficiency, productivity, and sustainability. It can be used in predictive maintenance, quality assurance, and process optimisation [3]. Combining lean management and AI can lead to significant cost and efficiency benefits in manufacturing [4].

AI technologies can be used as additional tools in the lean manufacturing toolkit, enhancing the effect. By using AI, manufacturing enterprises can collect, analyse, and structure production information, improve the quality of products, and increase overall efficiency. Implementing AI in lean manufacturing can help optimise production, reduce losses, and increase working productivity.

Predictive maintenance (PdM) is a key strategy in manufacturing, aiming to reduce costs and improve product quality. The integration of AI into the maintenance process has achieved the most significant progress in predicting the condition of rotating machinery based on the monitoring of vibration parameters such as vibration, acceleration, and displacement. On the other hand technological advancements such as Big Data and the Internet of Things have made PdM more effective. Machine learning models, such as Gradient Boosting (GB) and Support Vector Machine (SVM), have been implemented for PdM [5]. These models have achieved high recall and accuracy, demonstrating their effectiveness [6].

Integrating AI and machine learning in manufacturing can improve efficiency, productivity, and sustainability [3]. Challenges in using AI in manufacturing include data acquisition, security risks, and lack of trust or understanding [7]. However, AI has the potential to be extremely helpful in applications such as predictive maintenance, quality assurance, and process optimisation [7]. Innovative technologies, such as Unsupervised Learning (UL) algorithms, can automate specific parts of industrial processes, reducing costs and human error. Combining AI and predictive maintenance can enhance manufacturing performance and reduce downtime [8, 9].

For instance, AI has been utilised to develop a distributed system for predictive maintenance across manufacturing plants, significantly enhancing the response time of monitoring systems by processing data near sensors and reducing the need for central data transmission [10]. This approach aligns with the broader vision of IoT-based predictive maintenance, leveraging Big Data Analytics and Machine Learning to foster intelligent manufacturing practices that are more efficient and cost-effective [11].

In the context of Industry 4.0, machine learning, a subset of AI, plays a crucial role in predictive maintenance strategies to monitor industrial equipment's health status. These strategies are designed to minimise downtime, enhance utilisation rates, and prolong the useful life of machinery components [12]. Furthermore, AI and IoT technologies have been combined in low-cost frameworks for anomaly detection, offering a pragmatic solution to predictive maintenance challenges in real-world industrial settings, thereby improving maintenance efficiency and equipment lifespan [13].

Predictive maintenance powered by AI aims to enable real-time maintenance interventions and strives to lower operational costs, diminish downtime, and enhance production quality, contributing to manufacturing excellence [14]. This excellence is further bolstered by adopting cyber-physical systems in manufacturing, which utilise big data to enable a cost-oriented dynamic predictive maintenance strategy, offering a more economical alternative to conventional preventive maintenance methods [15].

The application of AI in predictive maintenance extends to developing business models in manufacturing, focusing on case studies that highlight the practical benefits of predictive maintenance technologies. Integrating deep learning and augmented reality into predictive maintenance further exemplifies AI's capability to enhance maintenance operations, making them smarter and more efficient, thereby paving the way for the future of IoT-enabled manufacturing [16].

In summary, AI's integration into predictive maintenance revolutionises manufacturing by offering smarter, more efficient, and cost-effective solutions. This technological advancement supports the broader goals of Industry 4.0 by enhancing production quality, reducing downtime, and ultimately leading to more sustainable manufacturing practices.

Supply Chain Management (SCM) in manufacturing is being transformed by integrating artificial intelligence (AI) technologies. AI, including machine learning (ML), is being used to improve various aspects of the supply chain, such as risk identification and management, material planning and forecasting, and optimisation of production processes [17, 18].

However, there are challenges to successfully integrating AI into supply chain management, including understanding and implementing responsible and ethical AI practices. Organisations must make an economic case for AI adoption, develop an implementation plan, and manage the coordination between humans and AI systems [19]. The use of AI in supply chain management offers the potential for increased efficiency and improved decision-making, but careful consideration must be given to the challenges and implications of its integration [20, 21].

Internal logistics in manufacturing can be improved by applying Lean Management tools, such as supermarkets and electric logistic trains [22]. These tools help eliminate waste and improve the flow of materials within a company, ultimately enhancing production processes [23]. Additionally, computer simulation and routing plans can optimise the transportation of goods and raw materials within a production plant [24].

Analysing and adjusting the internal logistics system increases factory efficiency and productivity [25]. Furthermore, artificial intelligence (AI) is playing an increasing role in the automotive industry's logistical aspects of production sites [26]. AI can assess disruption risks caused by natural disasters or social actions and propose countermeasures to ensure material availability. Overall, combining Lean Management tools and AI can significantly improve internal logistics in manufacturing [27].

Manufacturing planning is a crucial process in the manufacturing industry. It involves production planning, scheduling, and coordination among different locations. Artificial intelligence (AI) has been applied to improve manufacturing efficiency and productivity. AI can be used for intelligent production scheduling, considering project constraints, temporal and spatial characteristics, and production part distribution [28].

AI can also be used to design and operate manufacturing systems, including system layout, capacity planning, and control of material and information flows [29]. Additionally, AI planning approaches have been employed to simplify the process of production planning and scheduling, providing valuable guidance for real production manufacture [10, 3]. Despite challenges such as data acquisition, security risks, and trust issues, AI has the potential to bring significant cost and efficiency benefits to manufacturing, especially when combined with the ability to capture large amounts of data [30, 10].

The area of ***quality management*** can be enhanced by integrating artificial intelligence (AI) and machine learning (ML) techniques. These technologies can improve quality assurance processes by detecting and analysing deviations from quality specifications, as well as predicting and preventing problems at an early stage [30-33]. The application of AI and ML in quality control is part of the emerging field of Quality 4.0, which aims to drive innovation in the manufacturing industry [34].

Traditional quality control tools, such as the Six Sigma methodology, may have limitations in handling the complexity and dynamics of modern manufacturing processes, making AI and ML valuable tools for addressing these challenges [35]. AI-driven data science methods, such as machine learning, can identify complex relationships in large amounts of data, contributing to process improvement and failure management [36]. Overall, the integration of AI into quality management systems can lead to more efficient and effective manufacturing processes.

4 PRE-IMPLEMENTATION STUDY

This study focus on PCB manufacturing and assembly. A pre-implementation study is an essential step in ensuring the successful adoption and integration of new technologies, methods, or processes within an organization. This type of study helps identify potential challenges, estimate benefits, and refine implementation strategies before full-scale deployment. The results of the study are

The aim of the study is to evaluate and demonstrate the possibilities of using AI tools in a model example of printed circuit board (PCB) manufacturing. The reason is that this sector can be used as an excellent example in terms of the maturity of automation and robotics, including possible improvements. At the same time, this market is poised for future growth according to Mordor Intelligence's report and market analysis [37]. Printed Circuit Board (PCB) market highlights its projected growth and the current trends shaping the industry from 2019 to 2029.

The estimated market size is expected to increase from USD 76.12 billion in 2024 to USD 93.87 billion by 2029, demonstrating a growth rate of 4.28% CAGR (Compound Annual Growth Rate). This growth can be attributed to the indispensable role played by PCBs in the contemporary electronics sector, which is driven by advancements in technologies such as 5G, IoT, and AI, despite a temporary decline in consumer electronics demand by the end of FY 2023. The report underscores the significance of PCBs in the miniaturisation of electronic components, thereby facilitating the development of portable, wearable, and more resilient consumer electronics devices.

It underscores the substantial market share held by consumer electronics, which leverage PCBs to achieve compactness and efficiency. The analysis also delves into the environmental concerns associated with PCB disposal and the impact of the COVID-19 pandemic on the semiconductor market. With Asia-Pacific expected to witness noteworthy growth, the report provides detailed information about the dominance of Chinese manufacturers, who hold a 54.76% market share in the region. Furthermore, it outlines the strategic positions of Taiwan and South Korea in the global PCB industry. [37]

4. 1 PCB Manufacturing Process and Optimisation

The manufacturing and assembly of PCBs involve processes that determine the functionality and reliability of the final product. This part deals with PCB manufacturing and assembly processes [38].

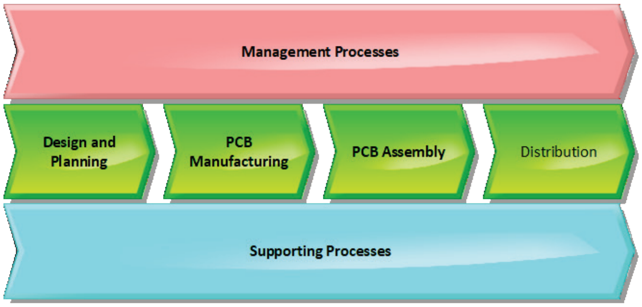


Figure 1 Process model of PCB Manufacturing and Assembly

Design and Planning: The PCB manufacturing process begins with design and layout, typically using CAD (Computer-Aided Design) software. The results are important for production planning.

This step is also crucial for determining the circuit functionality component layout and ensuring electrical efficiency and reliability.

Photolithography: Photolithography involves transferring the circuit design onto the PCB using a photosensitive film. This step is pivotal for pattern definition and precision.

Etching: The etching removes unwanted copper from the board, leaving the circuit pattern behind. Chemical or plasma etching techniques are commonly employed.

Drilling: Drilling creates vias and mounting holes for through-hole components. Precision is key to ensuring the functionality of multi-layer boards.

Plating and Copper Deposition: Plating enhances the electrical connectivity between layers through vias. Copper deposition adds a thin copper layer on the surface and within the drilled holes.

Solder Mask Application: The solder mask, usually green, protects the circuit from oxidation and prevents solder bridging during component soldering.

Silkscreen Printing: Silkscreen printing adds labels and component identifiers, facilitating manual assembly and inspection.

4.1.1 PCB Assembly Process

Solder Paste Stencilling: Solder paste is applied to the board in areas designated for component attachment. This step requires precision to ensure proper soldering and electrical connection.

Components Picking: Components are placed on the board using automated equipment, aligning with the solder paste deposits.

Reflow Soldering: The board passes through a reflow oven, where controlled heating melts the solder, securely attaching components to the board.

Inspection and Quality Control: Various inspection techniques, including AOI (Automated Optical Inspection), are employed to ensure the accuracy and quality of the assembly.

Through-Hole Component Insertion (if applicable): For boards requiring through-hole components, this manual or automated process inserts and solders components to the board.

Final Inspection and Functional Testing: The completed board undergoes final inspection and functional testing to ensure it meets all specifications and functional requirements.

4.2 Opportunities for Implementing AI tools for Improving Processes in Manufacturing and PCB Assembly

Fig. 2 presents a designed model of AI tools in the context of potential improvements in PCB manufacturing and assembly. The design and structure is based on a process mapping approach, dividing processes into control, primary and support processes. The management processes consist mainly of processes that are implemented by the management of the company, and AI tools can be an important support for decision-making here, especially in the creation of analyses and predictions of future developments.

The core processes are the actual process of designing, manufacturing, and assembling the PCBs, including the provision of lo. Within these processes, sub-processes are shown where AI tools can be implemented to improve the processes themselves. The last level consists of support processes, in particular, processes related to predictive maintenance and product planning.

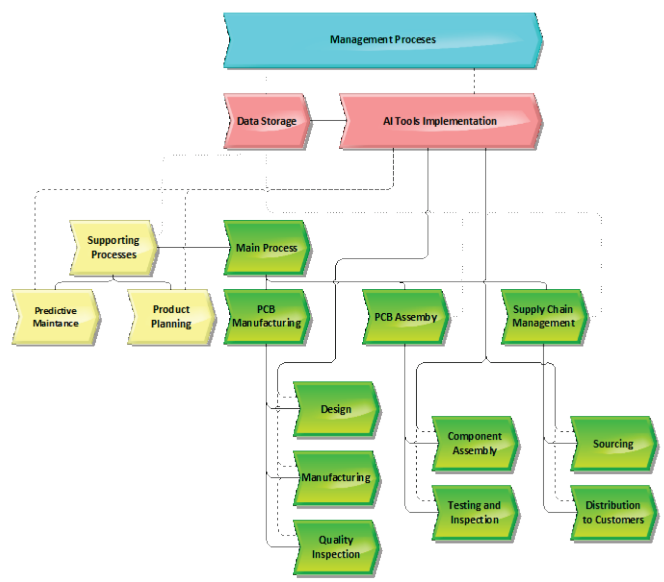


Figure 2 PCB Designed Manufacturing and Assembly with AI Tools

Specifically, the following tools can be used within these processes:

Machine Learning and Deep Learning: Machine learning algorithms can analyse data from previous designs and production cycles to identify the most efficient configurations and material use. Furthermore, deep learning models can predict equipment failures and maintenance needs by analysing patterns in operational data.

AI-driven Robotics: Robots equipped with AI can place components on PCBs with high precision and speed, adapting to different sizes and types of PCBs. At the same time, advanced robotic systems allowing for visual inspection using convolutional neural networks can detect defects and inconsistencies on PCBs during and after manufacturing.

Natural Language Processing (NLP): AI tools utilising NLP can analyse large volumes of textual data from internal reports, emails, and documents to provide management with summaries and decision-making materials. Simultaneously, AI tools can automate management's communication with suppliers, simplify ordering, and track inventory.

Predictive Analytics and Simulation: AI can simulate the entire manufacturing process, identify bottlenecks, and predict the impact of changes in the process. This also relates to possible process improvements, where algorithms can analyse historical data and demand forecasts to optimise production planning, improve inventory management, and minimise delivery times.

Image Recognition and Computer Vision: Computer vision systems can continuously monitor and analyse products on the production line, instantly identifying defects and ensuring that all PCBs meet the required quality standards.

Manufacturing Data Management Systems; AI tools can aggregate and analyse data from various parts of the manufacturing process, allowing for quick adjustments and process optimisation.

These are just a few examples of selected AI tools that can be implemented in the manufacturing and PCB assembly process. Integrating these technologies requires careful planning and consideration of each manufacturing operation's specific needs and goals.

A key element is the development of a suitable data storage. Data storage is crucial for AI tools for several reasons:

Source for learning and training: AI and machine learning (ML) models "learn" from vast volumes of data. The more quality data available for training, the more accurate and reliable they can be in their predictions and decisions. Data storage provides a base for storing these data, allows easy access for learning processes, and enables iterative improvement of models.

Support for real-time analysis: Many AI applications, such as decision support, real-time monitoring, and automation, require immediate access to data. Powerful data storage enables fast analysis and real-time data processing, essential for applications requiring immediate responses or updates.

Availability and scalability: AI systems often work with exponentially growing data volumes. Data storage must be able to scale to accommodate growing needs without losing performance or availability. Efficient storage ensures that AI systems have constant access to data regardless of their volume.

Data security and protection: Data used by AI systems may contain sensitive information. Secure data storage ensures protection against unauthorised access and attacks, which is key for complying with legal and ethical standards. In addition, data security mechanisms such as encryption and tokenisation play an important role in protecting data during storage and transmission.

Integration and interoperability: AI applications often work with data from various sources and must integrate data from different systems and technologies. Flexible and interoperable data storage allows easy integration and processing of data from various sources, which is key for comprehensive analyses and full utilisation of available data.

In the context of the demanding requirements of AI and ML, choosing the right data storage solution becomes a critical decision. An effective solution must support fast analysis of large data volumes, scalability for future growth, and provide advanced data security and protection features.

4.3 Discussion of Results

These tools can be introduced to improve the processes mentioned above. However, they are very costly technologies, and their implementation may bring certain negatives on the other side. Although the benefits of AI are undeniable, there are potential negatives that may arise from its use in this area. This text will examine in detail what these negatives may be and their impact on the industry, workforce, and society. The negatives detected in the literature mentioned in Chapter 3 are detailed below.

Loss of Jobs and Impact on Employees: One of the most discussed negatives of using AI is the potential loss of jobs. In the PCB manufacturing area, automation could mean that many traditional manual operations will be replaced by machines, which could significantly reduce the need for human labour. This trend could negatively impact individuals who may lose their jobs and entire communities where the manufacturing sector constitutes a considerable part of the economy.

Ethical Issues: The use of AI in PCB manufacturing processes also raises ethical issues, particularly concerning algorithm bias. AI systems are created and trained by humans to include human biases unconsciously. This bias can lead to inefficient decisions or discriminatory practices within the manufacturing process, negatively impacting the quality of production.

Loss of Personal Skills and Practical Experience: Automation and the integration of AI into PCB manufacturing could lead to the loss of human skills and practical experience. While AI can streamline processes, it can also cause us to lose valuable manual skills and practical experiences passed down and improved upon for generations. This loss can have long-term consequences for innovation and the ability to solve unexpected problems that may arise in the manufacturing process.

Increase in Social and Economic Inequality: Automation and the use of AI can also contribute to increasing social and economic disparities. Companies with access to the latest AI technologies can gain a significant competitive advantage, while smaller businesses may fall behind. This division can lead to a concentration of power and wealth in the hands of a few, while small and medium-sized companies may struggle to maintain competitiveness.

Improving manufacturing processes is also associated with introducing various methods and tools based on finding new or innovative solutions. These methods and tools also help address problems in the manufacturing cycle.

A key element here is people - individual teams who initiate these changes. Even though artificial intelligence tools can significantly assist in improving processes, it must be pointed out that the focus must always be on individuals. People provide creative thinking, empathy, and the ability to adapt, which AI and automation cannot fully replace. While AI can provide recommendations based on data analysis, final decisions often require human judgment that considers context, ethics, and long-term impact.

Effective implementation of changes also requires coordination between different departments and levels of management, which depends on human interactions, communication, and collaboration. While technology can provide tools and methods for improvement, the human factor plays an irreplaceable role in innovation and implementing changes in manufacturing processes. The engagement, experience, and creativity of people form the pillars for the successful improvement of manufacturing processes.

4.4 Summary

Through SWOT analysis, the conclusions and findings of the pre-implementation study are summarised. The SWOT analysis was chosen because it enables the identification of the Strengths, Weaknesses, Opportunities, and Threats associated with implementing AI in a manufacturing environment.

Strengths:

- Increased efficiency and productivity
- Potential for predictive maintenance
- Supply chain optimisation
- Quality control improvements.

Weaknesses:

- High investment and operational costs
- Acceptance of new technologies by workers
- Low initial knowledge of workers
- Complexity of implementation
- Dependence on data.

Opportunities:

- Innovation in products and services
- Personalisation of customer requirements
- Expansion into new markets
- Improvement in sustainability.

Threats:

- Risk of technology obsolescence
- Cybersecurity
- Regulatory and ethical challenges
- Dependence on technology suppliers.

In conclusion, introducing AI tools into manufacturing processes offers many significant opportunities for improving efficiency, quality, and innovation. However, it is also accompanied by challenges that require careful consideration and strategic planning.

5 CONCLUSION

This paper comprehensively examines the integration of Artificial Intelligence (AI) for process improvement in the manufacturing sector. It explores the potential and challenges of deploying AI technologies for process improvement in the manufacturing area. The authors analyse recent literature to emphasise AI's contributions to enhancing operational efficiency, productivity, and innovation in manufacturing. However, they also address critical challenges industries face when adopting AI.

The core of the paper is dedicated to an extensive literature review, which highlights the synergy between lean manufacturing and AI, the role of AI in predictive maintenance, the impact of AI on supply chain management, advancements in manufacturing planning through AI, and the enhancement of quality management with AI technologies.

A pre-implementation study on the production and optimisation of Printed Circuit Boards (PCBs) showcases the practical application and potential improvements AI tools can be implemented to manufacturing processes.

This paper critically supports the role AI should play in manufacturing and rallies for a future whereby AI and human expertise shall conjoin, driving the next industrial revolution.

As part of further research, we believe it is appropriate to look at the interaction between AI and human resource development. The reason is that human resources play, and will continue to play, an important role in improving productive processes.

6 REFERENCES

[1] Sharma, A. K., Pinca-Bretotean, C., & Sharma, S. (2023). Artificial intelligence in lean manufacturing paradigm: A review. *E3S Web of Conferences, 391*, 01163. https://doi.org/10.1051/e3sconf/202339101163

[2] Kuizheva, S. K., Zadorozhnaya, L., Ovsyannikova, T. A., & Zarubin, V. I. (2021). The tasks of using artificial intelligence in lean manufacturing. *Novie tehnologii/New technologies, 17*(6), 106-115. (in Rusian) https://doi.org/10.47370/2072-0920-2021-17-6-106-115

[3] Plathottam, S. J., & Iloeje, C. O. (2023). A review of artificial intelligence applications in manufacturing operations. *J. Adv. Manuf. Process., 5*(3), e10159. https://doi.org/10.1002/amp2.10159

[4] Jobin, M. (2021). Intelligent Prediction Model: Optimized Neural Network for Lean Manufacturing Technology. *J. Eng. Res., 11*(2A). https://doi.org/10.36909/jer.12747

[5] Netisopakul, P., & Phumee, N. (2022). AI-Enhanced Predictive Maintenance in Manufacturing Processes. *The 22nd International Conference on Control, Automation and Systems (ICCAS2022)*, Jeju, Republic of Korea, 1107-1112, https://doi.org/10.23919/ICCAS55662.2022.10003774

[6] Mota, B., Faria, P., & Ramos, C. (2023). Predictive Maintenance for Maintenance-Effective Manufacturing Using Machine Learning Approaches. In: García Bringas, P., et al. *17th International Conference on Soft Computing Models in*

Industrial and Environmental Applications (SOCO 2022). Lecture Notes in Networks and Systems, vol 531. Springer, Cham. https://doi.org/10.1007/978-3-031-18050-7_2

[7] Osmëni, T., & Ali, M. S. M. (2022). Contemporary Generation: Artificial Intelligence Contribution to Manufacturing. *International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA2022)*, Skopje, North Macedonia, 31-35. https://doi.org/10.1109/CoNTESA57046.2022.10011233

[8] Murugiah, P., Muthuramalingam, A. & Anandamurugan, S., (2023). A design of predictive manufacturing system in IoT-assisted Industry 4.0 using heuristic-derived deep learning. *Int. J. Commun. Syst., 36*(5). https://doi.org/10.1002/dac.5432

[9] Meddaoui, A., Hain, M., & Hachmoud, A. (2023). The benefits of predictive maintenance in manufacturing excellence: a case study to establish reliable methods for predicting failures. *Int J Adv Manuf Technol, 128*, 3685-3690. https://doi.org/10.1007/s00170-023-12086-6

[10] Liu, J., Liang, R., & Xian, J. (2022). An AI Planning Approach to Factory Production Planning and Scheduling. *International Conference on Machine Learning and Knowledge Engineering (MLKE2022)*, Guilin, China, 110-114. https://doi.org/10.1109/MLKE55170.2022.00027

[11] Nangia, S., Makkar, S., & Hassan, R. (2020). IoT based Predictive Maintenance in Manufacturing Sector. *Proceedings of the International Conference on Innovative Computing & Communications (ICICC2020)*. https://doi.org/10.2139/ssrn.3563559

[12] Cinar, E., Kalay, S., & Sariçiçek, İ. (2022). A Predictive Maintenance System Design and Implementation for Intelligent Manufacturing. *Machines, 10*(11), 1006. https://doi.org/10.3390/machines10111006

[13] Filios, G., Karatzas, S., Krousarlis, M., Nikoletseas, S., Panagiotou, S. H., & Spirakis, P. (2023). Realizing Predictive Maintenance in Production Machinery through Low-Cost IIoT Framework and Anomaly Detection: A Case Study in a Real-World Manufacturing Environment. *The 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT2023)*, 338-346. https://doi.org/10.1109/DCOSS-IoT58021.2023.00062

[14] Meddaoui, A., Hain, M., & Hachmoud, A. (2023). The benefits of predictive maintenance in manufacturing excellence: a case study to establish reliable methods for predicting failures. *Int. J. Adv. Manuf. Technol., 128*(7-8), 3685-3690. https://doi.org/10.1007/s00170-023-12086-6

[15] He, Y., Han, X., Gu, C., & Chen, Z. (2018). Cost-oriented predictive maintenance based on mission reliability state for cyber manufacturing systems. Adv. Mech. Eng., 10(1), p. 168781401775146. https://doi.org/10.1177/1687814017751467

[16] Liu, C. et al. (2022). Probing an intelligent predictive maintenance approach with deep learning and augmented reality for machine tools in IoT-enabled manufacturing. *Robot. Comput. Integr. Manuf., 77*, p. 102357. https://doi.org/10.1016/j.rcim.2022.102357

[17] Ghabak, V., & Seetharaman, A. (2023). Integration of Machine Learning in Agile Supply Chain Management. *The 15th International Conference on Computer and Automation Engineering (ICCAE2023)*, Sydney, Australia, 6-12. https://doi.org/10.1109/ICCAE56788.2023.10111340

[18] Brintrup, A., Baryannis, G., Tiwari, A., Ratchev, S., Martinez-Arellano, G., & Singh, J. (2023). Trustworthy, responsible, ethical AI in manufacturing and supply chains: synthesis and emerging research questions. *arXiv*, arXiv:2305.11581. https://doi.org/10.48550/arxiv.2305.11581

[19] Hendriksen, C. (2023). Artificial intelligence for supply chain management: Disruptive innovation or innovative disruption? *J. Supply Chain Manag., 59*(3). https://doi.org/10.1111/jscm.12304

[20] Dijmărescu, E. (2023). AI Trends: Salient Aspects for the Manufacturing Sector and Its Global Supply Chain. In: Pamfilie, R., Dinu, V., Vasiliu, C., Pleșea, D., & Tăchiciu, L. eds. *The 9th BASIQ International Conference on New Trends in Sustainable Business and Consumption*. Constanța, Romania, 8-10 June 2023. Bucharest: ASE, 168-175. https://doi.org/10.24818/BASIQ/2023/09/053

[21] Rana, J., & Daultani, Y. (2022). Mapping the Role and Impact of Artificial Intelligence and Machine Learning Applications in Supply Chain Digital Transformation: A Bibliometric Analysis. *Oper. Manag. Res.*, 1-26. https://doi.org/10.1007/s12063-022-00335-y

[22] Evangeliou, N., Stamatis, G., Bravos, G., Plorin, D., & Stark, D. (2022). AI for Inbound Logistics Optimisation in Automotive Industry. In Vermesan, O. ed. *Artificial Intelligence for Digitising Industry – Applications*, 1st Edition. River Publishers, New York, 11-19. https://doi.org/10.1201/9781003337232-3

[23] Dotoli, M., Fanti, M. P., Iacobellis, G., & Rotunno, G. (2014). An integrated technique for the internal logistics analysis and management in discrete manufacturing systems. *Int. J. Comput. Integr. Manuf., 27*(2), 165-180. https://doi.org/10.1080/0951192X.2013.802370

[24] Karwasz, A., & Skuza, D. (2019). Improving Internal Transport. In: Machado, J., Soares, F. & Veiga, G. (eds) *Innovation, Engineering and Entrepreneurship. HELIX 2018. Lecture Notes in Electrical Engineering, vol 505*. Springer, Cham. https://doi.org/10.1007/978-3-319-91334-6_53

[25] Phruksaphanrat, B., Sinsomsak, C., & Khaoyangyuen, P. (2020). Simulation Based AGVs Routing Analysis for Internal Logistics in Automotive Manufacturing. *Int. J. Simul. Syst. Sci. Technol., 21*. https://doi.org/10.5013/IJSSST.a.21.04.05

[26] Dubromelle, Y., Ounnar, F., & Pujo, P. (2012). Internal logistics as a service in a holonic and isoarchic control model. *CCCA12*, Marseille, France, 1-6. https://doi.org/10.1109/CCCA.2012.6417887

[27] Kavka, L., Dočkalíková, I., Čujan, Z., & Fedorko, G. (2020). Technological and Economic Analysis of Logistic Activities in Interior Parts Manufacturing. *Adv. Sci. Technol. Res. J., 14*(3), 204-212. https://doi.org/10.12913/22998624/122062

[28] Shi, S. et al. (2022). Research on Intelligent Planning and Scheduling Method Based on Manufacturing Big Data. *The 7th International Conference on Big Data Analytics (ICBDA2022)*, Guangzhou, China, 54-57. https://doi.org/10.1109/ICBDA55095.2022.9760327

[29] Chryssolouris, G., Alexopoulos, K., & Arkouli, Z. (2023). Artificial Intelligence in Manufacturing Systems. In: *A Perspective on Artificial Intelligence in Manufacturing. Studies in Systems, Decision and Control, vol 436*. Springer, Cham. https://doi.org/10.1007/978-3-031-21828-6_4

[30] Ding, J., Chen, M., Wang, T., Zhou, J., Fu, X., & Li, K. (2023). A Survey of AI-Enabled Dynamic Manufacturing Scheduling: From Directed Heuristics to Autonomous Learning. *ACM Computing Surveys, 55*(14), Article No. 307, 1-36. https://doi.org/10.1145/3590163

[31] Escobar, C. A., & Morales-Menendez, R. (2023). The decay of Six Sigma and the rise of Quality 4.0 in manufacturing innovation. *Quality Engineering, 36*(2), 316-335. https://doi.org/10.1080/08982112.2023.2206679

[32] Caglar, T. R., & Jochem, R. (2023). A System Approach for Creating Employee-Oriented Quality Control Loops in Production for Smart Failure Management System in SMEs.

Intelligent Human Systems Integration (IHSI 2023): Integrating People and Intelligent Systems. https://doi.org/10.54941/ahfe1002835

[33] Svigaris, Z. (2022). Management of Unplanned Changes in Production Processes: AI Control Systems. *Int. J. Artif. Intell. Appl., 13*(4), 89-98. https://doi.org/10.5121/ijaia.2022.13407

[34] Aichele, A., Sauer, A., Grundstein, S., & Schöck, J. (2022). Qualitätsmanagement- Methoden für Produkte mit KI-Funktionalitäten/Quality management methods in the development of mechatronic systems with AI functionalities - QM methods for mechatronic systems with AI. *wt Werkstattstech online, 112*(11-12), 802-806. https://doi.org/10.37544/1436-4980-2022-11-12-76

[35] Arora, A., & Gupta, R. K. (2022). A Comparative Study on Application of Artificial Intelligence for Quality Assurance in Manufacturing. *The 4th International Conference on Inventive Research in Computing Applications (ICIRCA2022)*, Coimbatore, India, 1200-1206. https://doi.org/10.1109/ICIRCA54612.2022.9985522

[36] Sundaram, S., & Zeid, A. (2023). Artificial Intelligence-Based Smart Quality Inspection for Manufacturing. *Micromachines, 14*(3), p. 570. https://doi.org/10.3390/mi14030570

[37] PCB Market Size & Share Analysis - Growth Trends & Forecasts (2024 - 2029) (2024). https://www.mordorintelligence.com/industry-reports/printed-circuit-board-market.

[38] Tupa, J., Basl, J., & Skocil, V. (2007). Application of process management tools for PCB manufacturing and diagnostics. In *ESTC 2006 - 1st Electronics Systemintegration Technology Conference, vol. 2*, 1166-1173. https://doi.org/10.1109/ESTC.2006.280157

Authors' contacts:

Jiri Tupa, Assoc. Prof.
(Corresponding author)
University of West Bohemia, Department of Materials and Technology,
Univerzitni 26, Pilsen, Czech Republic
tupa@fel.zcu.cz

Andrea Benesova, Ing., PhD
University of West Bohemia, Department of Materials and Technology,
Univerzitni 26, Pilsen, Czech Republic
benesov2@fel.zcu.cz

Farntisek Steiner, Assoc. Prof.
University of West Bohemia, Department of Materials and Technology,
Univerzitni 26, Pilsen, Czech Republic
steiner@fel.zcu.cz

Tomas Rericha, Ing., PhD
University of West Bohemia, Department of Materials and Technology,
Univerzitni 26, Pilsen, Czech Republic
trericha@fel.zcu.cz

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240523102857

Subject review
Received: 2024-05-23, Accepted: 2024-05-27

The German Case of Clean Energy Transition: How Grid Customers Perceive Regulatory Requirements, esp. Photovoltaic 'Obligation to Install'

Max Regenfelder*, André P. Slowak, Angela Werner, Marlene Weiblen, Josha Goldmann, Serpil Senger

Abstract: Whilst Germany's transition towards renewable energy sources is gaining momentum, such as, new photovoltaic installations numbers have multiplied over past three years, new policies' implementation is yet to be fully rolled out to all grid customer segments. First, this paper systematically lays out current developments in the areas of decentral renewable energies generation, electrification of heating & mobility, and supply grid operators. We carefully reflect on the institutional factors driving and shaping current developments. Second, the study focusses on grid customers' perception of new legislation that imposes the installation of photovoltaic systems onto roofs. We find that despite a majority supports the 'obligation to install', other pull factors may still be more important in shaping the decision-making process towards purchasing a photovoltaic system.

Keywords: distribution system operator; DSO; electricity grid; energy transition; photovoltaic obligation; renewables

1 INTRODUCTION - GERMANY'S CLEAN ENERGY TRANSITION IS GAINING MOMENTUM

Clean energy transition ranks on top of European Union's political agenda. Energy is a centrepiece of the EU Green Deal, which wants to reach carbon neutrality by year 2050. The policy sets ambitious goals for the pathway toward climate neutrality, and notably, the energy sector to date accounts for three quarters of the European greenhouse gas emissions [1, 2]. The sector is key to carbon neutrality, as it for instance provides the foundation for decarbonisation of mobility and heating.

The Renewable Energy Directive (revised 2023) defines targets on renewable energy sources usage, that is, to reach a minimum share of 42.5% of EU's gross final consumption of energy by 2030. EU member countries commit to contribute as specified in separate, national climate action plans. [3]
In the case of Germany, the action plan is embodied in a bundle of both federal and state-level legislations. The transformation of Germany's energy system toward clean energy started last decades but will be pushed further in the years to come: Renewable energies are required to cover 80% of year 2030 electricity demand, with full decarbonisation to be achieved by 2045. [4]

Tab. 1 depicts installed electricity generation capacity of renewables by end 2020, Germany. Renewables contributed already 45% to total German electric power generation for the year 2020. Which highlights how ambitious the German targets are, if benchmarked with EU targets.

Table 1 Installed capacity of renewable energy, 2020, Germany [5] (Tab. 19, p. 58, translated)

Technology	Installed generation capacity in GW at 31st Dec 2020
Onshore wind	54.4
Offshore wind	7.8
Photovoltaic	53.7
Biomass	8.8
Run-of-river	3.9
Other	0.6

Long-term forecasts of electric power generation for Germany can be found in the 'Grid Development Plan'. By 2045, the share of renewable energies shall rise to more than 91%, all scenarios. Note that this goes along with significantly risen gross electricity consumption – 953.5 or 1128.2 TWh respectively are forecasted scenarios for year 2045 [5].

The German Renewable Energies Act spells out incentives for the adoption of renewable energies, as well as it establishes a framework for customers' remuneration. The expansion of electricity grids is a government key priority and hence facilitated by this legislation [4].

The act does recognise the important role of distribution system operators. Namely, they integrate decentralized, renewable energy sources and batteries into the grid; they also make the grid absorb new sources of electricity demand, e.g., from electric vehicles. The distribution grid then is where 'smartness' will come in: Particularly, enabling flexibilisation of consumption respectively demand via smart meters and adjustable/ adaptable electr(on)ic appliances [6-9]. This paper is dedicated to analyse the transition from a distribution system operator' perspective.

2 ELECTRICITY DISTRIBUTION SYSTEM OPERATORS ARE FOCAL TO ENERGY TRANSITION

The German electricity grid consists of transmission and distribution grids.

Transmission grids carry large amounts of electric power over longer distance, for instance, from offshore and onshore wind turbines to the industrial centres in southern Germany. Such grids operate on extremely high voltage level. The latter, distribution grids run mostly on low or mid voltage. They distribute electricity to households and to industry. Distribution grids must cope with raising electricity demand triggered by electric vehicles or heat pumps, as well as integrating decentralised energy generation, e.g. from photovoltaic systems.

Distribution and transmission grids are operated by separate legal entities in Germany: If an energy company counts more than 100k customers, electricity suppliers and distribution system operators must become legally unbundled entities from each other and act as separate, for customers distinguishable companies [10]. Distribution system operators maintain, operate, and extend the electricity grid. They have an obligation to purchase and to remunerate for decentralised generation of electricity, e.g. from private households operating photovoltaic (PV) systems or from wind park operators. Remuneration is calculated by statutory feed-in tariffs, which are set at a level to promote renewables [4]. Distribution system operators consequently are a key contact for grid customers, electricians and planning companies because distribution system operators manage the grid connection process and keep records of the capacity of the local grids.

The electricity distribution system operators' perspective is exemplified by the company Netze BW GmbH. Alongside with TransnetBW GmbH for transmission grids, it is a legally unbundled subsidiary company of integrated energy supplier EnBW AG. Netze BW GmbH is the largest electricity distribution grid operator in the German state of Baden-Württemberg.

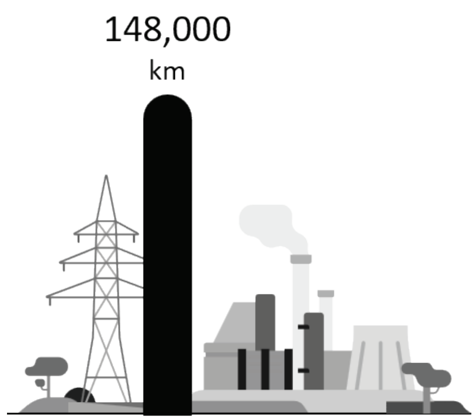


Figure 1 Transmission & distribution grid length of TransnetBW GmbH and Netze BW GmbH, dated 2023 [11] (adapted)

The German clean energy transition impacts distribution grids. The so-called "Network development plan 2037/2045" sets out some scenarios and development paths, projected for until year 2037 respectively climate neutrality reached year 2045. Key points are [5]:

- Compared with year 2020, renewable energies' share approximately to double – potentially up to 91% in the year 2045;
- Energy generation to decentralize and therefore will become more complex to manage for grid operators; and
- Clean energy transition to experience a major rebuild of distribution grids capacity ('once in a generation task').

For the state of Baden-Württemberg, PV systems' expansion matters most of all in lieu of fast increasing clean energy generation. This is due to favourable solar radiation in the southern as opposed to northern German states and, limited potential of wind turbine energy [13].

For an overview of the increase of German PV energy generation and different scenarios see Tab. 2.

Table 2 Overview of KPIs: Germany's forecasted electric power generation & consumption until year 2045 [5], (Tab. 1, p. 24, adapted, translated)

Scenario and year	Base at 31st Dec 2020	A	B	C	A	B/C
		y. 2037			y. 2045	
Electric power generation (installed capacity in GW)						
Photovoltaic	53.7	260	280	320	325	395
Renewable electric power generation	131.4	411.9	450.4	506.4	521.4	623.9
Electric power generation. Total	215.7	463.2	501.3	557.3	568.4	670.4
Electricity consumption(in TWh)						
Net electricity consumption (est.)	478.0	725.6	793.1	872.9	890.4	1064.5
Gross electricity consumption (est.)	532.8	778.7	846.4	926.4	953.5	1128.2
Share of renewables on gross electricity consumption (est.)	45%	>86%	>87%	>88%	>91%	>91%

Tab. 2 illustrates photovoltaics' role for renewable-based electric power supply across Germany: long term or until 2045, photovoltaics will by far become the most important renewable electric power source if accounted by installed capacity. Note that about half of the installed PV capacity stems from installations mounted on roofs, the other half from larger scale ground-mounted installations.

3 PHOTOVOLTAIC 'OBLIGATION TO INSTALL' IN SEVERAL GERMAN STATES

To accelerate the uptake of renewable energies from roof-mounted PV systems, a number of German states legislated obligations for property owners to install PV systems. Such 'Obligation to Install' applies to a) new built houses, b) when roofs are renovated, or c) where new parking lots are built. Aim is to unlock potential suitable surface areas for photovoltaics.

Our study focusses on above 'Obligation to Install' and for the state of Baden-Württemberg. The obligation has been applicable to different segments since different dates:

- since 1st January 2022 for non-residential builts and parking lots with more than 35 lots;
- since 1st May 2022 for residential, new buildings; and
- since 1st January 2023 for roof replacement/full roof refurbishment of existing buildings.

At least 60% of suitable roof surface shall be used for PV installation [13, 14].

In order to increase panel size of the study, we identified similar 'Obligations to Install' in other German states. Similarity assumed if obligation shows same timeline characteristics or scope. E.g., date of introduction, or scope residential vs. non-residential. Tab. 3 provides an overview of similar PV obligations across the German federal states.

It shall be noted that the sample is representative for the key decision-making criteria and property ownership status of participants albeit it is not fully representative for German population. We e.g. excluded states that introduced a PV

'Obligation to Install' for only public buildings. For private residential buildings, the obligation affects many grid customers, so that this focus of our study is most relevant.

Table 3 Overview of similar German PV obligations incl. closing dates

Scope	Non-residential, new buildings	Residential, new buildings	Roof replacement/full refurbishment	Parking lots	
Baden-Württemberg	YES 01.01.2022	YES 01.05.2022	YES 01.01.2023	YES 01.01.2022	[13, 14]
Berlin	YES 01.01.2023	YES 01.01.2023	YES 01.01.2023	NO	[15]
Hamburg	YES 01.01.2023	YES 01.01.2023	YES 01.01.2024	YES 01.01.2024	[16, 17]
Rhineland-Palatinate	YES 01.01.2023	NO	NO	YES 01.01.2023	[18]

Distribution system operators are the main point in policies' implementation. As far as distribution system operators are concerned, grid customers' attitudes and public perception de facto shapes policy communication. It determines what guidance and support becomes accessible to customers. And a good communication of distribution system operators could simplify grid connection processes as well as prevent additional respective follow-up inquiries by grid customers and by electricians.

4 REPRESENTATIVE SURVEY OF GRID CUSTOMERS: FACTORS IMPACTING UPON ATTITUDES AND PURCHASE DECISIONS FOR PHOTOVOLTAIC SYSTEMS

The survey was motivated by the industry's need to better understand the range of attitudes and motives of grid customers installing PV systems. This became a strategic need to respond to the state of Baden-Württemberg's new regulatory requirement of 'Obligation to Install' photovoltaics: Does the public perceive the obligation stipulated by the new regulatory requirements as positive, and fair?

Recent literature has showcased the relevance of fairness and justice to successfully establish and scale up new energy policies in modern societies [19, 20, 21]. More research is needed into the particulars for specific energies and countries. Our study makes a focussed contribution to that debate.

We wished to gain insight how relevant this obligation is to the decision-making process of grid customers. We argue that electricity distribution system operators need to know which factors have which impact on decisions to install PV systems, as well as the timings and reasons in the process. Findings can be used to design better approaches of informing grid customers and to address the demand of connecting PV systems to distribution grids.

4.1 Methodology and Sample

To document the state of knowledge and attitudes of the public in regard to upcoming or existing photovoltaic obligations, we selected German federal states that are similar and including Baden-Württemberg. Similar in terms of what regulatory requirements were to be implemented and when (as discussed in the previous section). The practical use of the data is to inform grid customers on their new PV 'Obligation to Install' more effectively.

Method of data collection was a large, representative set of short online interviews (approx. 10 minutes), conducted January 2023. The data was collected and compiled by the market research institute mindline energy GmbH.

The sampling targeted participants 18 years and older, from the states of Baden-Württemberg, Berlin, Hamburg, and Rhineland-Palatinate. The selection was then further narrowed down by these criteria:

- Survey participant is decision maker for energy supplier of their household, and
- Survey participant owns a residential or non-residential building, or s/he does plan to build one, and s/he
  - a) already owns a photovoltaic system or,
  - b) plans to install such within the next 12 months or,
  - c) is not currently mind-settled against new installations.

Sample size total n = 501, which consists of n = 200 owners of photovoltaic systems, n = 189 planning an installation within the next 12 months, and n = 112 non-rejecters. The sample is composed representative of gender.

We did not include straight opponents of the installation of Photovoltaic (PV) Systems into the panel, because we expected only minor insights from this group on how to improve communication to the public, respectively to grid customers.

4.2 Results: Majority Supports PV 'Obligation to Install' – Focus on Positive Perception and Fairness

The main reasons in favour of installing PV systems were safe energy supply, lower electricity costs, and environmental motivations. 'Obligation to Install' appears to be of secondary importance for acquiring a PV system.

Fig. 2 provides an overview of reasons for installing a PV system.

We also asked whether respondents are aware of PV obligations because the new regulatory requirements have only been implemented recently in Germany. However, most of the respondents were already familiar with implied obligations. Note that owners of non-residential buildings, for whom PV 'Obligation to Install' has already taken effect months earlier, do show a significantly higher level of awareness. Berlin, in general, records a higher level of awareness in our data set. Fig. 3 shows the results.

Information sources for respondents who stated being aware of the PV obligations are: press/ media (73%), internet incl. forums (41%), respondents' commissioned electricians (11%), as well as distribution system operators (9%). Meaning that both, electricians and distribution system operators are only involved to low extent into the information perceived by customers.

We further quizzed respondents about how they assess their level of information in PV obligations. 4 out of 10 feel that their information level is very good (14%) or good (24%). 38% at least state they're informed. 19% of the respondents perceive being badly informed or 4% not informed at all, respectively.

The level of information can be considered of particular importance when it comes to the perception of the PV obligations. It shapes the attitude of respondents. See Fig. 4 for details.

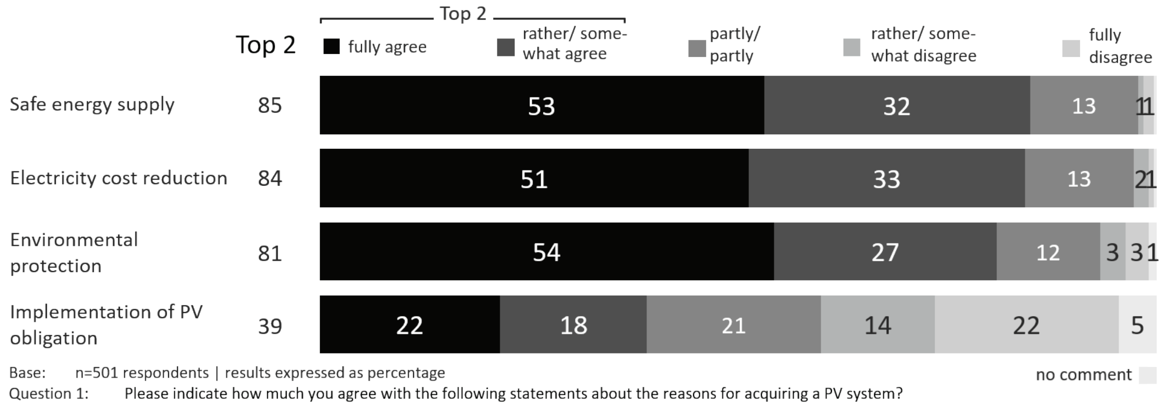


Figure 2 Reasons in favour of purchasing a PV system

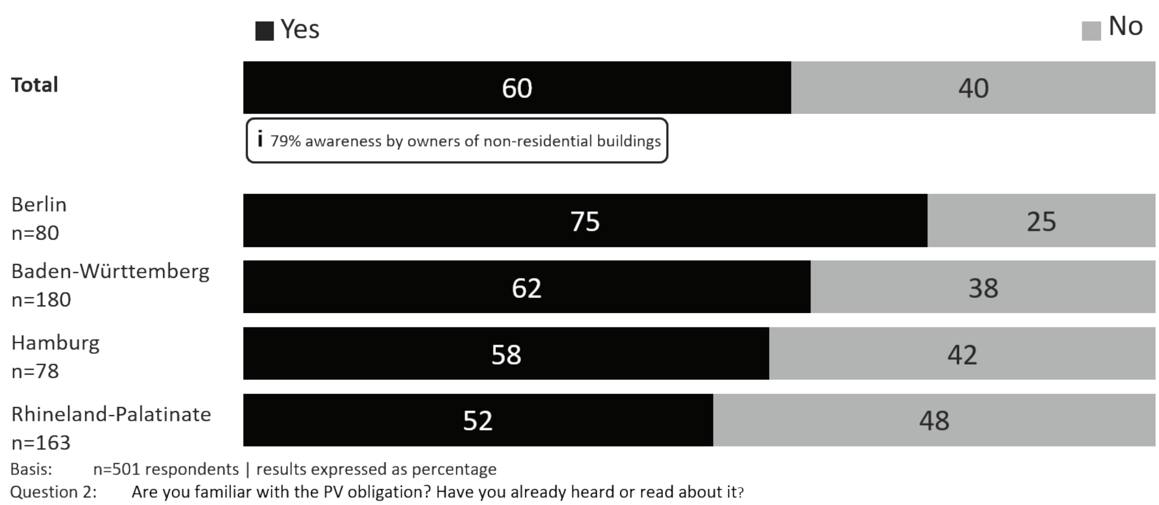


Figure 3 Awareness of PV 'Obligation to Install'

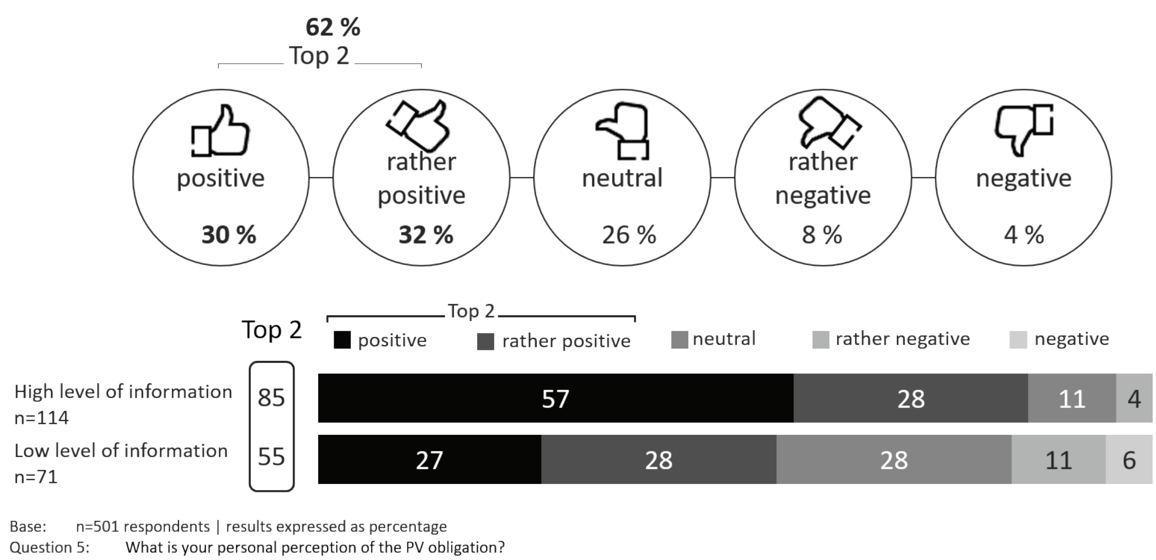


Figure 4 Individual perception of PV 'Obligation to Install'

Overall, we find a clearly positive perception of the 'Obligation to Install'. In addition, the more respondents are informed about it, the more positive appears to be their perception.

Besides individual perception, we asked the panel about fairness of the 'Obligation to Install'. This is a crucial point to any law makers: Legislation should be perceived as fair, for it to become widely accepted across society. Where an obligation is perceived as unfair, it may yield valuable lessons to be learned for future making of legislations.

54% of the respondents perceive PV obligations as fair – only 20% as unfair (26% do not know/ did not respond). This is again a clear majority in favour of the PV obligations. Fig. 5 sums up respondents' reasoning of perceived fairness vs. unfairness:

Figure 5 Top three reasons of perceived fairness vs. unfairness of PV obligations

An 'Obligation to Install' is perceived as fair, in principle. The survey participants particularly considered environmental protection and energy efficiency. Respondents who perceive it as unfair are mainly concerned about additional financial burden. In this context, also rejection of coercion itself is an issue of grid customers.

4.3 Discussion of Findings

Our study yields rich insight into the purchase decision of PV systems, and in consumers' context.

Key points are safe energy supply, energy costs, and environmental protection. Implementation of the newly introduced PV obligation as such is rather of secondary motivation. It shall be noted for context that during the data collection period of this study, energy prices went high and secure energy supply got challenged in the wake Ukraine war. However, we observe a strong support of the general public in favour of the PV obligation across selected German states. Concerns on the other side are dominated by fear of additional financial burden.

Data is from selected German states having introduced similar PV obligations. While environmental protection and secure supply priorities & objectives may be generalisable to other countries pushing clean energy transition, energy prices may differ. Therefore, results have to be read with a pitch of salt. Some are context specific for Germany, some may apply internationally. It was not the purpose of this paper to develop general theory. The sample composition implies some constraints: we do not have insight into a control group of customers rejecting PV systems at all, as such opposing customers were taken for not being interested to participate in interviews inquiring the new PV regulatory requirements.

5 CONCLUSIONS

Clean energy transition is clearly gaining momentum not just in policy making but also regarding its implementation. We observe ambitious plans and targets of decarbonisation of electricity generation both, on European and national levels. Our analysis did stress the importance of distribution grids if for the share of renewable energy sources to increase.

Especially, distribution systems operators are focal actors to manifold stakeholders: They manage the grid connection of decentralised electric power generation and, are key contact to grid customers and commissioned electricians. This fact highlights how the 'Obligation to Install' in selected German states serves as a device unlocking scaling-up the potential of PV systems and fostering renewable energy sources' share upon total electricity consumption.

We argue that it is crucial for legislative action to be widely accepted by the public, if to be effective long-term. For the PV 'Obligation to Install', we spot evidence of wider new policy acceptance: The higher the individual information level about the obligation is, the more positive its perception. Focussed, relevant information seems crucial for wider acceptance of policies enforcing PV obligations. However, in the individual perception of, and purchase decision for photovoltaic systems, some other motives seem to overshade the policy: Especially, cost savings for electricity, and safe supply. In addition, environmental protection is stated as motivation. The energy cost savings are a market-based mechanism incentivising installation of PV systems. Nevertheless, we want to shed light on the perception of 'fairness'. For policy makers it should much more guide the design of legislation in future. Since additional costs for installation of a photovoltaic system are the most prominent root cause of perceived unfairness, the costs should be addressed in their structure and allocation. This aligns with prior studies finding that installation costs are an important factor [22] respectively which highlight effectiveness of incentives in the promotion of (residential) PV systems (i.e. direct subsidies, tax deductions, feed-in tariffs) [23, 24, 25]. One could therefore establish smarter public grant schemes for property owners, to reduce perceived unfairness.

6 REFERENCES

[1] European Commission (2019). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions - The European Green Deal*. COM/2019/640 final, Brussels.

[2] Widuto, A. (2023). Energy transition in the EU. *Briefing for the European Parliament*, Brussels, www.europarl.europa.eu/ RegData/etudes/BRIE/2023/754623/EPRS_BRI(2023)754623 _EN.pdf.

[3] European Commission (2023). Directive (EU) 2023/2413 of the European Parliament and of the Council of 18 October 2023 amending Directive (EU) 2018/2001, Regulation (EU) 2018/1999 and Directive 98/70/EC as regards the promotion of energy from renewable sources, and repealing Council Directive (EU) 2015/652. *Official Journal of the European Union, L Series*, 1-77.

[4] Bundesministerium für Justiz (2014/ 2024). Erneuerbare-Energien-Gesetz vom 21. Juli 2014 (BGBl. I S. 1066), das zuletzt durch Artikel 1 des Gesetzes vom 5. Februar 2024 (BGBl. 2024 I Nr. 33) geändert worden ist. *Bundesgesetzblatt 1,* 1066/revised in *Bundesgesetzblatt 1*(151), 1-42. (in German)

[5] 50Hertz Transmission GmbH, Amprion GmbH, TenneT TSO GmbH, TransnetBW GmbH (2023). *Szenariorahmen zum Netzentwicklungsplan Strom 2037 mit Ausblick 2045*, *Version 2023 Entwurf der Übertragungsnetzbetreiber*. Berlin, Dortmund, Bayreuth, Stuttgart. (in German)

[6] Kerscher, S., & Arboleya, P. (2022). The key role of aggregators in the energy transition under the latest European regulatory framework. *International Journal of Electrical Power & Energy Systems, 134*, 1-16. https://doi.org/10.1016/j.ijepes.2021.107361

[7] Medina, J., Muller, N., & Roytelman, I. (2010). Demand Response and Distribution Grid Operations: Opportunities and Challenges. *IEEE Transactions on Smart Grid, 1*(2), 193-198. https://doi.org/10.1109/TSG.2010.2050156

[8] Stetz, T., von Appen, J., Niedermeyer, F., Scheibner, G., Sikora, R., & Braun, M. (2015). Twilight of the Grids: The Impact of Distributed Solar on Germany's Energy Transition. *IEEE Power and Energy Magazine, 13*(2), 50-61. https://doi.org/10.1109/MPE.2014.2379971

[9] Uzum, B., Yoldas, Y., Bahceci, S., & Onen, A. (2024). Comprehensive review of transmission system operators – Distribution system operators collaboration for flexible grid operations. *Electric Power Systems Research, 227.* https://doi.org/10.1016/j.epsr.2023.109976

[10] Bundesministerium für Justiz (2005/ 2024) Gesetz über die Elektrizitäts- und Gasversorgung (Energiewirtschaftsgesetz - EnWG). *Bundesgesetzblatt 1*(42), 1970-2018/ revised in *Bundesgesetzblatt, 1*(161), 1-12. (in German)

[11] EnBW AG (2024). *EnBW Company Portrait,* Stuttgart.

[12] Pape, C., & Geiger, D. (2023) *Regionalisierung des Ausbaus erneuerbarer Energien,* Fraunhofer IEE, Kassel. (in German)

[13] Landesumweltministerium Baden-Württemberg (2021). Verordnung des Umweltministeriums zu den Pflichten zur Installation von Photovoltaikanlagen auf Dach- und Parkplatzflächen (Photovoltaik-Pflicht-Verordnung-PVPf-VO) Vom 11. Oktober 2021. *Gesetzblatt für Baden-Württemberg, 31*, 847-850. (in German)

[14] Land Baden-Württemberg (2023). Klimaschutz- und Klimawandelanpassungsgesetz Baden-Württemberg (KlimaG BW) vom 7. Februar 2023. *Gesetzblatt für Baden-Württemberg, 2*, 26-48. (in German)

[15] Land Berlin (2021). Solargesetz Berlin vom 5. Juli 2021. *Gesetz- und Verordnungsblatt für Berlin, 77*(54), 837-839. (in German)

[16] Stadt Hamburg (2020). Hamburgisches Gesetz zum Schutz des Klimas (Hamburgisches Klimaschutzgesetz - HmbKliSchG) vom 20. Februar 2020. *Hamburgisches Gesetz- und Verordnungsblat*t, *10*, 148-179. (in German)

[17] Stadt Hamburg (2024). Gesetz zur Stärkung des Klimaschutzes und des Ausbaus der erneuerbaren Energien in Hamburg (Klimaschutzstärkungsgesetz). *Hamburgisches Gesetz- und Verordnungsblatt*, *1*(47), 443-445. (in German)

[18] Land Rheinland-Pfalz (2021). Landesgesetz zur Installation von Solaranlagen (Landessolargesetz - LSolarG) vom 30. September 2021. *Gesetz- und Verordnungsblatt für Rheinland-Pfalz*. (in German)

[19] Liu, L., Bouman, T., Perlaviciute, G., & Steg, L. (2020). Public participation in decision making, perceived procedural fairness and public acceptability of renewable energy projects. *Energy and Climate Change, 1*. https://doi.org/10.1016/j.egycc.2020.100013

[20] Bal, M., Stok, M., Bombaerts, G., Huijts, N., Schneider, P., Spahn, A., & Buskens, V. (2023). A fairway to fairness: Toward a richer conceptualization of fairness perceptions for just energy transitions. *Energy Research & Social Science, 103*. https://doi.org/10.1016/j.erss.2023.103213

[21] Liebe, U., & Dobers, G.M. (2020). Measurement of fairness perceptions in energy transition research: A factorial survey approach. *Sustainability, 12*(19). https://doi.org/10.3390/su12198084

[22] Zander, K., Simpson, G., Mathew, S., Nepal, R., & Garnett, S. (2019). Preferences for and potential impacts of financial incentives to install residential rooftop solar photovoltaic systems in Australia. *Journal of Cleaner Production*, 230, 328-338. https://doi.org/10.1016/j.jclepro.2019.05.133

[23] Lee, M., Hong, T., & Koo, C. (2016). An economic impact analysis of state solar incentives for improving financial performance of residential solar photovoltaic systems in the United States. *Renewable and Sustainable Energy Reviews, 58*, 590-607. https://doi.org/10.1016/j.rser.2015.12.297

[24] D'Adamo, I., Mammetti, M., Ottaviani, D., & Ozturk, I. (2023). Photovoltaic systems and sustainable communities: New social models for ecological transition. The impact of incentive policies in profitability analyses. *Renewable Energy, 202*, 1291-1304. https://doi.org/10.1016/j.renene.2022.11.127

[25] Du, Y., & Ma, T. (2023). From FIT to FIP: assessing the impact of feed-in policies on renewable development in Germany. *Applied Economics Letters*, *30*(18), 2597-2606. https://doi.org/10.1080/13504851.2022.2100048

Authors' contact:

Max Regenfelder
(Corresponding author)
Netze BW GmbH,
Schelmenwasenstraße 15, 70567 Stuttgart, Germany
m.regenfelder@netze-bw.de

André P. Slowak
University of Roehampton,
Business School,
Roehampton Lane, London, SW15 5PJ, United Kingdom

Angela Werner
Netze BW GmbH,
Schelmenwasenstraße 15, 70567 Stuttgart, Germany

Marlene Weiblen
Netze BW GmbH,
Schelmenwasenstraße 15, 70567 Stuttgart, Germany

Josha Goldmann
Netze BW GmbH,
Schelmenwasenstraße 15, 70567 Stuttgart, Germany

Serpil Senger
Netze BW GmbH
Schelmenwasenstraße 15, 70567 Stuttgart, Germany

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240523123116

Subject review
Received: 2024-05-23, Accepted: 2024-05-27

Artificial Intelligence System for the Digitalization of Information Distribution

Hrvoje Kober, Maja Trstenjak*, Tihomir Opetuk, Hrvoje Cajner, Goran Đukić

Abstract: In the realm of manufacturing, creativity and innovation are pivotal for gaining competitive edges and ensuring long-term viability. With technology evolving rapidly and market demands escalating, manufacturing firms strive to optimize processes to stay competitive globally. Artificial Intelligence (AI) emerges as a vital tool in this transformation, heralding Industry 4.0's focus on integration and automation. AI integration promises enhanced product consistency, cost reduction, and operational efficiency. By fostering collaboration between humans and machines, connected systems can gather, analyze, and optimize production data. This synergy lays the groundwork for Industry 5.0, prioritizing human-centricity and sustainability. This paper examines AI's role in ushering in Industry 5.0 at local glassware company. By reviewing pertinent literature and addressing implementation challenges, it delineates strategies for market leadership through AI adoption. The implementation unfolds in four phases, focusing initially on digitizing information flow to streamline management. Additionally, it offers insights into leveraging AI to gain competitive advantages, illustrating its transformative potential in the dynamic manufacturing landscape.

Keywords: artificial intelligence; AI; digitalization; human-centric; Industry 4.0; Industry 5.0; information distribution; organizational communication

1 INTRODUCTION

In the context of accelerated technological development, manufacturing companies face the imperative of integrating artificial intelligence (AI) into their operations to remain competitive within Industry 4.0 or 5.0. In the last three years, AI systems have proven that they can meet, and surpass, human performance in image recognition, speech transcription, and direct translation, have learned to drive, identify relevant information in text to answer a question, recognize people's faces (even if are blurry images) and human emotions, create their schemes and detect malicious programs, etc. [1].

The industry is increasingly focusing on improving product consistency and reducing overall costs. This can only be achieved through successful collaboration between humans and robots. Using artificial intelligence systems, machines can collect and process various data to further improve the production process. Industry can use artificial intelligence to process information from connected machines and IoT devices that allow companies to monitor all activities and processes from start to finish [2].

The fourth industrial revolution, Industry 4.0, introduced the concepts of the Internet, artificial intelligence (AI), and machine learning (ML) to manufacturing. Therefore, there is a need to understand the capabilities of AI and ML and how to successfully implement them in manufacturing areas to achieve the best possible results [3]. A detailed study of the given correlation and its improvement to the level of the machine - artificial intelligence - human cooperation would make a significant step towards Industry 5.0, which is characterized by a combination of human intelligence and cognitive computing, which results in the development of production with an emphasis on speed and efficiency.

Understanding the motivations and benefits of introducing artificial intelligence (AI) into production operations is key to achieving a competitive advantage within Industry 5.0. The analysis of technological innovations in the production sector emphasizes that the introduction of AI enables an increase in the efficiency and consistency of production processes. Using AI systems, companies can optimize resources, reduce production time, and achieve a high level of consistency in product quality. Also, it is important to emphasize that the introduction of AI enables advanced data analytics and making informed decisions. AI systems analyze large amounts of data in real-time, providing deeper insight into production performance, market needs, and industry trends.

Analyzing the evolution of AI in industry, it highlights the benefits of adaptability and reactivity that AI systems bring to manufacturing companies. The ability to adapt to changes in real-time enables a faster response to dynamic market conditions and production challenges [4, 5]. Also, concrete examples of improving work processes through the introduction of AI are explored in detail, including autonomous vehicle driving, facial recognition, content generation, and product personalization. These elements directly contribute to innovation and modernization of production operations, which all leaders in their categories strive for in the current market.

1.2 The Impact of Industries 4.0 and 5.0 on the Operations of Manufacturing Companies

Industry 4.0 marks the introduction of new technologies and paradigms that have a profound and transformational impact on the way work is done. It analyzes the increased connectivity between different parts of the production chain using the Internet of Things (IoT), artificial intelligence (AI) and machine learning (ML). The connection of these technologies enables the complete integration of production, from raw materials to the final product [3]. The appearance of the term Industry 5.0 refers to the concept of cooperation between humans and robots and smart machines in an industrial context. The central element of this approach lies in the ability of robots to support human work through the optimal use of advanced technologies. Industry 5.0 [6] builds on the automation and efficiency pillars of Industry 4.0, adding a human element. In the manufacturing sector, robots have historically performed tasks that were dangerous, monotonous, or physically demanding, such as welding and painting in car factories and handling heavy materials in warehouses. Because of increasingly connected and intelligent machines, Industry 5.0 aims to integrate the cognitive abilities of computer systems with human intelligence within the framework of cooperation. Also, it is crucial to investigate how industries 4.0 and 5.0 encourage the development of ethical frameworks and sustainable practices. Companies are setting new standards in the responsible use of technologies, taking into account the ethical implications of introducing AI into production operations [5]. The dynamics of cooperation between humans and AI are studied within the framework of Industry 5.0. Also, it is considered how artificial intelligence systems become an integral part of work teams, harmonizing with human abilities and optimizing joint performance [7].

1.3 Technical Aspects of AI in the Manufacturing Industry

The technical aspect of AI, to improve its performance over time is given below through a simplified overview of key elements [8]:

- Data collection: The collection of high-quality data is critical to a successful AI system. Data can be structured (e.g., database tables) or unstructured (e.g., images, texts).
- Databases: Data is often stored in databases to facilitate management and quick access. Different types of databases can be used depending on the needs of the project.
- Data processing: Data processing involves cleaning, transforming, and normalizing data to make it suitable for analysis and learning. This is an important step because the quality of the data directly affects the quality of the model.
- Machine learning algorithms: Choosing the appropriate algorithm depends on the type of task. For supervised learning, where labeled data is available, algorithms such as linear regression or deep neural networks are often used.
- Model Training: Model training involves feeding an algorithm with data to learn connections and patterns. The model is adjusted to minimize the error between its predictions and actual results. Training can take from a few minutes to a few days depending on the complexity of the model and the amount of data.
- Validation and testing: After training, the model must be validated and tested on a separate data set to assess its accuracy and ability to generalize to new data.
- Implementation and integration: Once the model has been successfully trained and tested, it is integrated into the system to perform the tasks it was designed for. Integration may involve creating interfaces, APIs, or other methods of communication with other parts of the system.
- Maintenance and monitoring: An AI system requires continuous monitoring and maintenance. Updating the model with new data, optimizing performance, and solving problems that may arise over time are key elements of the long-term success of an AI system.
- It should be emphasized that the details may differ depending on the specific task and the type of AI system being developed.

2 RELATED WORK

Although AI solutions such as ChatGPT are being in everyday use and continuously developed, so far in the scientific literature there is only little evidence for its use in manufacturing, according to papers indexed in the Web of Science database and related to the manufacturing industry. After the literature screening and the analysis of the related papers the gaps will be recognized and AI system for business support will be formed.

Industry 5.0 can be understood as a virtual-real interactive system with great potential which is in need for "safe in physical spaces, secure in cyberspaces, sustainable in ecology, sensitive in individual privacy and rights, service for all, and smartness of all" [9]. AI and big data in the industry have found the way due to the need for economic, safe, and sustainable manufacturing in the market. The integration of AI technologies enables dealing with dynamic processes [10].

AI assistants such as Alexa, Bibxy or Siri are multi-purpose tools, while the manufacturing industry demands more unique solution for the specific needs. Understanding of benefits is needed to ensure the adequate functioning inside the manufacturing system. Some authors claim that yet the benefits need more evidence collected in natural manufacturing environment. Disadvantages, limitations, and risks concern reduced worker autonomy, constrained language understanding, increased dependency on software, and harmful exploitation [11]. Some of the tasks cannot be carried out under robotization or automation; therefore, novel human-work support tools are expected. Support tools such as augmented reality (AR) and AI can be used then which leads to job simplification enabling inexperienced, unskilled, or less skilled employees to perform the work in the selected manual production processes [12].

AI is expected to be able to act autonomously, "support people through assistance systems, use resources more effectively, make processes more environmentally friendly and enable new working models with direct participation and greater transparency". The influence of AI on employment is controversial with many benefits but also with raised concerns about job losses, growing autonomy and control mechanisms towards human behavior. The participative leadership of the future conducts flexibly within the framework of self-organizing networks and interdisciplinary, democratically formed teams. Executives see themselves as coaches and moderators [13].

AI system supports internal business processes and workflows, it is addition to high degree of interactivity, transparency and reliability [14].

Intelligence assistance system enables helping employees and providing competence-related support which results in short- and long-term efficiency in problem-solving in companies [15].

AI systems assist workers in decision-making, resource coordination and task execution. Such models are iteratively optimized and verified through scenarios engineering and acquire new knowledge and refine its knowledge base, while security remains an issue because of which it is suggested to incorporate federated intelligence and smart contracts technologies in constructing and training such models. When both humans and AI act as decision agents, there is a need to emphasize the management of the impact of technological involvement on human performance and vice-versa. Therefore the different strategies of human-machine collaboration should be discussed to identify the most suitable strategy for organizational decision making in Industry 5.0, enhanced by AI [16].

To increase manufacturing system productivity, reduce human effort and avoid the possibility of injury the maintenance activities should be predicted by assessing the health of the machine. Data-driven prognostics rely on statistical and AI methods, including machine learning and deep learning models. Prognostics and health management is one of the key fields in which AI will be very useful [17].

Control and supervision mechanisms can be AI-AR-VR based and characterized by a large degree of autonomy to optimize processes, quality and maintenance operations. Remote cooperation, assistance and maintenance is also part of the environment for having no man on floor increasing safety and putting the focus on control and human centrality [18].

When there is a need for repeated changes in the schedule, AI can come in use. The principles of the artificial neural networks and IF-THEN scheduling rules simulate flexible manufacturing [19].

Artificial intelligence should maximize output performance to boost firm demand and supply. On the other hand, business intelligence enables the company to make financial and business decisions in which big data analytics is a crucial assistance. To maintain production efficiency and worker comfort, the human variable performance must be included for the production schedule. It is first to be used as a targeting tool for a global firm plan and time buffer allotment system and later for the job scheduling by the human resource which can outperform traditional models of performance improvement in speed, accuracy and human error [20]. The traditional manufacturing activities can be renovated with adoption of AI technologies which can lead to supply chain resiliency and sustainability. Real-time tracking of supply chain activities is of high importance [21].

Digital twins are highly compatible with AI, but at the same time utilization of digital twins in warehouse management has been relatively neglected over the years. AI cna improve warehouse management, supply chain optimization and operational efficiency in various industries [22].

Artificial Intelligence of Things (AIoT)-based automated picking system has been development for online shop and services for automated shipping systems. The systems divert consumers who are moved by AIoT, while robotic manipulators replace human tasks to pick. It enabled the increase of evaluation efficiency, speed, and convenience of the processes [23].

The demand for personalization remains high, and the work environment should be safe, efficient and personalized collaborative workplace. Frameworks such as DeFACT enable the different enterprises and parallel workers to be organized, coordinated and scheduled based on decentralized autonomous organizations and operations to promote mutual benefits among members. This enables the provision of higher-quality personalized products and services with the safety of data and knowledge. The use of generative artificial intelligence is important here [24].

In the transport sector, AI is being used for autonomous delivery, and self-driving trucks, which is still just an idea that should be implemented with high economic efficiency. AI is also integrated in special technical assistance systems in trucks and the human-AI combination of systems has proven to be the most efficient [25]. AI can play a substantial role in digital transformation, especially in the servitization process which is related to the creation of new services based on existing products that are already in the company's offer. Companies improve their offer and instead of only products, they offer products and services or only services created on the product. Selling services require different business models than in selling products, and advanced solutions should be considered. AI can here be a very useful tool [26].

Rules and standards for the use of IoT and AI should be established to minimize the consequences of technology misuse [27].

AI can be used for cyber threat detection to protect modern digital ecosystems. ML-based classifiers and ensembles detect anomaly-based malware and network intrusion. Their integration in the overall system is important [28].

AI can be used as a chatbot as part of the MES with a prediction system as a MES layer. Chatbot is an assistance in production coordination for the workforce to learn from their inputs. This kind of system provides live updates in natural language which enables easy information extraction in comparison to the traditional search techniques [29].

AI controller is implemented in business support systems by machine learning. It can be used in thermal power plant and chemical product manufacturing plant [30].

In implementing robotic AI-based assistance systems and exploiting all the potential the design methodology Value Sensitive Design (VSD) is proposed as a starting point to align this technology with human values [31].

AI is a key enabler of the transition between static, hard-coded algorithms and flexible and innovative ones. The automotive sector can benefit from the usage of such techniques. The special focus can be put on measuring system, customer satisfaction analysis, and demand prediction. But AI can also be implemented in the product (Advanced Driver Assistance Systems, as an in-vehicle IoT system) [17].

Development and use of AI in the manufacturing industry does not only bring opportunities but also challenges, but many claim that AI will be innovative and revolutionary assistance to the industry [32].

3 IMPLEMENTATION OF AI IN THE MANUFACTURING SYSTEM – CASE STUDY

The framework of AI implementation in the manufacturing system was developed on case study of a Croatian company, which is of the leading glassware companies in the region.

The simplest description of this solution is obtained by dividing the implementation process into two key parts, which are further divided into four phases, with the first two phases forming a virtual assistant, and the second two creating a virtual manager in business.

3.1 Phase I – Introduction (Kindergarten)

First phase consists of conceptualizing an idea and filling the database with existing information and documents and assigning the simplest tasks to artificial intelligence and monitoring their execution to solve problems that may arise during use and improving the overall functioning of the AI system. With the successful implementation of this phase, the project can safely be continued, this phase is almost completed and is currently in the process of adjusting the last small elements in order to successfully continue with phase.

3.2 Phase II – Product (High School Student)

It is currently the most important and probably the most complex phase of development - defining the more complicated tasks that artificial intelligence must perform. The goal is to reach the desired level and achieve that the result of this phase is a finished product that can be offered to the market. The main idea is to enable a maximally universal solution that can be successfully implemented in various industries without any problems, and that the finished product after the second phase be maximally automated with the main goal of enabling the company's management to monitor the situation in the company more simply and efficiently and to be able to, quickly and efficiently, without the need for the help of the engineer who developed the system and without knowledge of the code, assign tasks to the artificial intelligence that it should perform depending on the driver. The simplest way is to describe with an example what AI should look like in this phase: for example, if the AI assesses that an email or feedback from a client is written in an angry or dissatisfied tone, it will automatically react and forward that email to the management, who will then be able to investigate it more easily with its help. the background of the problem, i.e., through a simple conversation with the AI, they will be able to get all the necessary information related to the problem and its solution. Also, at this stage, it is crucial not to demand impossible tasks from artificial intelligence, but to be realistic and offer a product that meets all needs, and not to lose its value by expecting unachievable results.

3.3 Phase III – Almost Genius

It is almost the final stage, the main goal of which is the full implementation of AI in the entire business with an emphasis on production. In more detail, the ultimate goal of this phase of AI implementation is, after collecting a sufficient amount of information, to implement it in the production plant by acquiring sensor devices that record the working environment (with a precision of up to mm^2 within 2 minutes) and "empty" robots that they come unprogrammed intending to be connected to a system in which there is already an artificial intelligence that will program them together with the already mentioned sensors and determine their paths, tasks, time, speed and a large number of other factors key to successful production.

3.4 Phase IV – Complete Genius

This is the last stage of implementation where the artificial intelligence becomes so aware of its situation that it can do most of the work on its own and has acquired a huge amount of data from years of operation that it can give advice to the company's management. This phase is still quite far from being feasible, but it is developing day by day with new thoughts and attitudes of the company's management.

In this paper, emphasis will be placed on the implementation of the first two phases of this project. Namely, the already implemented first phase will be analyzed in detail, and the main part will be focused on the implementation of the second phase, the result of which is a finished product that can be offered to the market regardless of the industry.

3.5 A Snapshot of the Current State

This is a company that for the last 10 years has allocated a significant amount of funds to acquire the most modern technologies, with which it could produce and supply the highest quality glass products in an environmentally friendly manner. Multi-year investments indicate the long-term vision of the company, in which the company wants to be the leading Croatian brand in the glass industry, which will provide its consumers, employees, community, and business partners with a bright future through glass products.

Analysis of the company's past business activities with a clear focus on export activities indicates their key contribution to further growth and development. Through the systematic implementation of export strategies, the company would successfully expand its market, diversify revenues, and achieve competitive advantages on a global level.

Export activities proved to be an indispensable factor in achieving long-term business sustainability, contributing to strengthening financial stability and increasing market share. In addition, the increased global presence of the company would result in the establishment of strategic partner relationships, which would further expand the network of business opportunities.

The company concluded that the current level of information distribution within the company is not at a satisfactory level and that it needs to be significantly increased so that the company can deal equally with other multinational companies that have been operating on planned export markets for many years.

Namely, it is crucial to emphasize that the main problem with the distribution of information within the company, as well as the information that goes outside it, is precisely the insufficient amount of control over it. That is why the key idea of implementing an AI system is to increase control and at the same time make it as easy as possible for the company's management to access all the required information in the fastest possible time frame.

With the current way of distributing information outside the company, the management cannot react in time to the emergence of a potential problem because it does not have insight into every e-mail that is sent to existing or potential clients. This way of "control" is unacceptable for the company because it learns about a possible problem much too late, i.e., it only finds out about a possible problem now when it has arisen without the possibility of eliminating the possibility of its occurrence.

To successfully solve the mentioned problem, the company implemented the first part of the project, the so-called "virtual assistant", which plans to significantly increase the level of control and enable the possibility of eliminating potential problems before it is too late, i.e. before the problem even arises.

3.5.1 Virtual Assistant

As already stated earlier, the project of introducing AI into the entire business takes place in two key parts, each part of which consists of two phases. This paper plans to provide a detailed insight into the processes that took place in the first part and what the final product will look like, the so-called Virtual Assistant, after completing the first two phases of implementation.

The key goal is to create an independent system that is more advanced than all the "ready-made" solutions that have been offered multiple times on the market recently, and to respect the main role of the system itself: Monitoring the distribution of information within the company.

3.5.1.1 Phase I

The main role of the AI-assisted system "Virtual Assistant (VA)" is an assistant in business. The primary tasks of the VA, which were emphasized in the first part of the development, are the management of business email communication and an AI-based chatbot application through which authorized employees have insight into the business and can manage the data used when redirecting emails. Development began in May 2023 through two initial phases. Phase I was crucial in the long run for the successful implementation of the entire project. Namely, in the first phase, the most important thing was to define the master data that will be used to fill the AI and lay the foundations for the further development of the agents that are the main part of phase II. The master data that was a key part of AI development was as follows:

- Downloading data about customers, references and offers from company's existing systems (Microsoft Exchange Server, custom ERP system)
- Connecting clients (customers) with their mail contacts and mail domains
- Determining the primary and secondary references of company's by customers according to the criteria of the most frequent communication, i.e. according to the initially defined contacts in the ERP.

Also, it was crucial for AI to start learning how to manage business email messages, because the entire development of phase II is based on communication via emails, both incoming and outgoing.

The key items identified as the most important for development in this phase were:

- Interception of incoming external mails
- Forwarding of e-mails to the officers in charge
- Preparation for approval or stopping of outgoing emails (e.g., with complex offers)
- Support for multiple internal domains
- Processing of e-mails sent to a group, nested e-mails, invitations to meetings, junk e-mails, confidential e-mails.

As the second phase plans to establish control over employees within the company, it was necessary to develop the following aspects of AI as preparation for further implementation:

- Notification of employee absence - employees report absence from work via e-mail (annual vacation, sick leave, maternity leave, day off, etc.)
- Management reports the absence of employees via chat.

The final step in the development of phase I was the creation of the Chatbot application VA, with which the company's management could test the current progress of AI and, through work with it, see what improvements are still needed to make the system complete after phase II. The key items that were checked and required from the VA Chatbot application were:

- Inquiries into structured data - about customers, references, received and sent emails
- Rating of the response of the Virtual Assistant
- Employee absence record
- Change primary or secondary referrer via chat
- Placing standing orders for users
- Access to the application is limited to the Administration.

3.5.1.2 Phase II

Distribution of information, which is very difficult to specify exactly what is meant by this term, but to successfully implement the implementation of AI, the company currently emphasizes the distribution of information via email addresses, both those entering the company and those that are moving within it and eventually also those that go outside the company.

To successfully implement AI in each of these three processes, it was necessary to teach artificial intelligence about data that is key to identifying potential opportunities and threats. Although most of this process took place, as stated earlier, in the first phase, it is easiest to describe this part of learning as a final embellishment. Namely, the company does not want to rush through the stages of the entire process and in the end deliver a product that is 80% completed in each stage but wants to bring each stage to perfection so that the result is 100% correct and efficient.

Also, it was necessary to list the key parts of this phase, and the same was done by creating the so-called "AI agents", each of which is in charge of implementing one of the key items for the successful implementation of the second phase and the first part.

Within the project, five key agents are listed specifically:

Agent Block **–** this is an agent whose role is the most demanding, but also the most important for the successful control of information distribution. Namely, its role is to monitor all outgoing e-mails to potential or future partners and to stop those e-mails that contain unprofessional elements (rude tone, inappropriate words, simple words, violent tone, threats, etc.). After the sending of a specific email is blocked, the information about this is forwarded to the company's management, as well as to the person responsible for controlling a specific employee. Also, along with the information about the inappropriate e-mail, the entire e-mail is delivered, as well as information on why the AI made such an assessment and decided to block the output of the e-mail.

The graphical scheme of the functioning of the Block agent is shown in Fig. 1.

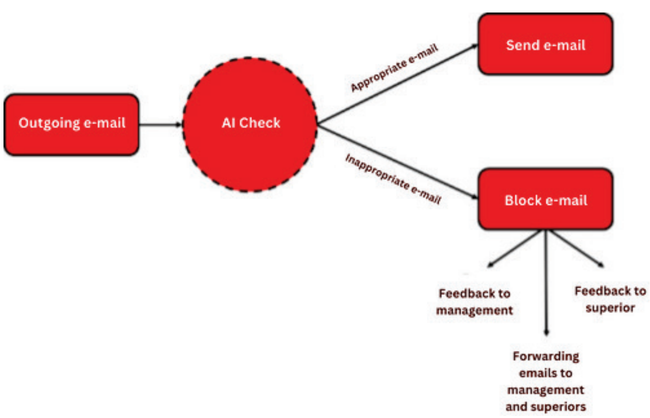


Figure 1 Agent Block functioning scheme

Agent Distribution - the main role of this agent, also called Agent Right, is the correct delivery of all incoming mail in the company. In more detail, every email that comes to the company's server is first accepted by AI, which studies the content of the email, and, depending on the content of the email, either delivers it to the written email address or changes the address to which the email should arrive. For example, if an e-mail related to a complaint about a certain product or delivery comes to the general contact e-mail address (which is available on the company's website), AI will receive it, study it, and not send it to the requested address, but will send it to a person in the company who is the representative of the company that sent the complaint. Likewise, if it is a matter of repeated complaints, that email will be sent to the management of the company, in addition to the person who is the representative, so that they can respond to the situation successfully and promptly. Also, as with the previous agent, he will be able to answer very simply why this information had to be found with a certain person.

The graphical scheme of the functioning of the Agent Distribution is shown in Fig. 2.

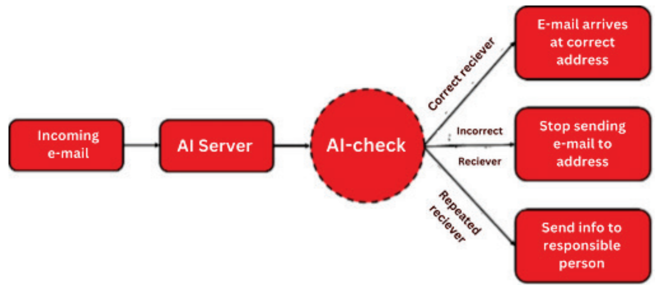


Figure 2 Agent Distribution functioning scheme

Agent E-mail – with, of course, a marketing goal in the form of information about the use of AI, this agent must facilitate communication in moments of increased demand. Namely, this agent can send an email from its personal email address. Of course, the e-mail is sent on request with a previous presentation of the planned for verification, but it makes it as easy as possible to respond to unnecessary e-mails that waste time unnecessarily. It functions as a "virtual secretary" that is a shield from the outside world and unsolicited emails.

The graphic scheme of the functioning of the E-mail agent is shown in Fig. 3.

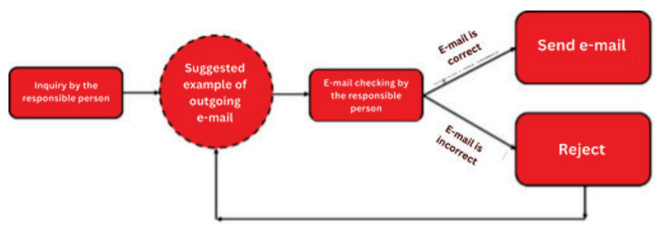


Figure 3 Agent E-mail functioning scheme

Agent Information – this agent must always be able to provide the company's management with information related to employees. More specifically, its main task is to store information and create a knowledge base about employees, adopt their work habits, and monitor their work. As employees are also expected to cooperate and communicate with AI and immediate feedback is more than satisfactory, AI must be able to report information about employees even when they have not reported it themselves. For example, an employee who forgot to report information about a sick leave did not even have to do the same because the AI collected enough information and was able to inform the management about the absence of a certain employee. This agent will also store personal data about employees, and one of the examples of its use will be congratulating on birthdays or other important moments.

The graphic scheme of the functioning of the Information agent is shown in Fig. 4.

Agent Help – this is very similar to the previous agent, but this agent has access to all information within the company, not only about employees. Everyone in the company, depending on their position within it, has access to a certain amount of information through conversation with

AI. The management of the company can in a very short period, or rather in a few sentences, get the exact information that interests them at that moment. In more detail, if an insufficiently high-quality result is noticed by a certain representative, a daily, weekly, or monthly analysis of the same can be reached in several queries to see where the problem arose and how to solve it most effectively.

The graphic scheme of the functioning of the Distribution agent is shown in Fig. 5.

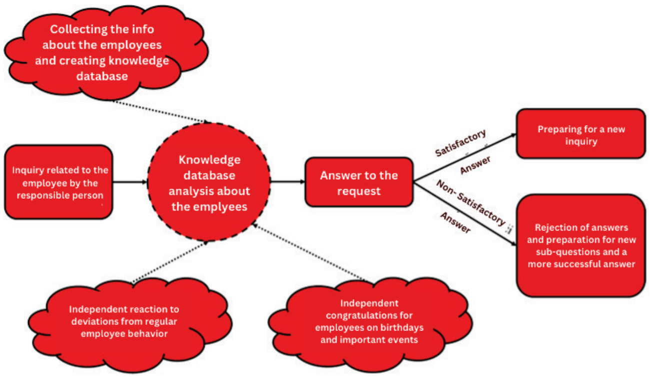


Figure 4 Agent Information functioning scheme

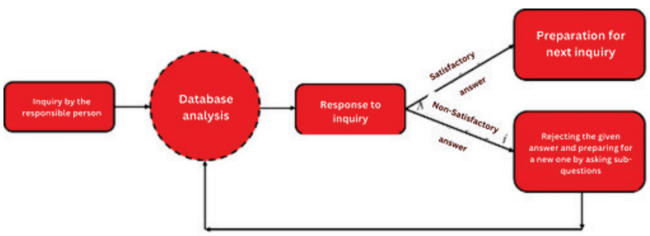


Figure 5 Agent Help functioning scheme

3.5.1.3 Virtual Host

After the successful implementation of the first part of the project, the plan is to further improve all acquired AI knowledge and implement it in other processes within the company. The emphasis of the virtual leader would naturally be on even better and more independent control of information, but also on its involvement in processes that it had only learned about until now. Of course, entering the process of production, processing, packaging, ordering, selling, etc. is not an easy task, but with the large amount of information that artificial intelligence should possess by then, this step is not too worrying.

Of course, it is important to note again that the company does not want to rush the entire project, and from this, it can be concluded that the second part is, for now, only an idea on paper, but taking into account the current speed of successful implementation and expectations related to the completion of the second phase by end of 2023, the beginning of the realization of the second part is no longer in such a distant future.

3.5.1.4 Phase III

Further development in phase III is based on the achieved results from phase II, for which additional improvement is planned. Most of the planned activities are based on the establishment of artificial intelligence communication and the existing ERP system. As in the previous cases, phase III consists of several key components that the desire is to realize, namely:

Communication with clients - in this step of implementation, the goal is to create an agent that checks the content of each order or offer, compares it with the request, and performs supervision in the form of a check written within the offer with the customer's price list located within the ERP system. Namely, the main task of this agent is to stop an improperly made offer and to let the management know that a mistake has been made with feedback. In this way, the possibility of human error will be minimized as much as possible, which will increase the efficiency and quality of the service.

Production - after AI is already connected to the existing ERP system in the company, the initial idea is to enable the implementation of AI in the production process using barcodes and readers. Namely, each order will have its own barcode, which, after reading it on the machine with a barcode reader, will give the machine insight into the order, and the machine, with the help of artificial intelligence, will perform the requested action by itself and thus further reduce the possibility of human error in the production process. The goal of this part is to enable artificial intelligence, with the help of unprogrammed robots and microsensors that record the entire plant at the highest level, to program robots and run the entire production plant with minimal to no human input. In more detail, the human role will mostly be monitoring the robot's work from a safe distance and final control of the finished products. As this is an industry where injuries at work are a very dangerous item because it is a raw material that breaks very easily during processing and can cause serious injuries, this step would minimize, if not eliminate, the danger to humans in this part of the production process.

Defective orders – the glass industry used to rely on selling glass that was not good enough to deliver to the customer, i.e. it had a defect. As time goes by, the trends in this industry change and there are no longer small trades that generate serious income by buying, processing and reselling defective used glass. In order to eliminate this cost, artificial intelligence will be taught to recognize the possibility of an incorrect order and will be able to react to it earlier than a human has done so far. Of course, there will still be defective glasses, but their time spent in the production process will be reduced and they will be separated from it with the desire to deliver finished products as quickly and efficiently as possible. Defective glasses will not be thrown away, but artificial intelligence will redirect these glasses to remelting.

3.5.1.5 Phase IV

It is about the last phase, but also about the phase that is currently the most complete. Namely, it is about the moment when artificial intelligence advances to the extent that it fully controls all the material it owns related to the company, but also all data related to current market trends and predictions of future trends. More precisely, artificial intelligence becomes so aware of the situation in which it finds itself that it can perform most tasks by itself, from the distribution of communications within the company to the process of procurement, production, packaging, sales, delivery, etc., and has acquired a huge amount of data from years of business that it able to advise the management of the company.

It is not necessary to emphasize too much that this phase is still really far away and will develop simultaneously with the increasingly rapid development of artificial intelligence. However, it should be emphasized that the development of elements from this field has always been rather rapid - more precisely, new innovative solutions have ensured immediate jumps that, after reaching their peak, would stagnate for a long time and wait for a new jump with the development of some new innovative solution.

4 CONCLUSION

Through a clear overview of its history and development, it is clear to conclude that AI represents a key factor in the transformation of traditional business models. Modern business requires innovative approaches, and the implementation of AI brings numerous advantages that can significantly improve the efficiency and competitiveness of the organization. It should be emphasized that not every company is ready to embark on a venture that many consider a step of the future, but those that decide to do so can develop a competitive advantage over less flexible companies.

Through the analysis of exact data and information, it has been shown that the application of AI in the distribution of information brings significant improvements in speed, accuracy and personalization of the process and enables easier control of communication processes for the company's management. Automating routine tasks frees up a significant amount of time that can be used more efficiently for more creative work and making strategic decisions. Also, the implementation of advanced analytical algorithms enables a deeper understanding of user needs, which results in better customization of products or services.

It is important to point out that the success of AI implementation depends on the cooperation of all relevant employees within the organization. Educating employees about the benefits and applications of artificial intelligence is essential to ensure a smooth transition to a digitized environment. Also, constant monitoring of ethical standards in the use of data and algorithms is necessary to avoid potential problems and preserve user trust.

Most importantly, this paper emphasizes the importance of timely integration of artificial intelligence into business processes to optimize the distribution of information. Accelerating digital transformation is not only a technological imperative but also a key strategy for long-term sustainability and creating a competitive advantage. By managing this transformation wisely, organizations can expect to improve operational efficiency, increase customer satisfaction, and open new opportunities for innovation. The implementation of AI is not only a step towards the future, but also a response to the requirements of the modern business environment, which requires agility, adaptability, and high-quality distribution of information.

5 REFERENCES

[1] Artificial Intelligence and the Future of Defense, HCSS. https://hcss.nl/report/artificial-intelligence-and-the-future-of-defense/ (Accessed on Apr. 22, 2024)
[2] AI and Robotics Leading Industry 4.0. IEEE Conference Publication, IEEE Xplore. https://ieeexplore.ieee.org/document/10061911 (Accessed on Apr. 22, 2024)
[3] A critical review on applications of artificial intelligence in manufacturing. Artificial Intelligence Review. https://link.springer.com/article/10.1007/s10462-023-10535-y (Accessed on Apr. 22, 2024)
[4] Waschull, S., & Emmanouilidis, C. (2022). Development and application of a human-centric co-creation design method for AI-enabled systems in manufacturing. *IFAC-PapersOnLine, 55*(2), 516-521. https://doi.org/10.1016/j.ifacol.2022.04.246
[5] Vyhmeister, E., Castane, G. G., Buchholz, J., & Östberg, P.-O. (2022). Lessons learn on responsible AI implementation: the ASSISTANT use case. *IFAC-PapersOnLine, 55*(10), 377-382. https://doi.org/10.1016/j.ifacol.2022.09.422
[6] Alves, J., Lima, T., & Gaspar, P. (2023). Is Industry 5.0 a Human-Centred Approach? A Systematic Review. *Processes, 11*, p. 193. https://doi.org/10.3390/pr11010193
[7] Haindl, P., Hoch, T., Dominguez, J., Aperribai, J., Ure, N. K., & Tunçel, M. (2022). Quality Characteristics of a Software Platform for Human-AI Teaming in Smart Manufacturing. In: Vallecillo, A., Visser, J., & Pérez-Castillo, R. (eds) Quality of Information and Communications Technology. QUATIC 2022. *Communications in Computer and Information Science, vol 1621*. Springer, Cham. https://doi.org/10.1007/978-3-031-14179-9_1
[8] A Guide to Artificial Intelligence in the Enterprise. Enterprise AI. Accessed: Apr. 22, 2024. https://www.techtarget.com/searchenterpriseai/Ultimate-guide-to-artificial-intelligence-in-the-enterprise
[9] Wang, X. et al. (2023). Steps toward Industry 5.0: Building '6S' Parallel Industries with Cyber-Physical-Social Intelligence. *IEEE/CAA Journal of Automatica Sinica, 10*(8), 1692-1703. https://doi.org/10.1109/JAS.2023.123753
[10] The Duo of Artificial Intelligence and Big Data for Industry 4.0: Applications, Techniques, Challenges, and Future Research Directions. IEEE Journals & Magazine. *IEEE Xplore*. Accessed: Apr. 18, 2024. https://ieeexplore.ieee.org/document/9667102
[11] Wellsandt, S., Hribernik, K., & Thoben, K.-D. (2021). Anatomy of a Digital Assistant. In *Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems*. Dolgui, A., Bernard, A., Lemoine, D., von Cieminski, G., & Romero, D. Eds., Cham: Springer International Publishing, 321-330. https://doi.org/10.1007/978-3-030-85910-7_34
[12] Szajna, A., & Kostrzewski, M. (2022). AR-AI Tools as a Response to High Employee Turnover and Shortages in Manufacturing during Regular, Pandemic, and War Times. *Sustainability, 14*(11), 1-17. https://doi.org/10.3390/su14116729
[13] Franken, S., & Wattenberg, M. (2019). The Impact of AI on Employment and Organisation in the Industrial Working Environment of the Future. *Proceedings of the 1st European Conference on the Impact of Artificial Intelligence and Robotics*, Oxford.
[14] Berger, P., & von Garrel, J. (2023). How to design a value-based Chatbot for the manufacturing industry: An empirical study of an internal assistance for employees. Künstl Intell. https://doi.org/10.1007/s13218-023-00817-6

[15] AI-Based Assistance System for Manufacturing. IEEE Conference Publication, *IEEE Xplore*. Accessed: Apr. 18, 2024. https://ieeexplore.ieee.org/document/9921520
[16] Chandel, A., & Sharma, B. (2023). Technology Aspects of Artificial Intelligence: Industry 5.0 for Organization Decision Making. In I*nformation Systems and Management Science*. Garg, L., Sisodia, D. S., Kesswani, N., Vella, J. G., Brigui, I., Xuereb, P., Misra, S. & Singh, D., Eds., Cham: Springer International Publishing, 79-90. https://doi.org/10.1007/978-3-031-13150-9_7
[17] El-Brawany, M. A., Adel Ibrahim, D., Elminir, H. K., Elattar, H. M., & Ramadan, E. A. (2023). Artificial intelligence-based data-driven prognostics in industry: A survey. *Computers & Industrial Engineering, 184*, p. 109605. https://doi.org/10.1016/j.cie.2023.109605
[18] Ometto, M. (2022). An innovative approach to plant and process supervision, Danieli Intelligent Plant. *IFAC-PapersOnLine, 55*(40), 313-318. https://doi.org/10.1016/j.ifacol.2023.01.091
[19] Li, D.-C., Chen, L.-S., & Lin, Y.-S. (2003). Using Functional Virtual Population as assistance to learn scheduling knowledge in dynamic manufacturing environments. *International Journal of Production Research, 41*(17), 4011-4024. https://doi.org/10.1080/0020754031000149211
[20] Song, L. (2023). Business Intelligence (BI) and Big Data Analytics (BDA) in Industry 5.0: Application of Adaptive Optimization Algorithms (AOA) to Improve Firm Performance. *Transformations in Business & Economics, 22*(2), 45-63.
[21] Ahmed, T., Karmaker, C. L., Nasir, S. B., Moktadir, Md. A., & Paul, S. K. (2023). Modeling the artificial intelligence-based imperatives of industry 5.0 towards resilient supply chains: A post-COVID-19 pandemic perspective. *Computers & Industrial Engineering, 177*, 109055. https://doi.org/10.1016/j.cie.2023.109055
[22] Drissi Elbouzidi, A., Ait El Cadi, A., Pellerin, R., Lamouri, S., Tobon Valencia, E., & Bélanger, M.-J. (2023). The Role of AI in Warehouse Digital Twins: Literature Review. *Appl. Sci., 13*, 6746. https://doi.org/10.3390/app13116746
[23] Muslikhin, M., Horng, J.-R., Yang, S.-Y., Wang, M.-S., & Awaluddin, B.-A. (2021). An Artificial Intelligence of Things-Based Picking Algorithm for Online Shop in the Society 5.0's Context. *Sensors, 21*, 2813. https://doi.org/10.3390/s21082813
[24] Yang, J., Wang, Y., Wang, X., Wang, X., Wang, X., & Wang, F.-Y. (2024). Generative AI Empowering Parallel Manufacturing: Building a '6S' Collaborative Production Ecology for Manufacturing 5.0. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 1-15. https://doi.org/10.1109/TSMC.2024.3349555
[25] Loske, D., & Klumpp, M. (2021). Intelligent and efficient? An empirical analysis of human–AI collaboration for truck drivers in retail logistics. *The International Journal of Logistics Management, 32*(4), 1356-1383. https://doi.org/10.1108/IJLM-03-2020-0149
[26] Nicoletti, B., & Appolloni, A. (2023). Artificial Intelligence for the Management of Servitization 5.0. *Sustainability, 15*(14), Art. No. 14. https://doi.org/10.3390/su151411113
[27] Hadzovic, S., Mrdovic, S., & Radonjic, M. (2023). A Path towards an Internet of Things and Artificial Intelligence Regulatory Framework. *IEEE Communications Magazine, 61*(7), 90-96. https://doi.org/10.1109/MCOM.002.2200373
[28] Schmitt, M. (2023). Securing the digital world: Protecting smart infrastructures and digital industries with artificial intelligence (AI)-enabled malware and intrusion detection. *Journal of Industrial Information Integration, 36*, p. 100520. https://doi.org/10.1016/j.jii.2023.100520
[29] Mantravadi, S., Jansson, A. D., & Møller, C. (2020). User-Friendly MES Interfaces: Recommendations for an AI-Based Chatbot Assistance in Industry 4.0 Shop Floors. In *Intelligent Information and Database Systems*. Nguyen, N. T., Jearanaitanakij, K., Selamat, A., Trawiński, B., & Chittayasothorn, S. Eds., Cham: Springer International Publishing, 189-201. https://doi.org/10.1007/978-3-030-42058-1_16
[30] Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion, 99*, p. 101896. https://doi.org/10.1016/j.inffus.2023.101896
[31] Vernim, S., Bauer, H., Rauch, E., Ziegler, M. T., & Umbrello, S. (2022). A value sensitive design approach for designing AI-based worker assistance systems in manufacturing. *Procedia Computer Science, 200*, 505-516. https://doi.org/10.1016/j.procs.2022.01.248
[32] Chen, W., He, W., Shen, J., Tian, X., & Wang, X. (2023). Systematic analysis of artificial intelligence in the era of industry 4.0. *Journal of Management Analytics, 10*(1), 89-108. https://doi.org/10.1080/23270012.2023.2180676

Authors' contacts:

Hrvoje Kober, mag. ing. mech.
Faculty of Mechanical Engineering and Naval Architecture,
Ivana Lucica 5, 10 000 Zagreb, Croatia
hrvojekober98@gmail.com

Maja Trstenjak, dr. sc.
(Corresponding author)
Faculty of Mechanical Engineering and Naval Architecture,
Ivana Lucica 5, 10 000 Zagreb, Croatia
maja.trstenjak@fsb.unizg.hr

Tihomir Opetuk, assoc. prof. dr. sc.
Faculty of Mechanical Engineering and Naval Architecture,
Ivana Lucica 5, 10 000 Zagreb, Croatia
tihomir.opetuk@fsb.unizg.hr

Hrvoje Cajner, assoc. prof. dr. sc.
Faculty of Mechanical Engineering and Naval Architecture,
Ivana Lucica 5, 10 000 Zagreb, Croatia
hrvoje.cajner@fsb.unizg.hr

Goran Đukić, prof. dr. sc.
Faculty of Mechanical Engineering and Naval Architecture,
Ivana Lucica 5, 10 000 Zagreb, Croatia
goran.dukic@fsb.unizg.hr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240624075046

Subject review
Received: 2024-06-24, Accepted: 2024-11-07

Generative AI in Education: Comparative Analysis of Free Presentation Tools for Teachers

TamaraReđep, Andrija Bernik*

Abstract: Since a significant amount of time and effort teachers spend on creating presentation, Generative Artificial Intelligence (GenAI) tools are used to automate many of the tasks associated with creating slides. Given the variety of GenAI presentation tools, the goal of this research was to identify those that are available online and offer free usage plans. Then compare and analyse them according to the identified features and recognized limitations, as well as the quality of the generated content. The obtained results indicate differences in the available free plans of the analysed tools, but the choice of a GenAI presentation tool depends on a number of different factors. However, teachers may be encouraged to use these tools mostly by the fact that they can create presentations in just a few minutes, possibly only by entering some additional options and instructions.

Keywords: Generative Artificial Intelligence; GenAI tools; Presentation; Slides; Teachers

1 INTRODUCTION

By embracing technological advances, education can remain relevant and effectively respond to the challenges of the digital world, thus preparing students for the demands of the 21st century. The field of artificial intelligence (AI) in education is focused on the research, development and evaluation of computer software that improves teaching and learning [1] and opens new possibilities such as personalized learning experiences and adaptation of educational materials to the individual needs of students [2]. Although teachers still have the main responsibility of teaching in any educational setting, artificial intelligence applications (AIA) are not only assisting education academically and administratively but also enhance their effectiveness [3].

Chen, Chen and Lin found that AI is already widely accepted and used in education. They confirm that with the help of AI, teachers were able to perform various administrative tasks more efficiently, such as reviewing and evaluating student assignments, and thus achieved a higher quality of their teaching activities. Also, the curriculum and teaching content can be adapted and personalized according to the needs of the students, thus improving the overall quality of learning [4]. Given this, for academic performance to be excellent, it is crucial to prioritize AI in education and implement its appropriate strategies to meet the needs and expectations of teachers and students through AI technologies [5].

However, since AI is still a relatively new field in education, it is necessary to encourage teachers to apply it by presenting them with the potential advantages that AI offers for both learning and teaching, as well as for their overall work in the classroom. Given that it is known that teachers often use presentations in their work, this paper will explore the possibilities of currently available AI tools that can generate presentations in just a few minutes, with the aim of making it easier for teachers to choose and use the tools.

2 GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS

The term "generative artificial intelligence" (GenAI) refers to computing techniques that are capable of generating seemingly new, meaningful content such as text, images or audio from the data they have been trained on. Already well-known tools such as Dall-E, Chat GPT, Copilot or Gemini are changing the way we work and communicate with each other due to their availability [6]. GenAI is quite different from other technologies of the last 20 years, because of its ability to generate original work that is almost indistinguishable from that of human authors. Given these capabilities, the question arises whether GenAI technology will destroy education, as we know it now or whether it will solve all the problems of education. However, although is unlikely to destroy education, it can destroy the legitimacy of some long-standing educational practices [7].

For the time being, available research on GenAI refers mostly to higher education. In his paper, Chiu examines the impact of GenAI, specifically how Chat GPT and Midjourney tools affect education in four domains – learning, teaching, assessment and administration. The results provide three suggestions for practice: know-it-all attitude, new prerequisite knowledge, interdisciplinary teaching, and three implications for policy: new assessment, AI education, and professional standards [8]. Ruiz-Rojas et al. also with their the results obtained for higher education show that GenAI tools have significant potential in education, and this especially applies to their use in combination with an instructional design matrix for the development of massive MOOC virtual classrooms [2]. In their work, Kaplan-Rakowski, Grotewold, Hartwick, and Papin show that teachers express positive attitudes toward GenAI (or GAI) regardless of their teaching style. In addition, they show that the more often teachers used GenAI, the more positive their attitudes were. They believe that GenAI can advance their professional development and be a valuable tool for students [9].

3 CREATING PRESENTATIONS USING GENAI TOOLS

In education, presentations are often used in teaching. Although the use of other programs such as the Google Slides platform should not be neglected, the PowerPoint program is still the most widely used. Previous research on the use of PowerPoint in the classroom refers to student acceptance of PowerPoint as a measure of its effectiveness, and the results have overwhelmingly shown that students like PowerPoint.

In addition, PowerPoint has the potential to enhance learning, but only if we first learn how to use it effectively [10].

Creating presentation materials requires complex skills to summarize key concepts and arrange them in a logical and visually appealing way, so the question is whether computers can mimic this process [11]. A considerable amount of time and effort is spent on preparing presentation slides, and an automated slide generator can help save time, effort, and consequently costs. Currently, tools such as Microsoft PowerPoint and Open Office help provide an outline and theme for the slides, but do not help select the content of the slides [12].

Although there is relatively little research on GenAI tools related to presentation creation, the available works describe new approaches in their creation. Ganguly and Joshi propose an automated technique to generate presentation slides from a text document (text or pdf) so that the original concepts in the input document are conveyed to the output slides. The paper focuses on the use of different aspects of machine learning and text mining. All slides extracted in this way were found to be appropriate and satisfactory according to the user's context [13]. Fu, Wang, McDuff, and Song, on the other hand, present a way of working that includes document summarization, image and text retrieval, and a slide structure to arrange key elements in a form suitable for presentation. They propose a hierarchical sequence-to-sequence approach that exploits the inherent structures within documents and slides and includes paraphrasing and layout prediction modules for slide generation. They show that such an approach produces slides with rich content and aligned imagery [11].

Several papers describe the generation of presentations from scientific papers. Thus, Hu and Wan propose a new system called PPSGen that uses regression methods to learn the importance scores of the sentence in a paper, and then uses an integer linear programming (ILP) method to generate well-structured slides by selecting and aligning key phrases and sentences. The results show that this method can generate slides with much better quality than traditional methods [14]. Shaj and John also propose an automated system that generates presentation slides with the help of summarization, but from scientific papers in PDF format. The papers were summarized using the Google BERT algorithm. The PDF of the scientific paper is uploaded into the system, which will allow only the important sentences to be included on the slides. Thus, the summarized content of a long scientific paper is obtained, which can then be used in presentation slides [12]. Wang, Wan, and Du propose a phrase-based approach to generate well-structured and concise presentation slides for academic papers. Phrases are extracted from the given paper, and then the salience of each phrase and the hierarchical relationship between a pair of phrases are learned. Finally, a greedy algorithm is used to select and align the salient phrase. The evaluation results confirm the effectiveness of this proposed approach [15].

It is possible to conclude that the main goal of applying GenAI presentation tools is to create a visually appealing and dynamic presentation by automating various aspects of the slideshow, including slide design, content suggestions and data analysis, thus making the whole process faster and more user-friendly [16].

4 RESEARCH METHODOLOGY

It has already been mentioned that teachers very often use presentations for teaching purposes to make the instructional content more accessible and clearer to students. However, creating presentations requires additional effort from teachers. Therefore, applying a new approach to creating presentations using GenAI tools for teaching purposes would certainly contribute to allowing the teacher to focus more on presenting the content rather than on the design and other interactive elements of the presentation.

Since there are currently no available scientific papers that study and compare the use of GenAI tools for creating presentations, the primary goal of this research is to identify those that are available online, compare them, and analyze them based on identified features and recognized limitations, as well as the quality of the generated content.

However, to avoid financial costs for teachers or schools due to the application of GenAI tools, only tools with available free usage plans will be analyzed. This will provide insight into the possibility of using the free versions of these tools. The free usage plan in this research does not refer to the availability of a free trial plan, as it has a limited duration. Additionally, another criterion to be considered is that the tools must be exclusively available online, meaning their application does not require any additional installation. For this reason, GenAI tools for creating presentations for Google Slides will not be analyzed, as their use requires the installation of an extension.

Given that the search for tools has encountered the most commonly defined and mentioned features (F) that make a good AI presentation tool, the analysis of the tools will be carried out according to these features. These are:

- **F1-User-friendly interface**: a user interface that allows users to easily navigate and use its features without additional training
- **F2-Media Integration**: the incorporate of various multimedia elements such as images, videos and animations that increase engagement and help convey complex ideas
- **F3-Customization**: customization of templates, layout, fonts and colours
- **F4-Collaboration**: things like real-time editing, comments, and version history
- **F5-AI Enhancements**: smart suggestions for content organization, automated design suggestions, and speech-to-text capabilities [17]
- **F6-Integration with popular software**: export to programs such as Microsoft PowerPoint, Google Slides
- **F7-Tutorials**: possibility of support [18]

Based on the previously established research goal, three research questions were defined:

- **RQ1:** What are the differences in the free versions of GenAI tools for creating presentations with regard to the most common features?
- **RQ2:** What limitations can be identified in the free versions of GenAI tools for creating presentations?
- **RQ3:** What are the differences in the content of presentations created using GenAI tools?

The research was conducted by the authors of this paper. The first author,Ređep T., is a primary school computer science teacher with more than 20 years of professional experience. She is also involved in scientific research, examining the effects of innovative teaching strategies and digital technologies in education, which she personally implements. The second author is Bernik A., who has a Ph.D. in the field of social sciences, and one of his research interests is precisely artificial intelligence, especially GenAI tools. The authors conducted their research in the period from April to May 2024.

The search for tools was conducted by entering the term "AI presentation maker" into a web search engine to narrow the search options. This step did not yield a comprehensive list of tools but instead displayed individual tool websites. During this step, tools were listed and identified based on their payment plans. Furthermore, the word "list" was added to the previously entered term, and the search was conducted using the keywords: AI presentation maker list. This method provided much clearer results, highlighting the most popular tools as well as websites listing these tools at the top of the search engine results. In this step, another verification was performed, and the list of tools was supplemented.

Regarding the creation of the presentations themselves and the analysis of the tools, it was first necessary to log into the tool's website, which was done via a Google account or by entering an email address. To verify the newly generated content, the content obtained by entering the phrase "artificial intelligence in education" was analysed in this study. Since the authors are from Croatia, the goal was to obtain a presentation in Croatian. Another goal was for the presentation to have up to 10 slides, so for those tools that had this option, the number of slides was set to fewer than 10. Furthermore, the tools were analysed based on the previously defined features and the limitations they possess, as well as the generated content.

5 RESULTS AND DISCUSSION

To enable further analysis, the first result of this research is a list of currently available online GenAI tools for creating presentations that have been selected based on whether they offer a free usage plan. The obtained results are shown in Tab. 1.

It is important to note that the list of tools is frequently updated and that current free plans are subject to change, not only regarding payment but also in terms of the features they offer. Therefore, these results are relevant only for the period during which the research was conducted.

Research results show that out of 31 available GenAI presentation tools, 14 meet the predefined criteria, indicating that there are slightly more tools that do not offer a free usage plan. The initial analysis of the selected 14 tools was conducted by checking whether the tool meets the set condition, i.e., whether it possesses a specific feature. The features according to which the analysis was conducted have already been listed and described, and they are F1-intuitive interface, F2-media integration (images, videos, animations), F3-slide customization, F4-collaboration, F5-AI enhancements, F6-software integration (ppt), and F7-tutorials. After the analysis, each tool was attributed with the mentioned feature if it possessed it.

Table 1 GenAI tools for creating presentations with paid and free usage plan

Paid	Free
Appy Pie	Canva
AutoSlide	Gamma
Beautiful AI	Pitch
DeckRobot	PopAI
Dectopus	Sendsteps
Design.AI	Simplified
Kroma	Slidecast
Plus AI	SlideMake
Presentation.AI	Slidesgo
Slidebean	SlidesGPT
SlideModel	SlidesPilot
Samllppt	SlideSpeak
Storydoc	Visme
Tome	WePik
TypeSet	
WeSlides	
Wonderslide	

We can conclude that all analysed tools have an intuitive interface that makes it easier for users to find their way in online platforms, and most of them also have the option of providing help and support if the user encounters a problem, which makes their use easier, therefore these tools are not intended only for teachers who have IT skills. In addition, most of the tools have the option of including various multimedia elements, as well as the further possibility of customization the slides, which makes the resulting presentations more interesting, which is important when stimulating students' interest in the teaching content. A small number of tools have the possibility of additional improvement of content and elements with the application of AI and collaborative activities on the generated presentation. These options should not be essential for the creation of quality presentations and the use of the tool by teachers, but only depend on their preferences. However, the least available option is the integration with the software, that is, the export of the presentation in a format that allows for further editing, such as ppt format. Unfortunately, most of the analysed tools have this option, but not in the free plan.

These results on the differences in GenAI presentation tools in their free versions regarding the most common identified features answer the first research question and are shown in Tab. 2.

Table 2 Comparison of GenAI presentation tools according to defined features

GenAI tool	F1	F2	F3	F4	F5	F6	F7
Canva	+	+	+	+	+	+	+
Gamma	+	+	+	+	+	+	+
Pitch	+	+	+	+		+	+
PopAI	+	+	+		+		
Sendsteps	+	+	+		+		+
Simplified	+	+	+	+	+		+
Slidecast	+	+	+				+
SlidesMake						+	
Slidesgo	+	+	+				+
SlidesGPT	+						
SlidesPilot	+		+		+		+
SlideSpeak	+						+
Visme	+	+	+	+			+
WePik	+	+	+				+

During the application of the tools and the generation of content, it was observed that the tools in their free versions, in addition to the mentioned features, also differ in some additional limitations that could affect the teacher's decision on their use. The ability to create a number of presentations proved to be particularly important. The vast majority of tools in the free plan have a limited number of presentations and slides or AI credits. However, a slightly bigger problem is that even after deleting the files, some tools do not allow the creation of new presentations, but refer to some of the paid plans. In addition, teachers could be interested in the possibility of analytics that certain tools have, because in this way they would monitor the activities when making their presentations in the classrooms. However, unfortunately, this possibility, if the tools have it at all, is fully available in the paid versions. Although the possibility of integration with the software and download in ppt format has already been analyzed previously, some tools allow downloading presentations in other formats as well, e.g. pdf, png, jpg, which can be important for teachers if they intend to edit the presentation additionally or just save it. In addition, it is limiting that only half of the analyzed tools allow the creation of content by uploading documents such as word, pdf documents, or even ppt presentations, considering that this option also reduces the time for teachers to create a presentation.

As for the actual presentation, there are also some limitations. As previously stated, although some tools enable the creation of content by uploading a document, this paper analyzed presentations created only by entering citations, considering that all tools have this possibility. But even here there are differences according to additional options that are entered before creating the presentation itself, such as the number of slides, number of words, type of audience, topic, style of presentation and presentation, duration and the like, which is visible in more than half of the analyzed tools. Additionally, some tools have the ability to modify the text before the final presentation is generated, so that the text can be checked, reducing corrections in the finished presentation.

Table 3 Comparison of GenAI presentation tools with regard to recognized limitations

GenAI	Limits	Upload	Detailed creation	Analytics	Export
Canva	10 slides	-	-	+	ppt, jpg, pdf, pdf, mp4
Gamma	400 credits, 10 (cards)	word, ppt, Google Slides/Docs	+	basic	ppt, pdf (Gamma branded)
Pitch	-	ppt	-	-	ppt, pdf (Pitch branded)
PopAI	1 project, 15 pages	word, pdf			
Sendsteps	2 ppt, 15 slides	pdf, pptx, docx, txt	+	-	-
Simplified	-	-	-	-	jpg, png, video, gif
Slidecast	-	-	-	-	pdf
SlidesMake	-	-	-	-	ppt
Slidesgo	14 slides	-	+	-	jpg, pdf
SlidesGPT	-	-	-	-	-
SlidesPilot	1 ppt	pdf, word	+	-	-
SlideSpeak	1 ppt	pdf, word, ppt, xls	+	-	-
Visme	10 AI credit limits	-	+	-	jpg, png, pfd
WePik	1 ppt	pdf, word, ppt, xls	+	-	-

Table 4 Comparison of GenAI presentation tools with regard to the resulting content of the presentation

GenAI tool	Extensive content	Croatian language
Canva	-	+ (translation)
Gamma	+	+
Pitch	-	-
PopAI	+	-
Sendsteps	+	+
Simplified	-	-
Slidecast	-	+
SlidesMake	+	-
Slidesgo	+	+
SlidesGPT	+	+
SlidesPilot	+	+
SlideSpeak	-	-
Visme	-	-
WePik	+	+

The identified limitations are the answer to the second research question and are shown in the Tab. 3.

In order to get an answer to the third research question, the content obtained in the presentations was analyzed. Despite the statement being in Croatian, half of the analyzed tools generated content in English. Due to the necessary translation, the application of these tools would be difficult for teachers in Croatia. In addition, differences were observed regarding the comprehensiveness and detail of the content. The results show that half of the tools provide very detailed content that hardly needs to be supplemented. This is very important because in this way teachers do not have to spend time on additional corrections and supplementing the text in presentations. No necessary connection was observed between the introduction of additional options when creating the presentation and the resulting content in terms of detail and extensiveness

Recognized differences with regard to the resulting content of the presentation are shown in the Tab. 4.

6 CONCLUSION

There is an increasingly urgent need for educators and researchers in the field of education to propose suitable ways and practices with the help of which students and teachers will be able to smoothly absorb the upcoming changes in education, especially with respect to the teaching-learning process [19].

One of the possibilities by which teachers can successfully respond to the challenges of the new age is using AI. As most teachers use presentations in their work, with the help of GenAI presentation tools they can create a visually attractive and dynamic presentation in such a way that various processes are automated, and then they have more time to focus on the content itself. However, choosing an AI presentation tool depends on a number of different factors such as existing presentation tools, typical presentation style, and how someone like to work with coworkers and colleagues [20].

The results of this research show that now there are a large number of tools on the market that enable the creation of presentations using AI. Almost half of the recognized tools listed in this paper offer a free usage plan, and all tools can be accessed via a Google account or by entering an email. The fact that presentations can be created in just a few minutes by uploading a document or just by entering the topic of the presentation and possibly some additional options and instructions could encourage teachers to use the tool. However, depending on the selected tool, the presentations will differ in terms of the extensive and detail of the content, as well as some possibilities with which the presentation can be supplemented, such as, for example, the integration of various multimedia elements. What could demotivate teachers the most to use these tools in the free usage plan is the limited number of presentations and slides, and the language of the content in the resulting presentations. However, the goal of the free usage plan is to show the basic capabilities of the tool itself so that users will be interested in purchasing it, but further decisions about use still depend on their preferences and needs. Considering that each tool has several pricing plans of usage, it is up to the teacher to evaluate which of the plans suits him best for application in the teaching process.

7 REFERENCES

[1] Woolf, B. P. (2015). AI and Education: Celebrating 30 years of Marriage. *CEUR Workshop Proceedings*. 1432. 38-47.

[2] Ruiz-Rojas, L. I., Acosta-Vargas, P., De-Moreta-Llovet, J., & Gonzalez-Rodriguez, M. (2023). Empowering education with generative artificial intelligence tools: Approach with an instructional design matrix. *Sustainability, 15*(15), 11524. https://doi.org/10.3390/su151511524

[3] Ahmad, S. F., Alam, M. M., Rahmat, M. K., Mubarik, M. S., & Hyder, S. I. (2022). Academic and administrative role of artificial intelligence in education. *Sustainability, 14*(3), 1101. https://doi.org/10.3390/su14031101

[4] Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8*, 75264-75278. https://doi.org/10.1109/ACCESS.2020.2988510

[5] Limna, P., Jakwatanatham, S., Siripipattanakul, S., Kaewpuang, P., & Sriboonruang, P. (2022). A review of artificial intelligence (AI) in education during the digital era. *Advance Knowledge for Executives, 1*(1), 1-9.

[6] Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering, 66*(1), 111-126. https://doi.org/10.1007/s12599-023-00834-7

[7] Hodges, C. B., & Kirschner, P. A. (2024). Innovation of Instructional Design and Assessment in the Age of Generative Artificial Intelligence. *TechTrends, 68*(1), 195-199. https://doi.org/10.1007/s11528-023-00926-x

[8] Chiu, T. K. (2023). The impact of Generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. *Interactive Learning Environments*, 1-17. https://doi.org/10.1080/10494820.2023.2253861

[9] Kaplan-Rakowski, R., Grotewold, K., Hartwick, P., & Papin, K. (2023). Generative AI and teachers' perspectives on its implementation in education. *Journal of Interactive Learning Research, 34*(2), 313-338.

[10] Jordan, L. A., & Papp, R. (2014). Powerpoint®: It's Not "Yes" or "No"--It's "When" and "How". *Research in Higher Education Journal, 22*.

[11] Fu, T. J., Wang, W. Y., McDuff, D., & Song, Y. (2022, June). Doc2ppt: Automatic presentation slides generation from scientific documents. In *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(1), 634-642. https://doi.org/10.1609/aaai.v36i1.19943

[12] Shaj, K., John, S. S. Roshney, P., & Anish. G. (2020). Learning Based Slide Generator. *International Journal of Engineering Research & Technology (IJERT), 9*(7). https://doi.org/10.17577/IJERTV9IS070415

[13] Ganguly, P., & Joshi, P. M. (2016). IPPTGen-intelligent PPT generator. *International Conference on Computing, Analytics and Security Trends (CAST2016)*, 96-99. https://doi.org/10.1109/CAST.2016.7914947

[14] Hu, Y., & Wan, X. (2013, June). Ppsgen: learning to generate presentation slides for academic papers. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

[15] Wang, S., Wan, X., & Du, S. (2017, February). Phrase-based presentation slides generation for academic papers. In *Proceedings of the AAAI Conference on Artificial Intelligence. 31*(1). https://doi.org/10.1609/aaai.v31i1.10481

[16] Top 13 AI Presentation Makers 2024: Key Features and Pricing (2023). Retrieved April 22, 2024, from https://slidecast.com/best-ai-presentation-makers/

[17] Pascual, K. (2024). 7 AI Presentation Tools to Design Your Slides in Minutes. Retrieved April 25, 2024, from https://penji.co/ai-presentation/

[18] McLean, D. (2024). 6 Best AI Presentation Tools in 2024 (Compared). Retrieved April 2, from: https://www.elegantthemes.com/blog/business/best-ai-presentation-tools

[19] Voskoglou, M. (2023). Artificial Intelligence and Digital Technologies in the Future Education. *Qeios, CC-BY, 4*. https://doi.org/10.32388/07VE29

[20] Dan. (2024). Best AI Presentation Makers of 2024 (with example outputs). Retrieved April 25, 2024, from https://www.plusdocs.com/blog/best-ai-presentation-makers

Authors' contacts:

Tamara Rеđep
1st Elementary School Varaždin,
Kralja P. Krešimir IV. 10, 42000 Varaždin, Croatia

Andrija Bernik
(Corresponding author)
University North,
104. brigade 1, 42000 Varazdin, Croatia
abernik@unin.hr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20...........

Categorization of paper
Received: y-m-d, Accepted: y1-m1-d1

14pt
14pt

Article Title Only in English (Style: Arial Narrow, Bold, 14pt)

14pt

Ivan Horvat, Thomas Johnson, Marko Marić (Style: Arial Narrow, Normal, 10pt)

14pt

Abstract: Article abstract contains maximum of 150 words and is written in the language of the article. The abstract should reflect the content of the article as precisely as possible. TECHNICAL JOURNAL is a trade journal that publishes scientific and professional papers from the domain(s) of mechanical engineering, electrical engineering, civil engineering, multimedia, logistics, etc., and their boundary areas. This document must be used as the template for writing articles so that all the articles have the same layout. (Style: Arial Narraw, 8pt)

8pt

Keywords: keywords in alphabetical order (5-6 key words). Keywords are generally taken from the article title and/or from the abstract. (Style: Arial Narraw, 8pt)

10pt
10pt

1 INTRODUCTION (Article Design)

(Style: Arial Narrow, Bold, 10pt)

10pt

(Tab 6 mm) The article is written in Latin script and Greek symbols can be used for labelling. The length of the article is limited to eight pages of international paper size of Letter (in accordance with the template with all the tables and figures included). When formatting the text the syllabification option is not to be used.

10pt

1.1 Subtitle 1 (Writing Instructions)

(Style: Arial Narrow, 10pt, Bold, Align Left)

10pt

The document format is Letter with margins in accordance with the template. A two column layout is used with the column spacing of 10 mm. The running text is written in Times New Roman with single line spacing, font size 10 pt, alignment justified.

Article title must clearly reflect the issues covered by the article (it should not contain more than 15 words).

Body of the text is divided into chapters and the chapters are divided into subchapters, if needed. Chapters are numbered with Arabic numerals (followed by a period). Subchapters, as a part of a chapter, are marked with two Arabic numerals i.e. 1.1, 1.2, 1.3, etc. Subchapters can be divided into even smaller units that are marked with three Arabic numerals i.e. 1.1.1, 1.1.2, etc. Further divisions are not to be made.

Titles of chapters are written in capital letters (uppercase) and are aligned in the centre. The titles of subchapters (and smaller units) are written in small letters (lowercase) and are aligned left. If the text in the title of the subchapter is longer than one line, no hanging indents.

10pt

Typographical symbols (bullets), which are being used for marking an item in a list or for enumeration, are placed at a beginning of a line. There is a spacing of 10pt following the last item:

- Item 1
- Item 2
- Item 3

10pt

The same rule is valid when items are numbered in a list:

1) Item 1
2) Item 2
3) Item 3

10pt

1.2 Formatting of Pictures, Tables and Equations

(Style: Arial Narrow, 10pt, Bold, Align Left)

10pt

Figures (drawings, diagrams, photographs) that are part of the content are embedded into the article and aligned in the centre. In order for the figure to always be in the same position in relation to the text, the following settings should be defined when importing it: text wrapping / in line with text.

Figures must be formatted for graphic reproduction with minimal resolution of 300 dpi. Pictures downloaded from the internet in ratio 1:1 are not suitable for print reproduction because of unsatisfying quality.

10pt

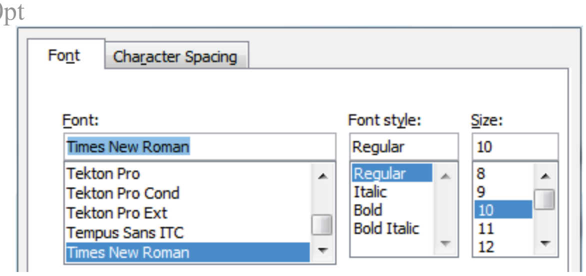


Figure 1 Text under the figure [1]
(Style: Arial Narraw, 8pt, Align Centre)

10pt

The journal is printed in black ink and the figures have to be prepared accordingly so that bright tones are printed in a satisfactory manner and are readable. Figures are to be in colour for the purpose of digital format publishing. Figures in the article are numbered with Arabic numerals (followed by a period).

Text and other data in tables are formatted - Times New Roman, 8pt, Normal, Align Center.

When describing figures and tables, physical units and their factors are written in italics with Latin or Greek letters, while the measuring values and numbers are written upright.

10pt

Table 1 Table title aligned centre
(Style: Arial Narraw, 8pt, Align Centre)

	1	2	3	4	5	6
ABC	ab	ab	ab	ab	ab	ab
DEF	cd	cd	cd	cd	cd	cd
GHI	ef	ef	ef	ef	ef	ef

10 pt

Equations in the text are numbered with Arabic numerals inside the round brackets on the right side of the text. Inside the text they are referred to with equation number inside the round brackets i.e. ".... from Eq. (5) follows ...." (Create equations with MathType Equation Editor - some examples are given below).

10pt

$$F_{\text{avg}}(t,t_0)=\frac{1}{t}\int_{t_0}^{t_0+t}F\left[q(\tau),p(\tau)\right]\mathrm{d}\tau, \tag{1}$$

$$\cos\alpha+\cos\beta=2\cos\frac{\alpha+\beta}{2}\cdot\cos\frac{\alpha-\beta}{2}, \tag{2}$$

$$(\boldsymbol{AB})^{\mathrm{T}}=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}, \tag{3}$$

$$AAMC=\frac{1}{n}\sum_{i=1}^{n}PVMC_i. \tag{4}$$

10pt

Variables that are used in equations and also in the text or tables of the article are formatted as *italics* in the same font size as the text.

10pt

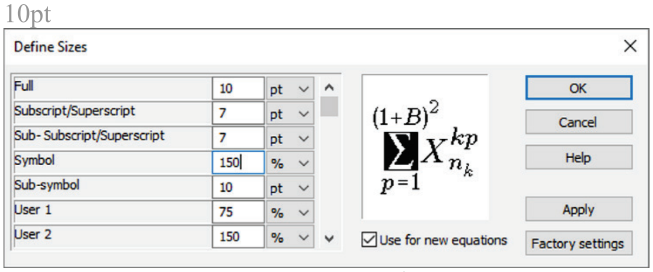


Figure 2 The texts under figures
(Style: Arial Narraw, 8pt, Align Centre)

10pt

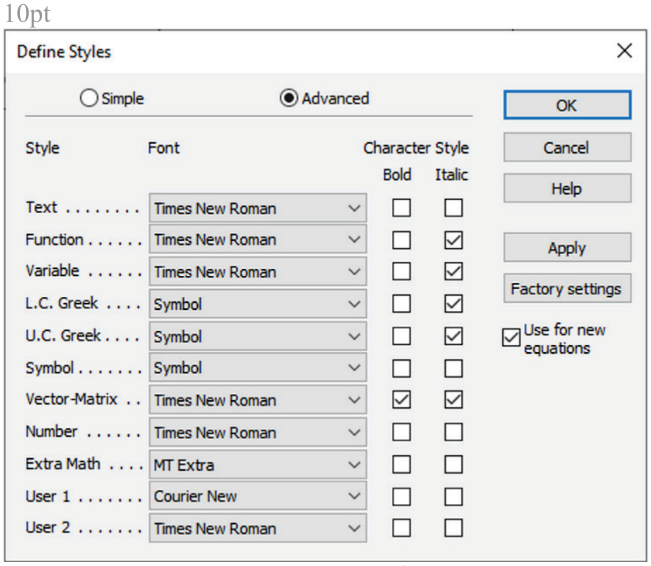


Figure 3 The texts under figures
(Style: Arial Narraw, 8pt, Align Centre)

Figures and tables that are a part of the article have to be mentioned inside the text and thus connected to the content i.e. " … as shown in Fig. 1…" or "data from Tab. 1…" and similar.

Latin or Greek characters in italics are used for physical symbols and normal characters for measuring units and numerical values. Text in figures is also written with normal letters (see obligatory Quantities and Units: ISO 80.000 - Part 1 and Part 2). Character size is to be chosen on the basis of the following criteria: after expected figure size reduction a capital Latin character should be about 2 mm high (no less than 6pt).

Tables are created with the word processing program. Each table is positioned in the desired place in the text. In the case of decimal numbers use comas (e.g. 0,253); use a small gap separating the thousands (e.g. 25 000, but not in the case of 1500).

10pt

2 PRELIMINARY ANNOTATION

10pt

Article that is offered for publication cannot be published beforehand, be it in the same or similar form, and it cannot be offered at the same time to a different journal. Author or authors are solely responsible for the content of the article and the authenticity of information and statements written in the article.

Articles that are accepted for publishing are classified into four categories: original scientific papers, preliminary communications, subject reviews and professional papers.

Original scientific papers are articles that according to the reviewer and the editorial board contain original theoretical or practical results of research. These articles need to be written in such a way that based on the information given, the experiment can be repeated and the results described can be achieved together with the author's observations, theoretical statements or measurements.

Preliminary communication contains one or more pieces of new scientific information, but without details that allow recollection as in original scientific papers. Preliminary communication can give results of an experimental research, results of a shorter research or research in progress that is deemed useful for publishing.

Subject review contains a complete depiction of conditions and tendencies of a specific domain of theory, technology or application. Articles in this category have an overview character with a critical review and evaluation. Cited literature must be complete enough to allow a good insight and comprehension of the depicted domain.

Professional paper can contain a description of an original solution to a device, assembly or instrument, depiction of important practical solutions, and similar. The article need not be related to the original research, but it should contains a contribution to an application of known scientific results and their adaptation to practical needs, so it presents a contribution to spreading knowledge, etc.

Outside the mentioned categorization, the Editorial board of the journal will publish articles of interesting content in a special column. These articles provide descriptions of practical implementation and solutions from the area of production, experiences from device application, and similar.

10pt

3 WRITING AN ARTICLE

10pt

Article is written in the English language and the terminology and the measurement system should be adjusted to legal regulations, standards and the International System of Units (SI) (Quantities and Units: ISO 80 000 - from Part 1 to Part 14). The article should be written in third person.

Introduction contains the depiction of the problem and an account of important results that come from the articles that are listed in the cited literature.

Main section of the article can be divided into several parts or chapters. Mathematical statements that obstruct the reading of the article should be avoided. Mathematical statements that cannot be avoided can be written as one or more addendums, when needed. It is recommended to use an example when an experiment procedure, the use of the work in a concrete situation or an algorithm of the suggested method must be illustrated. In general, an analysis should be experimentally confirmed.

Conclusion is a part of the article where the results are being given and efficiency of the procedure used is emphasized. Possible procedure and domain constraints where the obtained results can be applied should be emphasized.

AI and AI-assisted tools do not qualify for authorship under TG/TJ's authorship policy. Authors who use AI or AI-assisted tools during the manuscript writing process are asked to disclose their use in a separate section of the manuscript. The publishing agreement process works as usual, with the authors keeping the copyright to their own work.

10pt

4 RECAPITULATION ANNOTATION

10pt

In order for the articles to be formatted in the same manner as in this template, this document is recommended for use when writing the article. Finished articles written in MS Word for Windows and formatted according to this template must be submitted using our The Paper Submission Tool (PST) (https://tehnickiglasnik.unin.hr/authors.php) or eventually sent to the Editorial board of the Technical Journal to the following e-mail address: tehnickiglasnik@unin.hr

The editorial board reserves the right to minor redaction corrections of the article within the framework of prepress procedures. Articles that in any way do not follow these authors' instructions will be returned to the author by the editorial board. Should any questions arise, the editorial board contacts only the first author and accepts only the reflections given by the first author.

10pt

5 REFERENCES (According to APA)

10pt

The literature is cited in the order it is used in the article. No more than 35 references are recommended. Individual references from the listed literature inside the text are addressed with the corresponding number inside square brackets i.e. "… in [7] is shown …". If the literature references are web links, the hyperlink is to be removed as shown with the reference number 8. Also, the hyperlinks from the e-mail addresses of the authors are to be removed. In the literature list, each unit is marked with a number and listed according to the following examples (omit the subtitles over the references – they are here only to show possible types of references):

9pt

[1] See http://www.bibme.org/citation-guide/apa/

[2] See http://sites.umuc.edu/library/libhow/apa_examples.cfm

[3] (Style: Times New Roman, 9pt, according to APA)

[4] Amidzic, O., Riehle, H. J., & Elbert, T. (2006). Toward a psychophysiology of expertise: Focal magnetic gamma bursts as a signature of memory chunks and the aptitude of chess players. *Journal of Psychophysiology, 20*(4), 253-258. https://doi.org/10.1027/0269-8803.20.4.253

[5] Reitzes, D. C., & Mutran, E. J. (2004). The transition to retirement: Stages and factors that influence retirement adjustment. *International Journal of Aging and Human Development, 59*(1), 63-84. Retrieved from http://www.baywood.com/journals/PreviewJournals.asp?Id=0091-4150

[6] Jans, N. (1993). *The last light breaking: Life among Alaska's Inupiat Eskimos*. Anchorage, AK: Alaska Northwest Books.

[7] Miller, J. & Smith, T. (Eds.). (1996). *Cape Cod stories: Tales from Cape Cod, Nantucket, and Martha's Vineyard*. San Francisco, CA: Chronicle Books.

[8] Chaffe-Stengel, P., & Stengel, D. (2012). *Working with sample data: Exploration and inference*. https://doi.org/10.4128/9781606492147

[9] Freitas, N. (2015, January 6). People around the world are voluntarily submitting to China’s Great Firewall. Why? Retrieved from http://www.slate.com/blogs/future_tense/2015/01/06/tencent_s_wechat_worldwide_internet_users_are_voluntarily_submitting_to.html

(Style: Times New Roman, 9pt, according to APA)

10pt

10pt

Authors’ contacts:

8pt

Full Name, title
Institution, company
Address
Tel./Fax, e-mail

8pt

Full Name, title
Institution, company
Address
Tel./Fax, e-mail

Note: Gray text should be removed in the final version of the article because it is for guidance only.

ECCC 2026

7th International Creep & Fracture Conference
18/20 May, Aix-en-Provence (France)

SAVE THE DATE!

EUROPEAN CREEP COLLABORATIVE COMMITTEE IS HAPPY TO INVITE YOU TO PARTICIPATE TO THE 7TH EDITION OF ITS CONFERENCE ON CREEP & FRACTURE IN HIGH TEMPERATURE COMPONENTS

WHEN?
18th-20th May 2026

WHERE?
Aix en Provence, France

This 7th International ECCC Creep & Fracture Conference will bring together engineers and scientists from around the world to present and discuss research and developments in all aspects of creep and creep-fracture interaction behaviour of high temperature industrial materials and components. The overall aim is to disseminate knowledge and identify future work items requiring attention from the high temperature research, design and standardisation communities, and component application.

www.eccc2026.org

ECCC

TEHNIČKI GLASNIK / TECHNICAL JOURNAL – GODIŠTE / VOLUME 19 – BROJ / ISSUE 4

PROSINAC 2025 / DECEMBER 2025 – STRANICA / PAGES 509-682

SVEUČILIŠTE SJEVER / UNIVERSITY NORTH – CROATIA – EUROPE

ISSN 1846-6168 (PRINT) / ISSN 1848-5588 (ONLINE)

TEHNICKIGLASNIK@UNIN.HR – HTTP://TEHNICKIGLASNIK.UNIN.HR