TEHNIČKI GLASNIK

TECHNICAL JOURNAL

TEHNIČKI GLASNIK / TECHNICAL JOURNAL – GODIŠTE / VOLUME 19 – BROJ / ISSUE 3

RUJAN 2025 / SEPTEMBER 2025 – STRANICA / PAGES 341-508

/

SVEUČILIŠTE SJEVER / UNIVERSITY NORTH – CROATIA – EUROPE

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

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

TEHNIČKI GLASNIK - TECHNICAL JOURNAL

Scientific-professional journal of University North

Volume 19 Issue 3
Varaždin, September 2025 Pages 341-508

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):
June 6, 2025	June 16, 2025	September 15, 2025

Legend:
(1) University North, (2) University of Slavonski Brod, (3) Faculty of Graphic Arts Zagreb, (4) Faculty of Civil Engineering Osijek, (5) Faculty of Engineering Rijeka, (6) Faculty of Mechanical Engineering and Naval Architecture Zagreb, (7) Faculty of Metallurgy Sisak, (8) Tomas Bata University in Zlín, (9) Department of Physics of the University of Josip Juraj Strossmayer in Osijek, (10) Faculty of Humanities and Social Sciences Osijek, (11) Karlovac University of Applied Sciences, (12) University of Applied Sciences Velika Gorica, (13) Department of Polytechnics - Faculty of Humanities and Social Sciences Rijeka, (14) Faculty of Electrical Engineering and Computer Science - University of Maribor, (15) Faculty of Civil Engineering - University of Maribor, (16) University College of Teacher Education of Christian Churches Vienna/Krems, (17) Faculty of Mechanical Engineering - Poznan University of Technology (Poland), (18) Mechanical Engineering Faculty Sarajevo, (19) University of Travnik - Faculty of Technical Studies, (20) Higher Education Technical School of Professional Studies in Novi Sad, (21) University of Novi Sad - Faculty of Technical Sciences, (22) Faculty of Mechanical Engineering - University of Montenegro, (23) Brno University of Technology, (24) Odessa State Academy of Civil Engineering and Architecture, (25) Faculty of Civil Engineering - University of Mostar, (26) Faculty of Manufacturing Technologies with the seat in Prešov - Technical University in Košice, (27) Faculty of Mechanical Engineering - University of Maribor, (28) College of Engineering, IT & Environment - Charles Darwin University, (29) Universite Libre de Bruxelles, (30) Vishwakarma Institute of Information Technology (Pune, India), (31) AISSMS Institute of Information Technology (Pune, India), (32) Permtech Research Solutions (India), (33) University of Belgrade, (34) National Dong Hwa University - Taiwan, (35) Faculty of Mechanical Engineering - Opole University of Technology (Poland), (36) TU Berlin - Germany, (37) Shahid Chamran University of Ahvaz - Iran, (38) University of Bologna - Italy, (39) University of Defence in Belgrade - Military Academy – Serbia, (40) Obuda University Budapest - Hungary

January 13th - 15th 2026

FIRE DEPARTMENT SAFETY EST. FDSOA 1989 HEALTH SAFETY OFFICERS ASSOCIATION

SAVE THE DATE!

FDSOA 2026 CONFERENCE

JANUARY 2026 | SCOTTSDALE DOUBLETREE RESORT BY HILTON

FDSOA Takes Safety to the Streets of Scottsdale!
Get ready for an unforgettable experience at the FDSOA Health, Safety & Apparatus Conference in Scottsdale! We're bringing the latest in safety innovations, dynamic discussions, and invaluable networking opportunities – all in the heart of one of Arizona's most scenic and exciting destinations.

Don't miss out on this exciting event! Mark your calendar and join us for an experience like no other.

Why Attend?

- Cutting-edge Safety information and solutions
- Expert-led sessions
- Connect with industry leaders and peers
- Explore Scottsdale's vibrant atmosphere

Stay in the loop and get the latest updates by visiting our website: fdsoa.org

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20230526232609

Original scientific paper
Received: 2023-05-26, Accepted: 2024-01-19

Parametric Study of Orthotropic Masonry Walls under Static and Dynamic Loading

Marija Smilović Zulim*, Marina Nikolić, Maša Ercegovac, Jure Radnić

Abstract: The paper presents a parametric study of unreinforced and confined masonry walls with orthotropic properties under in-plane static and dynamic loading. A previously developed FEM model was extended to simulate the behaviour of the orthotropic masonry and was used to perform a series of analyses. The orthotropic behaviour of masonry is simulated with a simplified constitutive model. The influence of the orthotropy of masonry on the behaviour of two-storey walls (with different lengths and qualities of the masonry) was investigated. The results of the analysis show that the bearing capacity and the displacements of the wall, as well as the stresses in the masonry, concrete and reinforcement depend significantly on the degree of orthotropy of the masonry. As the degree of orthotropy increases, the differences in the parameters compared to isotropic masonry increase. The influence of the orthotropy of the masonry is greater under dynamic loading than under static loading. The influence of orthotropy is greater for stiff masonry than for soft masonry, and for confined masonry than for unreinforced masonry.

Keywords: numerical analysis; masonry wall; orthotropy; static and dynamic loading

1 INTRODUCTION

Masonry walls are complex anisotropic composite structures composed of masonry and reinforced concrete elements (horizontal and vertical beams, foundation etc.). The masonry is a composite material composed of brick units and mortar. Therefore, masonry walls generally always have anisotropic properties, i.e. in-plane orthotropic properties. The orthotropy of the masonry is mainly caused by vertical holes in brick units. Due to variations in the stiffness and strengths of the horizontal and vertical joints between the masonry units, additional orthotropy of masonry is exhibited. The plaster of the masonry walls additionally influences their stiffness and orthotropy, especially if it is reinforced. They mostly have higher stiffness and resistance in the vertical direction than in the horizontal direction. Fortunately, the stress of the walls in the practice are usually much greater in the vertical than in the horizontal direction.

Many different numerical models have been proposed to simulate the behaviour of orthotropic masonry under static and dynamic loads. Some of them are listed in Refs. [1-15]. Among the oldest models are those of Page et al. [1] and Dhanaeskar [2]. Andreaus [3] gave the failure criteria for masonry walls under in-plane loading. Lorenco et al. proposed several different numerical models for the analysis of masonry structures [4-6]. An orthotropic damage model for masonry structures was proposed by Berto et al [7, 8]. Calderini and Lagomarsino presented a continuum model for in-plane elastic behaviour of masonry [9]. Lishak et al. developed 2D orthotropic failure criteria for masonry [10]. A simple model for analysing unreinforced masonry shear walls under combined axial, shear and bending loading developed by Ghiassi et al. [11]. Pela et al. proposed two models for the analysis of masonry structures [12]. Penava et al. investigated the resistance clay block masonry wall using a micromodel considering anisotropy of the masonry unit [13]. Bilko and Małyszko used the orthotropic failure criterion proposed by Hoffman) as an orthotropic elastic-plastic constitutive model for masonry walls [14]. Most recently, macro-modelling of orthotropic dam341age in masonry using combination of micro-mechanics and continuum FE analysis proposed Drougkas [15].

From the presented review of the literature, it is evident that the modelling of masonry with orthotropic properties is still an actual problem due to the complexity of such constitutive models. Furthermore, there are still no such numerical model that is simple and capable to simulate the primary nonlinear effects of the masonry walls, including their orthotropy, and appropriate for extensive engineering application. Therefore, further research in this area is still welcome.

The objective of this paper is to give an answer to question: How much the orthotropic properties of the masonry affect the behaviour of the masonry walls. This paper presents a parametric study of two storey unreinforced and confined masonry walls under in-plane static and dynamic loading, in which the influence of several parameters are considered (wall length, material parameters of masonry and degree of masonry orthotropy). The innovation of the work lies in the attempt to consider the influence of the orthotropy of masonry and masonry walls on their behavior and ultimate bearing capacity under static and dynamic (earthquake) loading. The aim of the paper is to stimulate further research in this field in order to quantify the effects of anisotropy using the presented numerical model, which is based on the results of parametric analyses on the considered unreinforced and bounded masonry walls, with different ratios of wall heights and lengths.

2 NUMERICAL MODEL

2.1 Basic Numerical Model

Adopted basic numerical model [16] is presented very shortly below.

Planar 2D and 1D finite elements for spatial discretization and explicit-implicit time integration for time history analysis were adopted. The main features of this model are its simplicity and the possibility of practical application It is possible to simulate the material and

geometric nonlinearities of planar concrete and masonry structures loaded in their plain. The constitutive material models are the same for static and dynamic analysis, i.e. the effect of strain rate on material characteristics is neglected.

The elastic-perfectly plastic constitutive model for concrete in compression is assumed. Modelling the opening and closing of cracks in concrete in tension-tension and tension-compression is feasible. The cracks are modelled as smeared, with fixed position. The effects of tensile and shear stiffness of cracked concrete are modelled. Reinforcement is modelled by bar elements within basic 2D elements with polygonal stress-strain relation, without slipping between the concrete and reinforcement. Using contact elements, it is possible to model penetration, separation, and slipping on the contact surface between the foundation and the ground. According to the adopted uniaxial stress-strain diagrams, contact elements are capable of transmitting normal and shear stresses.

The same constitutive model as for concrete was used for soil modelling, with appropriate material parameters.
Constitutive model for masonry in [16] does not include the effect of the orthotropy. The improved orthotropic masonry model in this paper is presented in continuation.

2.2 Ortotropic Masonry Model

A simplified orthotropic macro model of masonry is used. The main directions of orthotropy are vertical (v) and horizontal (h). Different compressive strengths (f_{mc}^{v}, f_{mc}^{h}), tensile strengths (f_{mt}^{v}, f_{mt}^{h},), Young's modulus (E_{v}, E_{h}), limit compressive strains (ε_{mc}^{v}, ε_{mc}^{h},) and limit tensile strains (ε_{mt}^{v}, ε_{mt}^{h},) of masonry are defined for direction v and h. Adopted orthotropic constitutive model of masonry for normal stresses σ_m^{v}, σ_m^{h}, is presented in Fig. 1.

The failure of the masonry is defined over normal strains ε_{mc}^{v}, ε_{mc}^{h}, ε_{mt}^{v}, ε_{mt}^{h}. A small number of masonry parameters were used as the model is intended for practical application.

A measure of degree of masonry orthotropy is defined by orthotropy coefficient c_o:

$$c_o = \frac{E_m^h}{E_m^v} = \frac{f_{mc}^h}{f_{mc}^v} = \frac{f_{mt}^h}{f_{mt}^v} = \ldots \tag{1}$$

where

$$E_m^v \cdot \nu_m^v = E_m^h \cdot \nu_m^h \tag{2}$$

Here ν_m^{v}, ν_m^{h} are Poisson's rations in v and h directions.
The shear modulus (G) of masonry is defined by:

$$G_m = \frac{1}{\frac{\left(1+\nu_m^h\right)}{E_m^h} + \frac{\left(1+\nu_m^v\right)}{E_m^v}} \tag{3}$$

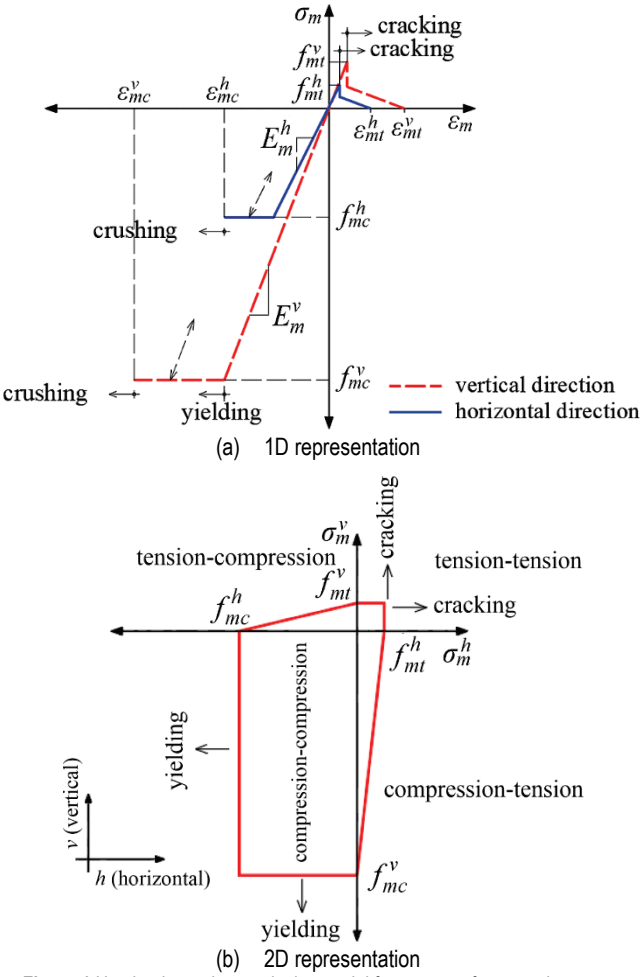


Figure 1 Used orthotropic constitutive model for masonry for normal stresses

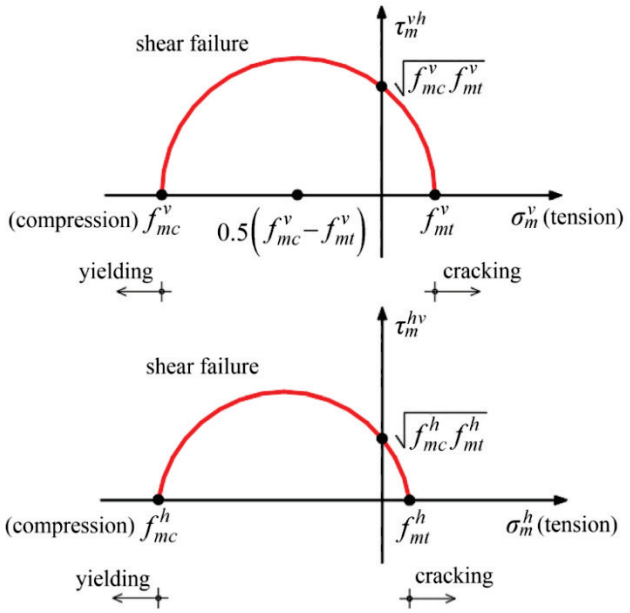


Figure 2 Used orthotropic constitutive model for masonry for shear stresses

In the model, it is assumed that the compressive strengths of the masonry in biaxial pressure are equal to its uniaxial compressive strength for a particular direction, i.e. that they

do not depend on the ratio of biaxial compressive stresses. Elastic-perfectly plastic behaviour is assumed for biaxial compression, with linear unloading. The elastic behaviour is also used in biaxial tension until the tensile strength of the masonry is reached. Then the model of smeared cracks is used, which can occur in the direction perpendicular to the tensile stresses. The opening and closing of the cracks as well as the tensile and shear strength of the cracked masonry are modelled as in the concrete model [16].

The failure of the masonry due to shear stresses is also modelled (Fig. 2). The model is simple and assumes a relationship between normal (σ_m^v, σ_m^h,) and shear (τ_m^v, τ_m^h,) stresses according to circle equation.

The improved numerical model is verified by means of experimental tests [17]. The comparison of the numerical results with the experimental results is shown in Fig. 3. As can be seen, the numerical results are in good agreement with the experimental results. A more detailed description of the structural geometry, material properties and other information about the experimental tests and the validation of the improved numerical model can be found in [17].

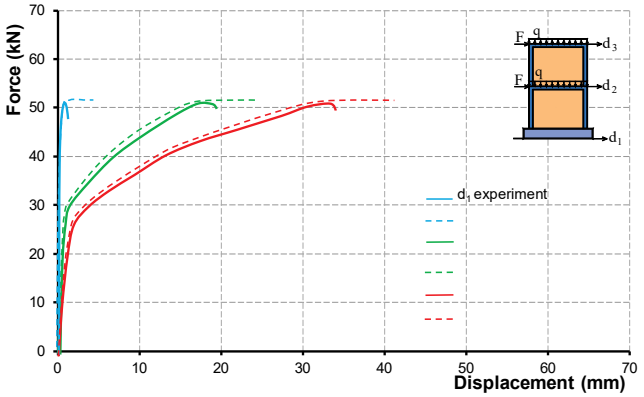


Figure 3 Validation of improved numerical model on experimental test [17]

Table 1 Scheme of the parametric study

Orthotropy coefficient c_o	Type of the wall	Material properties of the wall	Length of the wall (m)	Case
0,2 / 0,4 / 0,6 / 1	Unreinforced masonry (URM) wall	stiff masonry (S1)	3,0	/c_o/-URM-S1-3
			6,0	/c_o/-URM-S1-6
			12,0	/c_o/-URM-S1-12
		soft masonry (S2)	3,0	/c_o/-URM-S2-3
			6,0	/c_o/-URM-S2-6
			12,0	/c_o/-URM-S2-12
	Confined masonry (CM) wall	stiff masonry (S1)	3,0	/c_o/-CM-S1-3
			6,0	/c_o/-CM-S1-6
			12,0	/c_o/-CM-S1-12
		soft masonry (S2)	3,0	/c_o/-CM -S2-3
			6,0	/c_o/-CM -S2-6
			12,0	/c_o/-CM-S2-12

Table 2 Adopted material parameters of masonry for various orthotropy coefficient c_o

Parameter	Stiff masonry (S1)					Soft masonry (S2)			
	c_o = 0,2	c_o = 0,4	c_o = 0,6	c_o = 1,0		c_o = 0,2	c_o = 0,4	c_o = 0,6	c_o = 1,0
E_m^v (GPa)	5,0	5,0	5,0	5,0		1,0	1,0	1,0	1,0
E_m^h (GPa]	1,0	2,0	3,0	5,0		0,2	0,4	0,6	1,0
G_m (GPa)	0,794	1,316	1,685	2,174		0,159	0,263	0,337	0,435
f_{mc}^v (kPa)	5,0	5,0	5,0	5,0		1,0	1,0	1,0	1,0
f_{mc}^h (kPa)	1,0	2,0	3,0	5,0		0,2	0,4	0,6	1,0
f_{mt}^v (kPa)	0,15	0,15	0,15	0,15		0,03	0,03	0,03	0,03
f_{mt}^h (kPa)	0,03	0,06	0,09	0,15		0,006	0,012	0,018	0,03
Notes: *v* indicates the vertical direction; *h* indicates the horizontal direction.									

3 PARAMETRIC STUDY

The previously developed and validated numerical model was used to investigate the effects of masonry orthotropy on unreinforced (URM) and confined masonry (CM) walls with different geometries and material properties. Tab. 1 shows the schematic of the parametric study performed.

3.1 Description of the Performed Numerical Analyses

Two-storey unreinforced (URM) and confined masonry (CM) walls were considered in the analyses.

The geometry of the walls is shown in Fig. 4. The walls are 0.24 m thick and of different lengths (3.0 m, 6.0 m, 12.0 m). The reinforcement of the foundations, the horizontal and vertical ring beams and the other geometry of the walls are shown in Fig. 3.

Walls with two types of wall material properties were analysed: stiff masonry (S1) and soft masonry (S2). Soft walls have low values of modulus of elasticity and strength, while stiff walls have relatively high values. The assumed material parameters for so-called stiff and soft masonry are listed in Tab. 2. The assumed material parameters for stiff masonry are standard parameters for well-conditioned or new masonry made of clay bricks (as bricks) and single-component adhesive. The assumed basic masonry parameters for the considered orthotropy coefficient co are listed in Tab. 2. The material properties of concrete, reinforcement and contact elements used in this study are listed in Tab. 3.It is

assumed that the foundations are on solid ground, with the possibility of uplifting. In addition, contact elements are used on the contact surface between the ground and foundation, in order to simulate more realistically the lifting of the foundation from the ground and the shear sliding of the foundation in relation to the ground.

The finite element discretization of the URM and CM walls is shown in Fig. 5. Due to the effectiveness of the numerical analyses, a relatively coarse mesh of finite elements was used for the wall. This has no significant impact on the conclusions of the parametric study. All nodes of finite and contact elements are free to move, only the nodes of contact elements in contact with the ground are fixed.

Table 3 Adopted basic material parameters for concrete, reinforcement and contact elements

Parameter	Concrete	Reinforcement	Contact elements
E (GPa)	30,5	210	30,5
G (GPa)	13,26	-	-
f_c (MPa)	25	560	25
f_t (MPa)	2,5	560	0

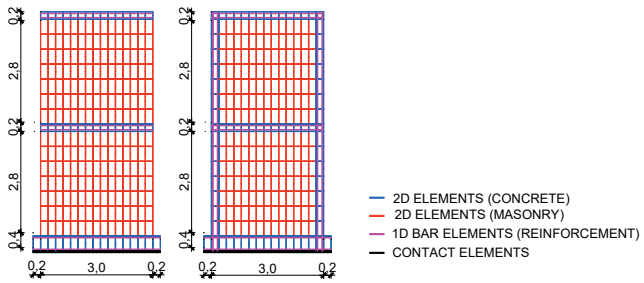


Figure 5 Finite element discretization of the walls with length of 3 m

3.2 Results

Some results of the performed parametric study of URM and CM walls are presented separately for static and dynamic loading.

3.2.1 Static analyses

In the performed static analyses the walls are firstly loaded with self - weight and constant vertical load q (q = 40 kN/m^2) and then incrementally with horizontal force F until walls failure (Fig. 4).

The ratio of ultimate load of the URM and CM walls to coefficient of orthotropy c_o is presented in Fig. 6. According to results it is evident that the ultimate load for URM walls with stiff masonry is greater than ultimate load of walls with soft masonry. It is also apparent that the longer walls have greater ultimate load than the shorter ones. The effect of orthotropy on the bearing capacity on the URM wall with soft masonry is greater than with stiff masonry.

Fig. 6 shows that the ultimate load decreases with the stiffness of the masonry. Even with a significant degree of orthotropy, the reduction in load bearing capacity is quite small compared to isotropic masonry. For walls with a coefficient of orthotropy c_o = 0.2 the reduction of ultimate load is about 10-12 %, compared to isotropic masonry (c_o = 1). The orthotropy effect of the masonry is also greater for shorter walls than for longer ones. In addition, the orthotropy effect is greater for soft masonry than for stiff masonry.

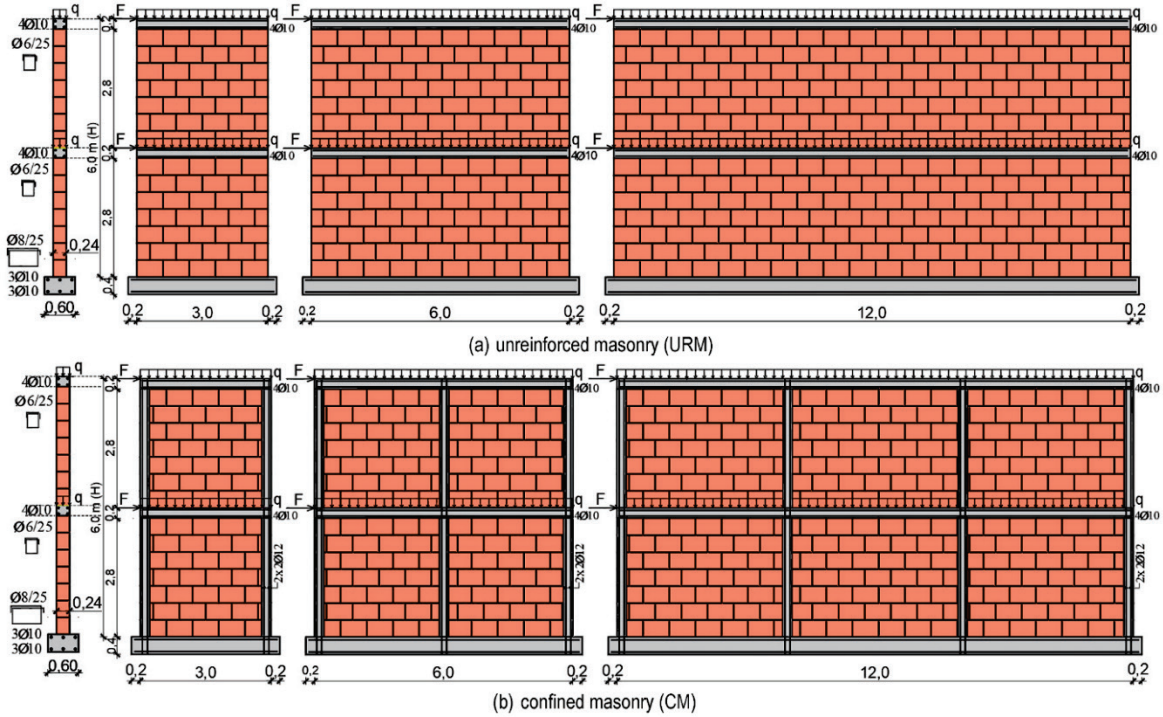


Figure 4 Geometry of considered masonry walls

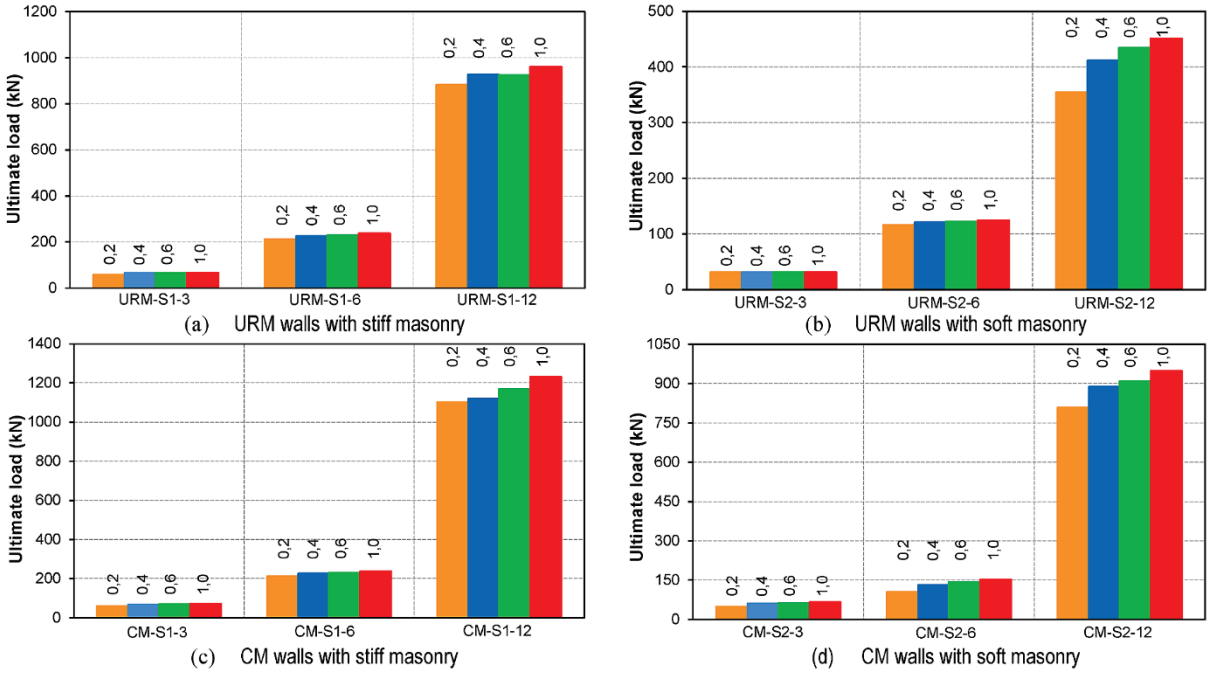


Figure 6 The ratio of ultimate load to coefficient of orthotropy c_o

The relationship between ultimate load and orthotropy coefficient c_o for CM walls is also shown in Fig. 6. Compared to URM walls, the stiffness of the masonry of CM walls has less influence on the ultimate load. Obviously, the ultimate load of the masonry is negligible, compared to the ultimate load of the horizontal and vertical concrete ring beams. It is also obvious that the CM walls have significantly smaller displacements than the URM walls.

This is to be expected, as the horizontal and vertical concrete ring beams, which form a quasi-concrete frame, make a major contribution to the overall stiffness of the wall. The effect of the orthotropy of the masonry on the displacement of CM walls is greater than for URM walls. This is due to the fact that the horizontal normal stresses in CM walls are greater than in URM walls. The effect of orthotropy is significantly greater in the shorter walls than in the longer walls because the shorter walls are more affected by shear forces (horizontal stresses) than the longer walls.

The relationship between the stresses in the masonry, in the concrete and in the reinforcement of the considered walls and the orthotropy of the masonry is analogous to the relationship between the ultimate load of the masonry and orthotropy coefficient c_o shown above.

3.2.2 Dynamic Analyses

The loading is applied in two steps, whereby in the first step the walls are loaded with the self-weight and the vertical load q, while in the second step the walls are loaded with the harmonic base excitation as shown in Fig. 7. T_1 is the first period of vibration of the elastic wall, which is initially calculated for all considered orthotropic walls. The adopted amplitude of the harmonic base excitation is 0.3g for CM walls and 0.1g for URM walls. The duration of the excitation is adapted to the stiffness of the wall and is $10 \cdot T_1$ and the total duration of the numerical analysis is $25 \cdot T_1$. In addition, a 2 % viscous damping is assumed.

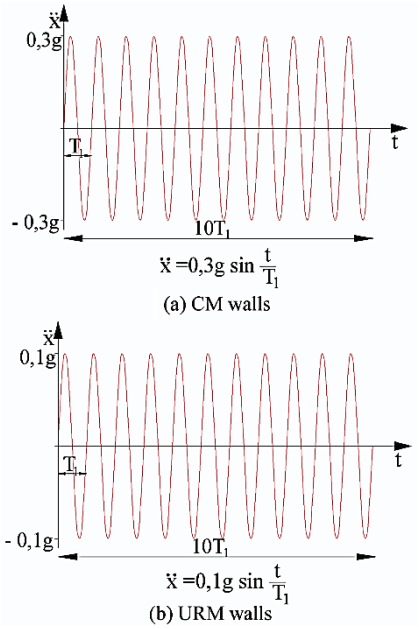


Figure 7 The adopted horizontal base excitation

Some numerical results of the dynamic analyses carried out are shown in Figs. 8-12. The horizontal displacements of the top of URM and CM walls are presented in Fig. 8. The relationship between the maximum horizontal displacement of top of the wall and the coefficient of orthotropy is presented in Fig. 9. The relationship between the maximum compressive stresses (σ_{yy}) in masonry and the coefficient of orthotropy c_o is presented in Fig. 10. The relationship between the maximum compressive stresses (σ_{yy}) in concrete and the coefficient of orthotropy c_o is presented in Fig. 11.

The relationship between the maximum tension stresses (σ_{yy}) in reinforcement and the coefficient of orthotropy c_o is presented in Fig. 12.

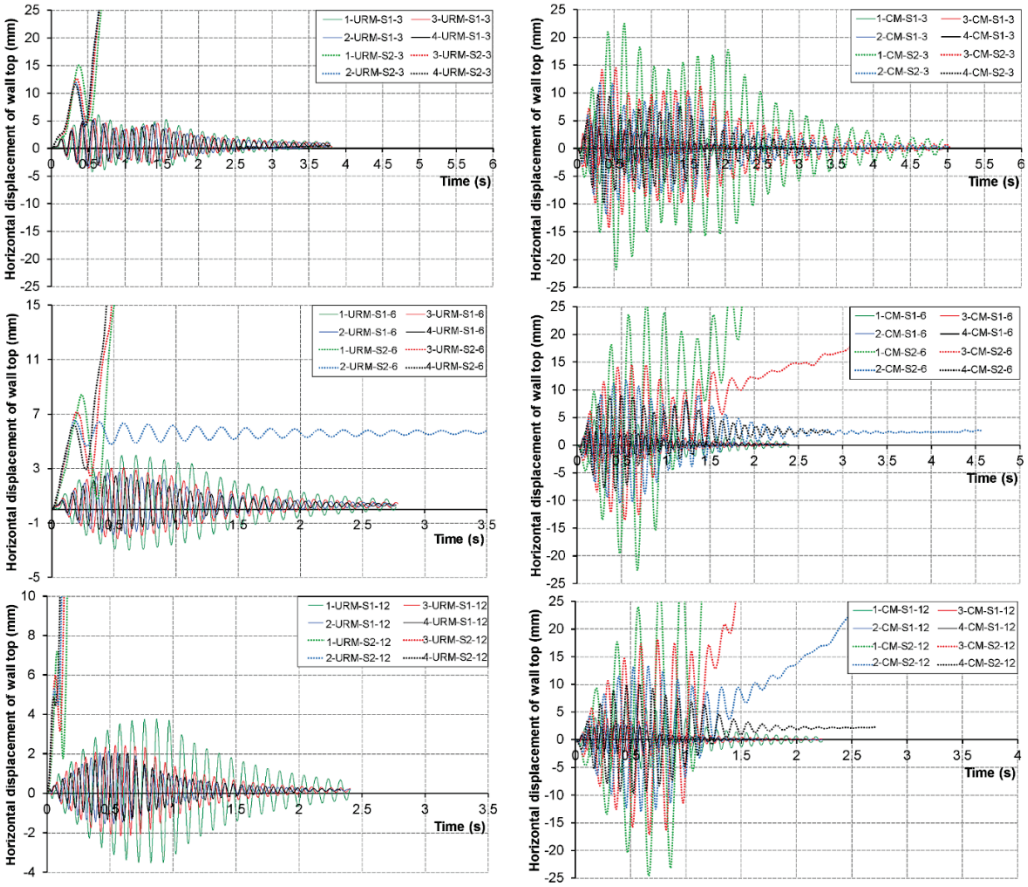


Figure 8 Horizontal displacement of the top of URM and CM walls

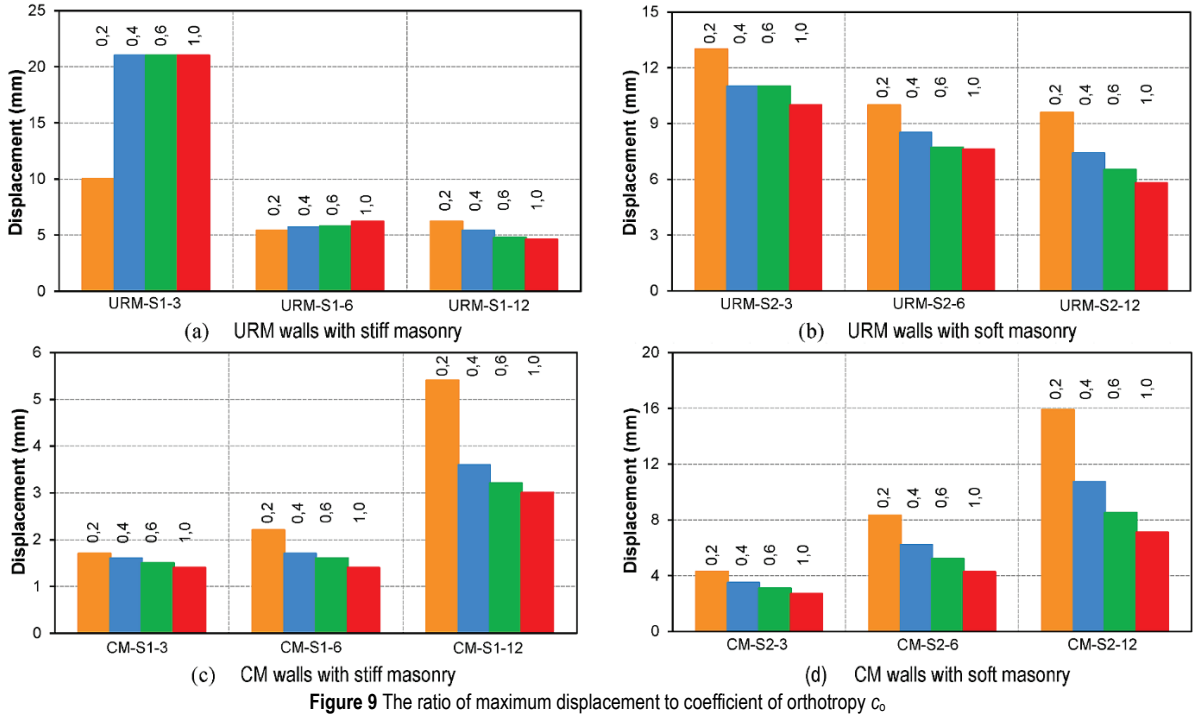


Figure 9 The ratio of maximum displacement to coefficient of orthotropy c_o

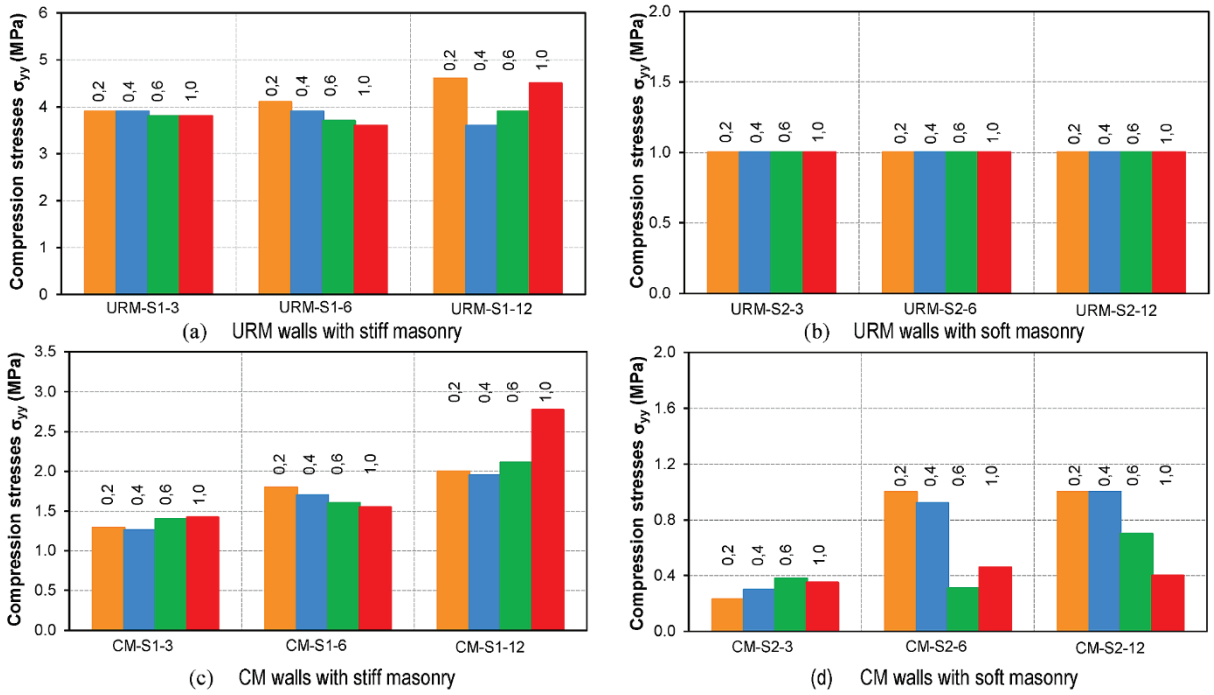


Figure 10 The ratio of maximum compressive stresses (σ_{yy}) in masonry to orthotropy coefficient c_o

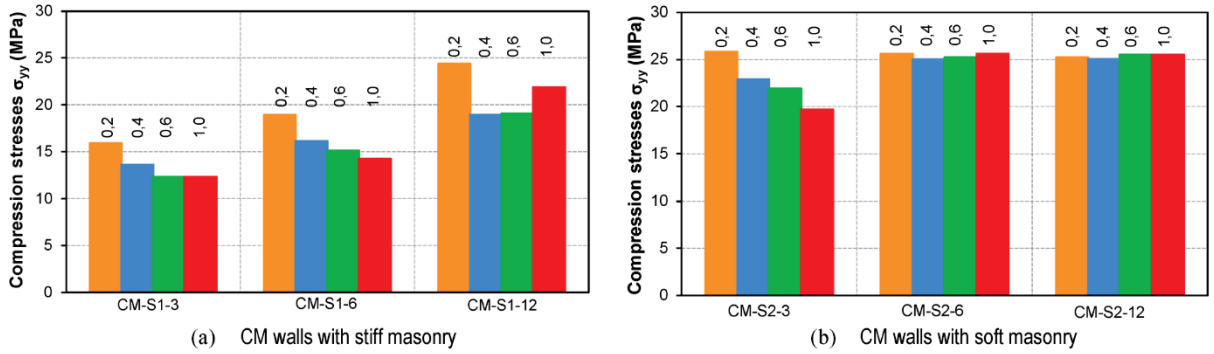


Figure 11 The ratio of maximum compression stresses (σ_{yy}) in concrete to orthotropy coefficient c_o

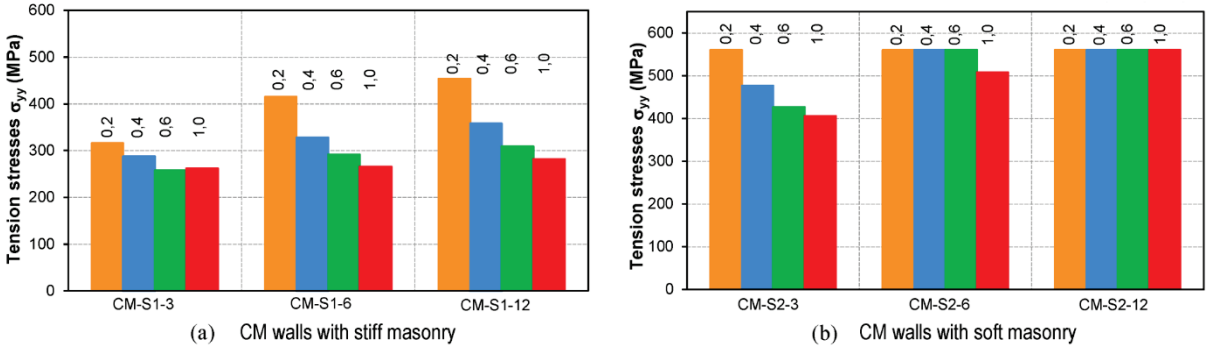


Figure 12 The ratio of between maximum tension stresses (σ_{yy}) in reinforcement to orthotropy coefficient c_o

As presented in Fig. 8-12, the displacements of the top wall, the stresses in masonry, in the concrete and in the reinforcement, as well as the ultimate load, depend on the coefficient of orthotropy c_o. With a high degree of orthotropy of the masonry (low c_o), the differences in the values of the above-mentioned parameters are greater than with isotropic masonry (c_o = 1). For the same c_o, the response of the masonry wall under dynamic loading is determined by its geometry, its type (unreinforced, confined) and the quality of the masonry (soft, stiff).

The orthotropy of the masonry also has a great influence on walls with stiff masonry, especially on longer walls. In the longer walls with soft masonry, yielding of concrete occurred in compression (the failure came through masonry). The orthotropy of the masonry has a major influence on the concrete stresses. In the longer walls with soft masonry, the

steel yielded. The orthotropy of the masonry also has a strong effect on the reinforcement stresses, especially in the longer walls.

As presented in Fig. 8-12, the displacements, the stresses in masonry, concrete and reinforcement as well as the load-bearing capacity depend on the coefficient of orthotropy c_o. With a high degree of orthotropy (low c_o), the differences in the values of the above-mentioned parameters are greater compared to isotropic masonry ($c_o = 1$). For the same c_o, the behaviour of the masonry walls depends on its geometry, its type (unreinforced, encased) and its masonry quality (stiff, soft).

The effect of orthotropy on the ultimate load of masonry walls is significant, especially for confined ones. Orthotropy has a greater effect on soft masonry than on stiff masonry. Most of the URM walls failed at a harmonic base excitation with an amplitude 0,1g.

4 CONCLUSION

The numerical FE model [16] for the nonlinear static and dynamic analysis URM and CM walls under in-plane loading, which is upgraded with an orthotropic constitutive masonry model, can simulate the main material nonlinearities. However, there is still a need for further verification of the model.

The displacements and the stresses in considered URM and CM walls under static and dynamic loading, as well as ultimate load, depends on the orthotropy of the masonry. As the degree of the orthotropy increases, differences in above parameters increase in comparison to isotropic masonry. The orthotropy effect is influenced by the wall geometry, the type of the masonry wall (unreinforced, confined) and the quality of the masonry (stiff, soft). The CM walls have a greater orthotropy effect than the URM walls. The greater influence of orthotropy in CM walls compared to URM walls is explained by the additional influence of the reinforced concrete elements (horizontal and vertical beams) on their stiffness and load-bearing capacity in the horizontal direction. The orthotropy effect is greater in case of stiff masonry than soft masonry. Longer walls have a greater influence of orthotropy than shorter walls. The greater influence of orthotropy in longer walls compared to shorter walls is explained by the greater influence of shear (compared to bending) than in longer walls. The influence of the orthotropy of the masonry in the masonry walls is greater for dynamic loading (harmonic base excitation) than for static loading.

The behaviour of the orthotropic masonry walls under static loading is almost identical to that of isotropic masonry, even if there is a great degree of orthotropy. It is shown in static analyses that walls with $c_o = 0,2$ have approximately 10-12 % less ultimate load, than the walls with $c_o = 1,0$. In some cases of dynamic loading, the masonry walls with $c_o = 0,2$ have over 50 % less ultimate load, than walls with isotropic masonry.

This parametric study holds practical interest, because codes provide directions on evaluating the isotropic properties of masonry. In order to achieve a more realistic description of the actual complex behaviour of masonry structures, a further development of the presented numerical model is required. In this sense, it is planned to improve the numerical model by improving the simulation of different factors influencing the wall (multiaxial stress state, cyclic loading, effect of plaster, crack modelling, etc.) and developing a suitable soil model. Therefore, further numerical and experimental investigations of the behaviour of the URM and CM walls with orthotropic masonry are necessary, especially under seismic loading.

Acknowledgments

This research is partially supported through project KK.01.1.1.02.0027, a project co-financed by the Croatian Government and the European Union through the European Regional Development Fund - the Competitiveness and Cohesion Operational Programme.

5 REFERECES

[1] Page, A. W. (1978). Finite element model for masonry. *Journal of the structural division – ASCE, 104(*8), 1267-85.

[2] Dhanaeskar, M., Kleeman, P. W., & Page, A. W. (1985). Biaxial stress-strain relations for brick masonry. *Journal of the structural division – ASCE, 111*(5), 1085-1100. https://doi.org/10.1061/(ASCE)0733-9445(1985)111:5(1085)

[3] Andreaus, U. (1996). Failure Criteria for Masonry Panels under In-Plane Loading. *Journal of structural engineering – ASCE*, *122*(1), 37-46. https://doi.org/10.1061/(ASCE)0733-9445(1996)122:1(37)

[4] Lourenco, P. B. & Rots, J. G. (1997). A multi-surface interface model for the analysis of masonry structures. *Journal of engineering mechanics, 123*(7), 660-668. https://doi.org/10.1061/(ASCE)0733-9399(1997)123:7(660)

[5] Lourenco, P. B., De Brost, R., & Rots, J. G. (1997). A plane stress softening plasticity model for orthotropic materials. *International journal for numerical methods in engineering, 40*(21), 4033-4057. https://doi.org/10.1002/(SICI)1097-0207(19971115)40:21<4033::AID-NME248>3.0.CO;2-0
Lourenco, P. B., Rots, J., & Blaauwendraad, J. (1998). Continuum Model for Masonry: Parameter Estimation and Validation. *Journal of structural engineering – ASCE, 124(*6), 642-652. https://doi.org/10.1061/(ASCE)0733-9445(1998)124:6(642)

[6] Berto, L., Saetta, A., Scotta, R., & Vitaliani, R. (2002). An orthotropic damage model for masonry structures. *International journal for numerical methods in engineering, 55(*2), 27-157. https://doi.org/10.1002/nme.495

[7] Berto, L., Saetta, A., Scotta, R., & Vitaliani, R. (2004). Shear behaviour of masonry panel: parametric FE analyses. *International Journal of Solids and Structures, 41*(16-17), 4383-4405. https://doi.org/10.1016/j.ijsolstr.2004.02.046

[8] Calderini, C. & Lagomarsino, S. (2008). Continuum model for in-plane inelastic behaviour of masonry. *Journal of structural engineering – ASCE, 134*(2), 209-220. https://doi.org/10.1061/(ASCE)0733-9445(2008)134:2(209)

[9] Lishak, V. I., Yagust, V. I., & Yankelevsky, D. Z. (2012). 2D orthotropic failure criteria for masonry. *Engineering structures, 36*, 360-371. https://doi.org/10.1016/j.engstruct.2011.11.033

[10] Ghiassi, B., Soltani, M., & Tasnimi, AA. (2012). A simplified model for analysis of unreinforced masonry shear walls under

combined axial, shear and flexural loading. *Engineering structures, 42*, 396-409. https://doi.org/10.1016/j.engstruct.2012.05.002

[11] Pela, L., Cervera, M., & Roca, P. (2013). An orthotropic damage model for the analysis of masonry structures. *Construction and building materials, 41*, 957-967. https://doi.org/10.1016/j.conbuildmat.2012.07.014

[12] Penava, A. et al. (2016). Three-dimensional micromodel of clay block masonry wall. *International Journal of Masonry Research and Innovation, 1(*4), 282-305. https://doi.org/10.1504/IJMRI.2016.081270

[13] Bilko, P. & Małyszko, L. (2020). An Orthotropic Elastic-Plastic Constitutive Model for Masonry Walls. *Materials, 13*(18). https://doi.org/10.3390/ma13184064

[14] Drougkas, A. (2022). Macro-modelling of orthotropic damage in masonry: Combining micro-mechanics and continuum FE analysis. *Engineering Failure Analysis, 14*. https://doi.org/10.1016/j.engfailanal.2022.106704

[15] Radnić, J. et al. (2011). Numerical Model for Static and Dynamic Analysis of Masonry Structures. *Gradevinar, 63(*6), 529-546. https://doi.org/10.1007/978-3-642-31497-1_1

[16] Smilovic, M. (2014). Behaviour and numerical model of masonary structures under static and dynamic load. *PhD Thesis*, Split (in Croatian).

Authors' contacts:

Marija Smilović Zulim, PhD, MEng, CE, Assistant Professor
(Corresponding author)
University of Split, Faculty of civil engineering, architecture and geodesy,
Matice hrvatske 15, 2100 Split, Croatia
marija.smilovic@gradst.hr

Marina Nikolić, PhD, MEng, CE, Assistant Professor
University of Split, Faculty of civil engineering, architecture and geodesy,
Matice hrvatske 15, 2100 Split, Croatia
marina.sunara@gradst.hr

Maša Ercegovac, MEng, CE
(This research was done while the author was a graduate student at University of Split, Faculty of civil engineering, architecture and geodesy)
Matice hrvatske 15, 2100 Split, Croatia
masa.ercegovic@gradst.hr

Jure Radnić, PhD, MEng, CE, Full Professor
University of Split, Faculty of civil engineering, architecture and geodesy,
Matice hrvatske 15, 2100 Split, Croatia
jure.radnic@gradst.hr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20230928070645

Original scientific paper
Received: 2023-09-28, Accepted: 2024-08-01

Thermo-Hydraulic Performance of Tubular Heat Exchanger with Opposite-Oriented Trapezoidal Wing Perforated Baffle Plate

Md. Atiqur Rahman*, S. M. Mozammil Hasnain, Rustem Zairov

Abstract: A detailed experimental study explored a new axial heat exchanger with swirling air over heated tubes. This heat exchanger features circular baffle plates with perforations and trapezoidal air deflectors at different angles. The tubes are aligned parallel to the airflow, and the deflectors create intense air turbulence, which improves heat transfer. The heat flux on the tubes is kept constant, and each baffle plate has eight deflectors positioned in reverse to induce swirling air within a circular duct that contains the hot water tubes. The baffle plates are spaced with varying pitch ratios, and the Reynolds number ranges from 16,000 to 30,000. The study found that the heat exchanger's performance highly depends on the pitch ratio and deflector angle, with the highest thermal enhancement factor (TEF) of 2.48 achieved at a 30° deflector angle and a pitch ratio of 1.2. These findings highlight the importance of optimizing design parameters to improve heat exchanger performance, offering valuable insights for better thermal management in industrial and environmental applications.

Keywords: deflector inclination angle; discontinuous swirl flow; flow recirculation; relative flow resistance; trapezoidal deflector; thermo-fluid performance

1 INTRODUCTION

Swirl flow is fluid motion along a helical or spiral path, where the particles possess tangential and axial velocity components. The resulting flow is called a helical/swirl when the fluid flows through a confined path, as in a shell or duct, either through suction or a forced flow generator [1].

Swirl flow is vital in various engineering and industrial applications for several reasons [2]:

1) Enhanced Mixing: Swirl flow enhances mixing efficiency by inducing tangential velocity components and axial flow. This is crucial in applications such as chemical reactors, where uniform mixing of reactants is essential for reaction efficiency.
2) Heat Transfer (HT) Improvement: In heat exchangers (HX) and cooling systems, swirl flow promotes better HT by creating turbulence and increasing the contact surface area between the air and the HT surface.
3) Reduction of Dead Zones: Swirl flow helps reduce stagnant or dead zones in fluid systems by imparting rotational motion. This is advantageous in preventing sedimentation or buildup of contaminants in pipelines or tanks.
4) Control of Flow Separation: In aerodynamics and fluid dynamics, swirl flow can help control flow separation over surfaces, which is crucial for maintaining aerodynamic efficiency and reducing drag in aircraft wings, turbines, and other applications.
5) Enhanced Combustion: In combustion chambers, swirl flow can improve combustion efficiency by ensuring better fuel-air mixing, leading to cleaner and more efficient combustion processes.
6) Vortex Formation: Swirl flow often leads to stable vortex structures, which can be utilized beneficially in various applications, such as vortex tubes for separating components of gas mixtures or cyclone separators for particle separation.
7) Energy Efficiency: By enhancing mixing and reducing flow losses, swirl flow contributes to overall energy efficiency in fluid systems, whether in industrial processes, HVAC systems, or environmental engineering applications.

Swirl flow in a confined path is generated using swirl generators. Swirl generators effectively enhance HT between the fluid and surfaces within the HX. These generators generate vortices or eddies that intensify fluid mixing, reducing boundary layer width and improving the convective heat transfer coefficient (h_m). Additionally, using swirl producers helps eliminate stagnant regions in the baffle spacing, ensuring an even heat distribution. Furthermore, these devices can regulate flow distribution, preventing fluid clogging or dead zone formation and ultimately enhancing the HX's overall efficiency (η) [3].

There are several ways to generate swirl flow characterized into three categories [4]:

- Use of fins or adjustable propellers tangentially deflecting the axial flow. Because of its simplicity, this device is generally used in industrial systems, particularly gas turbines. However, this type of device introduces significant head losses, and the swirl intensity is limited (design of fins) [5].
- Rotating mechanical devices generate a rotational movement of the fluid passing between them [6].
- Tangential injection of part or all fluid quantity into a main duct. The intensity of the swirl is then determined by the ratio between the flow injected tangentially and that injected axially [7, 8].

Swirl flows are highly valued for effectively mixing fluids and extending residence times, facilitating complete reactions. They are extensively employed in diverse engineering applications, including cyclone separators, separation processes, combustion chambers, turbo machinery, solar applications, heat exchangers, bubble

generators, electronics cooling, and environmental pollution mitigation systems [9].

There are two types of swirl generators: decaying and non-decaying. Decaying swirl generators [10, 11] create swirling or vortex-like motion that diminishes over time due to viscosity, friction, or turbulence, while non-decaying swirl generators sustain continuous swirling motion throughout the flow.

Recently, devices that create swirl have gained widespread use to enhance heat management across various industries. This adoption is due to their cost-effectiveness and straightforward installation procedures. The swirling patterns, known as vortices, are classified based on the orientation of their rotational axis [12]. Longitudinal vortices rotate perpendicular to the flow direction, whereas transverse vortices spin parallel. Creating these vortex structures has significantly enhanced heat transfer efficiency by optimizing flow characteristics within the stream [13, 14]. However, incorporating fins increases the system's flow resistance and friction losses, necessitating higher pumping power [15].

Many studies have focused on methods of generating swirls in the confined space. Inclined ribs, arc ribs, turbulators, twisted tape, guide vanes, blades, etc., generally do this despite the notable thermal enhancements resulting from implementing swirl generators; a significant Δp ensued, rendering the aforementioned thermal systems ineffective. Researchers have focussed on perforation as an alternate option, significantly reducing Δp and enhancing the thermal enhancement factor ($TEF = j/j_0/(f/f_0)^{1/3}$).

Perforations create openings or passages in a material, allowing for better airflow, heat dissipation, and improved thermal efficiency. Some key points highlighting the importance of perforation in heat transfer include [16]:

1) Improved Ventilation: Perforations enhance airflow through materials, which is crucial for efficient heat transfer by facilitating the removal of heat and preventing overheating.
2) Enhanced Surface Area: Perforations increase the material's surface area, promoting quicker heat dissipation and more effective thermal energy transfer.
3) Reduced Thermal Resistance: Perforations decrease thermal resistance by facilitating heat transfer across surfaces, allowing for faster cooling or heating depending on the application.
4) Improved Heat Dissipation: In applications requiring efficient heat dissipation, such as electronic devices or power systems, perforations aid in removing excess heat by promoting airflow through the material.
5) Customization and Flexibility: Perforations can be customized in various shapes and sizes to meet specific heat transfer needs. This flexibility enables optimization of heat transfer performance across different applications.

Researchers have explored modifying fins and baffles through openings and cavities to increase surface area and integrating slots and grooves to enhance fluid mixing and heat transfer efficiency. A few of the notable works have been discussed below.

Wang et al. [17] conducted experimental and numerical investigations on HT and friction (f) characteristics of vortex generators (VGs) using various hole shapes, including circular, equilateral triangle, square, rectangular, and rhombus holes. The study varied the pitch ratio (PR) from 1.63 to 4.89 while maintaining an angle of attack (α) of 45°. A comparison between planar VGs and VGs with drilled holes revealed that circular holes enhanced local heat dissipation by improving jet flow and promoting fluid rotation, which is particularly beneficial in recirculation zones. In contrast, hole shapes with sharp corners did not contribute positively to heat dissipation. A maximum TEF of 1.14 for Case A (circular holes) at Reynolds number (Re) = 9090 was noted.

In a separate study, Wang et al. [18] investigated the application of perforated VGs on mini-channel HX walls, an area with limited research. This study examined the influence of VG placement, *PR*, hole diameter, and VG configuration on the TEF of the HX. Results indicated that VG placement affects downstream vortex structures, where larger opening areas can suppress vortices or cause certain structures to vanish, potentially compromising HT performance. Optimal HT was observed at specific *PR*, highlighting the critical role of hole placement. Ajarostaghi et al. [19] introduced an innovative turbulator, showing an increase in Nusselt number (Nu) with the number of blades.

Mousavi et al. [20] numerically investigated the effects of incorporating a novel curved turbulator to enhance HT within a pipe. This turbulator includes multiple rows of flow directors designed to induce turbulent swirl flows. The study analyzed five geometric parameters—curvature angle, flow director diameter, cone angle, nozzle outlet diameter, and number of flow director rows—across the *Re* range of 10,000 - 35,000. The research findings indicate that the outlet diameter of the conical nozzle has minimal impact on thermal performance. Additionally, the TEF and overall heat transfer coefficient (U) improve as the number of flow director rows increases. Higher curvature angles lead to more pronounced secondary flows, resulting in higher average *Nu*. However, the curved turbulator configuration exhibits lower TPF than its uncurved counterpart due to significant Δp.

In their study, Wang et al. [21] investigated the effect of rotational motion on HT and fluid dynamics in a circular duct, incorporating a nozzle (conical) at the outlet. The researchers conducted experiments using a 1200 mm long pipe with an 80 mm diameter. They examined 3 different nozzle ratios (0.5 - 0.75) and tested seven distinct blade swirl generators under varied flow conditions. For the solid pipe experiments, flow *Re* ranged from 42,000 to 170,000. In contrast, the porous pipe experiments covered flow Re between 70,000 and 130,000, with BR varying between 0.002 and 0.050. The result indicated valuable information about h_m, overall Δp, and the characteristics of swirling flow, highlighting crucial phenomena, such as vortices forming along the pipe's central axis.

Hussein and Hameed [22] conducted an experimental study to assess the effectiveness of a double-pipe HX in facilitating HT between air and water. Segmental baffle plates (SBP) featuring semi-circular holes were employed to

enhance HT. The edges of these semi-circular holes were modified to act like fins. Air served as the HT fluid in the annular region between the shell and tube of the HX, while water flowed through the tube. The airflow *Re* ranged from 2700 to 4000 across seven distinct conditions, while the water flow maintained a constant *Re* of 34159. Three sizes of semi-circular holes (30 mm, 25 mm, and 20 mm in diameter) were investigated to evaluate their impact on the TPF(Thermal performance factor) of HX. Assessment criteria included *Nu*, *U*, and *TPF,* and configurations with and without baffles were compared. Results indicated significant enhancement in TPF with baffles, increasing the average h_m by 29.7%, 62%, and 80.6% for hole diameters of 30 mm, 25 mm, and 20 mm, respectively, with the best TPF achieved using baffles with 20 mm diameter holes.

Hassan et al. [23] investigated the impact of hole number and shape on the TEF of both solid and perforated conical rings. The study examined convergent and divergent ring configurations, each with 25 mm and 50 mm end diameters, resulting in a diameter ratio 0.5. The rings had a consistent thickness of 2 mm and a length of 60 mm. The distance between consecutive rings was fixed at 120 mm to maintain a pitch-to-diameter ratio (p/d) 1.875. A constant heat flux of 3000 W/m² was applied to the outer tube, while the *Re* varied from 6000 to 26000. Results showed that solid divergent rings achieved higher *Nu*, reaching 360.2 at *Re* = 26000. However, this configuration also exhibited the highest *f*, peaking at 5.04 at *Re* = 6000; perforating the rings with circular, square, or equilateral triangular holes reduced both the *Nu* and *f*, thereby enhancing TEF, especially with more holes. Circular holes demonstrated the highest TEF, while triangular holes showed the lowest. For solid divergent rings at *Re* = 6000, a maximum TEF of 1.1 was achieved. In contrast, circularly perforated rings achieved a TEF of 1.06 with lower pumping power requirements, showcasing improved thermal performance while maintaining efficiency.

A unique type of heat exchanger, referred to as an axial flow tubular heat exchanger, was developed by Rahman and Dhiman [2, 24, 25]. The purpose was to enhance the HT by creating a swirl airflow path across the tube bundle in a tubular HX. Perforated circular baffle plates with trapezoidal air deflectors(inclined at various angles) were used to generate swirl flow. Each plate has four deflectors arranged clockwise at identical angles, creating a swirling flow as air enters the HX. However, the tube configuration remained unchanged, maintaining a consistent heat flow.

Consequently, the swirling motion increased turbulence in the air, improving HT. The study was conducted by varying the baffle spacing under the *Re* range of 16000 to 28000. The influence of *PR* and α on the h_m and *f* of the HX was investigated. The result revealed that an HX with a deflector angle of 50° and a Pitch ratio (l/D) of 1.4 displayed an augmentation of 3.75 in TPF compared to others. This work was further extended by Rahman [26], who used an oriented trapezoidal deflector and found an increase in TEF of 2% compared to Rahman and Dhiman[24].

Rahman, M. A. [26, 27] further explored the consequence of rectangular punched holes with one or two flow deflectors. The influence of deflector orientation on heat transfer and fluidic performance was estimated through experiments. When using one deflector, a surge of 41.49% was seen in TPF. Employing two flow deflectors of opposing orientations decreased TEF by 0.19 times compared to an HX with an SBP. The results indicated that the opposite-oriented deflectors achieved a higher HT rate than inline ones, although this advantage came with higher Δp losses. Rahman [28, 29] further directed his work to see the effect of multiple rectangular deflectors' orientations (inline shutter type and opposite-oriented). The result indicated that the shutter type shows a higher TEF; however, the opposite orientation of the deflector shows a higher h_m and higher Δp. Going further, Rahman studied different deflector geometries, such as triangular [29] and sawteeth [30]. Upon changing the shape of the baffle plate from circular to conical [31-33], Rahman achieved a higher hm value nearly 1.5 times compared to the circular baffle plate at the expense of Δp.

Based on the analysis of the gathered data, the following insights have been observed regarding the HT properties of non-decaying swirl flow:

1) The *Nu* for swirl flow exceeded that of axial flow. This enhancement is due to higher local velocities, increased turbulence, and buoyant forces caused by the flow's curvature and centrifugal effects.
2) Significant Δp occur because baffles obstruct fluid flow, causing flow separation near the edges of the baffles. Consequently, greater pumping power is often required to compensate for the increased Δp while maintaining the same heat load.
3) A smaller pitch ratio typically results in higher heat transfer rates and greater Δp.

Additionally, swirl flow facilitates the HX's compaction, ultimately improving space efficiency. Swirlers can also aid in eliminating stagnation zones by promoting a more uniform temperature distribution over the tube surface, which proves essential in applications where a uniform temperature profile is critical, such as thermal processing or heating/cooling scenarios.

The recent work on swirl generation mentioned above in the literature indicates that the performance of an HX is not solely dependent on the *Re*/velocity/mass flow rate of the working fluid but is sensitive to the direction of flow (parallel/counter flow) but also a type of flow (axial/radial/swirl). Among these, swirl flow gives a maximum TEF value. Nonetheless, there is limited research on the thermal-fluid properties of HX featuring baffle plates as turbulators. It is also noted that the TEF is sensitive to the geometry of the swirler, as shown by Rahman [32-43]. Only a handful of geometry has been studied, requiring further exploration. Furthermore, using air as a working fluid has been insufficiently studied despite its numerous advantages, such as ease of disturbance introduction and lower fouling resistance than other liquids. Although air allows for better flow control due to its non-sticky nature, its hydrodynamic and thermal boundary layers resemble those of other liquids.

This study aims to develop and experiment with a novel swirl generator which employs trapezoidal deflectors (TDBP) positioned opposite each other on the baffle plates.

Through the experimental evaluation processes, the following objectives were planned:

- To develop setup and fabricate swirl-inducing perforated baffle plate designs with flow deflectors and its validation with the reported work.
- To investigate the effect of deflector's α on heat transfer for various *Re*.
- To investigate the baffle plate spacing (PR) effect on HT for various *Re*.
- To investigate how the Δp and f have been affected by varying the *Re* and PR of the HX.
- To estimate TEF of the HX by varying *PR*, *α*, *Re* based on *j*, *f*, Δp.

The parameter range is as follows:

- *PR* (*l*/*D*): 0.6, 0.8, 1, 1.2.
- *α* (inclination angle): 30°, 40°, 50°.
- *Re* (Reynolds number): 15,000 to 30,000.

The resultant *Nu*, *f* values are compared with those obtained from HX without a baffle plate working under similar operation parameters.

2 METHODOLOGY

2.1 Experimental Setup with HX Details

The experiment employs a variety of equipment, including the air intake unit, HX, pressure and temperature gauges, a water-circulating loop, and a data acquisition system. Detailed information about the equipment can be seen in references 3, 24, and 25. Three pressure ports are generated on both the inlet and exit of the HX to obtain the Δp. Ten RTD thermocouple probes are pasted on the tube surface carrying hot water to acquire surface temperature (using DAQ assistant NI-9213 thermocouple modules), and two additional thermocouples are used to acquire air temperature as it arrives and departs the HX. The velocity of air is calculated using the pitot tube. The 8 kW variac electrical heating element with thermostat (SSU 0-300 °C) controlled ensures a constant temperature of 60 °C for the heated water. A pump and Rotameter (LZS-25) control the hot water flow rate. The airflow rate can be adjusted while keeping the flow rate (water) constant during the experiment.

Table 1 Testing conditions used in experiments [32, 33]

Air-inlet temperature, °C	32.5 ± 0.5
Air-inlet velocity, m/s	7 - 10
Water-inlet temperature, °C	60 - 65
Water mass flow rate, kg/s	0.06

The experimentation conditions are outlined in Tab. 1, with the precision of the instruments is evaluated using the root mean square method, with precision in temperature measurement of ± 0.5, Δp of ± 0.1, and flow rate of ± 0.006. The approximate highest level of uncertainty connected with the Re = ± 3.25, v = ±5, f = ±5.34, and Q = ±5%.

The test section revealed in Fig. 1 is an acrylic duct with a length of 60 cm, a width of 19 cm, and a thickness of 0.5 cm, with a thermal conductivity (k_p) of 0.2. Parallel to the duct are copper tubes composed of Copper (C12200), which have a k_t = 300 (W/mK). The tubes possess an inner width of 8 mm with a thickness of 1mm, carrying hot water supplied by the distributor. They are supported by innovative baffles [32-40].

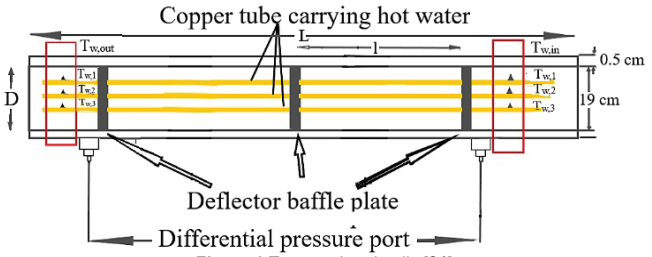


Figure 1 Test section details [24]

In contrast, the fluid on the other side of the duct is air drawn in from the surroundings, as the primary focus of this study is on air. Two groups of thermocouples are utilized to acquire the surface temperature of the copper tubes carrying the hot fluid: $T_{w,\,in}$, and $T_{w,\,out}$. Each group contains 5 thermocouples (t_{w1}-t_{w5}), one devoted to each tube. The arrangement of these thermocouples for capturing the temperature at the inlet and outlet of the copper tubes is illustrated by Rahman [32, 33]. Moreover, additional thermocouples ($T_{a,\,in}$ and $T_{a,\,out}$) are employed to record the air temperature as it passes through the test section.

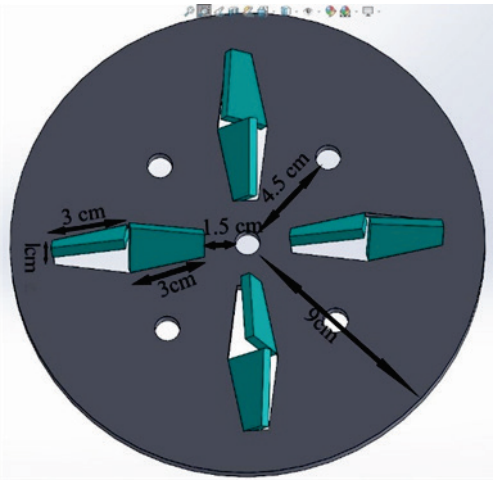


Figure 2 CAD model of Trapazoidal deflector baffle plate

The diagram in Fig. 2 illustrates the latest Trapezoidal Deflector Baffle Plate (TDBP) developed. It features a single central tube encircled by four supplementary tubes organized in a circular formation. These tubes are arranged 4.5 cm from the centre at an angle of 60° from each other. The TDBP has four openings to enable airflow into the ducts. The trapezoidal holes have been designed with 1*d* and 2*d* dimensions for the parallel sides and a side of 3*d*, where *d* represents the tube diameter (1 cm). Two identically sized trapezoidal deflectors have been pasted in reverse at a prearranged angle alongside the baffle plate (as shown in Fig. 3a). The deflector arrangement converts airflow from axial to swirl patterns. The deflector's presence initiates swirl motion in the axial air stream, transforming the airflow into a well-defined swirling pattern as it passes through the tube bundles. The angle α strengthens the axial flow and converts it into plug flow.

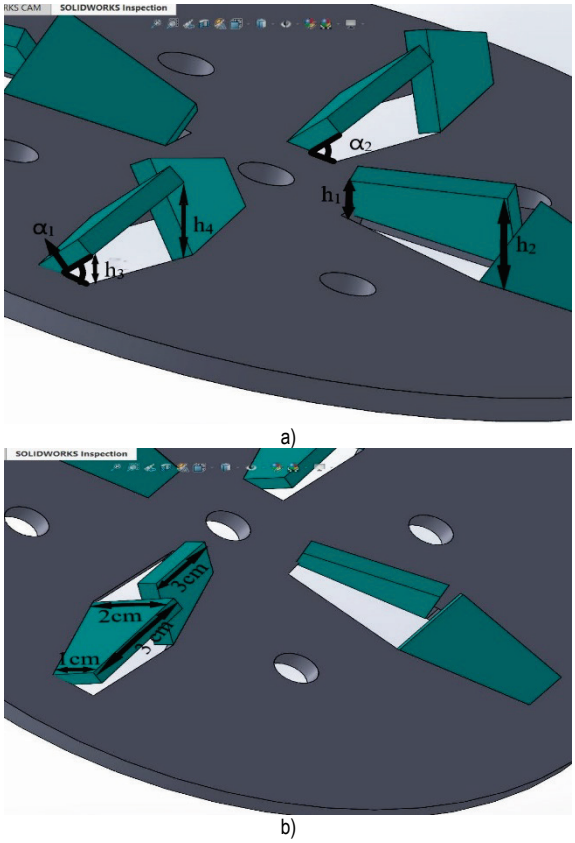


Figure 3 a) Baffle plate detail and b) deflector detail

Consequently, opposite swirl streams are generated in the baffle spacing, preventing the formation of stagnant zone. The swirling motion increases *Δp* and augments HT. The baffle plates spacing, known as the pitch of the baffle plate, is crutial in formation of vortices, recirculation, and turbulence within the baffle spacing. This study evaluated their impact by examining PR values of 0.6, 0.8, 1, and 1.2. As a result, turbulence and vortices are generated within the duct due to the rotational air structures, which constantly clean the tube wall and manipulate the thermal boundary layer. The research investigates three different angles: α = 30°, 40°, and 50°. This setup creates an opening similar to a vent, allowing air circulation and serving as the flow area. The deflectors are located 15 mm from the center. All deflectors have the constant height ratio (h_1/h_2 and h_3/h_4 = 0.5) irrespective of the α, as shown in Fig. 3b

2.2 Design Constraint and Data Reduction

Pitch ratio [24-28] is estimated as

$$PR = \frac{l}{D} \tag{1}$$

Blockage ratio [30-32] is estimated using

$$BR = \frac{U}{S} \tag{2}$$

Where U is the cross-sectional area of the baffle plate (4 times the cross-sectional area of the rectangular opening) and S is the cross-sectional area of the baffle plate.

This paper examined the impact of TDBP in a circular channel on thermo-hydraulic performance, with a fixed BR value of 0.7, under different *Re* conditions. To gather data, multiple tests were conducted on the channel using various baffle plates with different α (30°, 40°, and 50°). A separate experiment was also carried out on a channel within the same Re range without a baffle. The obtained results were evaluated to determine the effects of PR and α on the heat exchanger's thermo-hydraulic efficiency.

Ranman's [24] method calculates the mean convection heat transfer coefficient ($h_{c,\,m}$ in W/m^{2}K).

$$Re = \frac{\rho \cdot v \cdot D_h}{\mu} \tag{3}$$

Where v represents the average speed in meters per second and D_h as the hydraulic diameter of the circular tube in meters. In this situation, we can compute the thermal properties of air, such as ρ (density in kg/m^3) and μ (dynamic viscosity coefficient in kg/m-s), by using the average temperature values of the air entering and leaving the system.

$$v = \sqrt{\frac{2\Delta p_o}{\rho}} \tag{4}$$

$$h_m = \frac{Q}{A_p \cdot \Delta t_{lm}} \tag{5}$$

In the given scenario, Q represents the rate at which heat is transferred from the air side, measured in Watts. The A_p symbolizes the heat transfer area on the copper tube, which is expressed in square meters. ΔP_o indicates the pressure drop at the orifice plate, measured in Pascals, and Δt_{lm} refers to the logarithmic mean temperature difference between the air and the copper tube's wall. HT rate is calculated as.

$$Q = C_p \cdot \rho \cdot v \cdot A_c \cdot (T_{a,\,out} - T_{a,\,in}) \tag{6}$$

A_c is annulus crossectional area, calculated as

$$A_c = \frac{\pi}{4} \cdot (D - D_e) \tag{7}$$

Where: D_e is the equivalent diameter of the HX; $T_{a,\,in}$ and $T_{a,\,out}$ average air temperature at inlet and outlet, respectively, whereas Δt_{lm} is LMTD for parallel flow given as

$$\Delta t_{lm} = \frac{(T_{w,\,in} - T_{a,\,in}) - (T_{w,\,out} - T_{a,\,out})}{\dfrac{\ln(T_{w,\,in} - T_{a,\,in})}{(T_{w,\,out} - T_{a,\,out})}} \tag{8}$$

when $T_{w,\,in}$ and $T_{w,\,out}$ are the average copper surface temperature at inlet and exit, respectively, determined as

follows:

$$T_{\mathrm{w,\,in}} = \left(\frac{\sum_{1}^{5} t_{\mathrm{w},\,i} A_i}{A_{\mathrm{p}}} \right)_{\mathrm{in}}, \; T_{\mathrm{w,\,out}} = \left(\frac{\sum_{1}^{5} t_{\mathrm{w},\,i} A_i}{A_{\mathrm{p}}} \right)_{\mathrm{out}} \tag{9}$$

The heating module contains five temperature junctions, indicated as i, on the copper pipe. These junctions are positioned at the entry and exit points of the experimental segment, aligning with the airflow. A_i corresponds to the surface area responsible for heat transfer. The average Nusselt number (Nu) and friction factor (f) are employed to express the flow and thermal characteristics of the duct.

$$Nu = \frac{h_{\mathrm{c,\,m}} D_{\mathrm{h}}}{\lambda} \tag{10}$$

$$f = \frac{2\Delta P \cdot D}{\rho \cdot v^2 \cdot L} \tag{11}$$

Δp is the pressure drop in the HX.

From Eqs. (5) and (6) follows

$$h_{\mathrm{m}} = \frac{C_{\mathrm{p}} \cdot \rho \cdot v \cdot A_{\mathrm{c}} \cdot (T_{\mathrm{a,\,out}} - T_{\mathrm{a,\,in}})}{A_{\mathrm{p}} \cdot \Delta t_{\mathrm{lm}}} \tag{12}$$

Reative dimensionless quantities such as thermal transfer enhancement (j/j_0), relative flow obstruction (f/f_0), and a benchmark $TEF = (j/j_0)/ (f/f_0)^{(1/3)}$ were utilized. The j and f values of an HX with a Baffle plate, while f_0 and j_0 are for HX without a baffle plate [38-43].

2.3 Validation

To assess the precision of the experiment, the obtained Nu and f are compared with values derived from standard correlations. Nu is calculated using the Gnielinski, and Dittus equations, while the Colebrook-White and Blasius correlations are employed for f. The comparison reveals the following mean deviations from experimental values: –9.054% for Gnielinski, +8.195% for Dittus and Boelter, +0.73% for Blasius and –0.649% for Colebrook-White [32].

3 RESULTS AND DISCUSSION

3.1 HT Augmentation

Enhancing HT can be achieved by installing a deflector baffle, which induces turbulent flow and disrupts thermal boundary layer formation for the experimental Re range.

Fig. 4 depicts the variation of the relative colburn factor (j/j_0) as a function of Re. Initially, there is an increase in j/j_0 until it reaches its peak value, followed by a gradual decline with a rise in Re. This pattern is observed consistently for all tested samples of the turbulent deflector baffle plate (TDBP). It also shows that increasing α value leads to a decline in j/j_0.

In turbulent flow, the energy accompanying the chaotic fluid movements or eddies is distributed throughout the fluid due to molecular interaction, generating multiple convective cells that facilitate faster HT than laminar flow. The transport of energy through cell interaction is effective across a large area, expediting the diffusion process. The movement of these eddies generates areas of varying Δp, replenishing the energy required for convection and accelerating the HT rate. Therefore, turbulent flow promotes HT by enhancing convection and diffusion. Fascinatingly, as α declines, so does the flow area in this scenario deflector now acting as nozzle; further increase in Re value makes flow unstable.

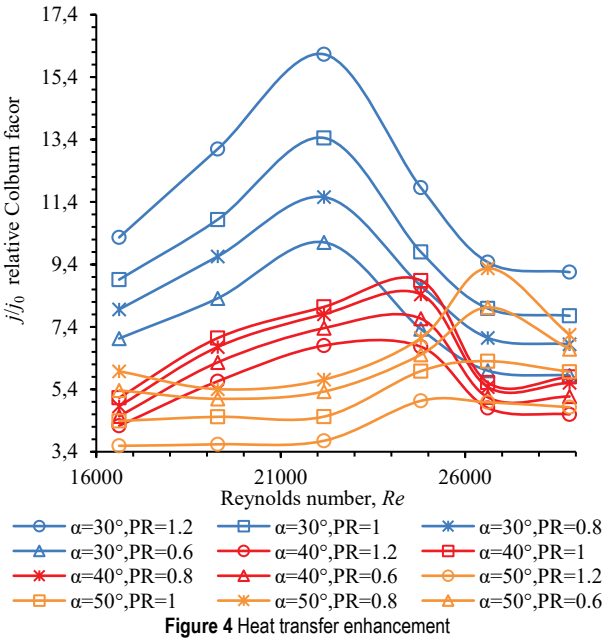


Figure 4 Heat transfer enhancement

For a deflector angle of 30°, the TDBP samples show the highest j/j_0 ratios compared to others. This indicates that a lower angle increases flow disruption and instability. When the deflector is angled towards the baffle plate, it narrows the flow area abruptly, boosting air velocity in the duct and creating a jet effect. This setup prevents boundary layer formation on the tube and wall, while the rotating fluid washes over the tube bundles, inducing a spiral flow that enhances mixing and HT.

Increased flow velocity improves the fluid's ability to clear the duct area, lowering dynamic air Δp near the walls of the duct and tube, enhancing heat transfer efficiency and raising Δp. The angle of the deflector significantly influences heat transfer rates: as the α declines, local air velocity and turbulence increase, leading to greater cavitation and enhanced air molecule interaction with the wall, which boosts heat transfer from the wall.

Research by Wang et al. [34] found that very close spacing between deflector baffles intensifies vortex interactions, causing them to break up and reduce heat transfer. Conversely, boundary layer separation occurs too early if the spacing is too broad, leading to higher pressure loss. Therefore, optimizing the spacing (PR) between

turbulent deflector baffle plates is crucial for maximizing heat transfer. Tab. 2 details the highest j/j_0 ratio for various *Re*, PR, and α.

Table 2 Maximum values

α	j/j_0			j/j_0		
	Max value			Max values		
	PR	j/j_0	*Re*	*PR*	*Re*	j/j_0
30	1.2	16.13	22200	0.6	29000	9.38
40	1	8.88	24800	0.6	29000	8.61
50	0.8	9.27	26600	0.6	29000	7.64

According to a study conducted by Wang et al. [34], smaller *PR* optimizes HT. This finding is sustained by the information provided in Tab. 2, which shows that the *PR* increases as the α increases.

The turbulence created in the baffle plate spacing affects the HT rate. When the *PR* is set at 1.2 and α is 30°, the average j/j_0 peaks at 16.13. On the other hand, for higher α values, the highest average j/j_0 is achieved at lower *PR* values such as 1 and 0.8. A smaller *PR* reduces the baffle spacing available for air and surface interaction. This promotes flow reversal and enhances the tube washing ability, resulting in better heat transfer performance ($h_{c,m}$).

Conversely, heat transfer diminishes as the *PR* increases due to a lower interface interaction between the fluids and wall, leading to lower $h_{c,m}$ values and slower heat transfer rates. Additionally, when the baffle spacing increases, the ΔT between the tube surface and the surrounding air decreases, resulting in a decline of HT. A decline in thermal contact (surface and air) causes this.

3.2 Relative Flow Resistance and Thermo-Fluid Performance

The data in Fig. 5 illustrates the relationship amongst f/f_0 and *Re*. All TDBP models exhibit a comparable trend, with f/f_0 starting low at lower *Re* and rising until reaching a peak. The maximum f/f_0 value was noted at $\alpha = 30°$, which declines with increasing α and *PR*. Consequently, at $\alpha = 30°$, it shows maximum HT and *f* losses, particularly at *PR* of 0.6. As α increased beyond 30°, the occurrence of blockages decreased, leading to reduced turbulent flow and limited eddies that typically cause pressure fluctuations along the inner side of the duct.

Previous research indicates that regions with higher turbulence, velocity, and extended fluid-surface interactions experience increased Δp. Consequently, it is anticipated that Δp will follow the sequence of α = 30°, 40°, and 50°. With smaller PR, lower α values lead to higher average f/f_0, with a peak of 9.38 at α = 30° and PR = 0.6. The addition of deflectors can increase Δp by disrupting the fluid's kinetic energy as it flows along the HX. The deflector having the smallest α generates a prominent swirling motion, resulting in substantial rotational interaction amongst the secondary flow and the wall. This interaction creates significant turbulence, causing a high Δp with elevated *f*.

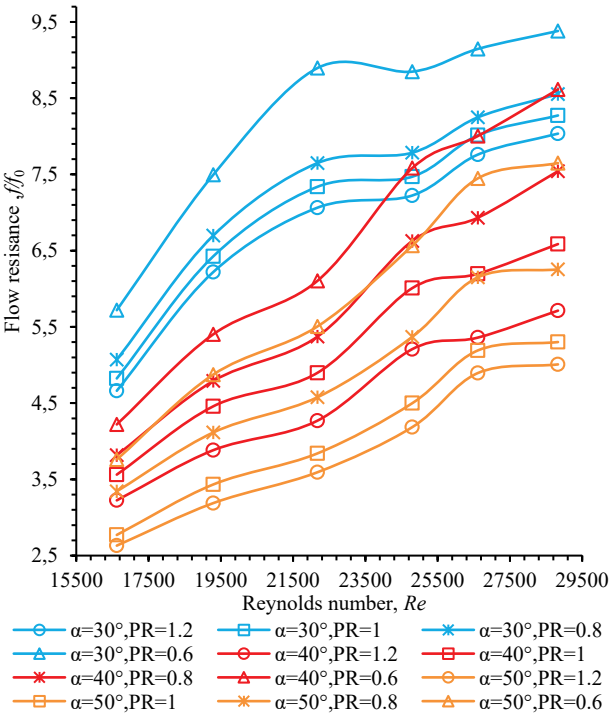


Figure 5 Flow resistance vs *Re*

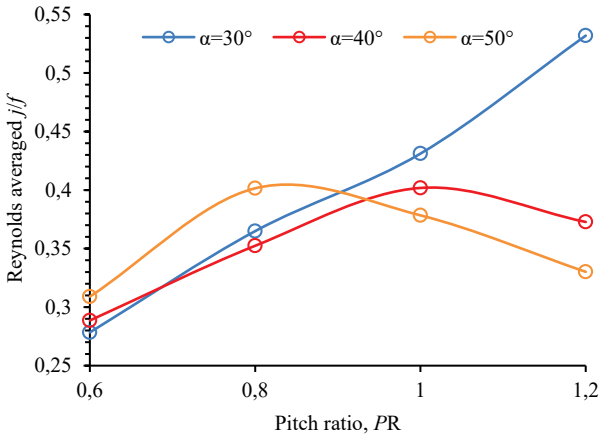


Figure 6 Area goodness factor vs *PR*

The term Area goodness factors defy the compactness of HX. Fig. 6 shows Reynonald's average *j/f,* which shows that higher *PR* works best for a given range of *Re*, i.e., 1.2, whose value reduces with a rise in α. Furthermore, the TEF of the heat exchanges has been analyzed and plotted in Fig. 7, which shows a similar trend as Fig. 5, displaying higher *TEF* values at lower α. The maximum *TEF* is seen at α of 30° because of immense turbulence due to swirl flow generated by low α and multiple baffles along the flow. This is because increasing *PR* increases the area for airflow, reducing Δp and lower velocity. The maximum *TEF* of 2.48 is seen at *PR* of 1.2 and an α of 30°.

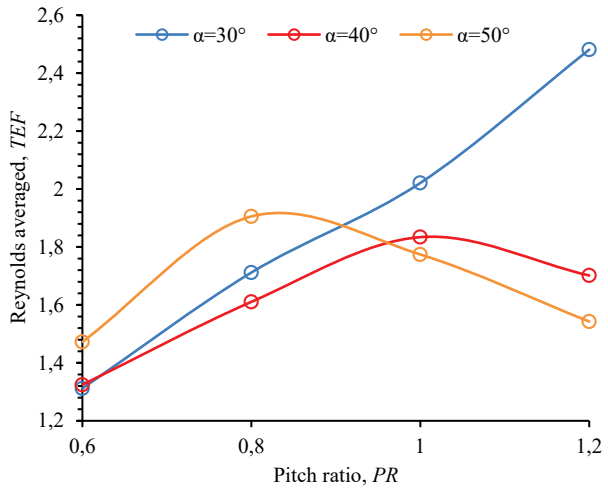


Figure 7 Thermal enhancement factor vs *PR*

4 CONCLUSIONS

The analysis of TDBP samples showed that increasing the duct α led to higher flow velocities, with increments of 29.6%, 21.7%, and 14.2% for α values of 30°, 40°, and 50°, respectively, in comparison to ducts without baffle plates. The HX gives optimal value for smaller α value and higher PR, with the TEF peaking at 2.48 for $\alpha = 30°$. Δp was greatest at smaller α values, showing reductions of 31.37%, 22.94%, and 17.31% for α = 30°, 40°, and 50°, respectively. The TDBP worked efficiently with *Re* below 25,000, providing good HT and *f* properties, although these diminished at higher *Re*. The impact of PR was minimal with air as the working fluid, suggesting that future studies should explore high-density fluids like slurry, water or oil. Different deflector arrangements should also be investigated to enhance swirl flow and TEF. The current study used a fixed baffle ratio (*BR*) of 0.7, which could be increased to lower Δp, and achieved a maximum velocity of 13 m/s, which could be further increased. The research focused on three baffle plates with four different PR values; exploring additional baffles and PR values and a wider range of α values is recommended. Finally, while the study used non-metallic baffle plates, future research should incorporate metallic plates to contribute actively to HT.

5 REFERENCES

[1] Rahman, M. A. & Dhiman, S. K. (2024). Study of Flow and Heat Transfer in Swirled Tubular Recuperator. http://hdl.handle.net/10603/576426

[2] Rahman, M. A. & Hasnain, S. M. M. (2024). Performance Improvement of Heat Exchanger with Perforated/Non Perforated Flow Modulator Producing Continous/Discontinous Swirl Flow. *Heat transfer*. https://doi.org/10.1002/htj.23135

[3] Rahman, M. A. & Dhiman, S. K. (2023). Performance evaluation of turbulent circular heat exchanger with a novel flow deflector-type baffle plate. *Journal of Engineering Research*, 100105. https://doi.org/10.1016/j.jer.2023.100105

[4] Boushaki, T. (2019). Introductory Chapter: Swirling Flows and Flames. IntechOpen. https://doi.org/10.5772/intechopen.86495

[5] Iyogun, C. O., Birouk, M. & Kozinski, J. A. (2011). Experimental investigation of the effect of fuel nozzle geometry on the stability of a swirling non-premixed methane flame. *Fuel, 90*(4), 1416-1423. https://doi.org/10.1016/j.fuel.2010.12.033

[6] Schmittel, P., Günther, B., Lenze, B., Leuckel, W. & Bockhorn, H. (2000). Turbulent swirling flames: Experimental investigation of the flow field and formation of nitrogen oxide. *Proceedings of the Combustion Institute, 28*(1), 303-309. https://doi.org/10.1016/S0082-0784(00)80224-6

[7] Boushaki, T., Sautet, J. C. & Labegorre, B. (2009). Control of flames by tangential jet actuators in oxy-fuel burners. *Combustion and flame, 156*(11), 2043-2055. https://doi.org/10.1016/j.combustflame.2009.06.013

[8] Elbaz, A. M. & Roberts, W. L. (2016). Investigation of the effects of quarl and initial conditions on swirling non-premixed methane flames: Flow field, temperature, and species distributions. *Fuel*, 169, 120-134. https://doi.org/10.1016/j.fuel.2015.12.015

[9] Rahman, Md. A., Hasnain, S. M. M., Pandey, S., Tapalova, A., Akylbekov, N. & Zairov, R. (2024). Review on Nanofluids: Preparation, Properties, Stability, and Thermal Performance Augmentation in Heat Transfer Applications. *ACS Omega*. https://doi.org/10.1021/acsomega.4c03279

[10] Sarac, B. A. & Bali, T. (2007). An experimental study on heat transfer and pressure drop characteristics of decaying swirl flow through a circular pipe with a vortex generator. *Experimental Thermal and Fluid Science, 32*(1), 158-165. https://doi.org/10.1016/j.expthermflusci.2007.03.002

[11] Jafari, M., Farhadi, M. & Sedighi, K. (2017). An experimental study on the effects of a new swirl generator on thermal performance of a circular tube. *International Communications in Heat and Mass Transfer*, 87, 277-287. https://doi.org/10.1016/j.icheatmasstransfer.2017.07.016

[12] Rahman, M. A., Hasnain, S. M. M. & Zairov, R. (2024). Assessment of improving heat exchanger thermal performance through implementation of swirling flow technology. *International Journal of Thermofluids*, 22, 100689. https://doi.org/10.1016/j.ijft.2024.100689

[13] Nair, S. R., Oon, C. S., Tan, M. K., Mahalingam, S., Manap, A. & Kazi, S. N. (2022). Investigation of heat transfer performance within annular geometries with swirl-inducing fins using clove-treated graphene nanoplatelet colloidal suspension. *Journal of Thermal Analysis and Calorimetry, 147*(2), 14873-14890. https://doi.org/10.1007/s10973-022-11733-6

[14] He, Y. L. & Zhang, Y. (2012), Advances and Outlooks of Heat Transfer Enhancement by Longitudinal Vortex Generators, *Advances in Heat Transfer*, 44, 119-185. https://doi.org/10.1016/B978-0-12-396529-5.00002-0

[15] Suja, S. B., Islam, M. & Ahmed, Z. U. (2023). Swirling jet impingements for thermal management of high concentrator solar cells using nanofluids. *International Journal of Thermofluids*, 19, 100387. https://doi.org/10.1016/j.ijft.2023.100387

[16] Rahman, Md. A. & Hasnain, S. M. M. (2024). Performance improvement of heat exchanger with perforated/non perforated flow modulator producing continous/discontinous swirl flow: A comprehensive review. *Heat transfer*. Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1002/htj.23135

[17] Wang, J., Fu, T., Zeng, L., Lien, F. S., Wang, H. & Deng, X. (2022). A Comparative study on thermo-hydraulic performance in a tube with different punched winglets, *International Journal of Thermal Sciences*, 181, 107772. https://doi.org/10.1016/j.ijthermalsci.2022.107772

[18] Wang, J., Zeng, L., Fu, T., Yu, S. & He, Y. (2024). Effects of the position and perforation parameters of the delta winglet vortex generators on flow and heat transfer in mini channels. *International Journal of Thermal Sciences*, 198, 108878. https://doi.org/10.1016/j.ijthermalsci.2023.108878

[19] Ajarostaghi, M., Zaboli, S. S., Kiani, B. M., Saedodin, S., Karimi, N. & Javadi, H. (2022). Hydrogen preheating in a PEMFC system employing a heat exchanger equipped with an innovative turbulator. *International Journal of Hydrogen Energy,* 47(85), 36264-36282. https://doi.org/10.1016/j.ijhydene.2022.08.204

[20] Mousavi Ajarostaghi, S. S., Aghanezhad, M., Davudi, H. & Mohammadzadeh Amiri, M. (2021). Numerical evaluation of the heat transfer enhancement in a tube with a curved conical turbulator insert. *International Journal of Ambient Energy*, 43(1), 5218-5231. https://doi.org/10.1080/01430750.2021.1945490

[21] Wang, D., Khalatov, A., Shi-Ju, E. & Borisov, I. (2022). Swirl flow heat transfer and flow characteristics in a solid and permeable pipe with exit nozzle. *International Journal of Thermal Sciences*, 173, 107425. https://doi.org/10.1016/j.ijthermalsci.2021.107425

[22] Hussein, M. A. & Hameed, V. M. (2022), Experimental Investigation on the Effect of Semi-circular Perforated Baffles with Semi-circular Fins on Air–Water Double Pipe Heat Exchanger. *Arabian Journal for Science and Engineering, 47*, pp. 6115-6124. https://doi.org/10.1007/s13369-021-05869-0

[23] Hassan, A. Md., Al-Tohamy, A. H. & Kaood, A. (2022), Hydrothermal characteristics of turbulent flow in a tube with solid and perforated conical rings. *International Communications in Heat and Mass Transfer, 134*. https://doi.org/10.1016/j.icheatmasstransfer.2022.106000

[24] Rahman, M. A. & Dhiman, S. K. (2023). Investigations of the turbulent thermo-fluid performance in a circular heat exchanger with a novel flow deflector-type baffle plate. *Bulletin of the Polish Academy of Sciences Technical Sciences*. https://doi.org/10.24425/bpasts.2023.145939

[25] Rahman, M. A. & Dhiman, S. K. (2024). Investigations on thermo-fluid performance of a circular heat exchanger with a novel trapezoidal deflector-type baffle plate. *Thermal engineering*, 10, 1-13. https://doi.org/10.56304/S0040363624700292

[26] Rahman, M. A. (2024). Study the effect of axially perforated baffle plate with multiple opposite-oriented trapezoidal flow deflectors in an air–water tubular heat exchanger. *World Journal of Engineering*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/WJE-10-2023-0425

[27] Rahman, M. A. (2023). Experimental Investigations on Single-Phase Heat Transfer Enhancement in an Air-To-Water Heat Exchanger with Rectangular Perforated Flow Deflector Baffle Plate. *International Journal of Thermodynamics*, 1-9. https://doi.org/10.5541/ijot.1285385

[26] Rahman, M. A. (2023). Effectiveness of a tubular heat exchanger and a novel perforated rectangular flow-deflector type baffle plate with opposing orientation. *World Journal of Engineering*. https://doi.org/10.1108/WJE-06-2023-0233

[27] Rahman, M. A. (2024). Thermo-hydraulic effect of tubular heat exchanger fitted with Perforated baffle plate with rectangular shutter-type deflector, *Korean Chem. Eng. Res., 62*(2), 1-9. https://doi.org/10.9713/kcer.2024.62.2.1

[28] Rahman, M. A. (2024). Thermo-Fluid Performance Comparison of an Inline Perforated Baffle with Oppositely Oriented Rectangular-Wing Structure in Turbulent Heat Exchanger. *International Journal of Fluid Mechanics Research*, *51*(1), 15-30. https://doi.org/10.1615/InterJFluidMechRes.2023051418

[29] Rahman, M. A. (2024). The effect of triangular shutter type flow deflector perforated baffle plate on the thermofluid performance of a heat exchanger. *Heat Transfer, 53*(2), 939-956. https://doi.org/10.1002/htj.22981

[30] Rahman, M. A. (2024). Thermal hydraulic performance of a tubular heat exchanger with inline perforated baffle with shutter type saw tooth turbulator. *Heat Transfer, 53*, 2234-2256. https://doi.org/10.1002/htj.23034

[31] Rahman, M. A. (2024). Thermo-Fluid Performance of a Heat Exchanger with a Novel Perforated Flow Deflector Type Conical Baffles. *Journal of thermal engineering, 10*(4), 868-879. https://doi.org/10.14744/thermal.0000846

[32] Rahman, M. A. (2024). The influence of geometrical and operational parameters on thermofluid performance of discontinuous colonial self-swirl-inducing baffle plate in a tubular heat exchanger. *Heat Transfer, 53*, 328-345. https://doi.org/10.1002/htj.22956

[33] Rahman, M. A. (2024). Thermal performance of tubular heat exchangers with the discontinuous swirl-inducing conical baffle with opposite-oriented flow deflectors. *Archives of Thermodynamics, 45*(2), 195-204. https://doi.org/10.24425/ather.2024.150865

[34] Wang, D., Khalatov, A., Shi-Ju, E. & Borisov, I. (2022). Swirl flow heat transfer and flow characteristics in a solid and permeable pipe with exit nozzle. *International Journal of Thermal Sciences,* 173, 107425. https://doi.org/10.1016/j.ijthermalsci.2021.107425

Author's contacts:

Md. Atiqur Rahman [a, b]
(Corresponding Author)
[a] Department of Mechanical Engineering,
Vignan's Foundation for Science, Technology & Research
(Deemed to be University),
Vadlamudi, Guntur-522213, Andhra Pradesh, India
rahman.md4u@gmail.com

[b] Department of Mechanical Engineering,
Birla Institute of Technology,
Mesra, Ranchi-835215, India

S. M. Mozammil Hasnain
Marwadi University Research Centre,
Department of Mechanical Engineering,
Faculty of Engineering & Technology,
Marwadi University,
Rajkot, Gujrat-360003, India

Rustem Zairov
Aleksander Butlerov Institute of Chemistry,
Kazan Federal University,
1/29 Lobachevskogo Str., Kazan 420008, Russian Federation

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240523122223

Original scientific paper
Received: 2024-05-23, Accepted: 2024-05-27

Factors Influencing the Purchase of Battery Electric Vehicles (BEVs): An Explorative Study Based on the Analysis of New Registrations and Expert Interviews in Germany

Johannes Hoffmann, Verena Szkudlarek, Daniela Ludin, Norbert Schreier*, Erika Mueller, Wanja Wellbrock

Abstract: This research aims to identify the key factors influencing the new registrations of battery electric vehicles (BEVs) in Germany. Focusing the topics economy, mobility change and climate sustainability a mixed-methods approach was used combining a regression analysis based on public databases and qualitative interviews with experts, represented by vehicle testing organizations, BEV leasing companies, car dealerships and automotive magazines. The study revealed similarities as well as discrepancies between expert opinions and the regression analysis results in the fields of real GDP, education rate, the average range of BEVs, the real estate price index, charging infrastructure and political conviction. Regarding practical implications, the results of this study help BEV manufacturers to enhance their marketing and product strategies. The findings of this study could help decision-makers to better understand the mind set of BEV customers.

Keywords: Battery Electric Vehicle (BEV); change in mobility; climate change; sustainable mobility

1 INTRODUCTION

Electro mobility has become increasingly important in recent years due to the reduction of Carbon Dioxide (CO_2) emissions in road traffic. The transport sector in Germany is a significant source of CO_2 emissions, to which road traffic contributes in particular. In 2018, road traffic accounted for around 18% of total CO_2 emissions in Germany [1]. Global warming, primarily caused by the greenhouse gas CO_2, is progressing at an increasing rate and is therefore one of the greatest challenges. In 2015, 195 countries including Germany agreed at the international climate conference in Paris to limit the temperature increase to a maximum of two degrees Celsius [2]. The proportion of BEVs (battery electric vehicles) in Germany is still low compared to conventional passenger cars with combustion engines. The research aim of this study is identifying measurable, socio-cultural influencing factors, which increase the number of new registrations of BEVs.

To better understand factors that influence new registrations of BEVs in contrast to combustion engines and to promote new registrations through this understanding, in this study a comprehensive analysis was carried out. The study consists of two parts, which are composed of a regression analysis with six variables as influencing factors and expert interviews with the aid of a supplementary questionnaire. The aim is to develop a statistical model and to identify and evaluate various influencing factors. The retrieved data is compared with the experts' interviews. The database used for this study is derived from a panel from 2018 to 2022 for the 16 German federal states.

2 LITERATURE REVIEW

Looking at the change of mobility towards electro mobility it quickly becomes clear how comprehensive this field is. To gain a complete and esteemed understanding of the existing literature, a systematic review based on the framework proposed by Zawacki-Richter (2019) [3].

2.1 Sustainability and Climate Change

The results show, that in Germany, the field of sustainability and climate change, especially in connection with electro mobility, has proven to be an extremely diverse and dynamic area that has become increasingly important in recent years [4]. With its ambitious climate targets and strong emphasis on sustainability in public policy and debate, Germany gives an example of how to deal with sustainability and climate change. The government has set itself the goal of being almost climate-neutral by 2050, which requires a drastic reduction in greenhouse gas emissions. This goal is manifested in a variety of political measures and initiatives to reduce emissions in various sectors of the economy [5]. However, other studies indicate that despite its ambitious climate targets, Germany will likely struggle to achieve them due to existing challenges and delays in implementing emission reduction measures. Implementing the necessary measures may require an increased political effort to achieve the targets set [6]. The German population's attitude towards these issues is generally positive and characterized by a high level of environmental awareness. This is reflected in the strong public support for renewable energies and the growing concern about the effects of climate change. Many Germans are willing to make personal changes to their lifestyle in order to live in a more environmentally friendly way, which is reflected in an increasing demand for sustainable products and services [7]. Contrary to this study, Klinger, Metag & Schäfer (2022) [8] emphasize scepticism about climate change and the demand for climate-neutral products and services, which is underlined by the statement that the spread of electro mobility in Germany is lagging behind the government's targets.

2.2 Change in Mobility

The transport sector contributes a significant share of total emissions in Germany, which is why the promotion of BEVs and the expansion of the corresponding infrastructure are key components of the country's sustainability strategy [5]. Although a high availability of an extensive repertoire of

technological expertise is available, Germany shows a very slow in-crease in the number of new registrations of BEVs, especially in international comparison with China [9]. In contrast to this, there is a slow growth in the acceptance and development of BEVs in Europe, also especially compared to China [10]. However, Germany is the largest and most advanced market in Europe [9].

In order to decelerate climate change, research and development has made considerable progress in Germany in the field of electro mobility. These range from improving battery technology and increasing the energy efficiency of electric vehicles all the way to developing sustainable production processes, which could increase BEV acceptance in Germany [11]. German car manufacturers and suppliers, supported by government funding programs and research initiatives, are investing heavily in this sector [12]. The willingness of the German population to switch to BEVs is encouraged by tax incentives, purchase premiums and exemption from vehicle tax. Cooperation between stakeholders from the public and private sectors is also crucial to debunking myths about electric mobility and raising awareness of its benefits [13]. Despite many benefits, incentives and subsidies on offer, acquisition costs remain high, which is holding back potential buyers in some cases and range anxiety is also a limiting factor for the purchase of a BEV [14, 15].

The development of a strong domestic market for electro mobility is seen in Germany as an impulse for the transformation to a green economy. Replacing the combustion engine with alternative drive systems can trigger a variety of macroeconomic effects. In addition, a scenario is considered that predicts six million electric vehicles by 2030, which shows that negative effects on vehicle production can be offset by positive effects in the production of energy technologies [16]. A further study emphasizes that the environmental benefits of electric vehicles depend on whether their additional electricity consumption is covered by renewable energy sources [17]. Furthermore, challenges for electro mobility in Germany are identified in the adaptation of vehicle technology, the value chain, the load on the electricity grid, electricity generation for the drive system, transportation and the charging infrastructure [18]. It has been shown that slow progress in the expansion of charging infrastructure is slowing down the transformation to a sustainable transport sector [19]. It is also illustrated that although BEVs have already been offered, promising economic incentives such as purchase premiums by the government [20], economic disadvantages are currently still preventing the widespread use of BEVs [21].

The change in mobility in Germany, increasingly characterized by the transition to electro mobility, is part of a comprehensive strategy to reduce CO_2 emissions and promote sustainable transport solutions. However, the introduction of BEVs meets with different reactions within the population [8]. While some citizens appreciate the advantages such as lower operating costs and environmental friendliness [15], there are concerns about charging times and the short range for BEVs [11, 22]. However, these can be overcome by the integration of charging stations in public areas, residential areas and workplaces [18]. The acceptance of BEVs is higher in urban areas in particular, as shorter driving distances and a better charging infra-structure prevail here [23]. Nevertheless, another study emphasizes that the integration of charging infrastructure in existing rural structures and areas, in particular can still be significantly expanded in order to generate a change in mobility and promote the acceptance of BEVs [24]. Among other things, the development of fast charging stations along the main transport axes is also underway to improve the long-distance suitability of BEVs [10]. It is emphasized that battery capacity and a well-developed charging infrastructure are decisive for long-distance suitability and thus for the acceptance of BEVs [22, 25]. Husarek et al. (2021) also address the importance of differences between urban and rural areas, charging demand on highways and the impact of energy demand and battery size on charging infrastructure [23]. In addition, an increasing average range of the vehicles increases their usability for longer distances, which can open up new customer groups [26, 27]. Following this, it is emphasized that the economic viability of electric cars depends largely on battery technology [28] and that the environmental benefits of electric cars can only be fully exploited if the electricity comes from renew-able sources [29]. Literature offers a few options for this. One is a combination of BEVs and private photovoltaic systems for private households, which is profitable above a mileage of 13,000 km per year [30].

2.3 Socio-economic Factors

The role of environmental awareness in the context of socio-economic factors reflects another commonality. Studies by Krause et al. (2016) [31], Mock et al. (2009) [32] and Letmathe & Suares (2020) [25] show that increased environmental awareness and macroeconomic conditions, such as an increase in gross domestic product, can often increase the willingness to buy BEVs. Sustainable mobility solutions play a relevant role in satisfying the need to reduce greenhouse gases and protect the environment [11]. It is also emphasized that educated groups of people tend to better understand the long-term economic benefits and environmental impacts [33, 34]. However, the relevance of individual influencing factors is not portrayed consistently in the literature. Some highlight the influence of education levels and attitudes as important factors for the acceptance of BEVs [34]. Others focus on charging infrastructure or product-related components [35], while Stockkamp et al. (2021) [36] present a whole range of influencing factors from product, politics and education to social influence and in-come as relevant. The spectrum is expanded by looking at specific areas of application and challenges [35, 37].

Here, the acceptance of BEVs in company fleets is examined, emphasizing the importance of product-related factors and environmental concerns [35]. In contrast, Hildermeier & Jahn (2021) [37] focus on the grid integration of electric mobility in Europe. In particular, they examine incentives for smart charging, which can reduce costs for consumers and promote the integration of renewable energies. The need for an integrated approach is further emphasized.

2.4 Research Gap

Overall, the literature review shows that the acceptance of BEVs is a complex interplay of various factors, ranging from technical and infrastructural aspects, environmental awareness and socio-economic conditions to education levels and safety concerns. An interesting observation is the lack of studies that directly address the influence of the real estate price index and the influence of the political orientation of buyers on the decision to purchase an electric vehicle. Although there are a number of studies that examine factors such as purchase price, range, environmental awareness, charging infrastructure and social influences on BEV purchase decisions, the specific role of the real estate price index and the political orientation of buyers remains largely unexplored. It is not clearly defined in literature whether and to what extent the real estate price index, which can be an indicator of the economic situation of an individual or a region, influences the decision to purchase an electric vehicle. In addition, there is also a lack of studies that investigate if support for or membership in a "green party" influences the propensity to purchase an electric vehicle.

These gaps in literature indicate that there is untapped potential to examine specific factors and to develop a more comprehensive understanding of the determinants behind the purchase of BEVs. Based on this research gaps the following research objectives and research questions emerge (Tab. 1).

RO1 analyses the main factors that influence the purchase of electric cars. Hereby, RQ1 looks at these factors in more detail. This is essential to understand the purchasing behaviour of BEV buyers. The research shows that socio-economic characteristics, attitudes, preferences, availability of charging infrastructure and energy costs all play a role [38]. Financial incentives, information about electric cars, convenience and environmental awareness also influence consumer attitudes and purchase intentions [39]. The availability of publicly accessible charging infrastructure and the visibility of electric cars are a further important factor [40]. RQ2 determines the relationship between individual variables and the acceptance of electric cars. Government measures such as financial incentives, traffic regulatory support and the expansion of charging infrastructure play a key role in promoting BEV acceptance [13]. In addition, studies also show that financial incentives and suitable business models can promote acceptance [41, 22].

RO2 examines the relationship between individual variables and their impact on BEV acceptance. With RQ3 the extent to which environmental awareness, as measured by green participation in the state government, has an influence on the decision in favour of BEVs will be investigated. It is shown that the acceptance of BEVs is related to individual differences in mobility needs and the acceptance of range extensions [43], which highlights the significance of environmental awareness and political orientation in shaping consumer preferences.

RQ4 focuses on the property price index and is based on the expectation that higher property prices are usually found in cities, not in rural areas. Generally, there are also more charging stations and services for BEVs in cities. This raises the question of whether this has an impact on the acceptance of BEVs. RQ5 focuses on the impact of the average range of BEVs as its influence on the purchase decision. Investigating whether longer ranges contribute to increased population acceptance, another study suggests that direct experience with BEVs can change evaluation and psychological factors influencing behavioural intention, potentially boosting acceptance [44].

Table 1 Research Objectives & Research Questions

	Research Objectives	Research Questions
Socio-economic Factors	1. Determine which of the factors analyzed are most relevant when purchasing BEVs. (RO1)	1. Which of the factors analyzed is the one, which most strongly promotes the registration of new BEVs? (RQ1)
		2. Which factors must governments pay attention to if they want to influence the mobility shift towards electro mobility? (RQ2)
Sustainability & Climate	2. Understand the relationship between individual variables and the extent of sustainable influence on these variables. (RO2)	3. Does environmental awareness, as measured by green participation in the state government; influence the decision in favor of BEVs? (RQ3)
		4. Does the property price index influence the population's willingness to buy BEVs? (RQ4)
		5. To what extent does the average range of BEVs influence purchasing decisions? (RQ5)
Change in mobility	3. Find out how electro mobility can be integrated in everyday life in various areas of society in order to achieve broad social acceptance. (RO3)	6. To what extent is the level of prosperity relevant to influence a mobility transformation with regard to electro mobility positively? (RQ6)
		7. What factors influence customer behavior of BEVs and how can these findings be used to improve their suitability for everyday use? (RQ7)

RO3 examines how electro mobility can be effectively integrated into society. RQ6 concerns the importance of domestic purchasing power in order to be able to achieve a change in mobility as a country, especially against the backdrop of even more expensive BEVs. It is examined to which factor the level of prosperity influences the purchase of a BEV. It is shown that socio-economic characteristics and attitudes vary considerably among BEV users, with early adopters having a higher average income, a higher level of education and more cars per household [33]. Further, it is shown, that suburban multi-person households are likely to be early adopters of BEVs. The study emphasizes that occupation, place of residence, and household size are considered socio-demographic purchase factors [45].

RQ7 focuses on factors influencing user behaviour in connection with BEVs and their impact on acceptance. Key factors such as symbolic attitudes, perceived barriers, mobility needs, personal norms and experiences with BEVs were identified [34]. Subjective norms and perceived behavioural control are also essential for users' intentions towards sustainable BEV consumption [46].

3 METHODOLOGY

A mixed-methods approach was chosen to conduct this study in order to collect both quantitative and qualitative data

and provide a more comprehensive perspective on the research questions [47, 48, 49]. For this purpose, a regression analysis was carried out in a panel as part of the quantitative research and qualitative expert interviews were conducted with specialists from the automotive industry. Article and the authenticity of information and statements written in the article.

3.1 Regression Analysis

The quantitative data on which this statistic is based was obtained from government-published analyses and surveys. These include publications by the Kraftfahrtbundesamt, the Bundesnetzagentur, the German Education Monitor Initiative Neue Soziale Marktwirtschaft (ISNM) and the statistical offices of the federal states. This ensures the sovereignty of the data. In order to provide a representative analysis with $n > 30$ [50] the methodology of a panel was applied. Consequently, the data collection for this study covered a period of five years (2018 to 2022) in each of the sixteen German Federal States ($n = 16*5 = 80$).

In this context, it was ensured that absolute figures were adjusted, making it possible to compare the data. The independent *Y* variable represents the starting point of the statistical analysis with the percentage of new BEV registrations compared to total new registrations. Respectively listed *X* variables are evaluated with *Y* by means of a correlation and regression analysis with regard to their relationship to each other. The variables of the underlying analysis are as follows:

Table 2 Variables

Variables	Importance of the variables
Y: New registrations of BEV passenger cars	Initial variable of the research topic.
X1: Real GDP	B have a higher price than conventional combustion engines. The study focuses on the potential purchasing power [51].
X2: Real charging infrastructure	Charging requires more time, which must be planned. The real charging infrastructure is crucial for evaluating concerns about the range of BEVs [14, 52].
X3: Average range of BEVs	The range of BEVs is steadily increasing. This variable influences the universal suitability of vehicles for customers [53].
X4: Education level	The level of education influences the attitudes and values of the population. A higher level of education can lead to greater awareness of environmental issues and innovative technologies [34].
X5: Real estate price index	Considering different infrastructural conditions and financial differences between cities and rural regions, a price gap can be identified, and possible conclusions can be drawn about new BEV registrations [45].
X6: Green participation in the state government	Green participation in the state government implies sustainability. It is questionable to what extent this is reflected in the purchase of new BEVs [43].

A correlation analysis carried out beforehand evaluates how strong the influence of one characteristic is on another characteristic [54]. The variables to be compared are treated equally. If two characteristics *X*, *Y* are at least interval-scaled, the correlation coefficient *r* is used to measure the strength of the linear relationship between the variables [54]. *r* is defined within the interpretation range of $-1 \leq r \leq +1$, where *r* must not be greater than 0.8 in order to avoid a linear correlation and thus an overlap in significance.

The final evaluation is carried out using a regression analysis created in Excel.

3.2 Interview

The second component involves conducting interviews with experts in the automotive industry in order to gain in-depth insights into the influence on the purchase of BEVs. Here, emphasis was placed on a broad knowledge and experience base among the selected experts, especially in customer contact. The interview partners (IPs) are presented below.

As managing director of a BEV leasing company, IP 1 has in-depth knowledge of a company that offers a wide range of leasing solutions for BEVs, thermal vehicles, hydrogen vehicles and bicycles. Under his leadership, the company focuses on the implementation of government programs for electric mobility in cooperation with federal ministries in order to facilitate customers' access to funding. The knowledge of the managing director is therefore of crucial importance for the investigation of the various factors influencing the approval of a BEV vehicle.

A large car dealership is used as an additional IP 2 for evaluating the factors influencing the purchase of BEVs. The car dealer-ship provides direct insights into customer preferences and purchasing decisions. As a sales point for various car models, including many different electric models, the car dealership generates valuable data on the various factors that influence why customers decide to buy an electric car. In particular, this information is based on practical aspects such as range and charging options. In addition, the dealership has experience with customer questions about BEVs and can identify trends and changes in demand over time, which is highly relevant for research into the boost of BEV purchases.

As an editor of an automotive magazine and in particular as a test driver, a high amount of vehicles passes through the desk of IP 3. The magazine is consumer-oriented and includes driving reports and vehicle evaluations. Trends must therefore be recognized at an early stage and customer wishes must be expressed. His expertise helps to make customer wishes even more explicit, especially concerning BEVs.

An expert from a vehicle testing organization will provide in-depth insights into several layers of the analysis as IP 4. On the one hand, through the certification of charging infrastructure in Germany and the subsequent current challenges and approaches to solutions, and on the other hand in the context of the frequency of new registrations, which must be approved by corresponding organizations. In addition, insights into possible concerns or progress in connection with electric vehicle batteries can be provided.
Before the interviews are carried out, a detailed guideline with relevant questions and criteria is prepared to ensure systematic coverage of the factors. The investigative process commences with questions of a general nature concerning the interviewee's professional background, including their role and relevance to BEV distribution. Subsequently, a specific

discussion of the variables mentioned above, evaluating their connection to the registration numbers. In addition, this research seeks to delineate strategic adjustments to the factors, which can increase the number of new BEV registrations. Where possible, the interviews will be conducted in person; if not virtually, for example utilizing platforms such as Microsoft Teams, recorded and transcribed to enable a precise evaluation afterwards.

After transcribing the interviews, recurring topics are identified as part of the analysis and, if possible, assigned to the *X* variables, otherwise additional categories are formed [55]. The key statements of the experts are summarized and categorized as described. This is followed by an analysis of the agreement between the experts' statements. In a subsequent section, the summarized statements are linked with the results of the regression. After combining both parts, an evaluation of the similarities and differences between the statements follows, thus answering the research objectives and questions with the existing database. It also examines whether and to what extent factors are justified or substantiated by the experts' statements.

4 RESULTS

The evaluation of the regression analysis (Tab. 3) shows that not all values are significant and can therefore be extended to the overall population. If the *P*-value exceeds the factor 0.1, the maximum significance level is exceeded. The coefficient indicates the strength and direction of the relationship between the independent *Y*-variable and dependent *X*-variables. The coefficient can be positive for a positive relationship or negative for a negative relationship. In this table, a positive or negative correlation can be recognized for each variable examined, which are defined below in their respective statements.

Table 3 Results regression analysis

	Coefficient	Standard error	St-statistics	*P*-Value	Upper 95%	Lower 95%
X1: Real GDP	–0,00149407	0,000902756	–1,6550165	0,10221458	–0,00329326	0,00030511
X2: Real charging infrastructure	–1,56301E–05	8,49914E–06	–1,8390206	0,06997898	–3,25689E–05	1,3087E–06
X3: Average range of BEVs	0,000720499	4,04615E–05	17,8070029	2,7942E–28	0,000639859	0,00080114
X4: Education level	–0,00115202	0,000442733	–2,6020781	0,0112139	–0,00203439	–0,0002697
X5: Real estate price index	6,81898E–06	4,34248E–06	1,57029788	0,12067048	–1,83556E–06	1,5474E–05
X6: Green participation in the state government	0,004164318	0,007032797	0,5921283	0,55559426	–0,00985202	0,01818066

The values obtained from the regression analysis reveal remarkable discrepancies between the experts' opinions and the empirical findings, especially in the context of new registrations of BEVs. It shows that GDP and the education rate do not have a significant influence on BEV registrations, which contradicts the results of some studies. The significant influence of the average range of BEVs and the property price index on registration figures should be emphasized, which clearly underlines the importance of technical developments and economic factors. Another particular observation is the divergence between the assessment of the charging infrastructure and the partial agreement regarding political participation, which reflects the complexity of the influence of socio-structural characteristics on the acceptance of BEVs.

Table 4 Consistency between the regression and expert interviews

Variables	Regression Analysis	Expert Interview	Consistency
X1: Real GDP	Increase in GDP has **no influence** on new BEV registrations (not significant)	Level of prosperity **influences** purchasing power	No
X2: Real charging infrastructure	Charging infrastructure has **no influence** on new BEV registrations (significant)	**Has an impact** on new BEV registrations	No
X3: Average range of electric vehicles	Range has an **influence** on new BEV registrations (significant)	Range has an **influence** on new BEV registrations	Yes
X4: Education level	Level of education has **no influence** on new BEV registrations (significant)	Level of education has **no influence** on new BEV registrations	Yes
X5: Real estate price index	Real estate prices have an **influence** on new BEV registrations (not significant)	Real estate prices have an **influence** on new BEV registrations	Yes
X6: Green participation in the state government	Tendency yes. **No clear statement** possible due to too high significance level	**No influence**. Possible connection due to high environmental awareness	Partially

4.1 Results Socio-economic Factors

Following the research questions (cf. Chapter 1.4) two factors can be identified based on the analysis carried out, which con-tribute essentially to the broadening of BEVs. These factors are:

- X3: Average range of BEVs
- X5: Real estate price index.

The results of the regression analysis and the overall opinion of experts matched for both variables examined. The high significance with regard to the range shows that the investigation can be clearly extended to the population as a whole. By confirming the high influence of the real estate price index through the expert survey, the corresponding variable can also be extended to the entirety. The disregard of variable 4 (education level) is due to the fact that this factor also produces a uniform result, but has no influence on new registrations.

In particular, the expansion of the range reduces the so-called range anxiety among buyers of BEVs, which leads to a significantly higher feeling of reliability. According to expert opinions, range is currently the decisive argument of BEV critics against buying an electric vehicle, which would

become progressively less important with the expansion. Despite the fact that charging times can be easily planned, there is a basic fear of uncertainty among the public regarding the range.

With regard to the real estate price index, less attention is paid to the difference between urban and rural areas, but even more to the type of housing, as measured by the real estate price index. The residential location plays a role with regard to the availability of a private charging space with a wall box, which is the case in a detached house compared to an apartment without a permanent parking space, which significantly increases the suitability of a BEV for everyday use. In addition, better living conditions imply increased purchasing power for BEVs that are even more expensive; however, there are also contrary opinions that the general level of prosperity has no influence on new registrations, which is discussed in more detail below.

One expert also refers to the reduction of the aforementioned range anxiety in cities, as the expansion of the charging infrastructure is usually better developed due to the higher population in cities and this subconsciously reduces range anxiety. This will be discussed in more detail below. Based on the analysis above, the German government must therefore significantly promote the technical expansion and development of batteries in order to increase new registrations. A state-subsidized combination of charging facilities in one's own home or in shared homes is also recommended in order to reduce the "exclusivity" of charging.

4.2 Results Sustainability and Climate Change

The scientific analysis of the factors influencing the registration of BEVs reveals a clear discrepancy between the results of the regression analysis and the expert opinions regarding green participation in the state government. This discrepancy manifests itself in particular in the inability of the regression analysis to make a definitive statement about the influence of green participation in the state government on BEV registrations due to an excessively high significance level. It underlines the need for a more careful consideration of the political dimensions and their potential influence on the acceptance and promotion of environmentally friendly technologies.

Experts, on the other hand, tend to assume that this political participation has no direct influence on BEV registrations. However, they suspect that there could be an indirect correlation due to increased environmental awareness in regions with stronger green participation in the state government. These differences of opinion and the only partial agreement in the results illustrate how complex and influenced by numerous factors the acceptance and the spread of BEVs is. They emphasize the need for an integrative view that includes both quantitative and qualitative perspectives in order to gain a more comprehensive understanding of the factors driving BEV adoption.

When asked about the influence of green participation in the state government on the purchase of a BEV, the experts' arguments suggest that this political variable has no direct influence. The experts argue that the decision to buy a BEV depends less on the state government and more on the long-term interest of the population who vote green. Another expert emphasizes that the purchase of BEVs is influenced more by decisions made by the federal government than by an individual party. These assessments illustrate that the promotion of BEVs is more strongly influenced by overarching national guidelines and the general environmental awareness of the population than by the specific involvement of a political party at state level.

For the X5 (Real estate price index) variable, the significance of property prices for new registrations of BEVs was examined. The regression analysis carried out indicates that property prices have an influence on new registrations of BEVs, but this influence is not statistically significant. This means that there is a correlation, but it is not strong enough to be considered decisive. In contrast, expert opinions confirm the high relevance of real estate prices for new BEV registrations, but experts offer contrary views. One expert argues that it is not the distinction between urban and rural locations that is decisive, but rather factors such as the type of housing, the size of the property and an affluent residential area. These factors could be more influential as they are directly linked to the financial capacity and lifestyle of residents. Other experts, however, emphasize the importance of urban residential locations, especially due to the denser charging infrastructure, which increases the acceptance and use of BEVs in urban areas. These insights underline the importance of socio-economic aspects in the purchase of BEVs. This consistency between the experts' opinions and the results of the regression analysis, despite the statistical non-significance, emphasizes the need to consider and further explore the influence of economic factors on BEV adoption.

The results of both the regression analysis and the expert opinions indicate a significant influence of the range of battery BEVs on their new registrations. The regression analysis identifies range as a significant factor, indicating its importance for user acceptance and the practical applicability of BEVs. Thus, this result confirms that an increased average range increases the acceptance of BEVs.

Experts confirm this finding and emphasize the importance of range for potential buyers of BEVs. The agreement in both analysis methods underlines the central role of vehicle range in the context of BEV adoption.

The expert statements regarding the range increase of BEVs emphasize the high relevance of this factor. Range anxiety is identified as a key factor that prevents potential buyers from purchasing a BEV. According to the experts, technological advances that increase the range can reduce this critical approach. Furthermore, range is a well-known argument used by BEV critics. Experts emphasize that the attractiveness of BEVs can be increased not only by technological improvements, but also by factors such as purchase price, vehicle selection and the expansion of the charging infrastructure. An interesting approach is the "store and charge" concept explained by one expert, which enables charging during other activities. These assessments make it clear that range plays a significant role in the market acceptance of BEVs compared to other factors. However, according to the experts, the factors mentioned, such as vehicle selection, must also be extended further, since with a

lower range, at least the charging infrastructure must be sufficiently available, according to another expert.

4.3 Results Change in Mobility

In terms of purchasing power, the GDP of Germany is used as the factor examined. According to the statistical survey, no influence of the level of prosperity on the number of new registrations of BEVs compared to conventional combustion vehicles can be identified. Due to the excessively high significance level, which is only just above 0.1, the statement is not significant, but can be considered generally valid in a general context. Accordingly, the analysis shows that purchasing power in the country has no effect. The negative coefficient supports this. This differs from the experts' statements. They cite the still very high price of BEVs, which in some cases is 50% higher than the price of combustion engines despite cheaper Chinese variants. A distinction must be made between private purchases and company cars. Expert opinions emphasize that the price of company cars, in contrast to a higher required level of prosperity in the private sector, plays barely any role. It is also emphasized that the current high inflation in Germany and cancelled subsidies have a negative impact on new registrations. According to the experts, the general level of prosperity is therefore relevant due to the available purchasing power.

As described, user behaviour is influenced to a greater or lesser extent by many factors. What is outstanding is the fact that, according to regression and experts, the level of education has no influence on the promotion of new registrations and can therefore be left out of further measures. Users still need to be introduced to BEVs. Many factors have a positive influence on achieving this. Systems and strategies, which are explained in more detail in Chapter 5, can sustainably and permanently increase behaviour and adoption. It can be said in advance that it will take a combination of several factors to achieve this development.

5 CONCLUSION

5.1 Summary

The analysis of factors influencing BEV registrations identifies two key factors: the range of the vehicles and the real estate price index. These significantly influence the acceptance and new registrations of BEVs. Extending the range reduces range anxiety and strengthens confidence in BEVs. It also shows that green participation in the state government does not have a direct influence on BEV registrations, but has an underlying indirect influence due to increased environmental awareness.

The real estate price index influences the suitability of BEVs for everyday use, as it indicates the availability of private charging stations and shows increased purchasing power for the purchase of BEVs. Germany's GDP has no direct influence on new BEV registrations, but the price of BEVs plays a decisive role, especially in the private sector.
Finally, the analysis of user behaviour shows that the education level has no influence on new BEV registrations. Instead, introducing potential users to BEVs through the development of various factors is crucial.

5.2 Discussion of the Results

As already mentioned, the results reveal significant discrepancies between the literature review and the empirical findings. What stands out is the fact that, according to the regression analysis and the expert interviews, the level of education has no influence on the promotion of new registrations, although educational measures were cited as relevant in the literature review. This shows the need to investigate the role of educational initiatives and the general level of education in the context of BEV adoption in more depth. Green participation in the state government is also a questionable factor. The connection here with more sustainable approaches is understandable, but the question arises as to whether more can be created from this context in terms of broad acceptance. The study shows that some factors have a greater or less influence on the acceptance of BEVs; a combination of technological improvements, an expansion of the charging infrastructure and political and economic measures is seen as the key to promoting BEV adoption.

5.3 Limitations and Practical Implication

There are also limitations to the research approach used. As the topic of electro mobility is growing rapidly, but is still quite new, some data sets are not yet mature enough to achieve workable results, as the comparative analysis of combustion engines, BEVs and the associated infrastructure is still too small. With regard to expert opinions, it must also be mentioned that such opinions are always subjective and experts do not necessarily answer purely objectively despite their extensive knowledge of the relevant sector.

With regard to practical implications, experts already provide relevant insights into measures that can be integrated into the German system for promoting BEVs. These generally include the promotion of technological progress with regard to battery development and the expansion of the charging infrastructure, but also area planning concepts for BEV suitability in everyday life, such as "store and charge" concepts at charging stations. New registrations are also expected to rise as the availability and variety of BEVs increases.

6 REFERENCES

[1] Umweltbundesamt (Hrsg.) (2020). Klimaschutz durch Tempolimit. Wirkung eines generellen Tempolimits auf Bundesautobahnen auf die Treibhausgasemissionen. https://www.umweltbundesamt.de/sites/default/files/medien/1410/publikationen/2020-06-15_texte_38-2020_wirkung-tempolimit_bf.pdf (in German)

[2] Bergk, F., Knörr, W. & Lambrecht, U. (2017). Klimaschutz im Verkehr: Neuer Handlungsbedarf nach dem Pariser Klimaschutzabkommen. https://www.umweltbundesamt.de/sites/default/files/medien/1410/publikationen/2017-07-18_texte_45-2017_paris-papier-verkehr_v2.pdf (Accessed on 23.11.2023) (in German)

[3] Zawacki-Richter, O., Kerres, M., Bedenlier, S., Buntins, K. & Bond, M. (Eds.) (2019). *Systematic Reviews in Educational Research: Methodology and Applications*. Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-27602-7

[4] Bringezu, S., Distelkamp, M., Lutz, C., Wimmer, F., Schaldach, R., Hennenberg, K., Böttcher, H. & Egenolf, V.

(2021). Environmental and socioeconomic footprints of the German bioeconomy. *Nature Sustainability, 4*, 775-783. https://doi.org/10.1038/s41893-021-00725-3

[5] Heimann, D. (2018). Unternehmensnetzwerke für nachhaltige Gewerbegebiete. *Standort, 42*, 223-228. (in German) https://doi.org/10.1007/s00548-018-0557-6

[6] Buchmann, M., Kusznir, J. & Brunekreeft, G. (2019). Assessment of the drafted German integrated National Energy and Climate Plan. *Economics and Policy of Energy and the Environment, 1*, 85-96. https://doi.org/10.3280/efe2019-001006

[7] Haas, T. (2020). From Green Energy to the Green Car State? The Political Economy of Ecological Modernisation in German. *New Political Economy, 26*(4), 660-673. https://doi.org/10.1080/13563467.2020.1816949

[8] Klinger, K., Metag, J. & Schäfer, M. (2022). Global Warming's Five Germanys – Revisited and Framed in an International Context. *Environmental Communication, 16*(8), 1108-1126. https://doi.org/10.1080/17524032.2022.2153897

[9] Zhao, Q. (2018). Electromobility research in Germany and China: structural differences. *Scientometrics, 117*, 473-493. https://doi.org/10.1007/S11192-018-2873-9

[10] Hecht, C., Figgener, J. & Sauer, D. (2022). Analysis of electric vehicle charging station usage and profitability in Germany based on empirical data. *iScience, 25*(12), 105634. https://doi.org/10.1016/j.isci.2022.105634

[11] Weigelt, M., Mayr, A., Böhm, R., Kühl, A. & Franke, J. (2018). Quo vehis, Elektromobilität?: Aktuelle Treiber und Hindernisse der Mobilitätswende in Deutschland. *Zeitschrift für wirtschaftlichen Fabrikbetrieb, 113*(1-2), 59-63. (in German) https://doi.org/10.3139/104.111863

[12] Ahmad, A., Khan, Z., Alam, M. & Khateeb, S. (2018). A Review of the Electric Vehicle Charging Techniques, Standards, Progression and Evolution of EV Technologies in Germany. *Smart Science, 6*(1), 36-53. https://doi.org/10.1080/23080477.2017.1420132

[13] Rietmann, N. & Lieven, T. (2019). How policy measures succeeded to promote electric mobility – Worldwide review and out-look. *Journal of Cleaner Production, 206*, 66-75. https://doi.org/10.1016/j.jclepro.2018.09.121

[14] Neubauer, J. & Wood, E. (2014). The impact of range anxiety and home, workplace, and public charging infrastructure on simulated battery electric vehicle lifetime utility. *Journal of Power Sources, 257*, 12-20. https://doi.org/10.1016/j.jpowsour.2014.01.075

[15] Weiss, M., Zerfass, A. & Helmers, E. (2019). Fully electric and plug-in hybrid cars - An analysis of learning rates, user costs, and costs for mitigating CO_2 and air pollutant emissions. *Journal of Cleaner Production, 212*, 1478-1489. https://doi.org/10.1016/j.jclepro.2018.12.019

[16] Ulrich, P. & Lehr, U. (2020). Economic effects of an E-mobility scenario – input structure and energy consumption. *Economic Systems Research, 32*(1), 84-97. https://doi.org/10.1080/09535314.2019.1619522

[17] Tena, D. L. de & Pregger, T. (2018). Impact of electric vehicles on a future renewable energy-based power system in Europe with a focus on Germany. *International Journal of Energy Research, 42*(8), 2670-2685. https://doi.org/10.1002/er.4056

[18] Burkert, A., Fechtner, H. & Schmuelling, B. (2021). Interdisciplinary Analysis of Social Acceptance Regarding Electric Vehicles with a Focus on Charging Infrastructure and Driving Range in Germany. *World Electric Vehicle Journal, 12*(1), *e*010025. https://doi.org/10.3390/wevj12010025

[19] Anderson, J., Lehne, M. & Hardinghaus, M. (2018). What electric vehicle users want: Real-world preferences for public charging infrastructure. *International Journal of Sustainable Transportation, 12*(5), 341-352, https://doi.org/10.1080/15568318.2017.1372538

[20] Vilchez, J., Smyth, A., Kelleher, L., Lu, H., Rohr, C., Harrison, G. & Thiel, C. (2019). Electric Car Purchase Price as a Factor Determining Consumers' Choice and their Views on Incentives in Europe. *Sustainability, 12*(22), *e*226357. https://doi.org/10.3390/su11226357

[21] Fournier, G., Baumann, M., Gasde, J. & Kilian-Yasin, K. (2018). Innovative mobility in rural areas - the case of the Black Forest. *International Journal of Automotive Technology and Management*, 18(3), 247-269. https://doi.org/10.1504/IJATM.2018.10013851

[22] Halbey, J., Kowalewski, S. & Ziefle, M. (2015). Going on a Road-Trip with My Electric Car: Acceptance Criteria for Long-Distance-Use of Electric Vehicles, Marcus, A. (Ed.) *Design, User Experience, and Usability: Interactive Experience Design*, DUXU 2015. Lecture Notes in Computer Science, 9188, 473-484, Springer, Cham. https://doi.org/10.1007/978-3-319-20889-3_44

[23] Husarek, D., Salapic, V., Paulus, S., Metzger, M. & Niessen, S. (2021). Modeling the Impact of Electric Vehicle Charging Infrastructure on Regional Energy Systems: Fields of Action for an Improved e-Mobility Integration, *Energies, 14*(23), *e*237992. https://doi.org/10.3390/en14237992

[24] Elangovan, P., Lust, D., Gökdemir, E., Silberer, J., Pietruschka, D. & Mrso, M. (2021). Smart2charge: smart grid enabled BEV charging infrastructure for rural areas. *The 5th E-Mobility Power System Integration Symposium*, 169-174. https://doi.org/10.1049/icp.2021.2520

[25] Letmathe, P. & Suares, M. (2020). Understanding the impact that potential driving bans on conventional vehicles and the total cost of ownership have on electric vehicle choice in Germany. *Sustainable Futures, 2*, *e*100018, https://doi.org/10.1016/j.sftr.2020.100018

[26] Zeng, Y., Schmitz, H. & Madlener, R. (2018). An Econometric Analysis of the Determinants of Passenger Vehicle Sales in Germany. *FCN Working Papers, 6*. https://doi.org/10.2139/ssrn.3239373

[27] Kölbl, R., Bauer, D. & Rudloff, C. (2013). Travel behavior and Electric Mobility in Germany. *Transportation Research Record, 2385*(1), 45-52. https://doi.org/10.3141/2385-06

[28] Ajanovic, A. & Glatt, A. (2020). Wirtschaftliche und ökologische Aspekte der Elektromobilität. *Elektrotechnik und Informationstechnik, 137*, 136-146. https://doi.org/10.1007/s00502-020-00812-x

[29] Pons-Seres de Brauwer, C. (2022). The Politics of Market Change towards Sustainability: Revisiting Germany's Policy Support Framework for Renewables. *Energies, 15*(11), *e*113898. https://doi.org/10.3390/en15113898

[30] Berndorfer, J. M. & Neudorfer, H. (2023). Simulation und Wirtschaftlichkeitsbewertung von Photovoltaikanlagen in Kombination mit Elektrofahrzeugen. *Elektrotechnik und Informationstechnik, 140*, 407-414. https://doi.org/10.1007/s00502-023-01140-6

[31] Krause, J., Small, M., Haas, A., & Jaeger, C. (2016). An expert-based bayesian assessment of 2030 German new vehicle CO_2 emissions and related costs. *Transport Policy, 52*, 197-208. https://doi.org/10.1016/j.tranpol.2016.08.005

[32] Mock, P., Hülsebusch, D., Ungethüm, J. & Schmid, S. (2009). Electric vehicles - A model based assessment of future market prospects and environmental impacts. *World Electric Vehicle Journal, 3*(1), 172-185. https://doi.org/10.3390/wevj3010172

[33] Trommer, S., Jarass, J. & Kolarova, V. (2015). Early adopters of electric vehicles in Germany unveiled. *World Electric Vehicle Journal, 7*(4), 722-732. https://doi.org/10.3390/wevj7040722

[34] Haustein, S. & Jensen, A. (2018). Factors of electric vehicle adoption: A comparison of conventional and electric car users based on an extended theory of planned behaviour.

International Journal of Sustainable Transportation, 12(7), 484-496. https://doi.org/10.1080/15568318.2017.1398790
[35] Roemer, E. & Henseler, J. (2019). The dynamics of electric vehicle acceptance in corporate fleets: Evidence from Germany. *Technology in Society, 68*, *e*101938. https://doi.org/10.1016/j.techsoc.2022.101938
[36] Stockkamp, C., Schäfer, J., Millemann, J. A. & Heidenreich, S. (2021). Identifying Factors Associated with Consumers' Adoption of e-Mobility-A Systematic Literature Review. *Sustainability, 13*(19). *e*10975. https://doi.org/10.3390/su131910975
[37] Hildermeier, J. & Jahn, A. (2021). Regulierungsansätze zwischen Markt und Staat bei der Netzintegration von Elektromobilität in Europa. *WSI Mitteilungen*, *74* (3), 226-233. https://doi.org/10.5771/034230020213226
[38] Nazari, F., Rahimi, E. & Mohammadian, A. (2019). Simultaneous estimation of battery electric vehicle adoption with endogenous willingness to pay. *eTransportation, 6*, *e*100088. https://doi.org/10.1016/j.etran.2019.100008
[39] Wang, X., Cao, Y. & Zhang, N. (2021). The influences of incentive policy perceptions and consumer social attributes on battery electric vehicle purchase intentions. *Energy Policy, 151*, *e*112163. https://doi.org/10.1016/j.enpol.2021.112163
[40] Silvia, C. & Krause, R. (2016). Assessing the impact of policy interventions on the adoption of plug-in electric vehicles: An agent-based model. *Energy Policy, 96*, 105-118. https://doi.org/10.1016/j.enpol.2016.05.039
[41] Breetz, H. & Salon, D. (2018). Do electric vehicles need subsidies? Ownership costs for conventional, hybrid, and electric vehicles in 14 U.S. cities. *Energy Policy, 2*, 238-249. https://doi.org/10.1016/j.enpol.2018.05.038
[42] Hagman, J., Ritzén, S., Stier, J. & Susilo, Y. (2016). Total cost of ownership and its potential implications for battery electric vehicle diffusion. Re*search in transportation business and management, 18*, 11-17. https://doi.org/10.1016/j.rtbm.2016.01.003
[43] Schneidereit, T., Franke, T., Günther, M. & Krems, J. (2015). Does range matter? Exploring perceptions of electric vehicles with and without a range extender among potential early adopters in Germany. *Energy research and social science, 8*, 198-206. https://doi.org/10.1016/j.erss.2015.06.001
[44] Schmalfuß, F., Mühl, K. & Krems, J., 2017. Direct experience with battery electric vehicles (BEVs) matters when evaluating vehicle attributes, attitude and purchase intention. *Transportation Research Part F: Traffic Psychology and Behaviour, 46*, 47-69. https://doi.org/10.1016/j.trf.2017.01.004
[45] Plötz, P., Schneider, U., Globisch, J. & Dütschke, E. (2014). Who will buy electric vehicles? Identifying early adopters in Germany. *Transportation Research Part - Policy and Practice, 67*, 96-109. https://doi.org/10.1016/j.tra.2014.06.006
[46] Dutta, B. & Hwang, H. (2021). Consumers Purchase Intentions of Green Electric Vehicles: The Influence of Consumers Technological and Environmental Considerations. *Sustainability, 13*(21), *e*12025. https://doi.org/10.3390/su132112025
[47] Fielding, N. G. (2012). Triangulation and Mixed Methods Designs: Data Integration with New Research Technologies. *Journal of Mixed Methods Research*, 6(2), 124-136. https://doi.org/10.1177/1558689812437101
[48] Timans, R., Wouters, P. & Heilbron, J. (2019). Mixed methods research: what it is and what it could be. *Theory and Society*, 48, 193-216. https://doi.org/10.1007/s11186-019-09345-5
[49] Wibisono, E. (2022). The Expansion of Qualitative Research Methods in Innovation Policy Studies, *STI Policy and Management Journal, 7*(1), 63-75. https://doi.org/10.14203/STIPM.2022.322
[50] Koh, K. & Ahad, N., 2020. Normality for Non-Normal Distributions. *Journal of Science and Mathemathics Letters, 8*(2), 51-60. https://doi.org/10.37134/jsml.vol8.2.7.2020
[51] Koengkan, M., Fuinhas, J. A., Belucio, M., Alavijeh, N. K., Salehnia, N., Machado, D., Silva, V. & Dehdar, F. (2022). The Impact of Battery-Electric Vehicles on Energy Consumption: A Macroeconomic Evidence from 29 European Countries. *World Electric Vehicle Journal, 13*(2), 36. https://doi.org/10.3390/wevj13020036
[52] Neaimeh, M., Salisbury, S., Hill, G., Blythe, P., Scoffield, D., & Francfort, J. (2017). Analysing the usage and evidencing the importance of fast chargers for the adoption of battery electric vehicles. *Energy Policy*, 108, 474-486, https://doi.org/10.1016/j.enpol.2017.06.033
[53] Barter, G., Tamor, M., Manley, D. & West, T. (2015). Implications of Modeling Range and Infrastructure Barriers to Adoption of Battery Electric Vehicles. *Transportation Research Record, 2502*(1), 80-88. https://doi.org/10.3141/2502-10
[54] Bourier, G. (2022). *Beschreibende Statistik*. Springer Gabler, Wiesbaden. https://doi.org/10.1007/978-3-658-05916-3
[55] Clarke, S., Sushil, S., Dennis, K., Lee, U., Gomoll, A. & Gates, Z. (2023). Developing Shared Ways of Seeing Data: The Perils and Possibilities of Achieving Intercoder Agreement. *International Journal of Qualitative Methods, 22*, *e*160973, https://doi.org/10.1177/16094069231160973

Authors' contacts:

Hoffmann Johannes, B.Sc.
Heilbronn University of Applied Sciences,
Faculty of Economics,
Bildungscampus, 74076 Heilbronn, Germany
johannes.hoffmann@hsgbr.com

Verena Szkudlarek, B.Sc.
Heilbronn University of Applied Sciences,
Faculty of Economics,
Bildungscampus, 74076 Heilbronn, Germany
vreni.szkudlarek@gmail.com

Daniela Ludin, Prof. Dr.
Heilbronn University of Applied Sciences,
Faculty of Economics,
Bildungscampus, 74076 Heilbronn, Germany
daniela.ludin@hs-heilbronn.de

Norbert Schreier, Prof. Dr.
(Corresponding author)
Esslingen University of Applied Sciences,
Faculty of Mobility and Technology,
Kanalstrasse 33, 73728 Esslingen, Germany
norbert.schreier@hs-esslingen.de

Erika Mueller, M.Sc.
Heilbronn University of Applied Sciences,
Sustainability Department,
Bildungscampus, 74076 Heilbronn, Germany
erika.mueller@hs-heilbronn.de

Wanja Wellbrock, Prof. Dr.
Heilbronn University of Applied Sciences,
Faculty of Economics,
Bildungscampus, 74076 Heilbronn, Germany
wanja.wellbrock@hs-heilbronn.de

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240206125322

Original scientific paper
Received: 2024-02-06, Accepted: 2024-07-08

Palm Print Recognition using Deep Learning

Ruaa Sadoon Salman, Mauj Haider AbdAlkreem*, Qaswaa Khaled Abood

Abstract: In recent decades, numerous studies have focused extensively on biometric palmprint recognition. Palm print recognition has gained significant popularity and importance across various domains owing to its exceptional efficiency and accuracy in personal identification. The biometric characterization of a person's palm print is unique. However, a way to enhance the image is needed in order to produce a better and clearer image. Recently, palm print recognition methods based on features acquired using a series of convolutional neural networks have been introduced, among which DenseNet-121 has a densely connected structure, unlike other structures. This paper presents a scheme for palm print recognition by image enhancement. Contrast-limited adaptive histogram equation (CLAHE) is one of the image enhancement methods that can provide bounded segment and region size and is based on deep learning using DenseNet-121. To measure performance, the CASIA dataset was used. Experimental results on the DS show that the palm print features of Denes 21 achieve a recognition accuracy of 99 %, demonstrating the effectiveness and reliability of the proposed palm print.

Keywords: biometric; CLAHE; deep learning; DenseNet-121; ROI

1 INTRODUCTION

Biometrics is a system that uses distinct behavioral patterns to automatically verify the identity of individuals. Physiological traits are genetic characteristics acquired during the embryonic stages of human development. The emergence of print-based biometric technology provides a promising alternative to traditional identification methods such as fingerprints, iris scanning, and facial recognition. Print recognition method datasets can be obtained easily and inexpensively due to their global and dynamic nature [1]. Biometric systems that use the palm of the hand have become of interest to researchers because they are systems with individual features that cannot be replicated among humans [2]. The importance of personal identification technology has increased in recent years, especially in authentication methods, which has led to increased research in this specialty [3, 4]. Traditional security technologies, such as passwords and magnetic cards, are no longer considered secure enough due to their susceptibility to being stolen or forgotten by the owner. Hence, in order to achieve a high level of security in interactions, biometric technologies have been developed for application in different system categories such as smart device logins, home security systems, and other control systems [5]. A palm print is a distinct biological pattern that is unique to each individual, as no two prints are identical, even among identical twins. The global use of this pattern for identification purposes is due to the fact that palm prints are not affected by external factors during the capture of the fingerprint and are similar to vein fingerprints in terms of the accuracy of the results, but the only factor that affects is the cutting of the hand or the burning of the hand and the disappearance of all lines [6, 7]. Palm or hand recognition is a key element of the biometric approach, which is widely recognized as one of the most efficient and successful means of identifying individuals. Basically, it refers to physical characteristics or behavioural traits that can be used to identify an individual [8, 9]. Palmprint-based personal authentication systems have been developed to exploit the distinctive lines, wrinkles, and features in the palm. These characteristics create a distinct and stable pattern that remains unchanged throughout a person's lifespan [10, 11].These systems use a camera or scanner along with its software to examine the acquired images and compare them to existing data sets in the system. It is worth noting that palm prints are equivalent to fingerprints [12]. The palm recognition system will use several techniques, including thermal methods, optical methods, and other methods, to enhance palm edges and other distinctive features [13, 14].The goal of this research paper is to confirm the identities or recognition of individuals by analysing their Palmprint by relying on deep learning through a pre-trained model and preliminary image processing operations.

1.1 Palm Print Recognition

Print recognition employs an infrared light source to detect the presence of hemoglobin in the blood. Deoxygenated hemoglobin manifests as a dark pattern when observed with the hand or finger. The equipment thereafter records an image of the distinct patterns formed by the ridges and lines on the wrist, palm, back of the hand, finger, or face. This is analogous to the methodology employed for capturing retinal patterns [15]. According to the observation, the vascular patterns on the backs of hands and palms are more intricate compared to those on fingers. This makes them more suitable for recognition matching and authentication purposes. Similar to other methods of biometric identification, print recognition is seen as being unaffected by the passage of time and distinct enough to accurately identify an individual [16, 17]. It has been discovered that the physiological biometric can be easily counterfeited by medical procedures, leading to the loss of the person's medical identity. Therefore, it is possible to analyze and implement a powerful and resilient biometric characteristic like palm print for the purpose of identifying individuals. The presence of vitality can be determined by observing the variations in blood flow inside the print while the heart beats [18]. Fig. 1 displays the distinct regions and their corresponding labeling on the palm [6].

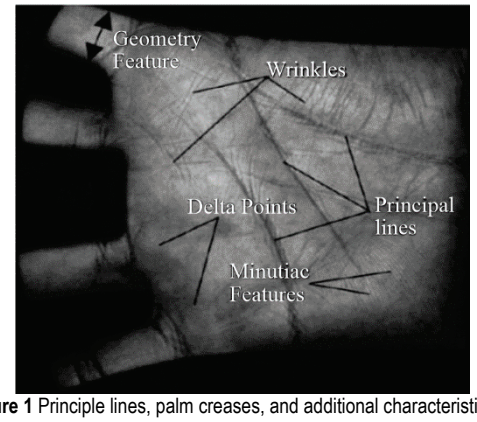


Figure 1 Principle lines, palm creases, and additional characteristics [6]

Palmprint recognition offers numerous benefits when implemented on consumer devices:

- Palmprints have comparable characteristics to fingerprints, however encompassing a far greater area. As a result of this reasoning, they are often considered to be more durable than fingerprints [19].
- Palmprints exhibit more resistance to spoofing in comparison to faces, which are readily accessible, or fingerprints, which can be inadvertently deposited on diverse flat surfaces.
- There are no additional costs required to acquire the gadget, as long as it comes with a camera (optical sensor) and a flash output (LED or screen) [20].
- The system is capable of performing multi-biometric recognition by integrating with other hand-based features, such as fingerprints [21], finger knuckles [22], wrist [23].
- It may be easily incorporated into the functionality of various consumer gadgets, such as AR/VR headsets [20], smartphones [24], gesture control systems, driver monitoring systems, etc.

Table 1 Summarized of previous studies

Year & Ref.	Data set	Employed method	Aim of paper	Limitation	Result of System
2021 [30]	CASIA	Gabor filters block-wise histograms triple-type feature descriptor	Extracting three types of palmprint features without the need for any raining samples.	Investigate other handcrafted traits to enhance the accuracy of fingerprint recognition.	Accuracy = 88.55 %
2022 [31]	CASIA	Convolutional Neural Network VGG16	VGG16 was chosen as the central network because it replaces large kernel filters (11 and 5 in the first and second convolutional layers, respectively) with many 3×3 filters. The images are then fed into layers of convolution and max pooling until features are extracted and then classified using the SoftMax function.	Parameters. Due of its profoundness and the abundance of fully integrated layers. The model has a substantial size of 500 MB.	Accuracy = 97.32 % EER = 0.0268
2022 [32]	CASIA	Return of Investment (ROI) - Mean Robust Extended Local Binary Pattern (MRELBP) *k* - nearest neighbor classifier	A method for identity verification and image normalization is presented. Features are extracted in the ROI extraction step from the input photomicrograph and then reduced through dimensionality reduction and classified using a nearest neighbor classifier	This research faces some drawbacks and challenges, such as computational complexity of classification, slow speed, memory and storage issues for large data sets, and sensitivity to the choice of k and distance measure.	Accuracy = 96.6 %
2023 [33]	CASIA	Convolution Neural Network Siamese Neural Net (SNN)	An approach to handprint identification that utilizes a Siamese network. The suggested methodology involves utilizing two convolutional neural networks with shared weights to extract characteristics from handprint images. These extracted features are subsequently compared using a contrast loss function to ascertain whether the two photos originate from the same individual. Consequently performance of the Siamese can also be explored , the extracted features become more distinct and conspicuous	Feature space. The original person's template can be securely preserved in a third-party authentication system, whether trusted or untrusted, to prevent theft It requires longer training time than regular networks, because Siamese networks include quadratic pairs to learn from. They are slower than other types. In addition, it does not extract all probabilities because the training is binary learning, so it will not produce prediction probabilities. The possibility of combining it with other deep learning techniques to improve the	Accuracy = 95.6% ERR = 0.044
2023 [34]	CASIA	Hetero-Associative Memory Encoder (HAMTE) neural network	The suggested network ensures the protection of the individual's Palmprint template by converting it into an irreversible template in a distinct	This type of memory network is commonly used in applications such as data compression and data retrieval to produce higher results and better identify people.	Accuracy = 90.2 % ERR = 0.02

2 LITERATURE REVIEW

Several studies in the literature have employed artificial intelligence (AI) to identify individuals through the utilization of Palm Print methods. The majority of palm print identification research in existing literature has employed either one or two variations of printed biometric photographs of the palm, back, or wrist. These images are often taken using infrared or near-infrared cameras [25, 26]. For instance, the utilization of innate patterns seen in palm prints has been suggested. The camera and near-infrared illumination system of a charge-coupled device (CCD) were used to take images. Subsequently, diode and thinning techniques were employed to extract the feature, resulting in the identification of connected palm print lines and details [27, 28]. A variety of feature extraction techniques have been

employed, spanning from manual approaches to deep learning (DL) [29].

This section aims to emphasize recent advancements in biometric recognition systems utilizing multispectral imaging technology. Specifically, it will focus on the evaluation of works conducted using the CASIA database. Tab. 1 summarized the related work that mention above.

3 PROPOSED SYSTEM

The proposed system for recognition palm print is based on the pre-training model of deep learning and preprocessing images using the global stander dataset. Fig. 2 illustrates the sequential stages of the implemented system.

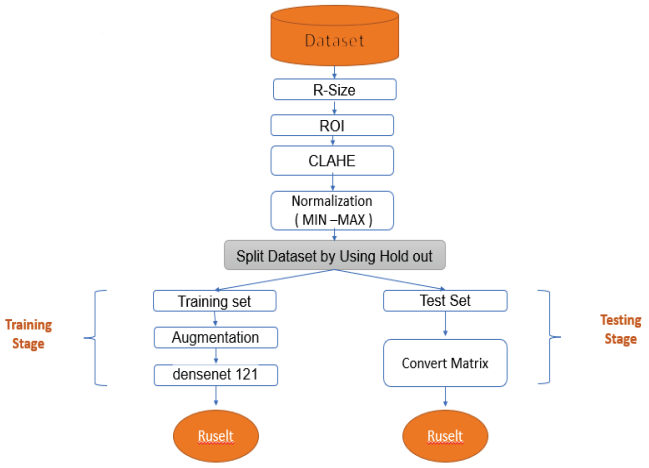
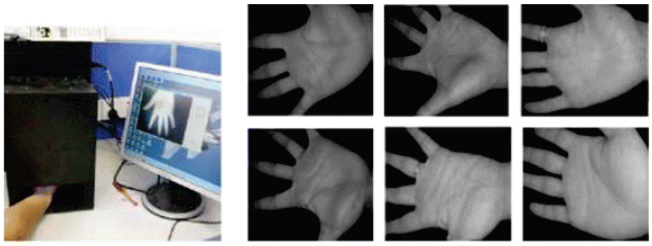


Figure 2 Proposed system

3.1 Dataset

The process of recognition is a complex and dynamic phenomenon that evolves alongside the ever-changing landscape of the electronic and digital realm. Despite facing obstacles related to cost, time, and precision, the organization effectively established dominance in the required market and devised optimal devices to facilitate efficient recognition. The Multispectral Palmprint Image Database, upheld by the Chinese Academy of Sciences (CASIA), is the subject of inquiry. CASIA is a comprehensive database dedicated to multispectral recognition, which includes the acquisition of palm vein images as well as the advancement of numerous biometric modalities [35]. This database encompasses a large collection of palm vein images, employing optical and spectral devices to capture the images with utmost precision. It consists of 7,200 high-resolution jpg images of palms, both right and left, obtained from a sample of 100 individuals. Each image has dimensions of 768×576 and is captured using custom-designed multiple spectral imaging devices as detailed in the description. The photos of each hand in this collection are gathered during two distinct sessions. The duration between two sessions exceeds one month. Three samples are presented during each session. Six palm photographs were captured concurrently using six distinct electromagnetic spectrums for each sample. In addition to white light, the illumination emanates light at wavelengths of 460, 630, 700, 850, and 940 nanometers. Fig. 3 shows that device that used in dataset an example of palm [36].

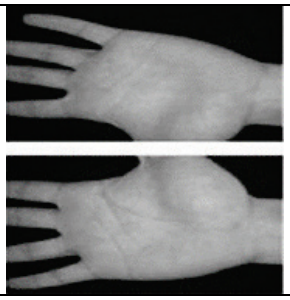


(a) (b)

Figure 3 (a) Palmprint Image Scan of "CASIA-Palmprint" dataset in the capture system (b) Some examples of the palmprints [36].

However, the objective of this variety in duration, postures, and light frequency is to attain a state of diversity. The CCD camera is configured to autonomously capture six images of each hand placed in front of a uniformly colored backdrop, without placing any restrictions on the user of the palm vein identification system [37, 36]. Tab. 2 shows the CASIA dataset [38].

Table 2 CASIA multispectral palm print image dataset

Palm number	7200
Sample number	6
Wavelength or light	940 nm, 850 nm,700 nm, 630 nm, 460 nm and white
Camera type	CCD camera
Image size (pixel)	768×576
Example	

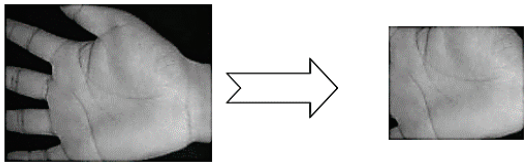
3.2 Pre-processing

Image resizing refers to changing an image's dimensions, making it larger or smaller without cutting off any part of it. In the proposed system, the work was based on changing the size to 256 × 256 in proportion to the model used. That is, the measurement for this process was done according to experience and which size was the most efficient due to training. The following steps were ROI, which is the context of the palmprint image and stands for the region of interest. The term refers to the precise image region with the most critical palm print identification or examination data. This region typically encompasses the salient characteristics of the palm, such as:

- Palm lines refer to the primary lines on the palm, including the life, heart, and fate lines.
- Palm edges refer to elaborate designs created by elevated and indented skin regions.

- Palm landmarks refer to distinct points of reference on the palm, including the finger base, the palm's midpoint, and the wrist.

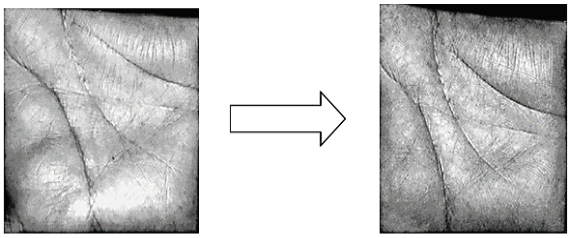
The following Fig. 4 shows the steps of resizing and ROI.



(A) Original image +Resizing (256*256) (B) ROI

Figure 4 Resizing and ROI

The proposed system implements the CLAHE (Contrast Limited Adaptive Histogram Equalization) as the primary method for adjusting image luminance and enhancing contrast. It effectively addresses issues related to variations in illumination and prevents excessive optimization, which can occur with the conventional histogram equation [39]. The proposed system initially utilized this technique to enhance low-contrast medical photographs. In contrast to processing the entire image, the CLAHE algorithm operates on squares, which are localized regions of the image. Contrast enhancement utilizes contrast-limited adaptive histogram equalization (CLAHE), a particular application of the histogram equalization method [40] that exhibits adaptive behavior in enhancing the image. The rank of the pixel's intensity in the local intensity histogram is directly proportional to the value by which the pixel's intensity is converted to a value that falls within the display range. CLAHE is an enhanced iteration of Adaptive Histogram Equalization (AHE) that modifies the enhancement procedure through the application of a clip level, which is a user-defined maximum. This clip level restricts the height of the local histogram and, as a result, the maximum contrast enhancement factor. The enhancement is consequently diminished in very homogeneous regions of the image, so preventing excessive amplification of noise and reducing the undesired edge-shadowing effect caused by unrestricted AHE. The CLAHE technique was first devised for medical imaging to mitigate the noise and edge shadowing phenomenon that occurs in uniform regions. The proposed approach applies the Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm to squares, which are localized regions of the image, rather than the entire image. In this approach, the histogram of each region is calculated as an initial phase. Next, the clipping threshold value is determined based on the required width of the contrast window. In the subsequent stage, every histogram value is reallocated while ensuring it does not beyond the pre-established threshold value. The grayscale mapping process involves determining the Cumulative Distribution Function (CDF) of the histograms in the last stage. The CLAHE approach employs pixel mapping using their immediate four neighbors. The lower sections are combined using bi-linear interpolation. The regions are categorized into three types, namely IR (interior Region), CR (corner region), and BR (border region), based on their adjacent conditions. Fig. 5 illustrates the sequential resizing process performed prior to applying the ROI and subsequent inclusion of the CLAHE contrast enhancer.



(A) Resizing (256*256)+ROI (B) CLAHE

Figure 5 Resizing and ROI+CLAHE

The final step in the proposed system is normalization by using (min-max) in rate (0-1) to measure the standardization of the value of the image. This method scales the un-normalized data to predefined lower and upper bounds. The equation is given as follows:

$$v = \frac{v - \min(A)}{\max(A) - \min(A)} \cdot \left(\text{newmax}(A) - \text{newmin}(A) + \text{newmin}(A)\right). \quad (1)$$

The attribute data, represented as A, is defined by its lowest value ($\min(A)$) and maximum value ($\max(A)$). The variable v denotes the revised value for each individual element in the dataset. v denotes the preceding value of every element in the dataset. $\text{newmax}(A)$ refers to the maximum value within the range, while $\text{newmin}(A)$ represents the minimum value within the range (i.e., the boundary values necessary).

3.3 Split Dataset

Hold out technical methods used in the proposed system for split dataset to three sets (training, testing and validation) in a proposed system using (80 % training and 20 % testing), In the Tab. 3 show the division of proportions for training, testing and validation from data set in the proposed system, Fig. 6 illustrates the percentage of each part.

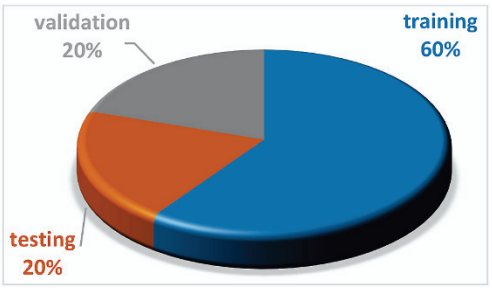


Figure 6 Splitting dataset

Table 3 The percentage of splitting dataset in the proposed system

Training	60 % = 4320
Testing	20 % = 1440
Validation	20 % = 1440

Table 4 Summarizes the layer of dense net 121 in the proposed system

Layer Type	Number of Layers	Function
Conv2D	1	Perform a 7×7 convolution operation with 64 filters, using a stride of 2. Then, apply batch normalization and ReLU activation.
Max Pooling	1	3×3 max pooling with stride 2
Dense Block 1	6	Each layer consists of: * 1×1 convolution with four growth rate filters * 3×3 convolution with growth rate filters * Batch normalization and ReLU activation * Concatenation with all preceding layers in the block.
Transition	1	1×1 convolution with 32 filters, followed by 2×2 average pooling
Dense Block 2	12	Similar to Dense Block 1, but with 32 filters in the 1×1 convolution
Transition	1	1×1 convolution with 64 filters, followed by 2×2 average pooling
Dense Block 3	24	Similar to Dense Block 2, but with 64 filters in the 1×1 convolution
Transition	1	1×1 convolution with 128 filters, followed by 2×2 average pooling
Dense Block 4	16	Similar to Dense Block 3, but with 128 filters in the 1×1 convolution
Global Average Pooling	1	Averages the output of the last Dense block over all spatial dimensions
Dense	1	Fully-connected layer with 1000 neurons (for ImageNet classification)
Softmax	1	Outputs the probability of each class

3.4 Training in the Proposed System

DenseNet-121 is a convolutional neural network architecture created in 2017 by Gao Huang and his colleagues. The design is characterized by its dense interconnections between layers, which enhance feature reuse and accuracy compared to other architectures such as VGG and ResNet. Here is a comprehensive analysis of its primary characteristics:

- **Dense Connections:** Unlike conventional CNNs, where each layer only gets input from the previous layer, DenseNet-121 establishes connections between each layer and all prior layers. This facilitates the integration of previously acquired properties into the network, hence improving the transmission of information and the dissemination of distinctive characteristics.
- **DenseNet-121** incorporates bottleneck layers composed of 1×1 and 3×3 convolutional layers. These layers decrease computational complexity and ensure efficient memory usage, all while retaining the ability to extract features effectively.
- **Growth Rate:** The overall output of a Dense Net block is increased by a specific number of feature maps for each layer. The growth rate of the model determines the balance between its complexity and accuracy. The growth rate 12 is utilized in DenseNet-121, which is why it is named as such.

DenseNet-121 incorporates transition layers between Dense blocks to regulate the feature map's size and avoid excessive memory consumption. These layers utilize 1×1 convolutional filters to decrease the number of feature maps and conduct down sampling, resulting in a more condensed and efficient network. The following Tab. 4 shows the layer used in the training stage.

The following Fig. 7 illustrates the result of training in the proposed system.

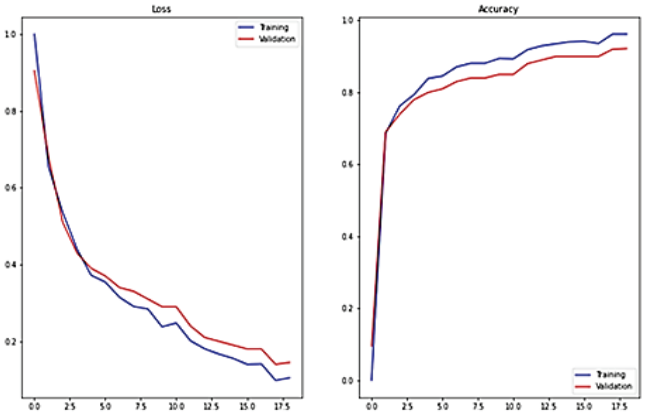


Figure 7 The result of loss and accuracy in the training stage

3.5 Evaluation of the Proposed System

A confusion matrix is a tabular representation where the rows correspond to the actual subjects and the columns reflect the predicted subjects. A line's presence indicates the quantity of matches. If the row corresponds to the column, meaning that the anticipated value equals the actual value, a mark or dot will be placed [4]. In the absence of such a result, a point will be appended beyond the boundary, signifying the presence of a diagonal line under conditions of high precision. The line signifies the true positive (*TP*) values, which occur when the predicted and actual values coincide ($i = j$). False positive (*FP*) values are denoted by any point above the line ($i \geq j$), while false negative (*FN*) values are represented by any point below the line ($i \leq j$).

The findings were obtained for each true positive (*TP*), false positive (*FP*), false negative (*FN*), and true negative (*TN*) in the confusion matrix for many subjects. The totals for each row and column were extracted and displayed in Fig. 8.

Understanding the calculation process in the deployed system yields eight outputs. The result is this: "The metrics used to evaluate the performance of a classification model are True Positives (*TP*), False Positives (*FP*), True Negatives (*TN*), False Negatives (*FN*), Precision, Recall, F1 score, and accuracy." The scales employed in the system are detailed in Tab. 5. Applicable to the fusion system, the scales' corresponding results are presented alongside an explanation of the measurements.

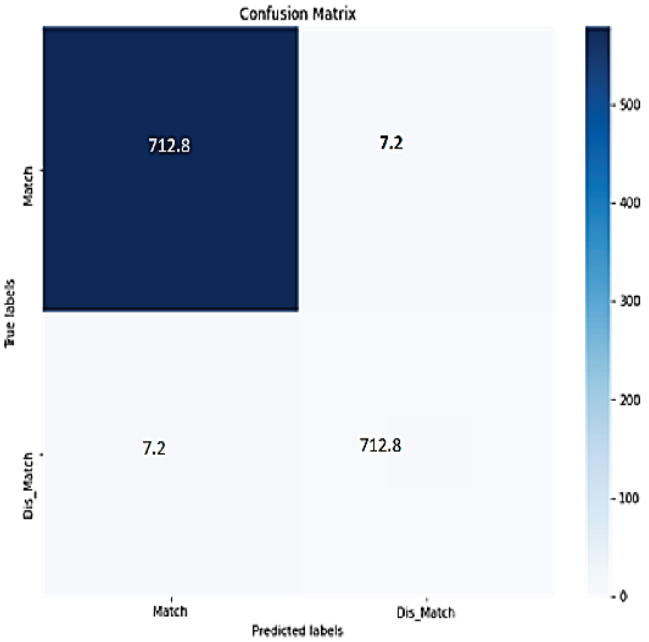


Figure 8 Total confusion matrix results

Table 5 Summarizes the result of testing in the proposed system

TP = 712.8	*FP* = 7.2	*TN* = 712.8	*FN* = 7.2
Precession	0.99 %		
F1	0.97 %		
Recall	0.98 %		

Table 6 Performance comparison through identification time for CASIA database

Arthur / Reference	Year	Methods	Accuracy
Lian Wu et al. [30]	2021	• Triple-Type Feature Descriptors (TFD) method for Triple-Type Feature Extraction (Texture, Gradient, Direction) then Feature Matching Fusion	88 %
Fatima A Ameen et al. [31]	2022	• VGG16 network, • Features extracted classified using the SoftMax function.	97.32 %
Amjad Rehman et al. [32]	2022	• Return Of Investment (ROI) • MRELBP • *k* - nearest neighbor classifier.	96.6 %
Mohamed Ezz et al. [41]	2023	• Siamese network • VGG-16	91.8 % on the left images 91.7 % on the right side of dataset
Ebtesam N. AlShemmary et al. [33]	2023	• CNNs	95.6 %
Eslam Hamouda et al. [34]	2023	• HAMTE • Siamese network	90.2 %
Proposed System			

Tab. 6 shows a comparison between research in previous works and what was achieved by the proposed system, what methods were used in previous works, and the accuracy that was achieved.

4 CONCLUSION

Identifying the nature of the palm is one of the topics that has become important in biometric systems because it possesses unique features that cannot be replicated among humans. In the research paper, an approach based on deep learning is proposed using the DenseNet-121 model and through pre-processing of the image by using the CLAHE filter and the normalization process, where the filter had an impact on the accuracy of the results because it worked to clarify the images by working to lighten the areas. Dark areas and blurring the light areas, where he worked on balancing the images of the palm print in the data set CASIA and the normalization process in order to avoid the model entering into a state of overfitting. Standardization of the data was performed and the results obtained were good, with accuracy reaching 99%, precession 0.99, f1 0.97 and recall 0.98.

5 REFERENCES

[1] Zhong, D., Du, X. & Zhong, K. (2019). Decade progress of palmprint recognition: A brief survey. *Neurocomputing, 328*, 16-28. https://doi.org/10.1016/j.neucom.2018.03.081

[2] Zhang, D., Zuo, W. & Yue, F. (2012). A comparative study of palmprint recognition algorithms. *ACM Comput. Surv., 44*(1), 1-37. https://doi.org/10.1145/2071389.2071391

[3] Haider AbdAlkreem, M., Sadoon Salman, R. & Khiled Al-Jibory, F. (2024). Detect People's Faces and Protect Them by Providing High Privacy Based on Deep Learning. *Tehnički glasnik, 18*(1), 92-99. https://doi.org/10.31803/tg-20231210183347

[4] Al-Tamimi, M. & AL-Khafaji, R. (2022). Finger vein recognition based on PCA and fusion convolutional neural network. *International Journal of Nonlinear Analysis and Applications, 13*(1), 3667-3681. https://doi.org/10.22075/ijnaa.2022.6145

[5] Genovese, A., Piuri, V. & Scotti, F. (2014). *Touchless palmprint recognition systems, vol. 60*. Springer. https://doi.org/10.1007/978-3-319-10365-5

[6] Al-Taie, S. A. M. & Khaleel, B. I. (2023). Palm Print Recognition Using Intelligent Techniques: A review. *J. Ilm. Tek. Elektro Komput. dan Inform.*, 9(1), 156-164.

[7] Hamidi, A., Khemgani, S. & Bensid, K. (2021). Transfer learning using vgg based on deep convolutional neural network for finger-knuckle-print recognition. In *Proceedings of the 2nd International Conference on Computer Science's Complex Systems and their Applications, Oum El Bouaghi, Algeria*, pp. 25-26.

[8] Brown D. & Bradshaw, K. (2019). Improved palmprint segmentation for robust identification and verification. *The 15th IEEE International Conference on Signal-Image Technology & Internet-Based Systems (SITIS2019)*, 1-7. https://doi.org/10.1109/SITIS.2019.00013

[9] Li, Q., Dong, P. & Zheng, J. (2020). Enhancing the security of pattern unlock with surface EMG-based biometrics. *Appl. Sci., 10*(2), p. 541. https://doi.org/10.3390/app10020541

[10] Meena, G. & Choudhary, S. (2019). Biometric authentication in internet of things: A conceptual view. *J. Stat. Manag. Syst., 22*(4), 643-652. https://doi.org/10.1080/09720510.2019.1609722

[11] Abood, Q. K. (2023). Predicting Age and Gender Using AlexNet. *TEM J., 12*(1). https://doi.org/10.18421/TEM121-61

[12] Fei, L., Zhang, B., Xu, Y., Guo, Z., Wen, J. & Jia, W. (2019). Learning discriminant direction binary palmprint descriptor. *IEEE Trans. Image Process., 28*(8), 3808-3820. https://doi.org/10.1109/TIP.2019.2903307

[13] Rida, I., Al-Maadeed, N., Al-Maadeed, S. & Bakshi, S. (2020). A comprehensive overview of feature representation for biometric recognition. *Multimed. Tools Appl., 79*, 4867-4890. https://doi.org/10.1007/s11042-018-6808-5

[14] Khaled, F. & Al-Tamimi, M. S. H. (2021). Plagiarism detection methods and tools: An overview. *Iraqi J. Sci.*, 2771-2783. https://doi.org/10.24996/ijs.2021.62.8.30

[15] Miura, N., Nagasaka, A. & Miyatake, T. (2007). Extraction of finger-vein patterns using maximum curvature points in image profiles. *IEICE Trans. Inf. Syst., 90*(8), 1185-1194. https://doi.org/10.1093/ietisy/e90-d.8.1185

[16] Fei, L., Zhang, B., Tian, C., Teng, S. & Wen, J. (2021). Jointly learning multi-instance hand-based biometric descriptor,. *Inf. Sci. (Ny)., 562*, 1-12. https://doi.org/10.1016/j.ins.2021.01.086

[17] Younis, M. A. & Al-Tamimi, M. S. H. (2022). Preparing of ECG Dataset for Biometric ID Identification with Creative Techniques. *TEM J., 11*(4).

[18] Raut, S. D. & Humbe, V. T. (2014). Review of biometrics: palm vein recognition system. *IBMRD's J. Manag. Res., 3*(1), 217-223.

[19] Jain, A. K., Ross, A. & Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Trans. circuits Syst. video Technol., 14*(1), 4-20. https://doi.org/10.1109/TCSVT.2003.818349

[20] Ungureanu, A.-S., Salahuddin, S. & Corcoran, P. (2020). Toward unconstrained palmprint recognition on consumer devices: A literature review. *IEEE Access, 8*, 86130-86148. https://doi.org/10.1109/ACCESS.2020.2992219

[21] Genovese, A., Piuri, V., Scotti, F. & Vishwakarma, S. (2019). Touchless palmprint and finger texture recognition: A deep learning fusion approach. In *The IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA2019)*, 1-6. https://doi.org/10.1109/CIVEMSA45640.2019.9071620

[22] Meraoumia, A., Chitroub, S. & Bouridane, A. (2011). Fusion of finger-knuckle-print and palmprint for an efficient multi-biometric system of person recognition. In *The IEEE International Conference on Communications (ICC2011)*, 1-5. https://doi.org/10.1109/icc.2011.5962661

[23] Matkowski, W. M., Chan, F. K. S. & Kong, A. W. K. (2019). A study on wrist identification for forensic investigation. *Image Vis. Comput., 88*, 96-112. https://doi.org/10.1016/j.imavis.2019.05.005

[24] Ungureanu, A.-S., Thavalengal, S., Cognard, T. E., Costache, C. & Corcoran, P. (2017). Unconstrained palmprint as a smartphone biometric. *IEEE Trans. Consum. Electron.*, 63(3), 334-342. https://doi.org/10.1109/TCE.2017.014994

[25] Meitram, R. & Choudhary, P. (2018). Palm vein recognition based on 2D Gabor filter and artificial neural network. *J. Adv. Inf. Technol., 9*(3). https://doi.org/10.12720/jait.9.3.68-72

[26] Al-Juboori, R. A. L. (2017). Contrast enhancement of the mammographic image using retinex with CLAHE methods. *Iraqi J. Sci.*, 327-336.

[27] Toygar, Ö., Babalola, F. O. & Bitirim, Y. (2020). FYO: a novel multimodal vein database with palmar, dorsal and wrist biometrics. *IEEE Access, 8*, 82461-82470. https://doi.org/10.1109/ACCESS.2020.2991475

[28] Al-Khafaji, R. S. S. & Al-Tamimi, M. S. H. (2022). Vein Biometric Recognition Methods and Systems: A Review. *Adv. Sci. Technol. Res. J.*, 16(1), 36-46. https://doi.org/10.12913/22998624/144495

[29] Ma, X., Jing, X., Huang, H., Cui, Y. & Mu, J. (2017). Palm vein recognition scheme based on an adaptive Gabor filter. *Iet Biometrics, 6*(5), 325-333. https://doi.org/10.1049/iet-bmt.2016.0085

[30] Wu, L., Xu, Y., Cui, Z., Zuo, Y., Zhao, S. & Fei, L. (2021). Triple-type feature extraction for palmprint recognition. *Sensors, 21*(14), p. 4896. https://doi.org/10.3390/s21144896

[31] Ameen, F. A. & AlShemmary, E. N. (2022). Palmprint Recognition Using VGG16. *Int. J. Tech. Phys. Probl. Eng. IJTPEJournal, 14*(53), 65-74.

[32] Rehman, A., Harouni, M., Karchegani, N. H. S., Saba, T., Bahaj, S. A. & Roy, S. (2022). Identity verification using palm print microscopic images based on median robust extended local binary pattern features and k-nearest neighbor classifier. *Microsc. Res. Tech.*, 85(4), 1224-1237. https://doi.org/10.1002/jemt.23989

[33] AlShemmary, E. & Ameen, F. A. (2023). Siamese Network-Based Palm Print Recognition. *J. Kufa Math. Comput.*, 10(1), 108-118. https://doi.org/10.31642/JoKMC/2018/100116

[34] Hamouda, E., Ezz, M. M., Mostafa, A. M., Elbashir, M. K., Alruily, M. & Tarek, M. (2023). Innovative Hetero-Associative Memory Encoder (HAMTE) for Palmprint Template Protection. *Comput. Syst. Sci. Eng.*, 46(1), 619-636. https://doi.org/10.32604/csse.2023.035830

[35] Hao, Y., Sun, Z., Tan, T. & Ren, C. (2008). Multispectral palm image fusion for accurate contact-free palmprint recognition. In *The 15th IEEE International Conference on Image Processing*, 281-284.

[36] Trabelsi, S., Samai, D., Dornaika, F., Benlamoudi, A., Bensid, K. & Taleb-Ahmed, A. (2022). Efficient palmprint biometric identification systems using deep learning and feature selection methods. *Neural Computing and Applications*, *34*(14), 12119-12141. https://doi.org/10.1007/s00521-022-07098-4

[37] Datwase, S. S., Deshmukh, R. R. & Gupta, R. S. (2022). Modern Available Palmprint Databases: A Review. *Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskiye nauki, 3*(227), 27-37. https://doi.org/10.18522/2311-3103-2022-3-27-37

[38] Al-Jaberi, A. S. & Al-Juboori, A. M. (2020). Palm vein recognition, a review on prospects and challenges based on CASIA's dataset. In *The 13th IEEE International Conference on Developments in eSystems Engineering (DeSE2020)*, 169-176. https://doi.org/10.1109/DeSE51703.2020.9450241

[39] Zuiderveld, K. (1994). *Graphics Gems IV*. Academic Press, San Diego, CA.

[40] Kim, J.-W., Kim, S.-B., Park, J.-C. & Nam, J.-W. (2015) Development of crack detection system with unmanned aerial vehicles and digital image processing. *Adv. Struct. Eng. Mech., 33*(3), 25-29.

[41] Ezz, M., Alanazi, W., Mostafa, A. M., Hamouda, E., Elbashir, M. K. & Alruily, M. (2023). Improved Siamese Palmprint Authentication Using Pre-Trained VGG16-Palmprint and Element-Wise Absolute Difference. *Comput. Syst. Sci. Eng., 46*(2). https://doi.org/10.32604/csse.2023.036567

Authors' contacts:

Ruaa Sadoon Salman
Ministry of Education, Karkh Three Directorate of Education,
Baghdad, Iraq
ruaa.s.alkhafaji@gmail.com

Mauj Haider AbdAlkreem
(Corresponding author)
Ministry of Education/Administrative Affairs,
Baghdad, Iraq
maujhader7@gmail.com

Qaswaa Khaled Abood
University of Baghdad, College of Science-Computer Science Department,
Baghdad Governorate, Baghdad, Iraq
qaswaa.k@sc.uobaghdad.edu.iq

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240229092505

Original scientific paper
Received: 2024-02-29, Accepted: 2024-05-21

Optimizing Scene Transitions for Sustained Narrative Immersion in Virtual Reality Films

Haein Yoon, Jin Wan Park*

Abstract: This paper investigates the challenges involved in adapting traditional film editing techniques for Virtual Reality (VR) films, with a particular focus on developing effective scene transitions that sustain narrative flow and enhance viewer immersion. It analyzes conventional editing methods and juxtaposes them against the unique demands of VR, leading to the proposal of solutions tailored to the immersive nature of VR. These solutions employ techniques such as the Dramatic Covenant, Long Take, and Field of View (FOV) adjustments, which are designed to improve spatial continuity and boost audience engagement in VR environments. The findings reveal that, although traditional techniques lay a fundamental groundwork, the unique characteristics of VR require a specialized approach that honors the viewer's immersive experience and their interaction within the narrative space. By developing practical strategies for filmmakers, this paper makes a contribution to the evolving field of VR films, thereby deepening our understanding of its unique narrative capabilities.

Keywords: film; narrative immersion; optimization; scene transition; virtual reality

1 INTRODUCTION

Literacy, traditionally understood as the ability to read and write, has continually evolved with new media forms. Media literacy extends this concept, encompassing the skills required to effectively understand, analyze, and interact with different forms of media [1]. This type of literacy is dynamic, expanding with each new media technology, from print to digital platforms.

The introduction of film brought a unique dimension to media literacy. It required audiences to learn to interpret visual narratives and understand cinematic techniques. This form of literacy involves decoding symbols, themes, and messages conveyed through the visual medium, which differs significantly from textual interpretation [2].

Virtual Reality (VR) marks a significant shift in media consumption, demanding a new form of literacy. Furthermore, VR literacy transcends traditional media engagement, requiring users to navigate and interact within a three-dimensional, immersive environment [3]. It challenges users to not only interpret content but to actively participate in it. In VR storytelling, scene transitions become pivotal in maintaining narrative flow and immersion [4, 20]. These transitions cannot simply replicate traditional film techniques; they require rethinking how stories unfold and how audiences engage with the narrative space.

This study explores the feasibility of applying traditional film grammar, such as "Invisible Editing", in VR [5]. Invisible Editing, which aims to create seamless transitions in film, faces unique challenges in VR's immersive environment, potentially causing disorientation or disrupting user immersion. Therefore, this study examines alternative approaches to VR scene transitions, such as the Dramatic Covenant, Long Take, and Field of View (FOV) techniques. Furthermore, this study contributes to a broader understanding of VR as a frontier in media. As VR technology continues to evolve, so will how we understand and interact with media. This ongoing transformation highlights the need for continuous research and adaptation in media literacy and VR storytelling.

Residual sections of the paper are organized as follow: In Section 2, reviews the required literature and cases, and in Section 3, discusses the proposed VR scene transitions. A summary and conclusion are provided in Section 4.

2 FILM EDITING AND ISSUES OF ADOPTING FILM GRAMMAR IN VR

The emergence of film as a distinct medium can be traced back to the time when filmmakers (directors) began selectively capturing reality and presenting it from their unique perspectives [6]. A prime example of this early phase in film is the world's first film, *The Arrival of the Train* (1896) (Fig. 1). This 50-second work, consisting of a single shot of a train arriving at La Ciotat station in France, demonstrated a straightforward portrayal of reality. While initially simple in its approach, focusing on real-time depiction without scene transitions or complex editing, this film laid the groundwork for the evolution of cinematic language. Its influence was instrumental in transitioning from static, single-shot scenes to the dynamic, multi-shot sequences that define contemporary filmmaking. *The Arrival of the Train* thus stands as a fundamental piece in the development of film, particularly highlighting the progression in scene transitions and editing techniques, marking a significant shift in how stories are told in cinema.

In the United States, Edwin S. Porter's *Life of an American Fireman* (1903) is a landmark in the history of film editing, showcasing early experimentation with narrative structure and editing techniques (Fig. 2). This film is particularly notable for its pioneering use of cross-cutting, a technique where two separate scenes are edited together to unfold simultaneously, creating a sense of urgency and narrative complexity [3]. Porter's innovative editing not only depicted the actions of the fireman and the unfolding drama inside the burning building in parallel but also revolutionized the way stories were told in cinema. This film played a crucial role in the evolution of cinematic language, moving it towards the sophisticated storytelling methods we see in modern cinema and solidifying the importance of editing as a key element of filmic expression.

During the era of silent cinema, approximately from the

mid-1910s to the late 1920s, the cinematic medium experienced prolific advancements in expressive techniques, with filmmakers globally exploring and refining the language of film from multifaceted cultural and artistic perspectives [6]. Among the notable works of this period, Sergei Eisenstein's *The Battleship Potemkin* (1925) emerges as an iconic silent Soviet film, renowned for its pioneering montage theory and intellectual montage sequences (Fig. 3). Contrasting with this, Edwin Porter's *The Great Train Robbery* (1903) demonstrated the nascent stages of cross-cutting, skillfully alternating between the narratives of the bandits' escape and the ensuing chase, thus innovatively presenting simultaneous time and space (Fig. 4). As the art of filmmaking advanced towards the works of D.W. Griffith, the language of editing crystallized into an essential tool for directorial storytelling, transforming from a mere technique into a core component of cinematic narrative [7]. This evolution brought about the refinement of temporal compression and the crafting of dramatic climaxes, heralding the maturation of cinematic storytelling language.

Figure 1 *Arrival of a Train*, 1896

Figure 2 *Life of an American Fireman*, 1903

Figure 3 *The Battleship Potemkin*, 1925

In the realm of VR, the aim is to foster a sense of physical presence within a computer-generated environment [8]. This contrasts with traditional film editing, which relies on the seamless division and recombination of time and space to create a narrative flow — a practice commonly referred to as "Invisible Editing" [5]. The direct application of filmic cuts

in VR may lead to a jarring 'teleportation effect,' disrupting the user's sense of immersion and continuity. Furthermore, cinematic techniques like gradual camera movements, which in film serve to guide the audience's gaze and emotions subtly, can provoke discomfort in VR, as they contradict the natural perception of motion [9]. The central challenge for VR storytelling, then, is to devise scene transitions that preserve the user's sense of spatial continuity and personal presence within the virtual environment. Neglecting this consideration can result in abrupt, disorienting transitions, akin to forced teleportation, which disrupt the natural flow and coherence from a content creation standpoint [10, 21].

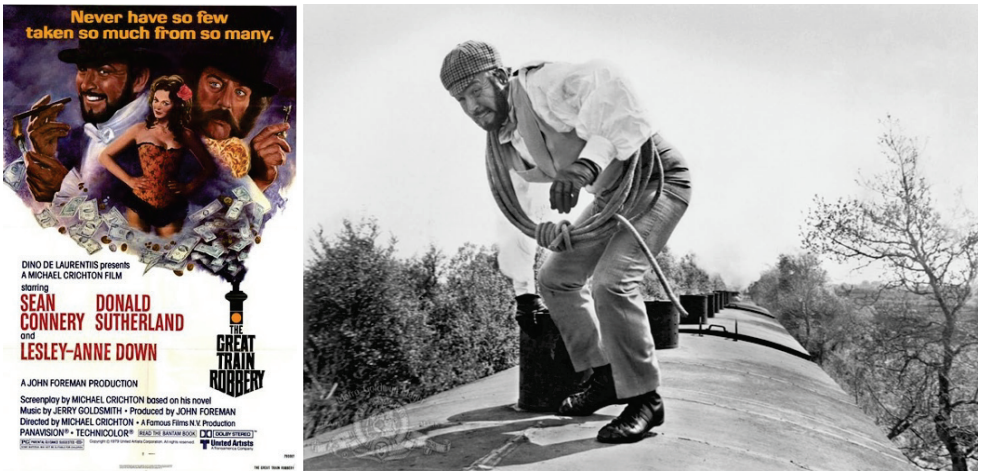


Figure 4 *The Great Train Robbery*, 1978

3 POSSIBLE PROPOSITIONS FOR VR TRANSITIONS

This section discusses alternative approaches that VR can adopt, distinct from the forceful and violent scene transitions that may result from adopting film grammar.

3.1 Dramatic Covenant – Explicit & Make Believe

In theatrical performances, audiences engage in a form of voluntary suspension of disbelief, recognizing the non-reality of the performance space [10]. This understanding is crucial in VR scene transitions as well. Here, as in theater, the audience implicitly agrees to accept the virtual space as a part of the narrative reality. This agreement, akin to a covenant, allows for a seamless transition between different spaces in the narrative. In films, a similar concept is evident, as seen in the film *Dogville* (2003), where the mere drawing of lines or placement of boxes serves as a covenant, signifying a transition in space (Fig. 5). This adaptation is promising as it uses familiar storytelling devices, with the audience's voluntary consent, to bridge different media forms.

Figure 5 *Dogville*, 2003, a view of the film set from above

To illustrate further, consider various examples of real-time scene transitions in theater, such as blackout-induced set changes, rotating stages, house curtains (left, right, up, and down), changes in lighting, and background transformations through projection mapping. When applied to VR, these methods allow for seamless scene transitions that create a sense of continuity for the audience [11].

Furthermore, encouraging the audience to imagine spatial movement collaboratively, as seen in *Dogville* (Fig. 6), or employing more explicit methods, such as rapidly

showcasing the process of resetting for a spatial change, such as dismantling stage settings, constructing a new building, and arranging props, maintains a natural form of spectatorship. In essence, onstage expressions involve a 'covenant' with the audience, and applying this concept in VR aims to explicitly reveal these agreements, transitioning the audience into a state of voluntary emotional immersion.

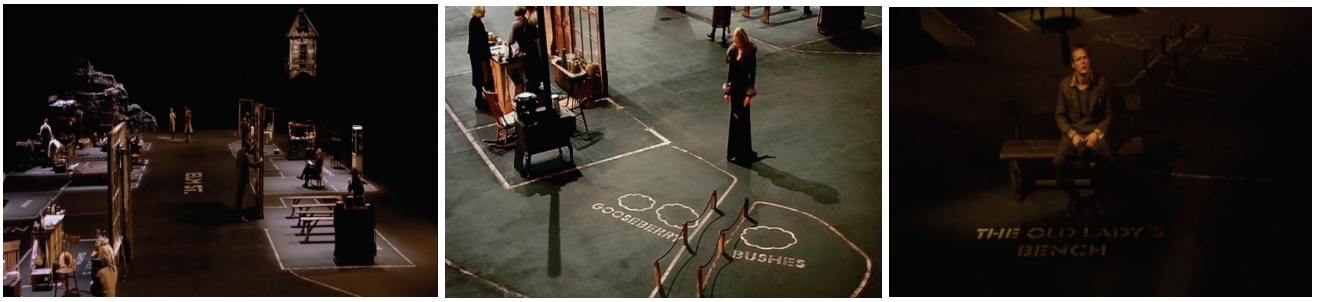


Figure 6 *Dogville* characters interacting with objects replaced by white signs and outlines

Figure 7 *The Player*, 1992, an opening sequence scene using the long take technique

Figure 8 *Chandelier*, 2014, a dance scene using the long take technique

3.2 Long Take – Intrinsic & Tricky Invisible Cut

The long take, a cinematographic technique characterized by extended, uninterrupted shots, is a pivotal method in film production [12]. This approach eschews traditional editing paradigms, opting for a continuous visual narrative. Such a technique not only challenges the technical prowess of filmmakers but also enriches narrative depth and viewer engagement [13].

Notable exemplars of this technique include Altman's *The Player* (1992) and the music video for Sia's *Chandelier* (2014). These works demonstrate the efficacy of the long take in fostering a fluid and immersive viewer experience. The technique's potency lies in its dual capacity to enhance narrative presence and skillfully mask its intricacies, presenting a seemingly effortless continuum to the audience.

First, the opening sequence in *The Player* exemplifies the long take, seamlessly extending for eight minutes without a discernible cut (Fig. 7). This uninterrupted shot ingeniously employs the camera movement to weave together multiple narrative threads within a single tracking scene. Next, in the *Chandelier*, creating a seamless and immersive experience as the camera unceasingly follows Maddie Ziegler's captivating dance performance through various rooms (Fig. 8). This long take technique, highlighting her expressive movements, not only showcases the synergy of cinematography and choreography but also amplifies the narrative and mood of the video. By doing so, it draws the viewer into a continuous,

unbroken flow, mirroring the song's themes of escapism and emotional turmoil.

In situations like these, the use of a long take can create the illusion of seamless transitions that are often imperceptible to the viewer. Such transitions often occur during natural occlusion, for instance, as the camera passes behind an object or during a swift movement within the frame. This methodology aligns with the principles of misdirection used in stage magic, where the audience's attention is so engrossed that the mechanics of the illusion are concealed.

Applying the technical editing methods of long takes to VR could involve seamlessly transitioning the environment by strategically obscuring the view with a smoothly moving object. This approach is reminiscent of portal zone loading, the opposite concept of seamless zon loading in 3D computer games. In essence, moving through an S-shaped passage could be likened to erasing data from the other side, providing a parallel to the concept of making data disappear by passing through a portal in VR transitions.

3.3 Gaze vs. Anti-Gaze of FOV

This section explores how the manipulation of the viewer's field of view (FOV) can be strategically used to guide or distract their attention. This manipulation is key to achieving smoother narrative transitions in virtual reality environments.

3.3.1 Out of FOV

In VR, FOV is an essential concept defining the scope and depth of the user's visual experience. FOV refers to the extent of the observable world visible through a VR Head-Mounted Display (HMD) at any given moment, measured in degrees [14, 26]. The maximum peripheral vision a person can see is about 180 degrees, and an example of each angle is shown in Fig. 9.

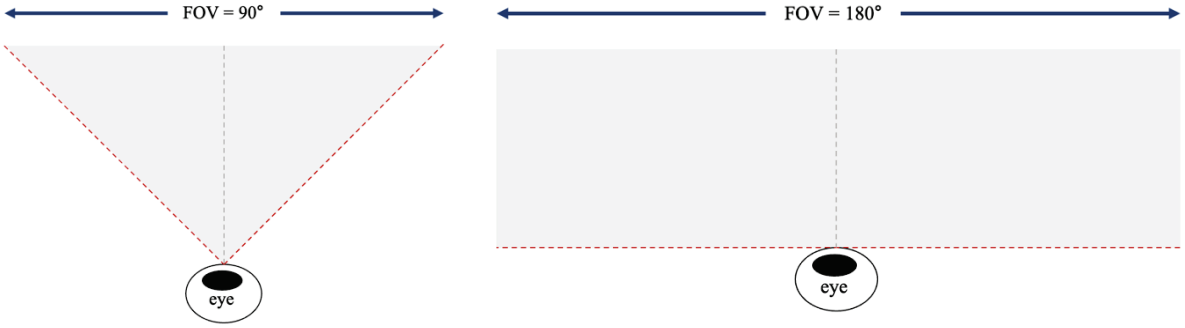


Figure 9 A Comparison of FOV by angle

A human can't see a full 360 degrees in a given situation. Leveraging this limitation to transform the unseen VR world allows natural scene transitions without compromising the user's sense of presence [15]. In such instances, the VR world can be perceived as a coexistence of a single shot and the next. Furthermore, to orchestrate a subtle shift in the virtual locale, one may either divert the user's attention from the target area or imperceptibly manipulate the digital environment itself. In this context, the transformation is induced externally, and the subject of the experience is passively led through the narrative. Such manipulation can engender a mysterious yet coherent sense of spatial progression. By emphasizing the newly introduced elements subtly, viewers are coaxed into accepting a "promised fiction", hence assimilating an illusion of natural spatial continuity. This practice not only harnesses the constraints of human vision to VR's advantage but also broadens the horizon for storytelling within the virtual scape.

3.3.2 Within FOV

The "Within FOV" approach pivots on engaging the viewer directly with the elements within their field of view [14]. Doing so fosters a sense of agency, allowing the audience to actively participate in and shape the unfolding narrative. This methodology significantly reduces reluctance towards scene transitions, as it is rooted in the viewer's control and interaction [16]. Such an immersive strategy is particularly effective in scenarios that require a more personal or introspective narrative approach, including flashbacks, where the viewer's engagement in the transition process can deeply enhance the emotional impact and authenticity of the experience.

The tactical implementation of this approach can take various forms, each designed to maximize the viewer's sense of involvement and control. Techniques such as zooming in on objects of the viewer's focus, transformative scene shifts, and the use of fantastical elements like particle effects are employed to create a seamless and engaging transition. These methods are reminiscent of traditional theatrical transitions, yet they are reinvented within the VR context to leverage the unique capabilities of this medium. In VR gaming, where user interaction and feedback are integral, these techniques align naturally with the design principles, enhancing the gameplay experience by making the player a co-creator of the virtual world [17]. Extending these methodologies to narrative VR content requires careful and thoughtful design. The director must carefully guide the viewer's gaze and decisions to steer the story along the desired path, creating the illusion of choice.

3.3.3 Out of Focus

While the two FOV methods described above determine the extent of the observable world within VR, "Out of Focus" techniques manipulate the sharpness of that world [18]. In other words, it gradually blurs the current scene before introducing the next, creating a natural flow that mirrors the way our mind processes shifts in attention or changes in environment. In VR storytelling, this method can be particularly effective. When there are changes in the narrative or location, the user may experience a gradual loss and regaining of focus. This can signal to the user that a transition is happening, and help them move smoothly from one scene to another, without abrupt changes.

The gradual transition achieved through the "Out of Focus" method not only aids in visual continuity but also enhances the narrative flow. By softening the edges of the scene before a transition, users are given a cue that the story is moving forward or shifting perspective [19, 23]. This method can be poignant in narrative-driven VR experiences where the flow of the story is paramount. The blur effect can be a storytelling device, indicating flashbacks, dream sequences, or shifts in the character's mental state.

Additionally, the method aligns with how human vision works, focusing and refocusing naturally, making it less taxing on the eyes. This subtle approach to scene transitions contributes to a more comfortable and prolonged VR experience, reducing the risk of visual fatigue or disorientation for the user.

3.4 Others

In addition, borrowing from traditional theater, VR can utilize 'dark changes' for transitions. This method involves momentarily plunging the scene into darkness before revealing the next scene. This can be particularly effective in VR as it gives the user's mind a brief pause, a blank canvas, momentarily free from visual stimuli, before introducing the new environment. This technique can also signify major narrative shifts or the passage of time, similar to how curtains close and open between acts in a theater.

Additionally, dramatic irony and foreshadowing elements, common in theatrical storytelling, can be adapted into VR. Subtle clues or motifs introduced early in one scene can foreshadow events or themes in the next, creating a cohesive narrative thread that guides the user through the VR experience [27].

Lighting, which directs user attention, can also be a subtle yet powerful tool for scene transitions in VR. By strategically altering lighting, such as spotlighting an object or character or gradually shifting the overall lighting tone, users can be guided naturally to the next focal point or scene. For instance, a gradual change from a brightly lit scene to a softer, dimly lit environment can indicate a transition to a more intimate or reflective part of the narrative [28]. This method not only guides the user visually but also sets the emotional tone for the new scene, enhancing the narrative impact.

Lastly, incorporating elements of deception and distraction offers another layer of depth to VR scene transitions. This approach can involve diverting the user's attention to one area of the scene while making changes in another. When the user's attention returns, they find the environment has transformed, creating a sense of surprise and wonder [29]. This method is akin to a magician's misdirection and can be particularly effective for introducing unexpected elements or shifts in the story. For instance, during an interactive VR experience, subtle changes in the surrounding environment can guide the user into a new narrative chapter. This method not only creates a smooth transition but also adds an element of discovery and playfulness to the VR experience.

4 CONCLUSION

This research has rigorously examined the adaptation of traditional film techniques to Virtual Reality (VR), focusing on scene transitions that are crucial for maintaining narrative flow and enhancing viewer immersion. Our research explored the practical application of cinematic techniques such as the Dramatic Covenant, Long Take, and Field of View (FOV) adjustments, which have proven pivotal in maintaining spatial continuity and enhancing audience engagement in immersive environments. We found that while these traditional techniques provide a solid foundation, the unique demands of VR require a reimagined approach to ensure seamless narrative transitions and to uphold the immersive quality that is signature to VR storytelling.

In particular, our analysis of VR scene transitions suggests that while these techniques can be adapted from traditional film, they must be modified to accommodate the three-dimensional and interactive nature of VR. The findings indicate that successful VR transitions hinge on the ability to not only guide but also manipulate the viewer's perception, which can be significantly more complex than in conventional filmmaking due to the viewer's ability to control their perspective within the virtual environment.

However, it's important to note that our study primarily relied on theoretical frameworks and conceptual analyses, which means the empirical evidence was limited. This lack of extensive empirical data might affect the generalizability of our conclusions, emphasizing the exploratory nature of this research. Future research is required to address this limitation by incorporating user-centered experiments to validate the effectiveness of the proposed VR transition techniques. These research could provide valuable insights into how viewers interact with and respond to these adapted film techniques within VR settings.

Ultimately, the insights garnered from this research contribute to the evolving landscape of VR film, highlighting the need for innovative approaches to integrate traditional film craft into new media forms. By continuing to explore and refine these techniques, filmmakers and content creators can better harness the potential of VR to tell stories in profoundly engaging ways. We hope that our findings will inspire further research and practical applications, advancing the field of VR storytelling and expanding its reach and impact in the media industry.

Acknowledgement

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A2A01061896).

5 REFERENCES

[1] Jeong, H.-S. (2007). *Media Education and Critical Literacy*. Communication Books.

[2] Bawden, D. (2008). Origins and concepts of digital literacy. *Digital Literacies: Concepts, Policies and Practices*, 30, 17-32, https://api.semanticscholar.org/CorpusID:158200341

[3] Sherman, W. R. & Craig, A. B. (1995). Literacy in virtual reality: a new medium. ACM SIGGRAPH Computer Graphics, 29(4), 37-42. https://doi.org/10.1145/216876.216887

[4] Marañes, C., Gutierrez, D. & Serrano, A. (2023). Towards assisting the decision-making process for content creators in cinematic virtual reality through the analysis of movie cuts and their influence on viewers' behavior. *International Transactions in Operational Research, 30*(3), 1245-1262. https://doi.org/10.1111/itor.13106

[5] Park, W. (2017). *Cinema Language*. AmorMundi.

[6] Joo, C. (1998). What is Cinema? Understanding History, Form, and Function. Georeum.

[7] Jeon, B. & Cha, M. (2018). VR & Changes in Cinematic Storytelling - Focusing on film composition unit, montage, space, mise-en-scène and perspective. Journal of Korea Multimedia Society, 21(8), 991-1001. (in Korean) https://doi.org/10.9717/kmms.2018.21.8.991

[8] Ding, N., Zhou, W. & Fung, A. Y. H. (2018). Emotional effect of cinematic VR compared with traditional 2D film. *Telematics and Informatics, 35*(6), 1572-1579. https://doi.org/10.1016/j.tele.2018.04.003

[9] Moss-Wellington, W., Sun, X. & Ch'ng, E. (2024). Going to the movies in VR: Virtual reality cinemas as alternatives to in-person co-viewing. *International Journal of Human-Computer Studies, 181*, 103150. https://doi.org/10.1016/j.ijhcs.2023.103150

[10] Mateer, J. (2017). Directing for Cinematic Virtual Reality: how the traditional film director's craft applies to immersive environments and notions of presence. *Journal of Media Practice, 18*(1), 14-25. https://doi.org/10.1080/14682753.2017.1305838

[11] Knorr, S. et al. (2018). Director's cut: a combined dataset for visual attention analysis in cinematic VR content. Proceedings of *the 15th ACM SIGGRAPH European Conference on Visual Media Production (CVMP'18)*. https://doi.org/10.1145/3278471.3278472

[12] Yoon, J. (2017). *Basic Concepts of Film Analysis*. Communication Books.

[13] Shafer, D. M., Carbonara, C. P. & Korpi, M. F. (2018). Exploring enjoyment of cinematic narratives in virtual reality: a comparison study. *International Journal of Virtual Reality*, 18(1), 1-18. https://doi.org/10.20870/IJVR.2018.18.1.2900

[14] Lin, J. J.-W. et al. (2002). Effects of field of view on presence, enjoyment, memory, and simulator sickness in a virtual environment. *Proceedings IEEE Virtual Reality 2002*. https://doi.org/10.1109/VR.2002.996519

[15] Duh, H. B.-L., Lin, J. W., Kenyon, R. V., Parker, D. E., & Furness, T. A. (2001). Effects of field of view on balance in an immersive environment. In *Proceedings IEEE Virtual Reality 2001*, 235-240. https://doi.org/10.1109/VR.2001.913791

[16] Slater, M., Usoh, M. & Steed, A. (1994). Depth of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 3(2), 130-144. https://doi.org/10.1162/pres.1994.3.2.130

[17] Liang, Z.-Q., Chen, M.-B., Wu, C.-X., Li, Y.-Q. & Lin, S.-H. (2021). The implementation and evaluation of the field of view in 3D PC game. *Int J Adv Appl Sci, 8*(12), 43-47. https://doi.org/10.21833/ijaas.2021.12.006

[18] Warren, W. H. & Kurtz, K. J. (1992). The role of central and peripheral vision in perceiving the direction of self-motion. *Perception & Psychophysics*, 51(5), 443-454. https://doi.org/10.3758/BF03211640

[19] Busselle, R. & Bilandzic, H. (2009). Measuring narrative engagement. *Media Psychology, 12*(4), 321-347. https://doi.org/10.1080/15213260903287259

[20] Dooley, K. (2017). Storytelling with virtual reality in 360-degrees: a new screen grammar. *Studies in Australasian cinema*, 11(3), 161-171. https://doi.org/10.1080/17503175.2017.1387357

[21] Aylett, R. & Louchart, S. (2003). Towards a narrative theory of virtual reality. *Virtual Reality, 7,* 2-9. https://doi.org/10.1007/s10055-003-0114-9

[22] Harris, R. J. & Cook, L. (2011). How content and co-viewers elicit emotional discomfort in moviegoing experiences: Where does the discomfort come from and how is it handled? *Applied Cognitive Psychology, 25*(6), 850-861. https://doi.org/10.1002/acp.1758

[23] Stanney, K. & Salvendy, G. (1998). Aftereffects and sense of presence in virtual environments: Formulation of a research and development agenda." *International Journal of Human-Computer Interaction*, 10(2), 135-187. https://doi.org/10.1207/s15327590ijhc1002_3

[24] Slater, M. & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments, 6*(6), 603-616.

[25] Kim, J., Jeong, Y., Stengel, M., Aksit, K., Albert, R. A., Boudaoud, B., Greer, T. et al. (2019). Foveated AR: dynamically-foveated augmented reality display. *ACM Trans. Graph., 38*(4), 1-15. https://doi.org/10.1145/3306346.3322987

[26] Ratcliff, J., Supikov, A., Alfaro, S. & Azuma, R. (2020). ThinVR: Heterogeneous microlens arrays for compact, 180 degree FOV VR near-eye displays. *IEEE Transactions on Visualization and Computer Graphics, 26*(5), 1981-1990. https://doi.org/10.1109/TVCG.2020.2973064

[27] Proferes, N. T. & Medina, L. J. (2017). *Film Directing Fundamentals: see your film before shooting*. Routledge.

[28] Nielsen, L. T., Møller, M. B., Hartmeyer, S. D., Ljung, T. C. M., Nilsson, N. C., Nordahl, R. & Serafin, S. (2016). Missing the point: an exploration of how to guide users' attention during cinematic virtual reality. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, 229-232. https://doi.org/10.1145/2993369.2993405

[29] Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1535), 3549-3557. https://doi.org/10.1098/rstb.2009.0138

Authors' contacts:

Haein Yoon
Department of Technology Arts, GSAIM of Chung-Ang University,
84 Heukseok-ro, Dongjak-gu, Seoul, Republic of Korea
salt9103@gmail.com

Jin Wan Park
(Corresponding author)
Department of Technology Arts, GSAIM of Chung-Ang University,
84 Heukseok-ro, Dongjak-gu, Seoul, Republic of Korea
jinpark@cau.ac.kr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240302201341

Original scientific paper
Received: 2024-03-02, Accepted: 2024-06-08

Optimizing Supply Chains: A Grey-DEMATEL Approach to Implementing LARG Framework

Hossein Talebzadeh*, Amirmohammad Fattahiamin, Mohammad Talebzadeh, Fariba Sanaei,
Parisa Khorashadi Moghaddam, Shervin Espahbod

Abstract: The performance of supply chains is directly impacted by overarching strategies and management paradigms. Success in the contemporary business landscape necessitates a comprehensive perspective that caters to the diverse needs of the market. The LARG (Lean-Agile-Resilient-Green) paradigm stands out as a versatile solution capable of addressing various concerns within the supply chain. This study introduces an innovative integrated framework for the implementation of the LARG supply chain, drawing upon insights from literature and expert knowledge, and encompassing 14 key factors. Subsequently, employing the grey-DEMATEL (decision-making trial and evaluation laboratory) method, we quantitatively measure the interrelations among these factors, culminating in the development of a structural model. The research findings underscore the significance of a technological approach as the most impactful factor in facilitating the LARG paradigm within the supply chain.

Keywords: Grey-DEMATEL Method; LARG Paradigm; supply chain performance; technological approach

1 INTRODUCTION

In the rapidly intensifying business landscape, the adoption of diverse strategies is imperative for gaining a competitive edge. Consequently, businesses are integrating various paradigms into their processes to enhance overall performance. Nevertheless, conflicts arising from the interplay of paradigms and strategies can disrupt planning and hinder performance. Furthermore, numerous global companies recognize the significance of effectively managing the linkages between supply sources and demand points. The escalating degree of globalization and interdependence among businesses has firmly entrenched the supply chain as an integral component of the business environment [1]. Consequently, upcoming competition is poised to shift from an organizational level to a supply chain level [2]. Hence, the application of a comprehensive perspective that addresses diverse concerns within a supply chain becomes imperative.

To formulate effective strategies aligning with the requirements of organizations and supply chains, the LARG (Lean - Agile - Resilient - Green) philosophy is suggested [3]. While intriguing, the multifaceted nature of the LARG paradigm within a supply chain introduces intricate trade-offs [4]. Consequently, navigating this complexity stands as a noteworthy accomplishment in supply chain management. Subsequently, this paper delves into a review of the four pillars of the LARG philosophy.

Lean: The concept of leanness originates from economic considerations within the production environment, with the primary objective of eliminating activities, resources, and processes that do not contribute to benefit. The overarching goal of the lean paradigm within a supply chain is to minimize the total costs across the entire supply chain [5]. The significance of this paradigm has been underscored in recent years, particularly during economic crises and their aftermath, prompting numerous businesses to embrace leanness as a survival strategy in the market [6]. Consequently, leanness assumes a pivotal role in organizational strategies and supply chain policies.

Agile: The ability to respond adeptly to unforeseen changes while maintaining consistent and acceptable performance is a crucial attribute in today's marketplace, referred to as agility [7]. Supply chain agility confers significant competitive advantages, given the escalating pace of changes in the business environment [8]. Nonetheless, supply chain agility is a multi-faceted concept that demands a comprehensive perspective [9].

Resilient: The response of a system, encompassing resistance and recovery, to the occurrence of disruptions, with the aim of preserving or restoring its original condition, defines resilience within that system [10]. The resilience engineering process, which involves designing or redesigning systems in line with resilience factors, plays a pivotal role in achieving supply chain resiliency [11]. However, the concept of resilience is not confined to specific guidelines; rather, it represents a widespread culture that endeavors to minimize the vulnerability of systems. Therefore, supply chain resiliency emerges as a contemporary and effective approach for safeguarding the supply chain against disruptions [12].

Green: The environmental and ecological challenges stand as among the most pressing issues faced by human societies, significantly influencing the business environment. Consequently, numerous studies have directed their attention to this domain, presenting various environmentally friendly approaches [13, 37]. Nevertheless, the intricate challenge of balancing economic benefits with environmental concerns persists globally. The emergence of diverse environmental crises has compelled decision-makers to intensify their focus on this realm [14, 38, 45].

The pursuit of integrating diverse paradigms to harness their respective advantages has been a focal point of research. While many previous studies have explored the integration of pairs of paradigms, the integration of Lean and Agile paradigms has garnered significant attention over an extended period [15, 16].

In this context, Agarwal et al. [17] conducted separate examinations of the outcomes associated with Lean and Agile paradigms. Subsequently, they delved into

investigating the outcomes stemming from the integration of Lean and Agile, referred to as the Leagile paradigm. They then proposed a comprehensive framework for achieving a Leagile Supply Chain (SC), encompassing various considerations relevant to SCs. The integration of Lean and Agile paradigms within an SC was scrutinized through the lens of strategic supplier partnerships by Qrunfleh and Tarafdar [18]. Their study underscored the pivotal role of SC strategies and practices in influencing the responsiveness and performance of the focal firm. To assess the level of Leagility in an SC, Rahiminezhad Galankashi and Helmi [19] introduced a novel measurement tool. Taking into account primary SC drivers, particularly logistic and cross-functional drivers, they outlined operational activities tailored for a Leagile SC and categorized them. The measurement tool was subsequently formulated based on these operational activities [35].

Presenting a comprehensive conceptual model that outlines the interrelations among Lean, Agile, Resilient, and Green paradigm practices in SCM, Carvalho et al. [20] position themselves as pioneers in integrating these four paradigms. Their emphasis lies in identifying efficient LARG practices and appropriate performance measures. Additionally, Cabral et al. [4] concentrated on LARG SCM to enhance supply chain competitiveness, recognizing it as a necessity in contemporary business environments. To this end, they introduced specific practices and utilized the Analytic Network Process method for optimal practice selection. However, it is noteworthy that the aspect of uncertainty in decision-making problems was overlooked in their approach.

A holistic framework for a Lean-Agile-Resilient-Green (LARG) supply chain amalgamates four pivotal paradigms in the supply chain, fortifying their interactions to align with the primary concerns of supply chain management. This paper presents a conceptual framework encompassing 14 factors aimed at facilitating the implementation of the LARG paradigm within the supply chain. These factors, derived from existing literature, have been prominently featured in prior studies. Subsequently, we employ the grey-DEMATEL (Decision-Making and Trial Evaluation Laboratory) method to discern and scrutinize the relationships between these factors, relying on expert judgment for a comprehensive analysis.

The subsequent sections of this paper are structured as follows: Section 2 conducts a thorough literature review and outlines a comprehensive framework for a Lean-Agile-Resilient-Green (LARG) supply chain. Section 3 details the grey-DEMATEL evaluation method proposed in this study. Section 4 consolidates and presents a summary of the obtained results. In Section 5, we delve into discussions and outline the managerial implications derived from the findings. Finally, Section 6 offers a synthesis of the paper's key discoveries and proposes potential directions for future research endeavors.

1.1 Literature Review

The exploration of integrating diverse paradigms to capitalize on their respective advantages has been a focal point of research. While many previous studies have examined the integration of a couple of paradigms, the combination of Lean and Agile paradigms has garnered considerable attention over an extended period [15-16].

Another well-explored pairing of paradigms for integration is Lean-Green Supply Chains (SCs) [20]. The integration of Lean and Green has gained considerable attention as a means to enhance both financial and environmental performance in SCs. Kainuma and Tawara [21] put forward a method for evaluating managerial and environmental performance in SCs, specifically considering re-use and recycling. Additionally, they delve into the significance of information sharing within a Lean-Green SC. A succinct review of prior research concurrently addressing Lean and Green SC management revealed its practical applicability. Dues et al. [22] specifically focused on trade-offs between Lean and Green practices in SCs, identifying potential synergies between the two paradigms. They introduced Lean and Green as complementary strategies in SCs. Carvalho et al. [23] put forth a strategic framework with a mathematical model to facilitate the integration of Lean and Green practices in SCs, emphasizing eco-efficiency. The proposed model was implemented in an automotive SC. Furthermore, Thanki and Thakkar [24] presented a qualitative assessment framework for Lean and Green SCs, based on Balanced Score Card perspectives. The proposed framework, anchored in causal relations between relevant factors, was implemented in a textile SC.

The integration of Lean, Green, and Resilient paradigms to improve the performance of an automotive SC has been explored by Govindan et al. [25]. Their study concentrates on specific practices and performance measures associated with each paradigm, employing the Interpretive Structural Modeling approach to analyze their interrelations. Additionally, Ruiz-Benitez et al. [26] investigated Lean, Green, and Resilient SC practices in the aerospace industry. They utilized the Interpretive Structural Modeling approach to scrutinize the interactions between these paradigms, complemented by the application of Importance-Performance Analysis to validate the results.

Ghazvinian et al. [3] proposed a comprehensive conceptual model elucidating the relationships among Lean, Agile, Resilient, and Green paradigm practices in supply chain management, asserting their leadership in integrating these four paradigms with suitability. They highlight the importance of identifying efficient LARG practices and appropriate performance measures. Additionally, Cabral et al. [4] concentrated on LARG Supply Chain Management (SCM) to enhance supply chain competitiveness, deeming it essential in contemporary business environments. To this end, they introduced specific practices and employed the Analytic Network Process (ANP) to select optimal practices. However, it is worth noting that their approach overlooked uncertainties in decision-making problems.

Additionally, Azevedo et al. [27] directed their attention to the LARG concept as a benchmarking tool for evaluating supply chain performance. In this regard, they employed the Delphi technique to assign weights to corresponding practices. Subsequently, they introduced a linear aggregated method tailored for automotive companies and their associated supply chains. Moreover, the application of LARG paradigms extends beyond the scope of supply chains. do Rosário Cabrita et al. [28] explored the integration of LARG principles into the Business Model Canvas as a strategy to enhance organizational business performance. Their objective was to refine the business model by incorporating LARG principles, striving to achieve an ideal business model for the organization.

2 RESEARCH METHODOLOGY

This study employs the grey–DEMATEL approach to uncover the relationships among LARG supply chain factors and identify the pivotal elements within this framework. DEMATEL is widely acknowledged as a comprehensive method for constructing and scrutinizing a structural model that unveils intricate causal relationships among diverse factors. Recognized for determining groups of influencing and influenced factors, DEMATEL has been frequently utilized in numerous articles where factors are considered as interconnected components [29-31]. This method was initially introduced by The Battelle Memorial Institute through its Geneva Research Centre. The DEMATEL technique relies on digraphs, which effectively illustrate the directed relationships among factors. Fig. 1 illustrates steps of current research.

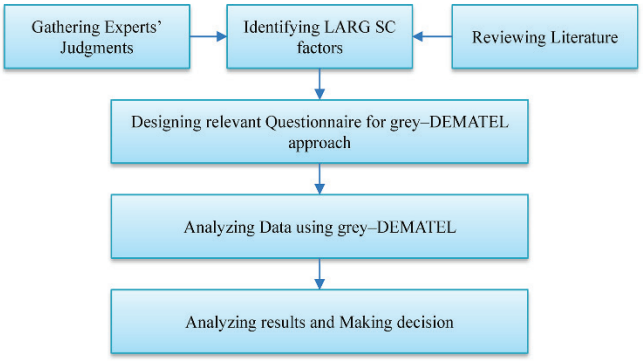


Figure 1 Research steps

The present research utilizes DEMATEL methodology to establish causal relationships among the alignment enablers that were identified, and to measure the magnitude of their impact. Nevertheless, the proposed framework relies on subjective factors, emphasizing the need to mitigate ambiguity arising from linguistic expression and the limited dataset used to evaluate factor relationships, as well as the inherent uncertainty in subjective assessments.

To address the inherent ambiguity in decision-making, the Grey System Theory or fuzzy sets theory could be employed [32, 38]. Grey numbers, forming the foundation of Grey systems, do not specify an exact value but rather determine an interval that contains the value. The primary advantage of grey systems lies in their capacity to generate plausible outcomes with incomplete information [33]. Consequently, numerous studies have applied grey systems theory across various domains.

The necessity for expressing results in precise numerical terms is evident. Therefore, we employ the modified CFCS method [34]. In the case of having a group of K evaluators, and $\otimes x_{ij}^k = \left[\otimes x_{ij}^k, \bar{\otimes} x_{ij}^k\right]$ representing a grey number used for evaluating the impact of the i^{th} factor on the j^{th} factor by the k^{th} evaluator, we can derive the crisp value through the following steps.

2.1 Modified CFCS Method

Step 1: The normalization process utilizing Eqs. (1) and (2).

$$\underline{\otimes}\bar{x}_{ij}^k = (\underline{\otimes} x_{ij}^k - \min_j \underline{\otimes} x_{ij}^k) \Big/ \Delta_{\min}^{\max} \tag{1}$$

$$\bar{\otimes}\bar{x}_{ij}^k = (\bar{\otimes} x_{ij}^k - \min_j \underline{\otimes} x_{ij}^k) \Big/ \Delta_{\min}^{\max} \tag{2}$$

Where $\Delta_{\min}^{\max} = \max_j \underline{\otimes} x_{ij}^k - \min_j \underline{\otimes} x_{ij}^k$.

Step 2: The computation of a total normalized crisp value, as outlined in Eq. (3).

$$Y_{ij}^k = \frac{\underline{\otimes}\bar{x}_{ij}^k (1 - \underline{\otimes}\bar{x}_{ij}^k) + \bar{\otimes}\bar{x}_{ij}^k \times \bar{\otimes}\bar{x}_{ij}^k}{1 - \underline{\otimes}\bar{x}_{ij}^k + \bar{\otimes}\bar{x}_{ij}^k} \tag{3}$$

Step 3: Computing the final crisp value by employing Eq. (4).

$$z_{ij}^k = \min_j \underline{\otimes} x_{ij}^k + Y_{ij}^k \Delta_{\min}^{\max}. \tag{4}$$

Step 4: Aggregating the crisp scores resulting from the defuzzification of K evaluations and generating an averaged score as illustrated in Eq. (5).

$$c_{ij} = \frac{1}{K} \sum_{i=1}^{k} z_{ij}^k. \tag{5}$$

2.2 DEMATEL

Step 1: Generating the normalized direct-influence matrix $\boldsymbol{D}$, denoted as $[d_{ij}]_{n\times n}$

$$\boldsymbol{D} = k \times \boldsymbol{A} \tag{6}$$

$$k = \min\left\{1 / \max_i \sum_{j=1}^{n} a_{ij},\ 1 / \max_j \sum_{i=1}^{n} a_{ij}\right\}, \tag{7}$$

$$i, j \in \{1, 2, ..., n\}$$

Step 2: Computing the total-influence matrix $\boldsymbol{T}$, denoted as $[t_{ij}]_{n\times n} \times \boldsymbol{I}$, where $\boldsymbol{I}$ is an $n \times n$ identity matrix.

$$\boldsymbol{T} = \boldsymbol{D}(\boldsymbol{I} - \boldsymbol{D})^{-1} \tag{8}$$

Step 3: Calculate *R* as the sum of rows and *J* as the sum of columns.

$$R = [r_i]_{n\times 1} = \left[\sum_{j=1}^{n} t_{ij}\right]_{n\times 1} \tag{9}$$

$$C = [c_j]_{n\times 1} = \left[\sum_{i=1}^{n} t_{ij}\right]_{1\times n} \tag{10}$$

Step 4: Creating a causal diagram by plotting (R+J, R–J).

3 RESULTS

In this study, following an extensive review of existing literature, we formulated a grey-DEMATEL questionnaire, which was then distributed to five experts in the petrochemical supply chain domain. In crafting the questionnaire and addressing potential uncertainties in judgments, we incorporated the term "influence" with five linguistic descriptors: Very High, High, Low, Very Low, and No. These descriptors were represented using grey numbers, as detailed in Tab. 1.

Table 1 The grey linguistic scale

Linguistic terms	Gery numbers
No Influence (NI)	(0 , 0)
Very Low influence (VL)	(0 , 0.25)
Low influence (L)	(0.25 , 0.5)
High influence (H)	(0.5 , 0.75)
Very High influence (VH)	(0.75 , 1)

Applying Grey system theory, we processed the data collected from the questionnaire. The transformation of Grey-DEMATEL values is elucidated in Tab. 2. Employing the specified equations, we implemented the DEMATEL method. The input data for the DEMATEL technique is presented in Tab. 3, and the resultant outcomes are detailed in Tab. 4.

Table 2 The initial direct-relation matrix

Grey value	Normalized value	Total normalized crisp value	Final crisp value
[0, 0]	[0,0]	0	0
[0.5, 0.75]	$[\frac{0.5-0}{1-0}, \frac{0.75-0}{1-0}]$	$(\frac{0.5\times(1-0.5)+0.75\times 0.75}{1-0.5+0.75})$	0 +0.65
[0.25, 0.5]	$[\frac{0.25-0}{1-0}, \frac{0.5-0}{1-0}]$	$(\frac{0.25\times(1-0.25)+0.5\times 0.5}{1-0.25+0.5})$	0 +0.35
[0, 0.25]	$[0, \frac{0.25-0}{1-0}]$	$(\frac{0\times(1-0)+0.25\times 0.25}{1-0+0.25})$	0 +0.05
[0.25, 0.5]	$[\frac{0.25-0}{1-0}, \frac{0.5-0}{1-0}]$	$(\frac{0.25\times(1-0.25)+0.5\times 0.5}{1-0.25+0.5})$	0 +0.35
[0, 0.25]	$[0, \frac{0.25-0}{1-0}]$ر	$(\frac{0\times(1-0)+0.25\times 0.25}{1-0+0.25})$	0 +0.05
[0, 0.25]	$[0, \frac{0.25-0}{1-0}]$	$(\frac{0\times(1-0)+0.25\times 0.25}{1-0+0.25})$	0 +0.05
[0, 0]	[0 , 0]	0	0
[0, 0]	[0 , 0]	0	0
[0, 0.25]	$[0, \frac{0.25-0}{1-0}]$	$(\frac{0\times(1-0)+0.25\times 0.25}{1-0+0.25})$	0 +0.05
[0, 0.25]	$[0, \frac{0.25-0}{1-0}]$	$(\frac{0\times(1-0)+0.25\times 0.25}{1-0+0.25})$	0 +0.05
[0, 0]	[0 , 0]	0	0
[0, 0.25]	$[0, \frac{0.25-0}{1-0}]$	$(\frac{0\times(1-0)+0.25\times 0.25}{1-0+0.25})$	0 +0.05
[0.5, 0.75]	$[\frac{0.5-0}{1-0}, \frac{0.75-0}{1-0}]$	$(\frac{0.5\times(1-0.5)+0.75\times 0.75}{1-0.5+0.75})$	0 +0.65

Table 3 The initial direct-relation matrix

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
C1	0	0.95	0.275	0.35	0.187	0.262	0.71	0.175	0.187	0.25	0.425	0.337	0.262	0.725
C2	0.725	0	0.5	0.5	0.187	0.425	0.425	0.187	0.262	0.187	0.337	0.425	0.425	0.275
C3	0.35	0.262	0	0.475	0.562	0.425	0.25	0.35	0.1	0.575	0.325	0.65	0.337	0.375
C4	0.1	0.275	0.262	0	0.412	0.425	0.412	0.575	0.087	0.5	0.5	0.875	0.175	0.18
C5	0.1	0.012	0.337	0.575	0	0.725	0.35	0.204	0.187	0.8	0.337	0.637	0.325	0.725
C6	0.35	0.426	0.575	0.65	0.5	0	0.337	0.337	0.487	0.35	0.65	0.412	0.725	0.65
C7	0.2	0.5	0.25	0.65	0.65	0.575	0	0.325	0.575	0.575	0.175	0.575	0.487	0.4
C8	0.012	0.112	0.65	0.65	0.487	0.575	0.725	0	0.637	0.725	0.8	0.725	0.575	0.175
C9	0.1	0.337	0.425	0.575	0.575	0.575	0.425	0.725	0	0.5	0.725	0.5	0.725	0.325
C10	0.1	0.337	0.637	0.725	0.725	0.412	0.637	0.337	0.262	0	0.65	0.875	0.575	0.35
C11	0.175	0.275	0.487	0.725	0.8	0.5	0.5	0.412	0.425	0.5	0	0.8	0.5	0.4
C12	0.175	0.187	0.187	0.725	0.65	0.35	0.237	0.325	0.4	0.875	0.8	0	0.5	0.25
C13	0.35	0.5	0.575	0.425	0.725	0.725	0.8	0.637	0.575	0.725	0.875	0.502	0	0.575
C14	575	0.65	0.487	0.25	0.8	0.8	0.875	0.337	0.325	0.8	0.725	0.65	575	0

C1: Green design, C2: Environmental management, C3: Supplier relationship, C4: Customer relationship, C5: Just in time, C6: Quality integration, C7: Flexibility, C8: Information sharing, C9: Visibility, C10: Velocity, C11: Information management, C12: Responsiveness, C13: Internal management, C14: Technological approaches

Table 4 The total-relation matrix

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
C1	0.076	0.212	0.155	0.197	0.180	0.176	0.145	0.127	0.119	0.188	0.206	0.211	0.163	0.199
C2	0.160	0.108	0.185	0.225	0.190	0.202	0.186	0.137	0.134	0.191	0.205	0.231	0.189	0.157
C3	0.118	0.140	0.135	0.235	0.245	0.212	0.177	0.161	0.122	0.250	0.215	0.271	0.188	0.174
C4	0.083	0.133	0.161	0.174	0.222	0.204	0.189	0.182	0.119	0.235	0.228	0.288	0.165	0.144
C5	0.096	0.121	0.188	0.262	0.199	0.262	0.204	0.157	0.143	0.292	0.234	0.288	0.202	0.225
C6	0.143	0.192	0.242	0.303	0.290	0.211	0.231	0.197	0.199	0.275	0.303	0.299	0.275	0.243
C7	0.115	0.186	0.190	0.286	0.286	0.260	0.173	0.182	0.197	0.281	0.231	0.296	0.233	0.200
C8	0.103	0.158	0.260	0.323	0.307	0.289	0.285	0.168	0.227	0.334	0.333	0.352	0.272	0.196
C9	0.112	0.179	0.230	0.303	0.305	0.282	0.246	0.245	0.147	0.298	0.317	0.316	0.280	0.208
C10	0.111	0.178	0.248	0.318	0.319	0.261	0.264	0.198	0.175	0.239	0.303	0.355	0.259	0.209
C11	0.118	0.169	0.229	0.314	0.323	0.268	0.246	0.204	0.191	0.294	0.226	0.343	0.248	0.213
C12	0.105	0.143	0.179	0.289	0.282	0.227	0.198	0.178	0.172	0.308	0.294	0.226	0.228	0.177
C13	0.163	0.229	0.282	0.333	0.368	0.341	0.325	0.264	0.243	0.369	0.376	0.366	0.236	0.272
C14	0.188	0.245	0.265	0.305	0.367	0.342	0.325	0.222	0.209	0.368	0.350	0.371	0.296	0.202

Table 5 The impacts exerted and received by each factor.

	R	J	R+J	R–J
C1	2.362	1.699	4.061	0.663
C2	2.507	2.399	4.906	0.108
C3	2.648	2.955	5.604	-0.307
C4	2.531	3.873	6.405	-1.342
C5	2.879	3.888	6.768	-1.008
C6	3.409	3.544	6.953	-0.135
C7	3.121	3.200	6.322	-0.078
C8	3.614	2.629	6.243	0.985
C9	3.476	2.402	5.878	1.073
C10	3.443	3.928	7.372	-0.485
C11	3.393	3.828	7.221	-0.435
C12	3.013	4.220	7.234	-1.206
C13	4.173	3.239	7.413	0.933
C14	4.061	2.825	6.886	1.236

The prominence of each factor is indicated by its (R+J) value, which denotes its significance. In addition, the (R–J) value is assigned as the relation value and classifies factors into two categories: cause group and effect group. The cause group includes factors with positive (R–J), such as (C1), (C2), (C8), (C9), (C13), and (C14), while the effect group comprises factors with negative (R–J), including (C3), (C4), (C5), (C6), (C7), (C10), (C11), and (C12). A detailed summary of the cause group and effect group is presented in Tab. 5. The causal diagram illustrated in Fig. 2 is generated based on the (R+J) and (R–J) values. This graphical representation effectively illustrates the distinctions between the cause group and the effect group.

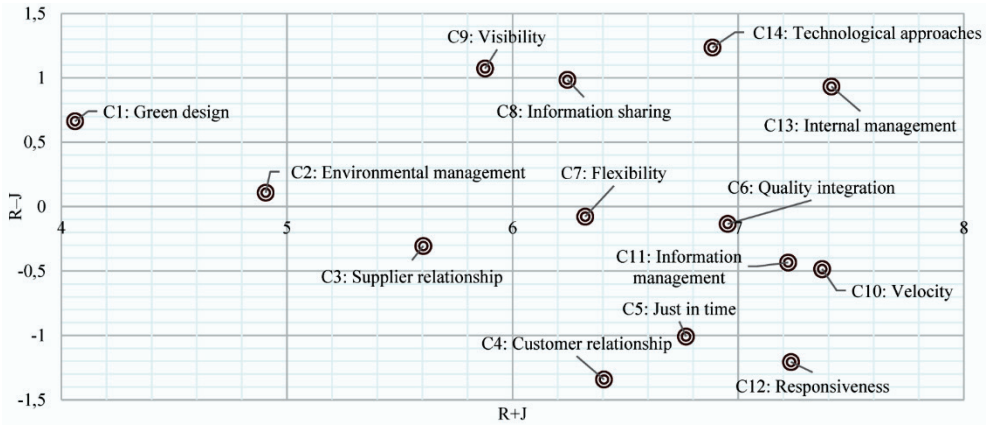


Figure 2 DEMATEL casual diagram

As depicted in Fig. 2, the disparity in the (R+J) index among factors is not highly significant. However, internal management exhibits the most substantial interconnections with the entire system, while green design displays fewer relationships.

In contrast to the previous index, the (R–J) index reveals notable differences among some factors. A detailed examination of the (R–J) index underscores the significance of cause group and effect group factors. Illustrated in Fig. 2, cause group factors encompass green design, environmental management, information sharing, internal management, and technological approaches [46]. Notably, technological approaches emerge as the pivotal factor for enhancing the LARG concept. Conversely, the effect group factors comprise supplier relationship, customer relationship, just in time, quality integration, flexibility, velocity, information management, and responsiveness.

4 MANAGERIAL IMPLICATIONS

Factors in the cause group act as catalysts for improvement since they instigate changes. Therefore, they demand heightened attention due to their significant impact on other factors. This paper underscores the importance of prioritizing focus on green design, environmental management, information sharing, velocity, quality integration, flexibility, internal management, and technological approaches as key elements in the cause group.

These factors are instrumental in facilitating the LARG concept within a supply chain. However, Fig. 2 highlights variations in the influence levels of each factor within this group, emphasizing the need for appropriate prioritization.

Firstly, the integration of modern technologies tailored to the particular requirements of the commercial setting has the potential to significantly enhance overall supply chain performance. The increasing complexities within the business landscape further underscore the paramount importance of adopting technological approaches. Thus, prioritizing a focus on technological advancements is deemed essential for all business entities, with managers playing a unique and pivotal role as drivers of business success.

Secondly, succeeding in today's dynamic business environment necessitates an accurate understanding of the situation and informed decision-making. In this context, information sharing emerges as a critical enabler, expediting the dissemination of information and enhancing the supply chain's capabilities to navigate challenges. The strategic emphasis of managers on fostering information sharing through infrastructure improvement and cultural development becomes imperative.

Thirdly, while paradigms and strategies are employed in supply chains to enhance overall performance, incorrect implementation poses a significant threat to the relationships across different layers in a supply chain. The relationships with customers and suppliers remain perpetual concerns in the supply chain domain. Therefore, during the initial stages of implementing new paradigms, managers must safeguard these crucial relationships from vulnerability.

Fourthly, all the interrelations prove to be effective for implementing the LARG paradigm in supply chains.

5 CONCLUSION

This paper endeavors to present a conceptual framework aimed at facilitating the implementation of a Lean-Agile-Resilient-Green supply chain and exploring the intricate interactions among its components. A comprehensive review of the literature culminated in the development of a reference framework comprising 14 factors. Subsequently, the grey-DEMATEL approach was employed to pinpoint critical factors and scrutinize their causal relationships. Given the inherent nature of our data, the grey system theory emerged as the most suitable method for deriving meaningful results. The findings highlight technological approaches as the most influential factor in the context of LARG supply chain implementation. Additionally, factors such as green design, environmental management, information sharing, and internal management are identified within the cause group, warranting heightened attention. Extensive exploration of the interactions among these enablers has been conducted and the results are thoroughly discussed. In conclusion, future research endeavors may involve the application of these findings across diverse industries in the new digital era [41], tailoring those to specific contexts and requirements. Future research on the LARG supply chain paradigm can be improved by combining qualitative [47, 48] and quantitative data analysis techniques. Interviews with supply chain professionals can give insights on implementing the LARG framework, while statistical methods and machine learning [36, 49] can identify trends and correlations among key factors. Data mining [39, 43], deep learning [44], and Artificial Intelligence [40, 42, 48, 50] techniques can further extract valuable information from supply chain data, providing actionable insights for optimization. By integrating these approaches, research can offer a comprehensive understanding of the LARG framework's effectiveness in different industries.

6 REFERENCES

[1] Christopher, M. & Holweg, M. (2011). Supply Chain 2.0: Managing supply chains in the era of turbulence. *International journal of physical distribution & logistics management*, *41*(1), 63-82. https://doi.org/10.1108/09600031111101439

[2] Mahdavimanshadi, M., Anaraki, M. G., Mowlai, M. & Ahmadirad, Z. (2024). A Multistage Stochastic Optimization Model for Resilient Pharmaceutical Supply Chain in COVID-19 Pandemic Based on Patient Group Priority. In *IEEE Systems and Information Engineering Design Symposium (SIEDS2024)*, 382-387. https://doi.org/10.1109/SIEDS61124.2024.10534683

[3] Ghazvinian, A., Feng, B., Feng, J., Talebzadeh, H. & Dzikuć, M. (2024). Lean, Agile, Resilient, Green, and Sustainable (LARGS) Supplier Selection Using Multi-Criteria Structural Equation Modeling under Fuzzy Environments. *Sustainability, 16*(4), 1594. https://doi.org/10.3390/su16041594

[4] Cabral, I., Grilo, A. & Cruz-Machado, V. (2012). A decision-making model for lean, agile, resilient and green supply chain management. *International Journal of Production Research*, *50*(17), 4830-4845. https://doi.org/10.1080/00207543.2012.657970

[5] Farahani, R. Z., Rezapour, S., Drezner, T. & Fallah, S. (2014). Competitive supply chain network design: An overview of classifications, models, solution techniques and applications. *Omega*, *45*, 92-118. https://doi.org/10.1016/j.omega.2013.08.006

[6] Prajogo, D., Oke, A. & Olhager, J. (2016). Supply chain processes: Linking supply logistics integration, supply performance, lean processes and competitive performance. *International Journal of Operations & Production Management*, *36*(2), 220-238. https://doi.org/10.1108/IJOPM-03-2014-0129

[7] Dokhanian, S., Sodagartojgi, A., Tehranian, K., Ahmadirad, Z., Moghaddam, P. K. & Mohsenibeigzadeh, M. (2024). Exploring the impact of supply chain integration and agility on commodity supply chain performance. *World Journal of Advanced Research and Reviews, 22*(1), 441-450. https://doi.org/10.30574/wjarr.2024.22.1.1119

[8] Mehregan, E., Sanaei, S., Manna, M., Bozorgkhou, H., & Heidari, S. (2023). The role of SCM practices in competitive advantage and firm performance: a mediating role of supply chain innovation and TQM. *Tehnički glasnik, 17*(4), 516-523. https://doi.org/10.31803/tg-20221223200658

[9] Gligor, D. M. (2014). The role of demand management in achieving supply chain agility. *Supply Chain Management: An International Journal*, *19*(5/6), 577-591. https://doi.org/10.1108/SCM-10-2013-0363

[10] Kamalahmadi, M. & Parast, M. M. (2016). A review of the literature on the principles of enterprise and supply chain resilience: Major findings and directions for future research. *International journal of production economics*, *171*, 116-133.

https://doi.org/10.1016/j.ijpe.2015.10.023
[11] Scholten, K., Sharkey Scott, P. & Fynes, B. (2014). Mitigation processes–antecedents for building supply chain resilience. *Supply Chain Management: An International Journal*, *19*(2), 211-228. https://doi.org/10.1108/SCM-06-2013-0191
[12] Hosseini, S., Barker, K. & Ramirez-Marquez, J. E. (2016). A review of definitions and measures of system resilience. *Reliability Engineering & System Safety*, *145*, 47-61. https://doi.org/10.1016/j.ress.2015.08.006
[13] Zhao, R., Liu, Y., Zhang, N. & Huang, T. (2017). An optimization model for green supply chain management by using a big data analytic approach. *Journal of Cleaner Production*, *142*, 1085-1097. https://doi.org/10.1016/j.jclepro.2016.03.006
[14] Fahimnia, B., Sarkis, J. & Davarzani, H. (2015). Green supply chain management: A review and bibliometric analysis. *International Journal of Production Economics*, *162*, 101-114. https://doi.org/10.1016/j.ijpe.2015.01.003
[15] Soni, G. & Kodali, R. (2012). Evaluating reliability and validity of lean, agile and leagile supply chain constructs in Indian manufacturing industry. *Production Planning & Control*, *23*(10-11), 864-884. https://doi.org/10.1080/09537287.2011.642207
[16] Ciccullo, F., Pero, M., Caridi, M., Gosling, J. & Purvis, L. (2018). Integrating the environmental and social sustainability pillars into the lean and agile supply chain management paradigms: A literature review and future research directions. *Journal of cleaner production*, *172*, 2336-2350. https://doi.org/10.1016/j.jclepro.2017.11.176
[17] Agarwal, A., Shankar, R. & Tiwari, M. K. (2006). Modeling the metrics of lean, agile and leagile supply chain: An ANP-based approach. *European journal of operational research*, *173*(1), 211-225. https://doi.org/10.1016/j.ejor.2004.12.005
[18] Qrunfleh, S. & Tarafdar, M. (2013). Lean and agile supply chain strategies and supply chain responsiveness: the role of strategic supplier partnership and postponement. *Supply Chain Management: An International Journal*, *18*(6), 571-582. https://doi.org/10.1108/SCM-01-2013-0015
[19] Galankashi, M. R., Helmi, S. A. & Hashemzahi, P. (2016). Supplier selection in automobile industry: A mixed balanced scorecard–fuzzy AHP approach. *Alexandria Engineering Journal*, *55*(1), 93-100. https://doi.org/10.1016/j.aej.2016.01.005
[20] Carvalho, H., Azevedo, S. G. & Cruz-Machado, V. (2010). Supply chain performance management: lean and green paradigms. *International Journal of Business Performance and Supply Chain Modelling*, *2*(3-4), 304-333. https://doi.org/10.1504/IJBPSCM.2010.036204
[21] Kainuma, Y. & Tawara, N. (2006). A multiple attribute utility theory approach to lean and green supply chain management. *International Journal of Production Economics*, *101*(1), 99-108. https://doi.org/10.1016/j.ijpe.2005.05.010
[22] Dües, C. M., Tan, K. H. & Lim, M. (2013). Green as the new Lean: how to use Lean practices as a catalyst to greening your supply chain. *Journal of cleaner production*, *40*, 93-100. https://doi.org/10.1016/j.jclepro.2011.12.023
[23] Carvalho, H., Govindan, K., Azevedo, S. G. & Cruz-Machado, V. (2017). Modelling green and lean supply chains: An eco-efficiency perspective. *Resources, Conservation and Recycling*, *120*, 75-87. https://doi.org/10.1016/j.resconrec.2016.09.025
[24] Thanki, S. & Thakkar, J. (2018). A quantitative framework for lean and green assessment of supply chain performance. *International Journal of Productivity and Performance Management*, *67*(2), 366-400. https://doi.org/10.1108/IJPPM-09-2016-0215
[25] Govindan, K., Azevedo, S. G., Carvalho, H. & Cruz-Machado, V. (2015). Lean, green and resilient practices influence on supply chain performance: interpretive structural modeling approach. *International Journal of Environmental Science and Technology*, *12*, 15-34. https://doi.org/10.1007/s13762-013-0409-7
[26] Ruiz-Benitez, R., López, C. & Real, J. C. (2017). Environmental benefits of lean, green and resilient supply chain management: The case of the aerospace sector. *Journal of cleaner production*, *167*, 850-862. https://doi.org/10.1016/j.jclepro.2017.07.201
[27] Azevedo, S. G., Carvalho, H. & Cruz-Machado, V. (2016). LARG index: A benchmarking tool for improving the leanness, agility, resilience and greenness of the automotive supply chain. *Benchmarking: An International Journal*, *23*(6), 1472-1499. https://doi.org/10.1108/BIJ-07-2014-0072
[28] do Rosário Cabrita, M., Duarte, S., Carvalho, H. & Cruz-Machado, V. (2016). Integration of lean, agile, resilient and green paradigms in a business model perspective: theoretical foundations. *IFAC-PapersOnLine*, *49*(12), 1306-1311. https://doi.org/10.1016/j.ifacol.2016.07.704
[29] Su, C. M., Horng, D. J., Tseng, M. L., Chiu, A. S., Wu, K. J. & Chen, H. P. (2016). Improving sustainable supply chain management using a novel hierarchical grey-DEMATEL approach. *Journal of cleaner production*, *134*, 469-481. https://doi.org/10.1016/j.jclepro.2015.05.080
[30] Govindan, K., Khodaverdi, R. & Vafadarnikjoo, A. (2016). A grey DEMATEL approach to develop third-party logistics provider selection criteria. *Industrial Management & Data Systems*, *116*(4), 690-722. https://doi.org/10.1108/IMDS-05-2015-0180
[31] Kirkire, M. S. & Rane, S. B. (2017). Evaluation of success factors for medical device development using grey DEMATEL approach. *Journal of Modelling in Management*, *12*(2), 204-223. https://doi.org/10.1108/JM2-09-2015-0062
[32] Bai, C. & Sarkis, J. (2013). A grey-based DEMATEL model for evaluating business process management critical success factors. *International Journal of Production Economics*, *146*(1), 281-292. https://doi.org/10.1016/j.ijpe.2013.07.011
[33] Fu, X., Zhu, Q. & Sarkis, J. (2012). Evaluating green supplier development programs at a telecommunications systems provider. *International Journal of Production Economics*, *140*(1), 357-367. https://doi.org/10.1016/j.ijpe.2011.08.030
[34] Xia, X., Govindan, K. & Zhu, Q. (2015). Analyzing internal barriers for automotive parts remanufacturers in China using grey-DEMATEL approach. *Journal of cleaner production*, *87*, 811-825. https://doi.org/10.1016/j.jclepro.2014.09.044
[35] Khorsandi, H. & Bayat, M. (2022). Prioritizing operational strategies of saman bank. *International Journal of Health Sciences*, 6(S7), 1442-1453. https://doi.org/10.53730/ijhs.v6nS7.11548
[36] Larijani, A. & Dehghani, F. (2023). An Efficient Optimization Approach for Designing Machine Models Based on Combined Algorithm. *FinTech*, 3(1), 40-54. https://doi.org/10.3390/fintech3010003
[37] Bazmi, M., Gong, J., Jessen, K. & Tsotsis, T. (2024). Waste CO_2 capture and utilization for methanol production via a novel membrane contactor reactor process: techno-economic analysis (TEA), and comparison with other existing and emerging technologies. *Chemical Engineering and Processing-Process Intensification*, 109825. https://doi.org/10.1016/j.cep.2024.109825
[38] Anbari, M., Arıkan Öztürk, E. B. R. U. & Ateş, H. (2020). Evaluation of sustainable transport strategies for Tehran with thetheir urbanization rate criterion based on the fuzzy AHP

method. *Journal of Xi'xxan University of Architecture Technology*, 12(7), 867-881.
[39] Larijani, A. & Dehghani, F. (2023). A Computationally Efficient Method for Increasing Confidentiality in Smart Electricity Networks. *Electronics*, 13(1), 1-15. https://doi.org/10.3390/electronics13010170
[40] Sadeghi, S. & Niu, C. (2024). Augmenting Human Decision-Making in K-12 Education: The Role of Artificial Intelligence in Assisting the Recruitment and Retention of Teachers of Color for Enhanced Diversity and Inclusivity. *Leadership and Policy in Schools*, 1-21. https://doi.org/10.1080/15700763.2024.2358303
[41] Abbasihafshejani, M., Manshaei, M. H. & Jadliwala, M. (2023). Detecting and Punishing Selfish Behavior during Gossiping in Algorand Blockchain. In *IEEE Virtual Conference on Communications (VCC2023)*, 49-55. https://doi.org/10.1109/VCC60689.2023.10474784
[42] Vahdatpour, M. S. & Zhang, Y. (2024). Latency-Based Motion Detection in Spiking Neural Networks. *International Journal of Cognitive and Language Sciences, 18*(3), 150-155.
[43] Karabulut, E., Gholizadeh, F. & Akhavan-Tabatabaei, R. (2022). The value of adaptive menu sizes in peer-to-peer platforms. *Transportation Research Part C: Emerging Technologies, 145*, 103948. https://doi.org/10.1016/j.trc.2022.103948
[44] Talebzadeh, M., Sodagartojgi, A., Moslemi, Z., Sedighi, S., Kazemi, B. & Akbari, F. (2024). Deep learning-based retinal abnormality detection from OCT images with limited data. *World Journal of Advanced Research and Reviews, 21*(3), 690-698. https://doi.org/10.30574/wjarr.2024.21.3.0716
[45] Soltanianfard, M. A., Abuhishmeh, K., Jalali, H. H. & Shah, S. P. (2023). Sustainable concrete made with wastewater from different stages of filtration. *Construction and Building Materials, 409*, 133894. https://doi.org/10.1016/j.conbuildmat.2023.133894
[46] Eslamdoust, S., Lee, J. H. & Bohrani, T. (2024). Enhancing team performance in the digital age: impact of technologically moderated communication in the interplay of e-leadership & trust. *International Journal of Business & Management Studies, 5*(04), 56-67. https://doi.org/10.56734/ijbms.v5n4a5
[47] Ghorashi, S. M., Azkia, M. & Mahdavi, S. M. S. (2015). Sociological Redefinition of the Concept of Neighborhood from the Residents' Viewpoint: A Phenomenological Study of Kan Neighborhood in District 5 of Tehran. *Community Development (Rural and Urban Communities), 7*(2), 221-240.
[48] Darvishinia, N. (2023). AI in Education: Cracking the Code through Challenges: A Content Analysis of one of the recent Issues of Educational Technology and Society (ET&S) Journal. *Partners Universal International Innovation Journal, 1*(4), 61-71.
[49] Mirshekari, S., Moradi, M., Jafari, H., Jafari, M. & Ensaf, M. (2024). Enhancing Predictive Accuracy in Pharmaceutical Sales through an Ensemble Kernel Gaussian Process Regression Approach. *International Journal of Computer and Information Engineering, 18*(5), 255-260.
[50] Farhang, M. & Safi-Esfahani, F. (2020). Recognizing mapreduce straggler tasks in big data infrastructures using artificial neural networks. *Journal of Grid Computing, 18*(4), 879-901. https://doi.org/10.1007/s10723-020-09514-2

Authors' contacts:

Hossein Talebzadeh
(Corresponding author)
Department of Computer Engineering, Science and Research Branch,
Islamic Azad University, shohada Hesarak blvd, Daneshgah Square, Sattari Highway, Tehran, Iran
E-mail: hossein@hossein.biz

Amirmohammad Fattahiamin
Department of Management, Faculty of Management and Economics,
Sharif University of Technology, Tehran, Iran

Mohammad Talebzadeh
Department of Civil and Environmental Engineering,
Texas A&M University, 201 Dwight Look Engineering Building
College Station, TX 77843-3136, Texas, USA

Fariba Sanaei
Department of Marketing, College of Business Administration,
University of Central Florida, 12744 Pegasus Dr, Orlando, FL 32816, USA

Parisa Khorashadi Moghaddam
Industrial Management, Tagliatela College of Engineering,
University of New Haven, 300 Boston Post Rd, West Haven, CT 06516, USA

Shervin Espahbod
Department of Management Science, Shannon School of Business,
Cape Breton University, 1250 Grand Lake Rd, Grand Lake Road, NS B1M1A2, Nova Scotia, Canada

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240316121244

Original scientific paper
Received: 2024-03-16, Accepted: 2024-05-28

Load Propagation Balancing Strategy for Wireless Sensor Networks

Idris Afzal Shah*, Mushtaq Ahmed

Abstract: In this paper, we present a simplistic approach for many-to-one energy hole issue in Wireless Sensor Network (WSN) by load factoring sensor nodes. Load factor is a measure of link count i.e. the number of nodes a recipient node receives data from. We consider a strip-like distribution of nodes where leaf nodes have a load factor of zero as they only sense the surroundings while the factor count grows steadily as the data progress towards the root of the tree (base station). We present an "early load sharing" intervention strategy where a node spawns child process (in upward direction towards sink) for load balancing. Under network saturation, the node switches to advanced mode and transmit directly to Base Station/sink (BS) with residual energy. Extensive simulations demonstrate the effectiveness of proposed strategy.

Keywords: advanced mode; load balancing; load factor; load propagation; wireless sensor networks

1 INTRODUCTION

WSNs are networks that are made up of a large number of sensors spread out across a large region in order to collect data about the region of interest. Sensors sense the surrounding, processing the acquired data as per the application, and transfer the processed field information to the centralized base station (BS) [1]. Because sensors are powered by a limited number of battery units, power management becomes a major issue. The larger the strain a sensor must bear and the higher its power consumption, the more messages it must send. In most circumstances, the sensor nodes battery units are not changeable, and the sensors are deemed throwaway after they die. The network may fail prematurely if the load imposed on these sensor nodes is not adequately balanced. To improve network efficiency, solutions must be designed that address the challenges of energy, network longevity, and scalability all at the same time. Scalability, fault tolerance, data aggregation, and energy efficiency are some of the key objectives of clustering, hence it plays a significant role. Clustering saves energy by designating energy-efficient nodes as Cluster Heads (CHs), which collect, consolidate, and transmit data from other nodes. As a CH, an optimal node based on various factors can be chosen, and when the energy of CHs falls below a threshold value [2, 3]. Additionally, the emergence of an energy hole near the sink has been a major concern. In a multi-hop network, data transmission follows a many-to-one pattern, resulting in a higher traffic burden for nodes near the sink. This is due to relaying of additional task of data aggregation and relaying which increases rapidly as the data progresses towards the sink/BS. As a result, these nodes quickly run out of energy, resulting in an energy vacuum and network failure [4, 5]. Many related works have attempted to balance energy consumption using extra nodes/advanced nodes near the sink, but none have succeeded in achieving perfectly balanced energy consumption or avoiding energy hotspots far away from the sink. The following is the summary of the main contributions of the proposed methodology:

- Load factor is an early network-wide indication of all under-utilized nodes in the WSN.
- Second, it serves as an implicit request for not only transferring data load but also the amount of data that should be redirected.
- Third, it also acts as a minimum hop count path for directing the spawned tasks.
- Finally, under network saturation when the load factor of all nodes (in upward region towards sink) has reached its threshold value, nodes switch to advanced power mode and transmit directly to sink.

The rest of the paper is organized as follows: A review of related work in is included in Section 2. The terminologies and energy model used in the proposed algorithm are discussed in Sect. 3. Sect. 4 describes the proposed work. Simulation is carried out in Sect. 5, and Sect. 6 draws a conclusion based on the findings.

2 RELATED WORK

A node deployment approach based on Gaussian node distributions was reported by Halder and Ghosal [6]. The number of sensor nodes was shown to be an effective metric in determining the lifetime of a network. They discovered that the Gaussian distribution could be effective in balancing energy. In corona-based networks, Wei Kuang Lai et al. [7] presented two deployment strategies to maximize network lifetime and achieve load balancing in corona based networks. They discovered that by adhering to a few basic assumptions, nearly balanced load may be obtained. Liu et al in [8] have proposed a model based on super links for off-loading data from overcrowded places to less crowded locations. Determining the optimal position of such nodes is done based on traffic and distance from the sink. Khan et al. [9] have proposed a robust model that takes routing decisions on a hop basis for realizing load balancing. Using a slice-based model, Liu et al. [10] came up with the idea of balancing energy usage in all nodes in the network. The authors used a mixed transmission technique that was divided into two halves (i.e. intra-slice and inter-slice). Their proposed solution could provide certain benefits in terms of load balancing, delay reduction, and lifetime extension. Arya et al. [11] established the necessary circumstances for bandwidth usage that allow for a routing path as well as

connection bandwidth evaluation and energy optimization using the ant colony optimization technique. Khabiri et al. in [12] proposed a clustering routing strategy based on the cuckoo algorithm for optimizing CH (cluster head) selection. Significant improvement in network lifetime has been achieved. Li et al. [13] have proposed a method that is based on high energy efficiency called EBCNC (Enhanced-Balanced Compressed Network Coding) in which the amount of data collected is greatly reduced using the data compression technique, while the efficiency of data collection is maintained using the transmission mechanism. Mahdi et al. [14] suggested a routing method for balancing energy in multi-hop networks. Their scheme could boost energy efficiency by improving the quality of the link between the source and the destination. Lipare et al. [15] suggested the idea of energy-efficient routing using one-to-one connection method. With the goal of preventing energy holes, easing heavy traffic, and balancing network load, the multiple layer method was utilized in the routing structure. Smaragdakis et al. [16] have proposed SEP (stable election protocol) to extend the time interval before the first node dies (called as stability period), which is critical for many applications where the sensor network's feedback must be trustworthy. SEP is based on the weighted election probabilities of each node to become cluster head depending on its remaining energy. Qing et al. [17] have proposed distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks (DEEC) wherein cluster-heads are elected by a probability based on the ratio between residual energy of each node and the average energy of the network. Shah et al. [18] have proposed a novel fuzzy logic clustering routing protocol that handles the rotation of cluster heads (CHs) using fuzzy logic with metrics as residual energy, base station distance and neighbour nodes in a cluster. By utilizing the softmax function for CH election and introducing novel procedures based on graph theory concepts, Osamy et al. [19] have propounded a clustering technique for IoT based wireless sensor networks the algorithm that achieves efficient resource allocation and management. Authors in [20] proposed a routing protocol energy-saving called Bacterial foraging optimization routing protocol (BFORP). It decreases the routing of excessive messages that may result in severe energy waste by recycling the information that frequents the source node into the sink. The preferable node in the sending routes is chosen by prioritizing the lowest traffic load, the highest residual energy, and the shortest path to the sink.

3 ENERGY MODEL

We adopt the radio model as given in [21]. We assume that the radio hardware loses energy in two main ways: first, the transmitter, which powers the power amplifier and radio electronics, and second, the receiver, which also powers the radio electronics. The power loss model for the free space channel and the multipath fading model for the receiver were utilised in the tests detailed here, with the choice made based on the distance between the two devices [22]. To counteract this loss, power control can adjust the power amplifier such that it uses the free space (*fs*) model if the distance is below a certain threshold or the multipath (*mp*) model otherwise. Therefore, in order to send a message that is l bits long over distance d, energy dissipated is given by Eq. (1) and (2).

$$E_{T_x}(l,d) = E_{T_x} - elec(l) + E_{T_x} - mp(ld) = \\ = \begin{cases} l \cdot E_{elec} + l \cdot \varepsilon_{fs} \cdot d^2 \text{ if } d < d_0 \\ l \ldots E_{elec}^{T_x} + l \cdot \varepsilon_{mp} \cdot d^4 \text{ if } d \geq d_0 \end{cases} \quad (1)$$

To receive message, radio expends

$$E_{R_x}(l) = E_{R_x} - elec(l) = l \cdot E_{elec} \quad (2)$$

The electronics energy, E_{elec}, depends on factors such as the digital coding, modulation, filtering, and spreading of the signal, whereas the amplifier energy, $\varepsilon_{fs} \cdot d^2$ or $\varepsilon_{mp} \cdot d^4$ depends on the distance to the receiver and the acceptable bit-error rate.

We make following assumptions:

- Sensors are randomly and unevenly deployed in a region as shown in Fig. 1.
- Sensors are homogeneous in nature.
- BS has unlimited battery supply.
- Sensors can directly reach BS at higher power cost.
- Every sensor and BS have knowledge about the topology of the network.

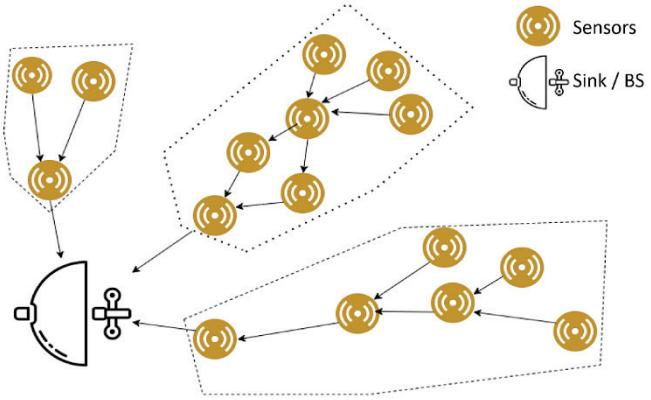


Figure 1 Network model

4 LOAD FACTOR MODEL

The load factor model is a type of localized load balancing in which each sensor only interacts with its neighbours towards the sink direction. Propagation is used to achieve global balancing and the subsequent refinement of local load data. The load balancing is achieved by spawning workload from a node with high load factor towards a low load factor node. Load factor is a measure of link count i.e. the number of nodes a recipient node receives data from.

4.1 Load Propagation

A two-tiered load balancing mechanism is used. The first step is to allow each sensor to determine it load factor. A sensor load status can be defined using fuzzy values light,

average, or heavy over time depending on their load factor. In common parlance, if a sensor is light, it wishes to be loaded more. It wants to get rid of some of its current burden if it is heavy. It is average if neither of these conditions are met.

Definition: The length of the shortest path between two sensors *m* and n in a wireless sensor network is referred to as *d* and the maximum distance between any two nodes of the network *N* is called its range.

$$range(N) = \max\{d_{m,n}\ for\ all\ m,n\ \varepsilon\ N\}$$

***Definition*:** If the node is lightly loaded, then its state is open. It is otherwise closed.

$$\begin{array}{ll} \textbf{\textit{if}}\ Sensor = Light & \textbf{\textit{THEN}} \\ STATE = 1(OPEN) & \textbf{\textit{ELSE}} \\ STATE = w_{\max}\ (\text{OFF}) & \textbf{\textit{WHERE}} \\ w_{\max} = range(N) + 1 & \end{array}$$

A sensor with light load factor allows the influx of fresh workloads. The second stage is to create a system-wide load factor propagation to make task migrations easier. Load factor propagation is represented by set of all *affinities.*

Definition: The least distance between a sensor *i* and a light loaded node in the system (in upstream region towards sink) is the defined as *affinity* w_i. If a system has no light nodes, w_i is equal to w_{max}. This implies that

$$w_i = \min\ \{d_{m,n}\ over\ k\ where\ STATE_k = 1\}$$
$$\textbf{\textit{if}}\ \ a\ light\ node\ exists\ towards\ sink$$
$$\textbf{\textit{else}}$$
$$w_i = w_{max}\ \ if\ no\ light\ node\ exists\ in\ the\ network$$

The affinity of a lightly loaded node is zero. The affinity of average or heavy sensor node is computed by adding one to the minimum affinity of neighbors. A system with no light nodes can be compared to one having a light node at a distance greater than the network's range.

Definition: A network's load propagation (LP) is the collection of all sensors affinities defined as:

$$LP = [w_1, w_2\ \ldots\ldots w_n]$$

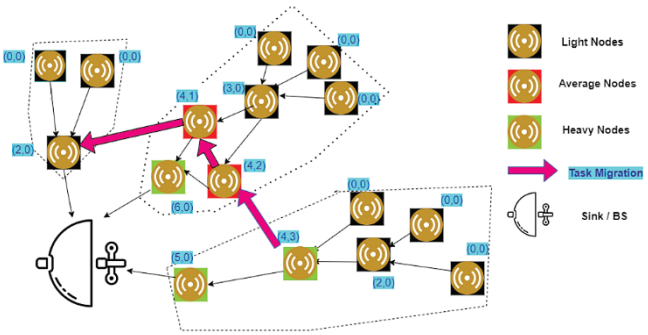


Figure 2 Load propagation model

Fig. 2 shows a WSN with BS at the corner along with load factor (*lf*) and the affinity (*a*) of each node as shown as (*lf* , *a*). Affinity values comprise the load propagation.

The load propagation has several characteristics. First, it is a network-wide indicator of all underutilized sensor nodes. Second, it contains an implicit work load request. Third, it directs unprocessed jobs using a minimum distance hops. However, knowing all affinities is a key requirement to formulate the load propagation model.

A load migration method will continue until one of the following requirements is met: The task arrives at the light node, or other tasks arrive at the light node and change sensor state to OFF. A new load propagation is reshaped if there is another underutilized node in the system. The task is then redirected to the new light sensor closest to it.

4.2 Saturation

There is no need to balance the load when all sensors are heavily loaded. During this time, any load balancing activities will just add to the system's overhead and impair performance.

Definition: If none of the sensors are lightly loaded the system is saturated. In other words, if all affinities are equal to w_{max}, the system is saturated.

When the network is saturated, any task movement is futile as it will further drain the nodes and cause energy hole issues. Under such scenario, nodes switch to advanced mode and transmit directly to BS with their leftover energy. When node death occurs due to any reason, its neighbors can set the affinity of the dead sensor to w_{max}, which prevents fresh jobs from being moved to the node.

4.3 Algorithm

Based on the load propagation model, a distributed and asynchronous load balancing mechanism can be designed for each node as:

$$\textbf{\textit{LOOP}}\ Sensor\ i\ determines\ its\ loading\ state$$
$$(Light, Average\ or\ Heavy)\ as\ per\ its\ load\ factor$$
$$\textbf{\textit{CASE Load State}} =$$
$$\textbf{\textit{light}}:$$
$$Set\ \ affinity\ w_i = 0$$
$$Ignore\ load\ \ information\ from\ neighbors$$
$$\textbf{\textit{Average}}:$$
$$w_i = 1 + \min\{w_j\}for\ all\ neighbors\ j$$
$$if\ w_i > w_{\max}\ then\ w_i = w_{\max}$$
$$(\textbf{\textit{Saturation}})\ Node\ i\ switches\ to\ advanced\ mode\ and$$
$$transmit\ directly\ to\ BS$$
$$\text{with}\ residual\ energy$$
$$\textbf{\textit{Heavy}}:$$
$$w_i = 1 + min\{w_j\}\ for\ all\ neighbors\ j$$
$$if\ w_i > w_{\max}\ then\ w_i = w_{\max}$$
$$(\textbf{\textit{Saturation}})\ Node\ i\ switches\ to\ advanced\ mode\ and$$
$$transmit\ directly\ to\ BS\ \text{with}\ residual\ energy$$
$$ELSE\ if\ \min\{w_j\} < w_i\ \textbf{\textit{THEN}}$$
$$Perform\ task\ migration\ to\ node\ j\ where\ w_j\ is\ minimum$$
$$\textbf{\textit{END CASE}}$$
$$Broadcast\ w_i$$
$$to\ all\ neighbors\ if\ modified\ since\ last\ update$$
$$\textbf{\textit{END LOOP}}$$

5 PERFORMANCE EVALUATION

The simulations are performed for a network scenario with 100 sensor nodes randomly distributed over 100 × 100 m^2 area. Comparative analysis of the proposed model is carried out with SEP, DEEC, Fuzzy model as proposed in [18] protocols considering uneven node distribution. Residual energy threshold are set as ≤ 30 % (light), ≤ 50 % (average) and up to 70 % (heavy) of the initial energy. Detailed parameters of simulation are listed in Tab. 1.

Table 1 Simulation parameters

Parameter	Value
Area Size	100 × 100 m^2
No of sensor nodes	100
Initial Energy	1J
E_{elec}	50 nJ/bit
Data Packet Size	400 bytes
ε_{fs}	10 pJ/bit/m^2
ε_{mp}	0.0013 pJ/bit/m^4

The performance of all algorithms has been evaluated by considering following metrics.

5.1 Number of Alive Nodes

Fig. 3 shows the comparison of DEEC, SEP, Fuzzy and proposed algorithm in terms of number of alive nodes. Proposed model performs better as the rounds increase than its counterparts. Fuzzy version initially performs better but as the network progresses it can be seen that proposed algorithm has a better network longevity. This shows that task migration has a positive impact on node balancing load, and that some of the data traffic owned by heavy-burden nodes is shifted to nodes with lower data traffic.

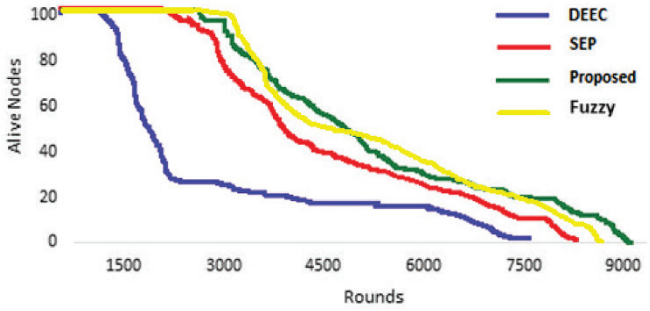


Figure 3 Number of alive nodes

5.2 Number of Dead Nodes

The proposed model has a better dead node ratio as compared to DEEC, SEP as shown in Fig. 4 owing to the fact that it's an early intervention strategy that spawns task migration as the load reaches the threshold value thereby relieving the heavily loaded nodes and increasing its lifetime. In comparison to fuzzy version, both algorithms perform good in patches with proposed slightly having a better node death ratio in the end.

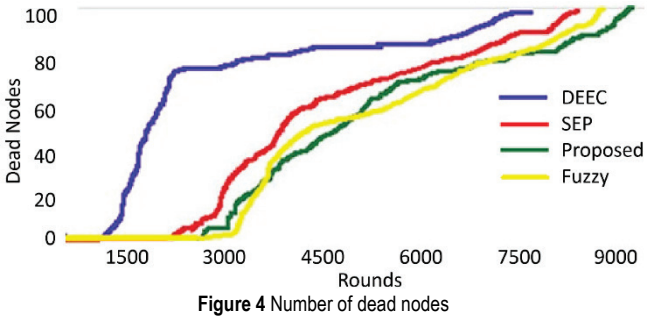


Figure 4 Number of dead nodes

5.3 Throughput

Fig. 5 compares these algorithms in terms of throughput, that is the total number of packets delivered to sink. The proposed model outperforms the other two algorithms (DEEC and SEP) as it's a active strategy rather than a late remedy approach and also owing to the fact that each node transmits directly to BS with leftover residual energy rather than transmitting on a hop by hop basis wherein packet loss due to energy hole issues in upstream region would take place. Comparing to fuzzy version, it performs better towards the end.

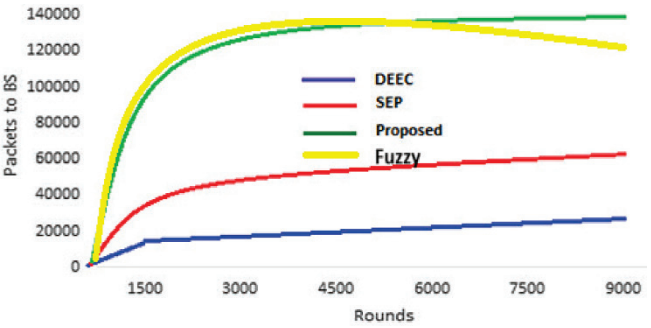


Figure 5 Throughput

5.4 First Node Death (FND), Half Node Death (HND) and Last Node Death (LND)

These statistics calculate the number of rounds until the network's first, half and last node dies. FND, HND is an indicator of overall network stability. LND is a measure of network lifetime. Fig. 6 draws the analysis. It can be seen that the suggested algorithm performs better overall.

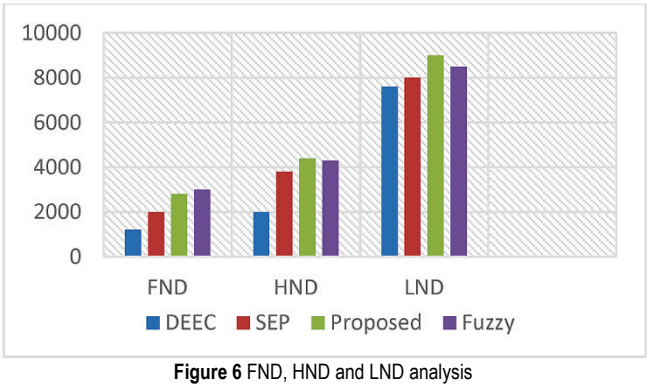


Figure 6 FND, HND and LND analysis

Finally, we summarize the advantages and disadvantages of our work as under:

- This early intervention strategy triggers task migration when the load reaches a threshold value, alleviating heavily loaded nodes and extending their lifespan.
- It is an active strategy instead of a late remedy approach, and because each node transmits directly to BS with leftover residual energy instead of transmitting hop by hop, this mitigates packet loss due to energy hole issues that may occur in the upstream region.
- The proposed model will suffer overhead when it comes to mobile wireless sensor networks wherein constantly changing network topology will affect the process of finding the best route and thereby the task migration scheduling.

6 CONCLUSION

In this paper, distributed load balancing system is proposed. The methodology is built on a demand-driven principle, in which underutilized nodes dynamically initiate load balancing requests. As a result of these requests, a system-wide load propagation model is created. Task migration takes place from overloaded sensors towards lightly loaded ones.

A global balance state is computed by approximating multiple localized balances in a series of steps. When the system is completely occupied, the idea of saturation is introduced to discourage pointless load migration whereby the node switched to advanced mode and transmit directly to BS with leftover residual energy. It is a proactive strategy which is more pragmatic in large scale WSNs. The efficacy of our model is mirrored through metrics viz. dead nodes, alive nodes and throughput.

7 REFERENCES

[1] Chaurasiya, S. K., Biswas, A., Bandyopadhyay, P. K., Banerjee, A. & Banerjee, R. (2022). Metaheuristic Load-Balancing-Based Clustering Technique in Wireless Sensor Networks. *Wireless Communications and Mobile Computing*. https://doi.org/10.1155/2022/8911651

[2] Madhu, S., Prasad, R. K., Ramotra, P., Edla, D. R. & Lipare, A. (2022). A Location-less Energy Efficient Algorithm for Load Balanced Clustering in Wireless Sensor Networks. *Wireless Personal Communications, 122*(2), 1967-1985. https://doi.org/10.1007/s11277-021-08976-1

[3] Kaur, S., Mir, R. N., Khamparia, A., Rani, P., Gupta, D. & Khanna, A. (2021). Heterogeneous load balancing clustering protocol for Wireless Sensor Networks. *Cognitive Systems Research, 70*, 10-17. https://doi.org/10.1016/j.cogsys.2021.07.001

[4] Behera, T. M. & Mohapatra, S. K. (2021). A novel scheme for mitigation of energy hole problem in wireless sensor network for military application. *International Journal of Communication Systems, 34*(11), e4886. https://doi.org/10.1002/dac.4886

[5] Asadollahi, H., Zandi, S. & Asharioun, H. (2021). Maximizing Network Lifetime in Many-to-One Wireless Sensor Networks (WSNs). *Wireless Personal Communications*, 1-13. https://doi.org/10.1007/s11277-021-09271-9

[6] Halder, S., Ghosal, A. & Bit, S. D. (2011). A pre-determined node deployment strategy to prolong network lifetime in wireless sensor network. *Computer Communications, 34*(11), 1294-1306. https://doi.org/10.1016/j.comcom.2011.01.004

[7] Lai, W. K. & Fan, C. S. (2017). Novel node deployment strategies in corona structure for wireless sensor networks. *IEEE Access, 5*, 3889-3899. https://doi.org/10.1109/ACCESS.2017.2681124

[8] Liu, X. & Zhang, P. (2017). Data drainage: A novel load balancing strategy for wireless sensor networks. *IEEE Communications Letters, 22*(1), 125-128. https://doi.org/10.1109/LCOMM.2017.2751601

[9] Khan, N. A., Saghar, K., Ahmad, R. & Kiani, A. K. (2016). Achieving energy efficiency through load balancing: A comparison through formal verification of two WSN routing protocols. In *The 13th IEEE International Bhurban Conference on Applied Sciences and Technology (IBCAST2016)*, 350-354. https://doi.org/10.1109/IBCAST.2016.7429901

[10] Liu, T., Gu, T., Jin, N. & Zhu, Y. (2017). A mixed transmission strategy to achieve energy balancing in wireless sensor networks. *IEEE Transactions on wireless communications, 16*(4), 2111-2122. https://doi.org/10.1109/TWC.2016.2642098

[11] Arya, R. & Sharma, S. C. (2018). Energy optimization of energy aware routing protocol and bandwidth assessment for wireless sensor network. *International Journal of System Assurance Engineering and Management, 9*(3), 612-619. https://doi.org/10.1007/s13198-014-0289-3

[12] Khabiri, M. & Ghaffari, A. (2018). Energy-aware clustering-based routing in wireless sensor networks using cuckoo optimization algorithm. Wireless Personal Communications, 98(3), 2473-2495. https://doi.org/10.1007/s11277-017-4983-8

[13] Li, Y., Yang, Z. & Zhang, Q. (2016). Efficient load balance data aggregation methods for WSN based on compressive network coding. *In IEEE International Conference on Electronic Information and Communication Technology (ICEICT2016)*, 111-115. https://doi.org/10.1109/ICEICT.2016.7879663

[14] Mahdi, O. A., Al-Mayouf, Y. B., Ghazi, A. B., Wahab, A. A. & Idris, M. Y. I. B. (2018). An energy-aware and load-balancing routing scheme for wireless sensor networks. *Indonesian Journal of Electrical Engineering and Computer Science, 12*(3), 1312-1319. https://doi.org/10.11591/ijeecs.v12.i3.pp1312-1319

[15] Lipare, A., Edla, D. R. & Dharavath, R. (2020). Energy efficient routing structure to avoid energy hole problem in multi-layer network model. *Wireless Personal Communications, 112*(4), 2575-2596. https://doi.org/10.1007/s11277-020-07165-w

[16] Smaragdakis, G., Matta, I. & Bestavros, A. (2004). SEP: A stable election protocol for clustered heterogeneous wireless sensor networks. In *Second international workshop on sensor and actor network protocols and applications (SANPA 2004), Vol. 3*.

[17] Qing, L., Zhu, Q. & Wang, M. (2006). Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks. *Computer communications, 29*(12), 2230-2237. https://doi.org/10.1016/j.comcom.2006.02.017

[18] Shah, I. A. & Ahmed, M. (2023). FK-means RA: Fuzzy K-Means Clustering Routing Algorithm for Load Balancing in Wireless Sensor Networks. *Wireless Personal Communications*, *130*(2), 1071-1083. https://doi.org/10.1007/s11277-023-10320-8

[19] Osamy, W., Alwasel, B., Salim, A., Khedr, A. M. & Aziz, A. (2024). LBAS: Load Balancing Aware Clustering Scheme for

IoT-based Heterogeneous Wireless Sensor Networks. *IEEE Sensors Journal*. https://doi.org/10.1109/JSEN.2024.3381852
[20] Abdulzahra, A. A., Khudor, I. B. A. Q. & Alshawi, I. S. (2023). Energy-efficient routing protocol in wireless sensor networks based on bacterial foraging optimization. *Indonesian Journal of Electrical Engineering and Computer Science*, *29*(2), 911-920. https://doi.org/10.11591/ijeecs.v29.i2.pp911-920
[21] Heinzelman, W. B., Chandrakasan, A. P. & Balakrishnan, H. (2002). An application-specific protocol architecture for wireless microsensor networks. *IEEE Transactions on wireless communications*, *1*(4), 660-670. https://doi.org/10.1109/TWC.2002.804190
[22] Rong, Z. & Rappaport, T. S. (1996). *Wireless communications: Principles and practice, solutions manual*. Prentice Hall.

Authors' contacts:

Idris Afzal Shah
(Corresponding author)
Department of Computer Science and Engineering,
Malaviya National Institute of Technology,
JLN Marg, Jaipur, Rajasthan 302017, India
E-mail: 2019RCP9183@mnit.ac.in

Mushtaq Ahmed, Dr.
Department of Computer Science and Engineering,
Malaviya National Institute of Technology,
JLN Marg, Jaipur, Rajasthan 302017, India
E-mail: mahmed.cse@mnit.ac.in

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240331030139

Original scientific paper
Received: 2024-03-31, Accepted: 2024-07-08

The Impact of Social Media Marketing on Digital Service Adoption in Educational Institutions: Exploring the Mediating Role of Brand Equity, Trust, and Word-Of-Mouth Advertising

Hosna Khorsandi*, Behzad Kazemi, Simin Zeynali, Mahsa Mohsenibeigzadeh, Pedram Zarei, Shahin Mirshekari

Abstract: Recently, educational institutions have turned to investing in new technologies to provide digital services to customers as a means of cost control, attracting new customers, and meeting customer expectations. The adoption of these new technologies has become crucial for these institutions as part of their strategy. Therefore, this research focuses on investigating the influence of social media marketing (SMM) on the intention to use digital services in educational institutions. The study also considers brand equity, trust, and word-of-mouth promotion as potential factors mediating this relationship. The method employed was descriptive correlational research, utilizing structural equation modeling for data analysis, with 368 students participating. The results indicate that SMM has a significant influence on brand equity, trust, word-of-mouth advertising, and the intention to utilize digital services in educational institutions. Furthermore, brand equity, trust, and word-of-mouth advertising exert a positive and significant impact on the intention to use digital services within educational institutions. Additionally, brand equity, trust, and word-of-mouth advertising act as vital mediators in the relationship between SMM and the adoption of digital services in educational institutions. Consequently, we can conclude that SMM contributes to an increased intention to use digital services in educational institutions by leveraging brand equity, trust, and word-of-mouth advertising.

Keywords: brand equity; educational institutions; intention to use digital services; social media marketing; trust; word of mouth advertising

1 INTRODUCTION

The rapid advancement and widespread adoption of communication technology have sparked a remarkable transformation across multiple aspects of human lives and organizational performance [1]. This technological revolution has revolutionized the behaviors and outlooks of individuals, organizations, and governments, while simultaneously giving rise to new industries, employment opportunities, and innovative ventures [2, 3]. The advent of digital educational services stands as a significant outcome stemming from the pervasive reach and growth of information technology within the educational realm. In light of this matter, educational institutions, much like many other service providers, have begun allocating resources towards incorporating new technologies into their operations in order to offer digital services to their clientele. This approach serves multiple purposes: cost management, customer acquisition, and meeting customer expectations. Consequently, the integration of these innovative technologies has emerged as a strategic imperative in the agenda of educational institutions. In today's competitive landscape, there is a growing emphasis on researching the features offered by educational institutions that affect individuals' intentions to use digital services. As the number of institutions increases, it is crucial to identify effective strategies that can attract more customers and increase sales [4]. To achieve this, it is essential to understand the internal and external factors that shape users' behavior towards digital services. Given the dynamic and constantly evolving nature of the digital environment, educational institutions must have a comprehensive understanding of their users to keep up with the changing trends. However, there is still a lack of acceptance among Iranian consumers towards purchasing digital educational products and services from institutions and companies. Therefore, the purpose of the present study is to investigate how social media marketing (SMM) affects individuals' intentions to use digital services offered by educational institutions, with a focus on the role of brand equity, trust, and word-of-mouth advertising in mediating actions.

2 LITERATURE REVIEW

2.1 Social Media Marketing and Usage Intention

The use of social media has revolutionized the way organizations interact with their customers on a global scale, thanks to technological advancements and increased internet usage [5]. Businesses are now leveraging social media as a means of communication with customers and develop successful branding strategies by integrating different channels. Social media has also become a platform for public discourse on various topics, from politics to entertainment [6]. As a result, marketers are increasingly relying on SMM to reach their target audiences, with studies showing its effectiveness in influencing customer intention to use services [7, 8]. It is therefore our belief that the following hypothesis should be tested:

H_1: SMM has a positive impact on individuals' intentions to use digital services offered by educational institutions.

2.2 Brand Equity, Social Media Marketing and Usage Intention

This term refers to the inherent and exceptional value of a brand, the desire for customers to pay more for the similar amount of quality through their strong attachment to the brand as well as the fact that they are highly attracted to the brand [5, 9-12]. For customers, brand equity is the essence of successful organizational activities, as it helps organizations understand and satisfy their needs and demands. The basis of brand equity is the brand power that is rooted in the minds of customers. In addition to assets, brands can have liabilities as well, including brand recognition, perceived quality, brand

associations, and other brand assets that can develop or diminish the value of goods and services. According to Aaker [13], a brand equity is an asset or liability that can grow or decline the value of goods and services.

Because of brand equity, customers perceive the brand favourably and make more purchases as a result. To enhance engagement opportunities and manage brand assets in a positive manner, it is essential that firms develop strategies that enhance and grow their brand equity in order to make sure this is achieved [14]. The effect of brand equity on the behavior of customers has also been demonstrated in studies [15]. Additionally, there is evidence that SMM is a crucial part of building brand equity, as has been demonstrated in research studies [14, 16, 17]. Therefore, the following hypotheses are therefore put forward as a result of this study:

H_2: SMM influences positively brand equity.

H_5: Brand equity influences positively individuals' intentions to use digital services offered by educational institutions.

H_8: The effect of SMM on individuals' intentions to use educational institutions' digital services is mediated by brand equity.

2.3 Trust, Social Media Marketing and Usage Intention

Throughout the history of business, trust has been a fundamental concept for transactions and exchanges. The trust that a customer has in a brand refers to their level of confidence in a brand's ability to perform the tasks that have been assigned to them [18]. Morgan and Hunt [19] argue that trust is achieved when one party is confident in the correctness of the other party. Berry [20] asserts that relational marketing relies on the principle of trust. Trust is also a crucial factor in creating and improving the quality of relationships as a result of the process of making and keeping commitments and promises. When trust is established, other parties will feel that they can be trusted and are reliable, and this will result in solid, honest, fair, and productive cooperation between them. A brand that is trustworthy is more likely to secure the loyalty of customers after the product has encountered unexpected problems, making it more likely that the product or service will be developed, sold, and promoted in the future [21, 22]. As well as trust, research has also shown that customer behavior intentions are influenced by trust [8, 23, 24]. Additionally, research has demonstrated the role of SMM in building brand trust [8]. The following hypotheses are therefore proposed as a result:

H_3: Trust in a brand is positively affected by SMM.

H_6: Trust has a positive impact on individuals' intentions to use digital services offered by educational institutions.

H_9: Brand trust mediates the effect of SMM on individuals' intentions to use digital services offered by educational institutions.

2.4 Word-of-Mouth Adverting, Social Media Marketing and Usage Intention

Word-of-mouth marketing refers to creating conditions that encourage people to talk about a product or service and facilitating these conversations [25]. Word-of-mouth marketing involves satisfying customers so that they become the best advertisers for the firm. It is about real consumers and why they want to talk about the firm and its products. Word-of-mouth marketing takes advantage of people's natural tendency to talk. Word-of-mouth advertising is the most powerful source of marketing. Despite the large amount of information presented by competitors through various marketing tools, advertisements, and salespeople, customers and potential customers will engage in word-of-mouth advertising by talking to each other and helping each other make decisions. Word-of-mouth advertising is beyond the control of marketers, but it is cheaper than other methods, making it important to understand how it works. Word-of-mouth advertising is widely used in the field of internet and social media [26-28]. Studies have shown that word-of-mouth advertising is effective in influencing customer behavioral intentions [29, 30]. Additionally, research has demonstrated the important role of SMM in word-of-mouth advertising [26-28]. The following hypotheses are therefore proposed as a result:

H_4: SMM has a positive impact on word-of-mouth advertising.

H_7: Word-of-mouth advertising has a positive impact on individuals' intentions to use digital services offered by educational institutions.

H_{10}: Word-of-mouth advertising mediates the effect of SMM on individuals' intentions to use digital services offered by educational institutions.

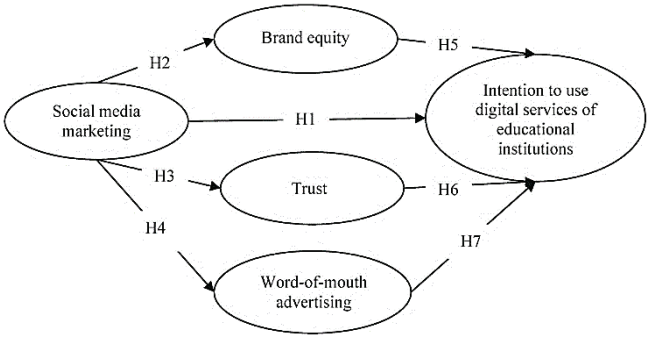


Figure 1 The conceptual model

There is a growing body of evidence in the theoretical literature that emphasizes the importance of SMM in the areas of brand equity, trust, word-of-mouth advertising, and intention to use services. Yet, few studies have developed a model for the effect of SMM on individuals' intentions to use digital services offered by educational institutions that stresses the role of brand equity, trust, and word-of-mouth advertising in mediating the effect of SMM on individuals' intentions to use digital services. Therefore, the main objective of this study is to show how SMM contributes to individuals' intentions to use digital services. This study uses brand equity, trust, and word-of-mouth advertising in educational institutions as a means of meditating the impact of SMM on individuals' intentions to use digital services. An overview of the conceptual model for the study can be found in Fig. 1 that has been developed based on the theoretical

literature and framework that have been derived from the literature.

3 METHODOLOGY AND RESEARCH METHODS

In this research, structural equation modeling (SEM) with partial least squares (PLS) is chosen to investigate the relationships between variables through structural equations.

3.1 Population and Sample

The methodology employed in this study involved selecting a population of Iranian students as the target group. A simple random sampling technique was utilized to choose participants who were available and willing to take part in the research. In order to ensure a representative sample, 450 questionnaires were distributed among students in various educational institutions, out of which 368 questionnaires were returned. Using Cochran's formula for determining sample size, we were able to determine the sample size in the present study based on previous research studies, which indicated that a sample size of 450 would be adequate to represent students who were utilizing digital services from educational institutions.

3.2 Instruments

It was determined that SMM measures could be measured by using a questionnaire that was developed by Seo and Park [31]. As part of the questionnaire, 11 items were included, which assessed entertainment, interaction, trendiness, customization, and perceived risk. The entertainment items were divided into two categories, and interaction was broken down into three categories. In order to determine the brand equity of a brand, Seo and Park [31] developed a questionnaire which consisted of 6 items, of which three items assessed the brand's awareness of the brand and three items assessed the brand's image. In order to measure trust, Shankar et al. [25] developed a questionnaire containing six items that was administered to all participants in the study. Besides this, a questionnaire developed by Kim and Ko [32] was used to measure word-of-mouth advertising, which included three items in the survey. There was a five-point Likert scale used to rate all of the items on a scale of 1 to 5, with 1 being completely disagreed and 5 being completely agreed with.

4 RESULTS

4.1 Testing of Measurement Model

Cronbach's alpha coefficient and composite reliability were used in order to assess the reliability of the measurement model, while factor loadings, average variance extracted and the Fornell-Larker test were used in order to assess the validity of the measurement model. The composite reliability index, proposed by AmirKhani and Borhani [33], was found to be more effective than Cronbach's alpha, as it does not assume that the observable variables of each measurement model have the same weights. Instead, composite reliability uses the factor loadings of the items when calculating, resulting in more accurate and better Cronbach's alpha ratios. The composite reliability index for the internal consistency of the measurement model was evaluated according to a criterion of 0.7 or higher. The confirmation factor analysis shows that the construct is well-defined when the factor loading of each item is 0.6 or higher in the confirmatory factor analysis [34, 35]. The factor loadings for the items of the variables in Table 1 were all above 0.6, confirming the coefficients of the factor loadings. If the factor loadings between the construct and its indicators are less than 0.6, those indicators should be modified or removed from the model. According to [36], whether or not the construct explains about 50% or more of the variance in its markers was tested by analyzing the average variance extracted (AVE). With an AVE value of 0.5 or higher, the construct is considered to be convergent, as it can explain about half or more of the variance [37]. A more detailed analysis of the design and reliability of the constructs can be found in Tab. 1, which demonstrates that the factor loadings, composite reliability, and AVE of the variables are adequate and appropriate.

Table 1 Factor loadings, composite reliability and AVE of variables

Variable	Item	Factor loading	Alpha	CR	AVE
Entertainment	1	0.852	0.69	0.832	0.711
	2	0.833			
Interaction	1	0.860	0.809	0.887	0.723
	2	0.843			
	3	0.847			
Trendiness	1	0.922	0.840	0.926	0.862
	2	0.935			
Customization	1	0.841	0.704	0.834	0.715
	2	0.817			
Perceived risk	1	0.875	0.698	0.869	0.768
	2	0.877			
Brand awareness	1	0.842	0.750	0.857	0.667
	2	0.820			
	3	0.787			
Brand image	1	0.726	0.681	0.825	0.612
	2	0.811			
	3	0.807			
Brand trust	1	0.862	0.908	0.929	0.688
	2	0.812			
	3	0.901			
	4	0.791			
	5	0.733			
	6	0.865			
Word of mouth	1	0.867	0.780	0.870	0.691
	2	0.797			
	3	0.829			
Intention to use digital services of educational institutions	1	0.728	0.869	0.90	0.601
	2	0.735			
	3	0.763			
	4	0.805			
	5	0.800			
	6	0.815			

We used the Fornell-Larker index as an indicator of discriminant validity in order to assess the constructs used in this study. The AVE of a construct must have a square root greater than the correlation between that construct and other constructs in order to meet the requirements for this index.

There is a higher correlation between the construct and its indicators in this case, highlighting the fact that it is more highly correlated with the construct than with other constructs. It is for this reason that Tab. 2 presents the results that relate to correlation and square root of the AVE, which is the second validity criterion, as well as their correlation. The correlation matrix also includes values below the diagonal to assess the relationships between variables, revealing that all variables have a positive and significant correlation coefficient.

Table 2 Correlation and square root of AVE of variables

Variable	SMM	Brand equity	Trust	Word of mouth	Usage intention
SMM	0.76				
Brand equity	0.68**	0.88			
Trust	0.63***	0.57**	0.83		
Word of mouth	0.41**	0.34**	0.42**	0.83	
Usage intention	0.60**	0.59**	0.59**	0.46**	0.77

*$p < 0.05$; **$p < 0.01$

4.2 Structural Model Testing

To forecast the intention to use digital services offered by educational institutions, a SEM based PLS approach was used to evaluate the proposed conceptual model. With the bootstrap method, 500 sub-samples were used to calculate the t-values for the path coefficients in order to determine the significance of the coefficients. The relationship between the variables can be illustrated in Fig. 2 by the model that has been tested. In accordance with the figure below, SMM has a significant and positive impact on brand equity, trust, word-of-mouth advertising, as well as on intention for students to make use of a digital service that is provided by educational institutions. Additionally, a positive and significant association is also found between brand equity, trust, and word-of-mouth advertising and the intention of students to use the digital services of educational institutions in the future. It is worth noting that the numbers inside the circles represent the variance that has been explained by the variables.

Table 3 Path coefficients and explained variance

Variable	β	t-value	p-value	Explained variance
on intention to use digital services of educational institutions via:				
SMM	0.195**	3.712	0.001	0.541
Brand equity	0.264**	5.545	0.001	
Brand trust	0.252**	4.513	0.001	
Word of mouth advertising	0.204**	4.641	0.001	
On brand equity via:				
SMM	0.697**	21.634	0.001	0.486
On brand trust via:				
SMM	0.64**	18.380	0.001	0.410
On word-of-mouth advertising via:				
SMM	0.42**	8.061	0.001	0.177

*$p < 0.05$; **$p < 0.01$

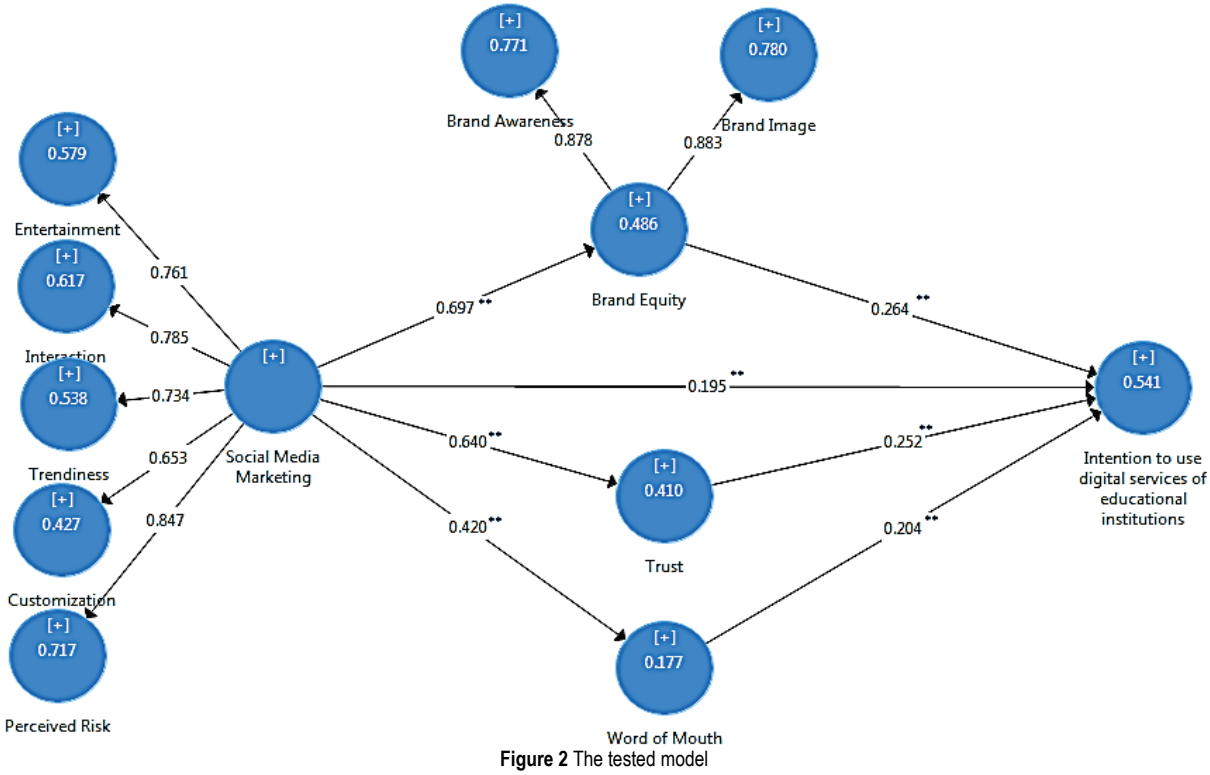


Figure 2 The tested model

The results of Tab. 3 suggest that SMM effectively improves brand equity, trust, word-of-mouth advertising, and the intention to use the digital services of educational institutions, as well as a positive and significant effect on the use of SMM. Moreover, the intention to use digital services of educational institutions being positively influenced by

brand, trust, and word-of-mouth advertising can also be observed to have a significant and positive impact on the intention to use digital services. There has been a substantial increase in the use of digital services by educational institutions, with 54% of the variance explained by models variables, 49% by those related to brand equity, 41% by those related to brand trust, and 18% by those related to word-of-mouth advertising. There are two types of indirect coefficients presented in Tab. 4.

Table 4 Indirect coefficients

Indirect paths	Indirect effects	*t*-value	*p*-value
SMM → Brand Equity → Intention to use digital services of educational institutions	0.184	5.260	0.000
SMM → Trust → Intention to use digital services of educational institutions	0.161	4.346	0.000
SMM → Word of Mouth → Intention to use digital services of educational institutions	0.086	3.937	0.000

As shown in the Tab. 4, the extent to which SMM has a positive effect on the intention of students in educational institutions to use digital services is mediated by brand equity, brand trust, and word-of-mouth marketing, all of which play an important role in mediating the positive effect. There are several factors that play a significant role in the decision to choose digital services in educational establishments, but brand equity, brand trust, and word-of-mouth advertising play a crucial role. In this study, the GOF index value of 0.60 suggests that the tested model has an adequate fit. Generally, values greater than 0.36 are considered acceptable and indicative of a good quality model.

5 DISCUSSION

In this study, it was examined what effect SMM had on the inclination to use digital services provided by educational institutions in terms of their inclination to use social media. Furthermore, it was investigated whether SEM could be used as a means of evaluating brand equity, trust, and word-of-mouth advertising in order to help to find the mediating effects of these factors. Based on the findings of the study, the proposed model was found to be a reasonably good fit to the data and could explain 54% of the variation in the intention of using digital services of educational institutions, 49% of the variation in brand equity, 41% of the variation in brand trust, and 18% of the variation in word-of-mouth advertising for educational institutions.

There was a significant and positive impact of SMM on the brand equity of the educational institution, trust, word-of-mouth advertising, and intention of using the digital services of the institution. This suggests that SMM can enhance brand equity, trust, word-of-mouth advertising, and the intention to use digital services of educational institutions. It is worth noting that this finding is in line with previous findings by Hanaysha [8], Garanti and Kissi [16], Zollo et al. [17], Darvishinia [27]. Therefore, the use of social media can increase brand equity, trust, word-of-mouth advertising, and the intention to use digital services of educational institutions if the social media usage is engaging for users, the content of social media is interesting, social media enables information sharing and exchange of opinions, users can easily express their opinions. The content of social media is frequently updated, social media allows users to customize their experience, social media allows users to share information about brands, products, and services with their friends, as well as they can share the content of institutional social media on their websites and blogs. As a result, the communication of people, groups, and products, as well as the sharing of information through social media and information sharing, has the potential to lead to brand equity, trust, word-of-mouth advertising, and an increase in the willingness to use educational institutions' digital offerings. The use of social media by educational institutions as a means of sharing information and knowledge transfer about their products and services, exchanging and transmitting information using social media, and enabling information searching through social media will therefore increase the likelihood that users of educational institutions' digital services will take advantage of their services in the future [39].

According to this model, the brand equity of educational institutions also influences positively the intention of students to take advantage of digital services provided by them. This implies that an increase in brand equity can lead to a higher intention to use digital services of educational institutions. This finding is consistent with previous studies conducted by Majeed et al. [38]. Therefore, if users of digital services of educational institutions are constantly aware of the institution, associate its characteristics quickly in their minds, remember its symbols or logo, perceive it as an experienced and customer-oriented institution, their behavior can be influenced, leading to a higher tendency to use digital services of educational institutions.

In addition, the model showed that brand trust has a significant influence on the likelihood that students will use the digital services offered by educational institutions in the future. This suggests that an increase in brand trust can lead to a higher intention to use digital services of educational institutions. There is significant agreement between the findings of this study and those of previous studies conducted by Kim et al. [23], Hanaysha [8], Ha and Nguyen [24]. Therefore, if the educational institution fulfills its obligations towards users of digital services, meets their expectations, considers their interests, and is perceived as fair and honest, users of digital services of the institution will trust its credibility, leading to a higher intention to use its digital services.

Because of the analysis, it has been found that word-of-mouth advertising significantly affects the intention of students to use digital services of educational institutions in a positive and significant manner. Consequently, it seems that an increase in word-of-mouth advertising may be associated with an increase in the intention to use digital

services offered by educational institutions. Therefore, if most users of digital services of the educational institution report positive recommendations for using its services, customers have mostly positive experiences, and users express and share their views about the institution on websites and social networks, the intention to use digital services of the educational institution will increase.

6 MANAGERIAL REMARKS

As an educational institution, it is recommended to take advantage of social media platforms as a potent tool for utilizing the power of SMM. This can be done by sharing information about your services and products, facilitating information exchange and transmission, keeping customers informed with up-to-date information about your services and products, introducing your services and products through social media, and enabling customers to search for information about your services and products on social media.

To capitalize on the role of brand equity, educational institutions are recommended to view brand equity as a crucial tool when it comes to reach out to new customers and retain existing ones. Brand equity fosters a positive attitude and a general positive effect, leading to an increase in the intention to use digital services of educational institutions. As brand equity is the foundation of product differentiation and brand recognition, it offers a convincing incentive to buy and generates favorable emotions towards the brand, thereby aiding in the process of intending to use the service.

To leverage the role of trust, educational institutions are advised to build customer trust by providing high-quality services that meet customer expectations, demonstrating commitment and humility in their interactions with customers, paying close attention to customer needs, responding promptly to customer inquiries and concerns, and guaranteeing their services.

To capitalize on the power of word-of-mouth advertising, educational institutions are recommended to provide high-quality services and encourage customers to share positive online recommendations for using their digital services. They should strive to generate mostly positive experiences, enabling customers to provide positive online recommendations for purchasing digital services from the institution.

7 CONCLUSION

The study concluded that the use of SMM, brand equity, word-of-mouth advertising, and brand trust are factors that are significantly associated with the intention of an educational institution to use digital services. The relationship between SMM and the intention to use digital services in educational institutions can also be enhanced by brand equity and brand trust, as well as word-of-mouth advertising, which may play a positive and significant mediating role. This suggests that SMM can increase the intention to use digital services in educational institutions through the enhancement of brand equity, word-of-mouth advertising, and brand trust. As a result, social media interactions between users of educational digital services and educational institutions facilitate the exchange of information and values, which can influence the intention to use digital services in educational institutions. Additionally, factors such as awareness of digital services and products, interesting content, information sharing, exchanging opinions, trendiness of information, and customized information search can also impact the intention to use digital services in educational institutions. Therefore, effective communication and information sharing between users of educational digital services and educational institutions through social media can foster confidence, attachment, and enthusiasm towards the institution, ultimately leading to a greater willingness to use digital services in educational institutions. The study's generalizability is limited as it only involved a sample of Iranian students. Additionally, the findings rely on self-reported data. To further explore the effects of SMM in educational institutions, future studies could consider using mixed research methods that incorporate qualitative data such as Ghorashi et al. [40] or Fusion Models [41]. It is important to note that the current study is correlational in nature, and therefore, causal inferences cannot be made.

8 REFERENCES

[1] Gholami, M. H., Asli, M. N., Nazari-Shirkouhi, S. & Noruzy, A. (2013). Investigating the influence of knowledge management practices on organizational performance: an empirical study. *Acta Polytechnica Hungarica, 10*(2), 205-216. https://doi.org/10.12700/APH.10.02.2013.2.14

[2] Abbasihafshejani, M., Manshaei, M. H. & Jadliwala, M. (2023, November). Detecting and Punishing Selfish Behavior During Gossiping in Algorand Blockchain. In *2023 IEEE Virtual Conference on Communications (VCC)*, 49-55. https://doi.org/10.1109/VCC60689.2023.10474784

[3] Nazari-Shirkouhi, S., Badizadeh, A., Dashtpeyma, M. & Ghodsi, R. (2023). A model to improve user acceptance of e-services in healthcare systems based on technology acceptance model: an empirical study. *Journal of Ambient Intelligence and Humanized Computing, 14*(6), 7919-7935. https://doi.org/10.1007/s12652-023-04601-0

[4] Nazari-Shirkouhi, S., Mousakhani, S., Tavakoli, M., Dalvand, M. R., Šaparauskas, J. & Antuchevičienė, J. (2020). Importance-performance analysis based balanced scorecard for performance evaluation in higher education institutions: an integrated fuzzy approach. *Journal of Business Economics and Management, 21*(3), 647-678. https://doi.org/10.3846/jbem.2020.11940

[5] Zandi, S. (2023). Revival of the Silk Road using the applications of AR/VR and its role on cultural tourism. *arXiv e-prints*, arXiv-2304. https://doi.org/10.48550/arXiv.2304.10545

[6] Li, F., Larimo, J. & Leonidou, L. C. (2021). Social media marketing strategy: definition, conceptualization, taxonomy, validation, and future agenda. *Journal of the Academy of Marketing Science*, *49*, 51-70. https://doi.org/10.1007/s11747-020-00733-3

[7] Darvishinia, N. & Clark, S. (2024). Empowering Rural Education: Exploring the Integration of Robotics and Remote Sensing Technologies. In *Society for Information Technology & Teacher Education International Conference* (pp. 2011-

2016). Association for the Advancement of Computing in Education (AACE). Las Vegas, Nevada, United States: Retrieved May 10, 2024 from https://www.learntechlib.org/primary/p/224251/

[8] Hanaysha, J. R. (2022). Impact of social media marketing features on consumer's purchase decision in the fast-food industry: Brand trust as a mediator. *International Journal of Information Management Data Insights*, *2*(2), 100102. https://doi.org/10.1016/j.jjimei.2022.100102

[9] Nazari-Shirkouhi, S. & Keramati, A. (2017). Modeling customer satisfaction with new product design using a flexible fuzzy regression-data envelopment analysis algorithm. *Applied Mathematical Modelling, 50*, 755-771. https://doi.org/10.1016/j.apm.2017.01.020

[10] Oliveira, M. O. R. D., Heldt, R., Silveira, C. S. & Luce, F. B. (2023). Brand equity chain and brand equity measurement approaches. *Marketing Intelligence & Planning*, *41*(4), 442-456. https://doi.org/10.1108/MIP-06-2022-0222

[11] Rezvani, S., Heidari, S., Roustapisheh, N. & Dokhanian, S. (2022). The effectiveness of system quality, habit, and effort expectation on library application use intention: the mediating role of perceived usefulness, perceived ease of use, and user satisfaction. *International Journal of Business Information Systems*, 1-18. https://doi.org/10.1504/IJBIS.2022.10049515

[12] Nazari-Shirkouhi, S., Keramati, A. & Rezaie, K. (2013). Improvement of customers' satisfaction with new product design using an adaptive neuro-fuzzy inference systems approach. *Neural Computing and Applications*, 23, 333-343. https://doi.org/10.1007/s00521-013-1431-x

[13] Aaker, D. A. (1996). Measuring brand equity across products and markets. *California Management Review, 38(3),* 102-120. https://doi.org/10.2307/41165845

[14] Haudi, H., Handayani, W., Musnaini, M., Suyoto, Y., Prasetio, T., Pitaloka, E., ... & Cahyon, Y. (2022). The effect of social media marketing on brand trust, brand equity and brand loyalty. *International Journal of Data and Network Science*, *6*(3), 961-972. https://doi.org/10.5267/j.ijdns.2022.1.015

[15] Boozary, P. (2024). The Impact of Marketing Automation on Consumer Buying Behavior in the Digital Space via Artificial Intelligence. *Power System Technology, 48*(1), 1008-1021.

[16] Garanti, Z. & Kissi, P. S. (2019). The effects of social media brand personality on brand loyalty in the Latvian banking industry: The mediating role of brand equity. *International Journal of Bank Marketing, 37*(6), 1480-1503. https://doi.org/10.1108/IJBM-09-2018-0257

[17] Zollo, L., Filieri, R., Rialti, R. & Yoon, S. (2020). Unpacking the relationship between social media marketing and brand equity: The mediating role of consumers' benefits and experience. *Journal of Business Research*, *117*, 256-267. https://doi.org/10.1016/j.jbusres.2020.05.001

[18] Atulkar, S. (2020). Brand trust and brand loyalty in mall shoppers. *Marketing Intelligence & Planning, 38*(5), 559-572. https://doi.org/10.1108/MIP-02-2019-0095

[19] Morgan, R. M. & Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. *The journal of marketing*, 20-38. https://doi.org/10.1177/002224299405800302

[20] Berry, L. (1991). Big ideas in services marketing. *Journal of Consumer Marketing, 3*(2), 47-51. https://doi.org/10.1108/eb008162

[21] Villagra, N., Monfort, A. & Sanchez Herrera, J. (2021). The mediating role of brand trust in the relationship between brand personality and brand loyalty. *Journal of Consumer Behaviour, 20*(5), 1153-1163. https://doi.org/10.1002/cb.1922

[22] Nazari-Shirkouhi, S., Keramati, A. & Rezaie, K. (2015). Investigating the effects of customer relationship management and supplier relationship management on new product development. *Tehnički vjesnik, 22*(1), 191-200. https://doi.org/10.17559/TV-20140623130536

[23] Kim, J., Leung, X. Y. & McKneely, B. (2023). The effects of Instagram social capital, brand identification and brand trust on purchase intention for small fashion brands: the generational differences. *Journal of Fashion Marketing and Management, 27*(6), 988-1008. https://doi.org/10.1108/JFMM-05-2021-0126

[24] Ha, N. & Nguyen, T. (2019). The effect of trust on consumers' online purchase intention: An integration of TAM and TPB. *Management Science Letters, 9*(9), 1451-1460. https://doi.org/10.5267/j.msl.2019.5.006

[25] Shankar, A., Jebarajakirthy, C. & Ashaduzzaman, M. (2020). How do electronic word of mouth practices contribute to mobile banking adoption? *Journal of Retailing and Consumer Services*, *52*, 101920. https://doi.org/10.1016/j.jretconser.2019.101920

[26] Tanhaei, H. G., Boozary, P. & Sheykhan, S. (2024). Analyzing the Impact of Social Media Marketing, Word of Mouth and Price Perception on Customer Behavioral Intentions through Perceived Interaction. *International Journal of Business and Social Science, 15*(1), 69-77.

[27] Darvishinia, N. (2023). AI in Education: Cracking the Code Through Challenges: A Content Analysis of one of the recent Issues of Educational Technology and Society (ET&S) Journal. *Partners Universal International Innovation Journal*, *1*(4), 61-71.

[28] Sanaei, F. (2024). How customers' satisfaction change with the use of AR shopping application: A conceptuall model. *arXiv* preprint *arXiv*:2401.10953. https://doi.org/10.48550/arXiv.2401.10953

[29] Hashemian, F., Maleki, N. & Zeinali, Y. (2024). From User Behavior to Subscription Sales: An Insight Into E-Book Platform Leveraging Customer Segmentation and A/B Testing. *Services Marketing Quarterly*, 1-29. https://doi.org/10.1080/15332969.2024.2313873

[30] Mirshekari, S., Moradi, M., Jafari, H., Jafari, M. & Ensaf, M. (2024). Enhancing Predictive Accuracy in Pharmaceutical Sales through an Ensemble Kernel Gaussian Process Regression Approach. *International Journal of Computer and Information Engineering, 18*(5), 255-260. https://doi.org/10.2139/ssrn.4860667

[31] Seo, E. J. & Park, J. W. (2018). A study on the effects of social media marketing activities on brand equity and customer response in the airline industry. *Journal of Air Transport Management*, *66*, 36-41. https://doi.org/10.1016/j.jairtraman.2017.09.014

[32] Kim, A. J. & Ko, E. (2012). Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand. *Journal of Business Research, 65*(10), 1480-1486. https://doi.org/10.1016/j.jbusres.2011.10.014

[33] AmirKhani, T. & Borhani, T. (2016). Public service motivation: A study of the impact of job design and employees subjective well-being in a public hospital. *Iranian journal of management sciences, 11*(41), 76-90.

[34] Noruzy, A., Dalfard, V. M., Azhdari, B., Nazari-Shirkouhi, S. & Rezazadeh, A. (2013). Relations between transformational leadership, organizational learning, knowledge management, organizational innovation, and organizational performance: an empirical investigation of manufacturing firms. *The International Journal of Advanced Manufacturing Technology, 64*, 1073-1085. https://doi.org/10.1007/s00170-012-4038-y

[35] Tavana, M., Nazari-Shirkouhi, S. & Farzaneh Kholghabad, H. (2021). An integrated quality and resilience engineering framework in healthcare with Z-number data envelopment analysis. *Health care management science, 24*, 768-785. https://doi.org/10.1007/s10729-021-09550-8
[36] Ghazvinian, A., Feng, B., Feng, J., Talebzadeh, H. & Dzikuć, M. (2024). Lean, Agile, Resilient, Green, and Sustainable (LARGS) Supplier Selection Using Multi-Criteria Structural Equation Modeling under Fuzzy Environments. *Sustainability, 16*(4), 1594. https://doi.org/10.3390/su16041594
[37] Alipour, N., Nazari-Shirkouhi, S., Sangari, M. S. & Vandchali, H. R. (2022). Lean, agile, resilient, and green human resource management: the impact on organizational innovation and organizational performance. *Environmental Science and Pollution Research, 29*(55), 82812-82826. https://doi.org/10.1007/s11356-022-21576-1
[38] Majeed, M., Owusu-Ansah, M. & Ashmond, A. A. (2021). The influence of social media on purchase intention: The mediating role of brand equity. *Cogent Business & Management, 8*(1), 1944008. https://doi.org/10.1080/23311975.2021.1944008
[39] Azimi Asmaroud, S. (2022). Preservice Elementary Teachers' Categorical Reasoning and Knowledge Transfer on Definition Tasks with Two Dimensional Figures. *Theses and Dissertations*. https://ir.library.illinoisstate.edu/etd/1588
[40] Ghorashi, S. M., Azkia, M. & Mahdavi, M. S. (2018). Phenomenological Study: (Kan Neighborhood, District 5, Tehran, Iran). *Quarterly Journal of Social Development (Previously Human Development), 12*(2), 29-54.
[41] Askari, M. & Karami, H. On the Relationship between Sensory Learning Styles and Reading Subskill Profiles: An Application of Fusion Model. *Language Related Research*, 117-0. http://dorl.net/dor/20.1001.1.23223081.1401.0.0.144.6

Authors' contacts:

Hosna Khorsandi
(Corresponding Author)
Department of Educational Leadership and Policy Studies,
University of Denver,
1999 E Evans Ave, Denver, CO 80210, USA
hosnakhorsandi.du@gmail.com

Behzad Kazemi
Department of Advanced Data Analytics,
Toulouse Graduate School, University of North Texas,
1155 Union circle #310930, Denton, TX 76203, USA

Simin Zeynali
Curriculum Planning Department,
Faculty of Humanities and Psychology, University of Tabriz,
29 Bahman Blvd., Tabriz, Iran 5166616471

Mahsa Mohsenibeigzadeh
Department of Management, College of Business,
University of Central Florida,
12744 Pegasus Dr, Alafaya, FL 32816, USA

Pedram Zarei, Ph.D. Student in Developmental Education
College of Education, Texas State University,
601 University Dr, San Marcos, TX 78666, USA

Shahin Mirshekari
Department of Marketing Science and Business Analytics,
Katz Graduate School of Business, University of Pittsburgh,
3950 Roberto Clemente Dr, Pittsburgh, PA 15260, USA

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240403134350

Original scientific paper
Received: 2024-04-03, Accepted: 2024-05-06

Variant Design of Modular Products Using Functional Modelling and Multi-Criteria Evaluation Method

Mirko Pastović, Mirko Karakašić*, Željko Ivandić, Ivan Grgić

Abstract: In this work, the function structure of a mobile machine for corn peeling was developed using the functional decomposition method. The function structure obtained by functional decomposition served as the basis for the definition of working principles, i.e. initial modules (function carriers) by which partial functions were solved. A multi-criteria evaluation of the working principles was used to select those working principles that received the highest rating. A technical and economic evaluation was carried out in accordance with VDI2225. This evaluation method was chosen because it allows working principles to be analysed and evaluated by applying different criteria and sub-criteria. The results of the overall quality of the compared working principles are presented graphically. This shows which working principle received the highest or lowest rating. By combining working principles with different physical effects and design forms, five conceptual variants were generated. In the further analysis through the other phases of the design process, the first conceptual variant was selected as the final product with a modular structure.

Keywords: conceptual design; functional decomposition; modular design; multi-criteria evaluation; technical systems; variant design

1 INTRODUCTION

The systematization of design knowledge [1-3] is extremely important in the early stages of the design process, such as the phase of the task clarification and definition and the conceptual phase. In the design process, these two phases have a decisive influence on the future product because it is necessary to make decisions during their implementation, for which not all information and data about the future product are still known [4]. It is therefore important that such decisions are correctly made and systematically written down. Other phases of the design process do not have such a problem because the technical description of the product is recorded in prescribed formal documents such as structural title blocks, technical drawings, computerized CAD models, control and service documentation. Therefore, it is necessary to apply different design tools that would enable the design team to make reliable and traceable decisions in the conceptual phase and the clarification task phase.

Application of functional modelling is an important tool that enables creation of product function models in the conceptual phase, creating connections between functions at different levels of the function structure [5]. By applying the functional decomposition method, the overall function is broken down into functions of less complexity of components and sub - assemblies, which enables a better understanding of the relationship between functions [6, 7]. The authors in [8] develop a method of hierarchical functional modelling for the analysis of complex systems. Since the function structure is limited to identifying auxiliary systems of individual components, the authors in [9] propose an integrated function structure and object-oriented design framework method. Papers [5, 7, 10], use graphic models of function flow diagrams. These models connect partial functions into function structures by monitoring the flow of energy, material and signal. Very often, such models can be very complex, and tracking the connections between functions is complicated and unclear. Also, the mentioned function models can hardly help the designer in analysing the functionality of technical systems in the conceptual phase [11]. Papers [12, 13] develop a model of the function and functionality matrix (MFF), which enables the connection of partial functions into function structures through the matrix form. Also, the MFF model enables searching for the principle of solutions for individual partial functions, which is not possible to achieve through the function flow diagram models. Such research is based on the application of the morphological matrix [14] in the generation of the conceptual variants in the process of the product design.

Modular design can be a good initial point for a design of variant products, whose modules form independent and connectable units [15]. Functional analysis of products with different or equal functions precedes to the design of function modules [16]. Different market requirements have influence on the product design that are created by combining and connecting modules with precisely defined functionalities into a unique structure of a new product. Certain specific user requirements, the need for mass production, shorter production time, assembly forms, improvement of production quality, increase of innovative products, costs reduction and simple maintenance, influenced on the development and application of modular design methods [16, 17]. In the paper [18], by applying a modular design method in the conceptual phase, the modular structure of the refrigerators was determined. The importance of the modular structure determination in the early stage of design, i.e. on the conceptual stage, was observed. Authors in [19] using Design Structure Matrix (DSM) and Modular Function Deployment (MFD), optimized product designs for automation. The products had a modular structure, which enabled them to increase the variety of a new products and higher degree of automation in the assembly line.

Modularity does not only represent products whose structure consists of exchangeable modules that form product variants, but also contains some other aspects of application [20]. The application of modular design methods is

significant in the development of the interface modules of the mobile APP [21], in the new design approach of the housing market [22] and in the design of smart furniture [16]. The connection between mechatronic solutions as the fulfilment of the functions of modular design structures in wheelchair construction is shown in [23]. The authors in [24] base modularity on the life - cycle processes such as manufacture, assembly, service, and recycling. By grouping components into modules, according to the way they are recycled, it has the effect of reducing the cost of withdrawing the product from use. According to research conducted on smartphones [25], it is necessary to take into account the economic and technological feasibility of modularization when designing new products. The authors introduce an ecological efficiency index that analyzes the environmental impact of upgradeable components such as batteries and motherboards. The proposed index can be helpful in determining the appropriate modular design of product variants. Also, multi - objective modular design methods are increasingly being developed. These methods include multiple objectives related to functionality, environmental and economic constraints. The authors in [26] develop a multi-objective green modular design method. This method uses atomic theory and fuzzy clustering to create module configurations.

It is extremely important to make the right decisions when choosing working principles and connecting them into the structure of the conceptual variant. It is also important, after several conceptual variants have been defined, to select those variants that make the most sense for further development in the other phases of the design process. The knowledge and experience of the designer is important here, but often the choice of the best working principle or conceptual variant differs from designer to designer [4]. Therefore, in the design process, as well as in the conceptual phase, different methods and tools are used that enable the designer to be reliable and repeatable in making decisions [27, 28]. Bad decisions in the conceptual phase affect costly rework and the demand for resources that could have been spent on innovative and new products [27]. Some of the more significant evaluation methods that use defined multi - criteria systems in the decision - making process in the conceptual phase are: Promethee method [29], technical and economic evaluation method according to VDI 2225 [30, 31], Analytic Hierarchy Process (AHP) [32, 33] and Quality Function Deployment (QFD) [34]. Regardless of the development of new methods and tools, and the desire to minimize the designer's subjective approach in decision - making, the need for his knowledge and experience is still an important part of the decision - making process in the conceptual phase [29].

2 METHODS AND METHODOLOGY

In the paper, a function model of a mobile machine for corn peeling was developed using the functional decomposition method. By the function flow diagrams, the overall function is decomposed into partial functions, which form the function structure at five functional levels. The connection between the functions is achieved through monitoring the flow of energy, material and signal. It is presented how the function model can serve as a basis for the generation of a modular structure, that is, the functional modelling of five independent modules (function carriers). For their interconnection, it is important to use the experience and knowledge of the designer as well as precisely defined interfaces. In the design phases of embodiment design and detailed design, the function carriers are formed into the following modules: stand with wheels, drive reducer, peeling table, fan and corn cob pressers. The function model did not prove to be sufficient for determining the carrier of functions. Therefore, this model is connected to the morphological matrix. The morphological matrix enabled the connection of partial functions from the function model with working principles, i.e. function solutions. The working principles served as a basis for determining the modules (function carriers). The selection of the best modules was achieved using the evaluation of working principles using the method of technical and economic evaluation according to VDI 2225 [30]. For this purpose, a system of criteria was determined. From this system, a set of goals was determined. For each goal, the goal importance, the goal importance factor and the goal rating were determined. After the evaluation process, the working principles that had the best values of overall goodness were connected in the structure of the first conceptual variant. Also, through further analysis, the remaining working principles are interconnected in next four conceptual variants. After further analysis, through the other stages of the design process, the first conceptual variant was selected as the final product, whose modular structure consists of five modules.

3 FUNCTIONAL DECOMPOSITION OF THE MOBILE MACHINE FOR CORN PEELING

The design task that resulted from the market research aims to design a prototype of a new mobile machine for corn peeling. The market survey was conducted on a sample of 30 family farms that harvest corn on the cob. The basic purpose of the machine would be the domestication of mercantile and seed cobs of corn after it has been harvested by corn pickers, and before the corn is stored. A detailed description of the requirements, resulting from the market analysis, is categorized into five categories listed in [4]. The more important design requirements, according to which the function structure was generated by the functional decomposition method, are extracted from [4] and listed below: the purity level of corn after passing through the machine should be 95 – 100 %, the possibility of driving with a tractor or an electric motor, ensure the mobility of machine with two or four wheels, two or more inputs for feeding the machine with the corn cob, one output for the exit of the peeled corn cob, easy use, modular structure, easy maintenance, satisfy the necessary safety regulations at work and reduce production costs.

Conceptual design, in the first step, aims to generate the function structure of the product. The overall function of a mobile machine for corn peeling is the carrier of the highest level of abstraction and is described at the first level of the function structure through the input - output flow of energy, material and signal (Fig. 1).

By analysing the requirements from the requirement list [4], the main function "*Completed corn cob peeling*", using the functional decomposition method, was divided into five levels. Thus, the level of its abstraction was reduced, that is, partial functions of lower complexity were determined. For partial functions, the carriers of these functions, i.e. the working principles, will be seeked later. By connecting the working principles, conceptual variants and modules of the mobile machine for corn peeling will be formed.

The function flow diagram of the second level of the function structure of the mobile machine for corn peeling (Fig. 2) was formed by decomposing the overall function from the first level of the function structure.

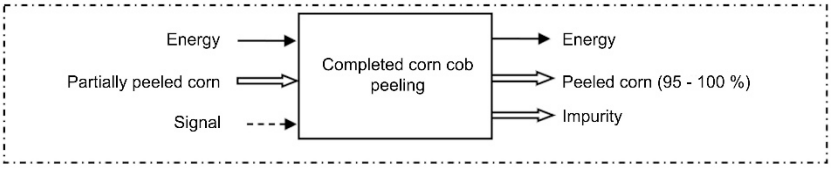


Figure 1 The first level of the function structure of the mobile machine for corn peeling

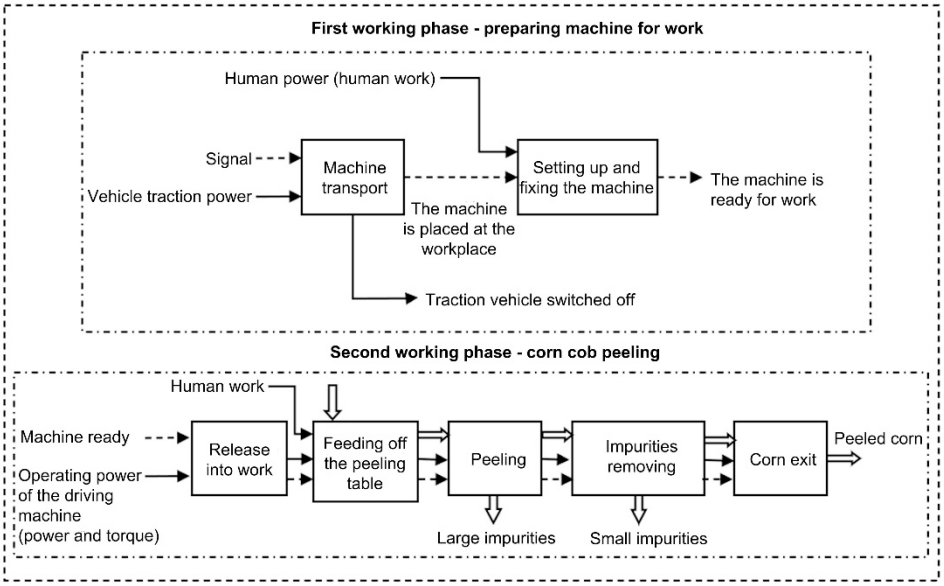


Figure 2 The second level of the function structure of the mobile machine for corn peeling

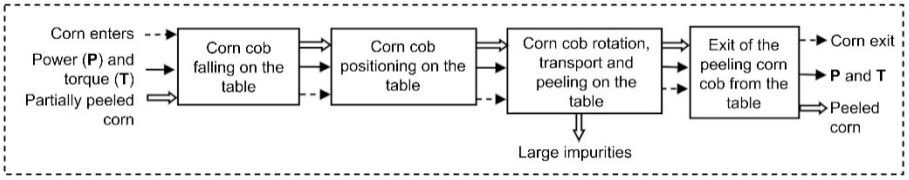


Figure 3 The third level of the function structure of the mobile machine for corn peeling

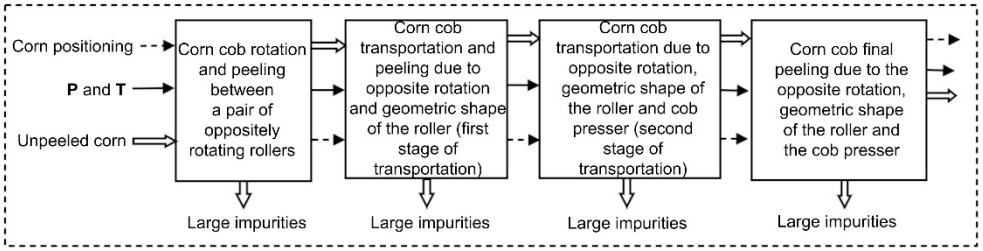


Figure 4 The fourth level of the function structure of the mobile machine for corn peeling

The second level of functional decomposition includes two phases of manipulating the machine. The first phase consists of preparing machine for work. To realize this phase, two functions "Machine transport" and "Setting up and fixing the machine" are required. After the first phase has been realized, the second phase follows. Five functions are needed to realize the second phase (Fig. 2).

Due to its complexity, the function "Peeling" is divided into four partial functions that form the third level of the function structure (Fig. 3).

Due to its complexity, the function "Corn cob rotation, transport and peeling on the table" was divided into four new functions that form the fourth level of the function structure (Fig. 4).

This complexity is also evident from its name, since it is necessary to functionally solve rotation, transport and peeling.

The function "*Corn cob final peeling due to the opposite rotation, geometric shape of the roller and the cob presser*" is also a complex function. Therefore, it is divided into two functions of less complexity. These two functions form the fifth level of the function structure of the mobile machine for corn peeling (Fig. 5).

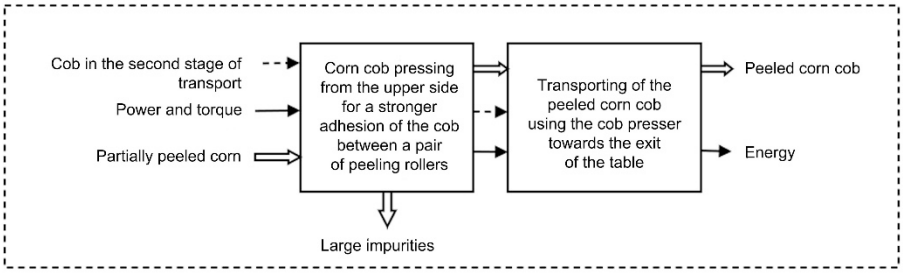


Figure 5 The fifth level of the function structure of the mobile machine for corn peeling

3.1 Function Carriers of the Mobile Machine for Corn Peeling

From the function structure, it follows that the mobile machine for corn peeling will have a modular structure, i.e. the machine itself will consist of several independent modules that will be interconnected via precisely defined interfaces.

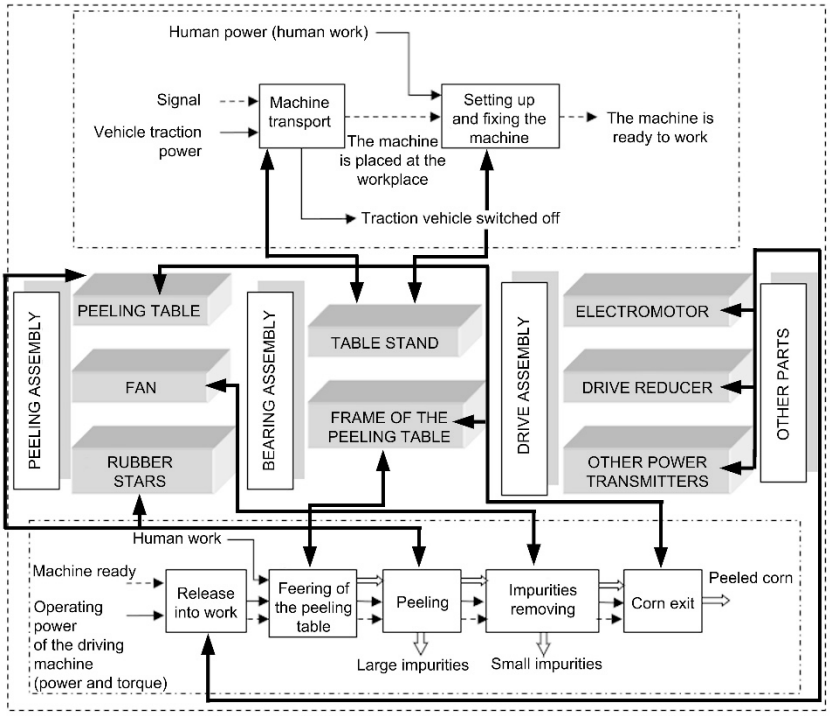


Figure 6 Integrated scheme of relations of partial functions and their carriers

Modules are functionally independent units, i.e. carriers of one or more functions, which have a minimum number of previously precisely defined connections with other modules, components and parts. From the function structure developed on five levels and shown in Fig. 1 to Fig. 5, the mobile machine for corn peeling can potentially have the following modules: stand with wheels, drive reducer, peeling table, fan and corn cob pressers. Connections between partial functions and their carriers are shown in Fig. 6. It follows from the above that with the mentioned modules the machine would be almost complete, if the equipping of the machine with electrical signalling is excluded. According to traffic laws, the machine, as a trailer, must have electric signalling, but it is not functionally analysed in this paper.

3.2 Determination of the Working Principles of the Partial Functions

The working principles represent technical systems of varying complexity that are possible solutions for partial

functions determined by the function structure of a mobile machine for corn peeling. They represent the function carriers and by their subsequent connection in the further course of the design process, they will form the structures of the construction variants (conceptual variants) and modules of the mobile machine for corn peeling. At the searching for a solution, priority is given to those partial functions that determine the principles of the overall solution, and the sequence is derived from the identification connections between the flow of energy, material and signal.

Partial functions and the principles of their solutions (working principles) are presented using the morphological matrix method (Fig. 7). The concretization of working solutions (principles of the solutions) is represented by solution sketches, solution schemes and geometric forms of potential modules of the mobile machine for corn peeling.

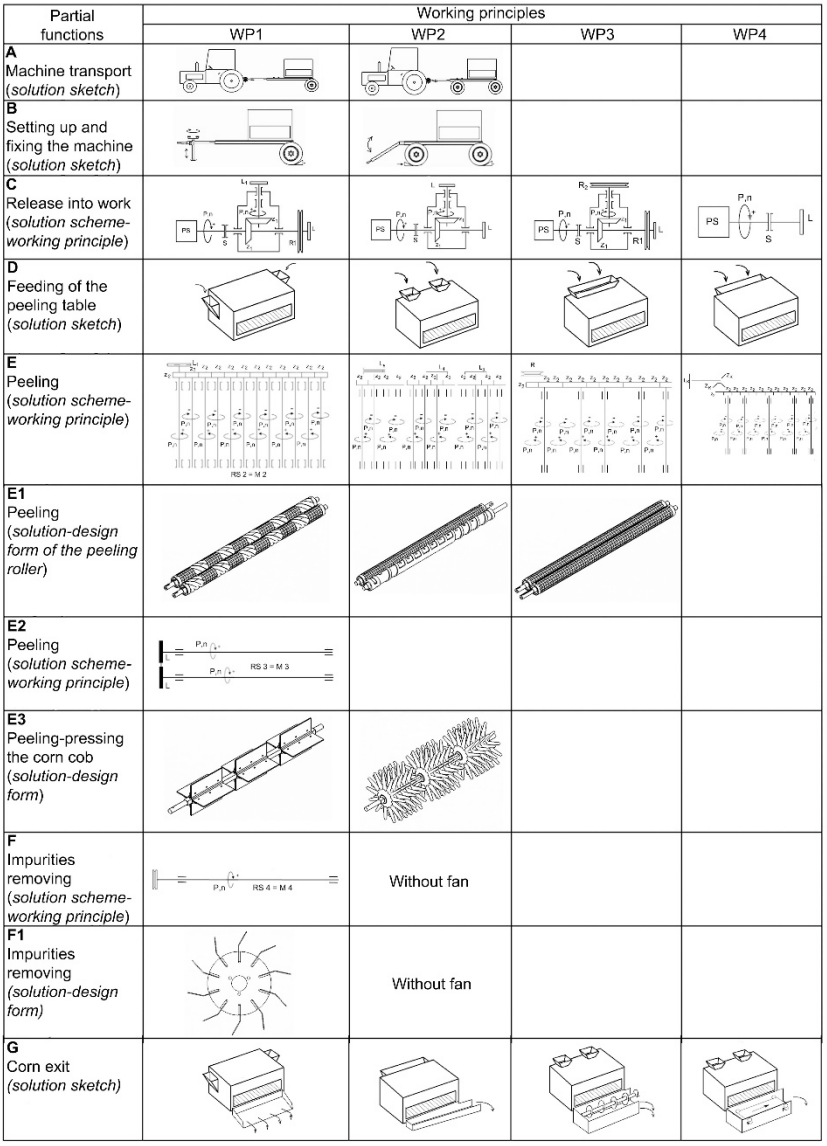


Figure 7 Morphological matrix of possible working principles of partial functions obtained from function structure

4 EVALUATION BY THE METHOD OF TECHNICAL AND ECONOMIC GOODNESS ACCORDING TO VDI 2225

4.1 Determination of the Criteria System and the Goal System

By analysing the working principles shown in the morphological matrix (Fig. 7) and combining them into working structures (conceptual variants), the basis for designing product variants was achieved. In order to select the best solution principles, a technical and economic evaluation was carried out according to VDI 2225. According to the evaluation, the best solutions were selected, which were then connected into working structures (conceptual variants).

The evaluation represents the phase in which all the proposed working principles of the solution (Fig. 7), in relation to the set of goals, are assigned appropriate ratings. The goals system is determined from the criteria system. It should be pointed out that during the evaluation in the conceptual phase, a lot of information about the product is

still missing. Therefore, after evaluation process, it is useful to reduce the choice of possible variant solutions (combination of working principles) to those variants that are the most promising. Then only they need to be further developed in the following design phases.

The criteria according to which the evaluation of working principles was carried out were defined based on the analysis of the most important requirements from the requirement list and the function structure of the mobile machine for corn peeling. The criteria system is divided into three levels (Tab. 1). On the first level there are three basic criteria, which on the second level are described by six sub-criteria. Finally, at the third level there are fourteen sub-criteria that describe the criteria from the second level.

Table 1 Evaluation criteria of working principles

First level	Second level	Third level
Safe function execution	Mechanical safety	Working principle reliability
		Working principle complexity level
	Basic function efficiency	Traffic operability
		Work operability
Technological design	Design complexity level	Parts complexity level
		Parts number
	Manufacturing and assembly complexity level	Machining share
		Casting technology share
		Welding device
		Assembly complexity level
Good exploitation properties	Adjustment and maintenance	Simple adjustment
		Simple maintenance
	Maintenance costs	Regular costs
		Extraordinary costs

According to the criteria from Tab. 1, a system of goals is determined, which for each evaluation cycle is divided into four hierarchical levels (Fig. 8). Two quantitative indicators are attached to each partial goal. The first indicator is goal importance at the associated level (G_{ijk}), and the second indicator is the goal importance factor relative to the ultimate goal (g_{ijk}). In this way, the advantage of certain partial goals is clearly suggested in relation to other goals that are on the same level. At the same time, the importance of individual partial goals is analyzed according to the degree of realization of the ideal solution (ultimate goal).

Due to the comprehensiveness of the evaluation of the working principles in the morphological matrix (Fig. 7), this paper presents the evaluation of the working principles WP1 and WP2, which solve the partial function A (*"Machine transport"*). Since they solve the function A, in the evaluation process they are marked with A WP1 and A WP2, that is, the working principle WP1 (two-wheel machine drive) which solves function A and the working principle WP2 (four-wheel machine drive) which also solves function A. For all other working principles from the morphological matrix (Fig. 7), the evaluation procedure is presented in [4].

For the evaluation process of all working principles, the same system of goals was used (Fig. 8). Different values are assigned only to individual goals for the indicators G_{ijk} and g_{ijk}. The assignment of the numerical values to quantitative indicators was achieved on the basis of the experience and knowledge of the designer. The final structure of the goal system with associated system elements and features, for the evaluation of working principles A WP1 and A WP2, is shown in Fig. 8.

According to the principle of consistency of the value of the importance of goals, the total value of the goal importance at the second hierarchical level (Fig. 8) corresponds to the value of the factor of the overall goal (ideal goal):

$$\mathrm{G}_i = \sum_{j=1}^{3} G_{ijk} = \mathrm{G}_{11} + \mathrm{G}_{12} + \mathrm{G}_{13}, \tag{1}$$

where $i = 1$ and $k = 0$.

The total value of the partial goal importance factor at the second hierarchical level, in the system of goals, must correspond to the value of the associated goal importance factor of the higher level, i.e. the value of the importance factor of the ideal goal:

$$g_i = \sum_{j=1}^{3} g_{ijk} = \mathrm{g}_{11} + \mathrm{g}_{12} + \mathrm{g}_{13}, \tag{2}$$

where $i = 1$ and $k = 0$.

The third hierarchical level of the system of partial goals with its goal importance values is determined by the expression:

$$\mathrm{G}_i = \sum_{j=1}^{3} \sum_{k=1}^{2} G_{ijk}, \tag{3}$$

where $i = 1$.

The third hierarchical level of the system of partial goals with its values of the goal importance factor is determined using the expression:

$$g_{ij} = \sum_{j=1}^{3} \sum_{k=1}^{2} g_{ijk}, \tag{4}$$

where $i = 1$.

The fourth hierarchical level of the system of partial goals with its goal importance values is determined by the expression:

$$\mathrm{G}_i = \sum_{j=1}^{3} \sum_{k=1}^{2} \sum_{l=1}^{4} G_{ijkl}, \tag{5}$$

where $i = 1$.

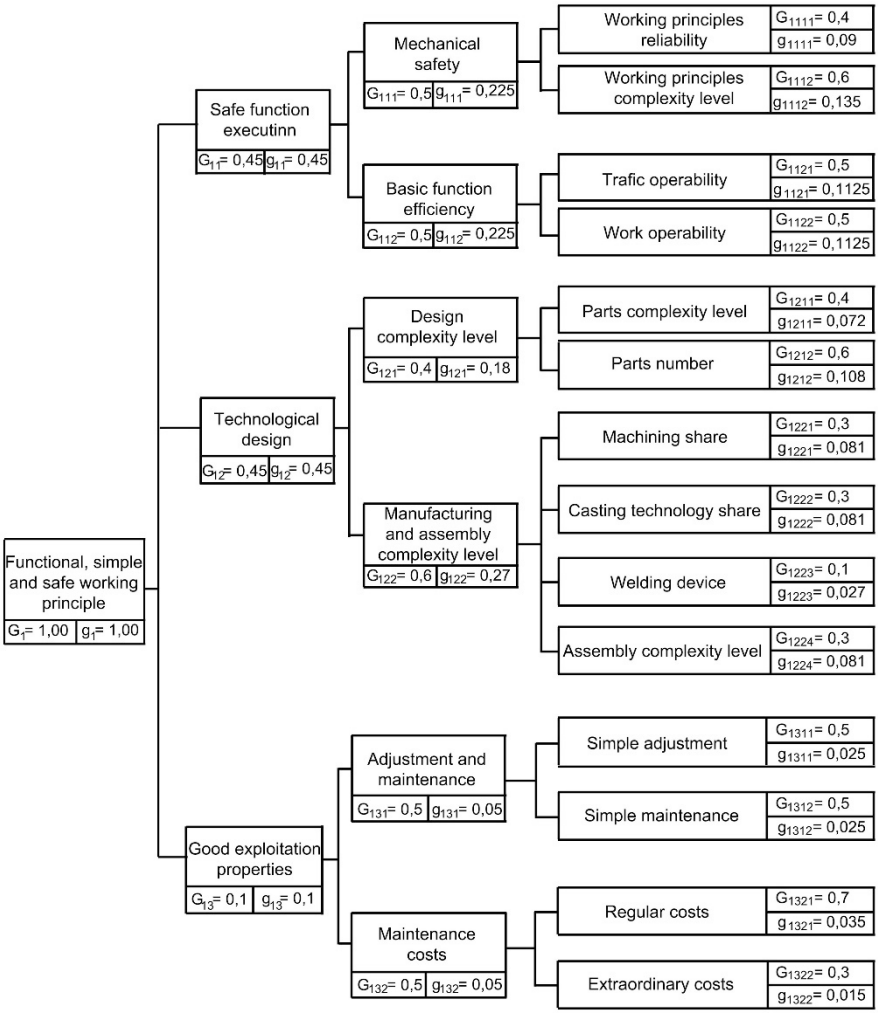


Figure 8 Goals system for working principles A WP1 and A WP2

Table 2 Calculated values of the goals system for working principles A WP1 and A WP2

Evaluation criterion	Goal mark	Goal description	Goal level	G_{ijk}	g_{ijk}
	C_1	Functional, simple and safe working principle	1	1,0	1,0
1	C_{11}	Safe function execution	2	0,45	0,45
2	C_{111}	Mechanical safety	3	0,5	0,225
3	C_{112}	Basic function efficiency	3	0,5	0,225
4	C_{1111}	Working principle reliability	4	0,4	0,09
5	C_{1112}	Working principle complexity level	4	0,6	0,135
6	C_{1121}	Traffic operability	4	0,5	0,1125
7	C_{1122}	Working operability	4	0,5	0,1125
8	C_{12}	Technological design	2	0,45	0,45
9	C_{121}	Design complexity level	3	0,4	0,18
10	C_{122}	Manufacturing and assembly complexity level	3	0,6	0,27
11	C_{1211}	Parts complexity level	4	0,4	0,072
12	C_{1212}	Parts number	4	0,6	0,108
13	C_{1221}	Machining share	4	0,3	0,081
14	C_{1222}	Casting technology share	4	0,3	0,081
15	C_{1223}	Welding devices	4	0,1	0,027
16	C_{1224}	Assembly complexity level	4	0,3	0,081
17	C_{13}	Good exploitation properties	2	0,1	0,1
18	C_{131}	Adjustment and maintenance	3	0,5	0,05
19	C_{132}	Maintenance costs	3	0,5	0,05
20	C_{1311}	Simple adjustment	4	0,5	0,025
21	C_{1312}	Simple maintenance	4	0,5	0,025
22	C_{1321}	Regular costs	4	0,7	0,035
23	C_{1322}	Extraordinary costs	4	0,3	0,015

The fourth hierarchical level of the system of partial goals with its values of the goal importance factor is determined using the expression:

$$g_{ijk} = \sum_{j=1}^{3}\sum_{k=1}^{2}\sum_{l=1}^{4} g_{ijkl}\,, \tag{6}$$

where $i = 1$.

Quantitative values of the partial goals importance factors (g_{111}, g_{112}, g_{121}, g_{122}, g_{131}, and g_{132}) of the third hierarchical level, are determined according to the rule of the product of all goals importance of higher level, following the descending index of the order of partial goals of the higher level:

$$g_{ijk} = \mathrm{G}_{ijk} \cdot G_{ij} \cdot G_i\,, \tag{7}$$

where $i = 1$, $j = 1, 2, 3$ and $k = 1, 2$.

Quantitative values of the partial goals importance factors of the fourth hierarchical level are calculated in the same way.

Calculated values of the goals importance and importance factors for working principles A WP1 and A WP2 are presented in Tab. 2

In this way, the distribution of the elements of the goals system with the corresponding values of the features of the partial goals at individual levels was prepared for the evaluation process.

4.2 Technical and Economic Evaluation of Working Principles

Each solution of the working principle of the mobile machine for corn peeling, after the evaluation process, will be ranked according to the relative rating of the degree of fulfilment of the goal. In this way, the advantage of a better solution compared to another less good solution is suggested.

The product is acceptable for further design and production if it meets the technical and economic criteria. The fulfilment of technical criteria enables the product to be suitable from a technical aspect, while the fulfilment of economic criteria ensures the product's economic justification. Therefore, during the evaluating process of the working principles A WP1 and A WP2, a technical and economic evaluation was carried out. According to [30, 35], the following rating scale was defined (Tab. 3).

Table 3 Rating scale for the implementation of the evaluation process

Solution quality	Rating
Unsatisfying	1
Partially satisfying	2
Satisfying	3
Very good	4
Excellent	5

Ratings were then added to each goal based on the designer's knowledge and experience (Tab. 4). Insufficient experience and knowledge of the designer can significantly reduce the implementation of the evaluation procedure and lead to unreliable evaluation results.

The total value of the criterion evaluation factor for working principles A WP1 and A WP2 (Tab. 4) is determined according to the expression:

$$Gw_k = \sum_{j=1}^{14} w_{jk}\,, \tag{8}$$

where w is the grade that is added to an individual goal from the system of goals and $k = 1, 2$ for two working principles A WP1 and A WP2 (Tab. 4).

The total value of the factor of real significance of the criteria for the working principles A WP1 and A WP2 (Tab. 4) is determined according to the expression:

$$Gwg_k = \sum_{j=1}^{14} wg_{jk}\,, \tag{9}$$

where $k = 1, 2$ for two working principles A WP1 and A WP2 (Tab. 4).

The utility factor of the evaluated solution for the working principles A WP1 and A WP2 (Tab. 4) is determined according to the expression:

$$W_k = \frac{Gw_k}{w_{\max} \cdot n}\,, \tag{10}$$

where $w_{\max}$ is the highest amount of the goal ratting from the set of selected goals that are evaluated for a particular working principle (Tab. 4), n is the number of selected criteria (Tab. 4) from the total set of criteria (Tab. 2) which are evaluated and $k = 1, 2$ for two working principles A WP1 and A WP2.

The technical goodness factor for the working principles A WP1 and A WP2 is determined according to the expression:

$$Wg_k = X_k = \frac{Gwg_k}{w_{\max} \cdot \sum_{j=1}^{14} g_{ijk}} = \frac{Gwg_k}{w_{\max} \cdot 1}\,, \tag{11}$$

where X_k is the overall technical goodness of a particular working principle and $k = 1, 2$ for two working principles A WP1 and A WP2.

In the economic evaluation of the goodness of the solution, only the manufacturing and assembly costs were used. Manufacturing costs consist of the costs of materials and their processing. Since the working principles of A WP1 and A WP2 are in the conceptual phase, the exact costs of the material and its processing cannot be determined at this stage of design. These costs can only be defined when the entire design process is completed and the optimal batch size is determined. For budget purposes, the authors of the paper estimated these costs based on many years of experience in the development and design of machines for corn peeling. Therefore, when conducting an economic evaluation in the conceptual phase, according to [35], the cost of the ideal solution is introduced, whose amount is $K_{\text{ideal}} = 1$. If the rating of the cost of manufacturing and assembly indicates how much this cost is above the cost of the ideal solution, then the value of the overall economic goodness of the solution can be determined using the expression [35]:

$$Y_k = \frac{K_{\text{ideal}}}{K_{\text{real}}} = \frac{1}{K_{\text{real}}} < 1\,, \tag{12}$$

where K_{real} is the real cost of production and assembly.

Table 4 List of technical and economic evaluation of working principles A WP1 and A WP2

Evaluation criterion	g_{ijk}	Evaluated goal	Variant 1 – A WP1		Variant 2 – A WP2	
			w_{j1}	wg_{j1}	w_{j2}	wg_{j2}
4	0,09	Working principle reliability	4,9	0,441	4,9	0,441
5	0,135	Working principle complexity level	4,9	0,662	4,0	0,54
6	0,1125	Traffic operability	4,0	0,45	4,2	0,473
7	0,1125	Working operability	4,5	0,506	4,2	0,473
11	0,072	Parts complexity level	4,0	0,288	4,0	0,288
12	0,108	Parts number	4,9	0,529	3,5	0,378
13	0,081	Machining share	4,2	0,340	3,8	0,308
14	0,081	Casting technology share	4,2	0,340	3,8	0,308
15	0,027	Welding devices	4,0	0,108	4,0	0,108
16	0,081	Assembly complexity level	4,6	0,373	3,5	0,284
20	0,025	Simple adjustment	4,2	0,105	3,8	0,095
21	0,025	Simple maintenance	4,4	0,11	4,0	0,1
22	0,035	Regular costs	4,8	0,168	4,4	0,154
23	0,015	Extraordinary costs	4,9	0,074	4,4	0,066
Σ	1	Total values	Gw_1	Gwg_1	Gw_2	Gwg_2
		Numerical amount of the total value	62,5	4,493	56,5	4,014
		Total value regarding the ideal solution	W_1	Wg_1	W_2	Wg_2
		Numerical amount	0,911	0,917	0,824	0,819
		Technical goodness X_k	$X_1 = 0,917$		$X_2 = 0,819$	
		Relative costs regarding the ideal solution	$K_1 = 1,4$		$K_2 = 1,8$	
		Economic goodness Y_k	$Y_1 = 0,714$		$Y_2 = 0,555$	

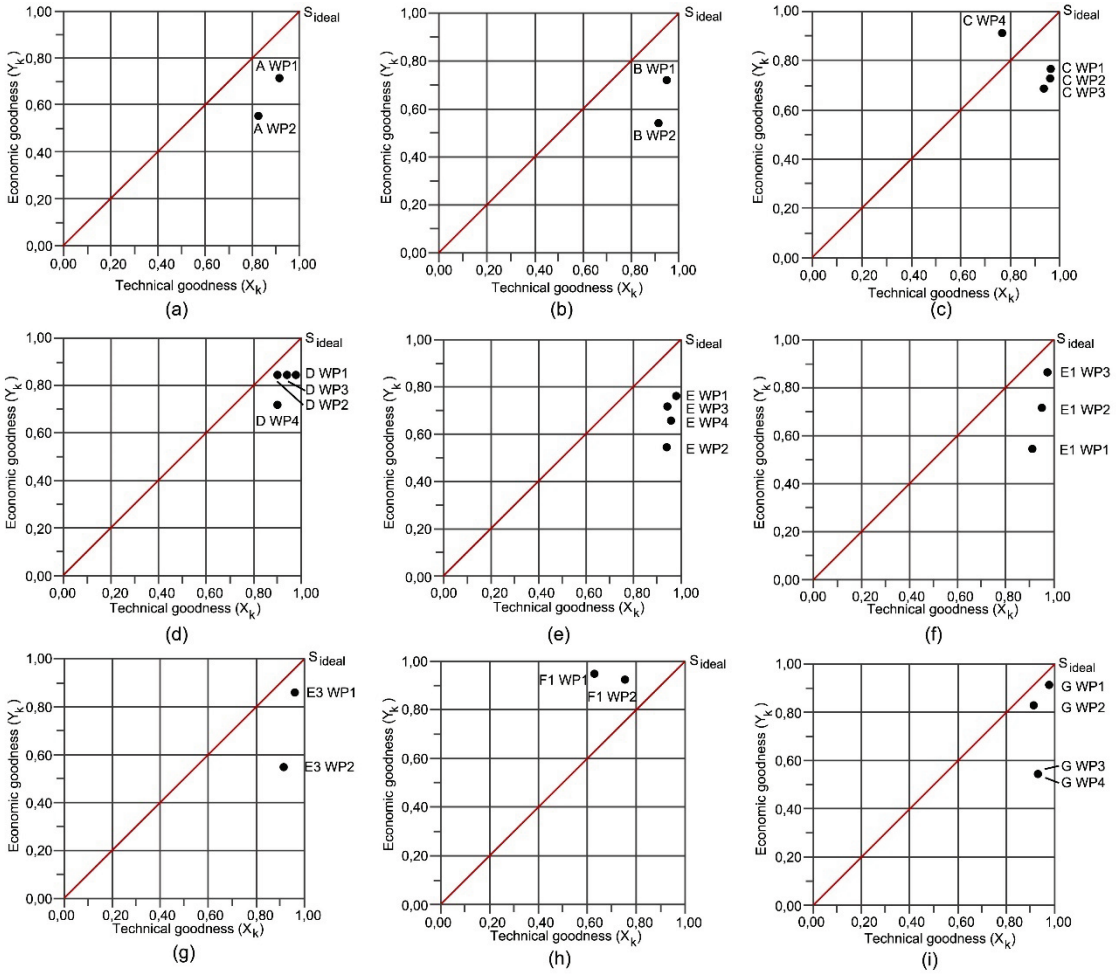


Figure 9 The overall technical and economic goodness of the working principles of the partial functions of the mobile machine for corn peeling

Solutions of technical and economic goodness for working principles A WP1 and A WP2 are presented in Tab. 4. For all other working principles shown in the morphological matrix (Fig. 7), solutions of technical and economic goodness are presented in [4].

The graphic representation of the solution of the performed technical and economic evaluation is shown in Fig. 9a. It is evident that the working principle of A WP1 has better overall goodness values compared to A WP2. Therefore, the working principle A WP1 represents the choice of the final solution of the partial function A (*"Machine transportation"*). The selection of the ideal solution is determined by the coordinates of the point Sideal with the amount X_{ideal} =1 and Y_{ideal} =1. The graphic representation of the obtained solutions for evaluating the working principles of the other partial functions, listed in the morphological matrix, is shown in Fig. 9b to Fig. 9i.

5 VARIANT DESIGN SOLUTIONS OF THE MOBILE MACHINE FOR CORN PEELING

After the evaluation process, field solutions was generated, i.e. five variant solutions (Fig. 10). The structure of the variant solutions is formed by varying the physical effects and design forms (working geometry, working movement and materials) from the morphological matrix (Fig. 7). In such a way, the synthesis of new technical systems (variant designs) was arranged. By combining the working principle of a partial function with the working principle of the following partial function, possible working structures, or variant designs, are formed. Due to the aforementioned, it is not possible every time to combine the working principles that received the highest rating into a unique variant solution. Therefore, the working principles E1 WP3 and F1 WP2, which received the highest rating in the evaluation process, are not included in the structure of the first variant solution (Fig. 9 and Fig. 10).

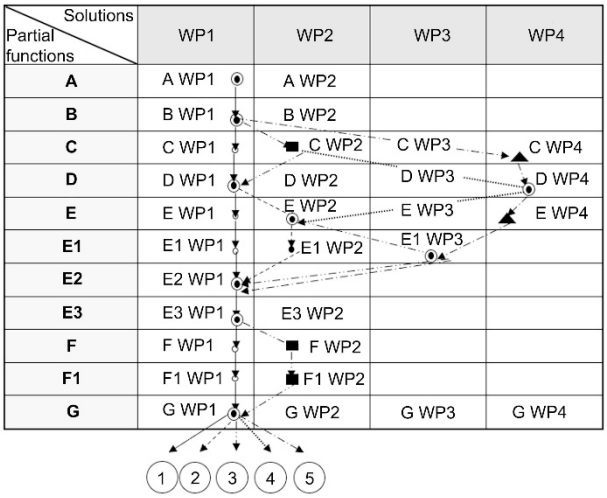


Figure 10 Combinations of working principles that generate variant solutions of a mobile machine for corn peeling

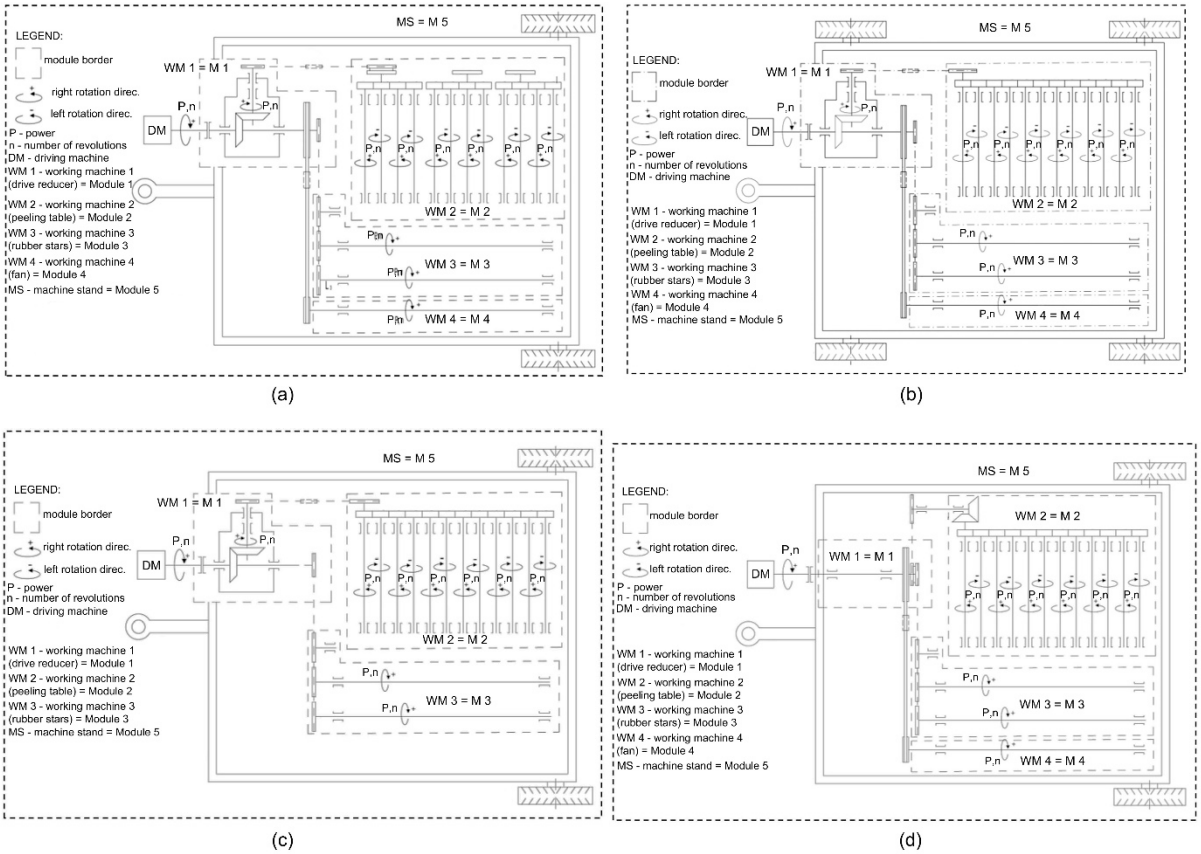


Figure 11 Schematic representation of function carriers and kinematics of design variants

After further elaboration of variant solutions, shown in Fig. 10, through analyses in other phases of the design process, the first variant solution was selected as the final design solution. This solution was further improved and

shaped as a final product in the detailed design phase. A schematic representation of the final design solution, its function carriers and kinematics is shown in Fig. 11a. Five modules, i.e. function carriers, form the modular structure of the first variant solution: module 1 represents working machine 1 (WM 1) i.e. drive reducer, module 2 represents working machine 2 (WM 2) i.e. peeling table, module 3 represents working machine 3 (WM 3) i.e. rubber stars, module 4 represents working machine 4 (WM 4) i.e. fan and module 5 represents machine stand (Fig. 11a).

The other four variants of the solution (Fig. 11b – Fig. 11d) were not developed in the other phases of the design process, but depending on the market needs, they leave the possibility of further development. Also, five design modules, defined as five function carriers, were implemented in the structure of these four variant solutions.

At the beginning of the design process (conceptual phase), using functional decomposition and morphological matrix, the initial modules, i.e. function carriers, are defined (Fig. 6). Due to its complexity, the function "*Peeling*" (Fig. 2) is divided into four functions on the third level of the function structure (Fig. 3). At the third level, the function "*Corn cob rotation, transport and peeling on the table*" contains a group of functions in its name. It is also equally possible to see in the names of the functions from the fourth and fifth level of the function structure. This points to the need for modularity, because those multiple functions are grouped into the one function name. By searching for the principle of solutions for functions from the function structure, by applying the morphological matrix, solutions were obtained in the form of more complex technical systems. These systems represent function carriers, or modules. The connection between the functions and their carriers is shown in Fig. 6.

By developing the initial modules, through the other phases of the design process, their final appearance and structure of the first variant solution shown in Fig. 11a. This modular structure, after refinement in the detailed design phase, was developed into the final design solution of the mobile machine for corn peeling shown in Fig. 12.

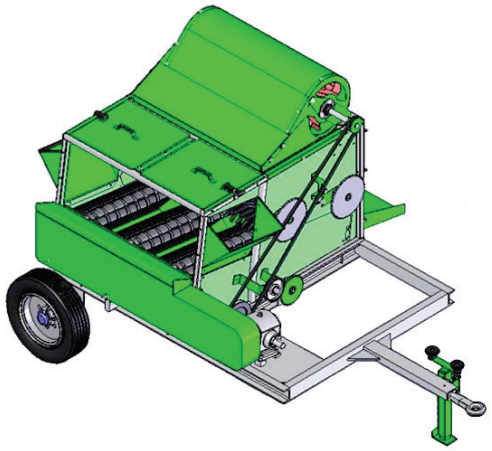


Figure 12 CAD model of the first design variant of the mobile machine for corn peeling

6 RECAPITULATION ANNOTATION

The application of the principles of functional modelling and modularity is extremely important for the development of variant products in the conceptual phase. After the analysis of requirements, using the method of functional decomposition, the function structure of the mobile machine for corn peeling was developed. The function structure proved as an important design tool that enabled the search for working principles that solved the functions who form the function structure. Through the relational relationship between the functions, which was realized by connecting the flows of energy, materials and signals, the connection between the functions was observed. Through the function structure, it is visible how certain functions in their name unite groups of less complexity functions. Therefore, by the function decomposition process, certain functions are broken down into the less complexity functions. Through the function structure, the connection of functions in their name indicates their grouping, that is, the need for modularity. Since the function structure shown by functional flow diagrams does not allow searching for the principle of the solution, i.e. the carrier of the functions, a morphological matrix was used for the solution search.

The function carriers are determined by working principles. Since it is evident from the morphological matrix that the working principles are complex technical systems, which combine multiple functions in their structure, they can represent modules that will be transformed and connected in the remaining stages of the design process into a modular structure of the machine for corn peeling. Five modules are defined through the function structure.

The principles of the solution, shown using sketches and schemes, are related to the partial functions in the morphological matrix. The selection of the best solution principles for each partial function was achieved by applying the method of technical and economic evaluation according to VDI 2225. By further combining them into five conceptual variants (working structures), the product variant was achieved.

The evaluation of working principles was achieved through a system of criteria, divided into three levels. From the system of criteria, a system of 23 goals, structured on four levels, was defined. Thus, each working principle was ranked according to the relative rating of the degree of goal fulfilment. The lack of this method was observed in the determination of the amount of the assessment, which evaluates the importance of a particular goal in the system of goals. The amount of the rating assigned results from the experience and knowledge of the designer, which leads to the conclusion that two different designers, depending on their knowledge and experience, can assign different values for the same goal they are analysing.

After the evaluation process, the following working principles achieved the best results, that is, they came close to the ideal solution: A WP1, B WP1, C WP1, D WP1, E WP1, E1 WP3, E3 WP1, F1 WP2 and G WP1. It is to be expected that these working principles should be connected in a single structure of the best conceptual variant. However, the principles of solutions E1 WP3 and F1 WP2 are not part of the structure of the conceptual variant that offers the most

possibilities for further design development (the first conceptual variant). The lack is reflected in the realization that it is not always possible to connect the working principles into a single structure according to the highest ratings achieved by evaluation process, but through precisely defined interfaces and geometric relationships. The experience and knowledge of the designer is crucial for this connection.

Further research aim is to develop a unique algorithm that will enable the application of repeatability of evaluation results, regardless of the level of experience and knowledge of the designer. In order to achieve such a goal, the development of a mathematical model and its implementation in the aforementioned algorithm of technical and economic evaluation would be approached.

7 REFERENCES

[1] Pahl, G., Beitz, W., Feldhusen, J. & Grote, K.-H. (2007). *Engineering Design: A Systematic Approach*. Springer, London. https://doi.org/10.1007/978-1-84628-319-2

[2] Ullman, D. G. (2010). *The Mechanical Design Process*. McGraw-Hill, New York.

[3] Dym, C. L., Little, P. & Orwin, E. J. (2014). *Engineering Design: A Project-Based Introduction*. 4th Edition, Wiley.

[4] Pastović, M. (2020). Procedura za razvoj konstrukcija modularnih i varijantnih proizvoda u fazi koncipiranja. *Doktorska disertacija*, Strojarski fakultet u Slavonskom Brodu. (in Croatian)

[5] Karakašić, M., Svalina, I., Novoselović, D., Samardžić, I., Glavaš, H. & Đokić, R. (2023). Application of the Functional Flow Diagrams in a Design of the Level Crossing Hydraulic Barrier Drive. *Tehnicki Glasnik, 17*(4), 554-565. https://doi.org/10.31803/tg-20230616191742

[6] Mao, X. & Sen, C. (2022). Toward Formal Qualitative Reasoning to Support Functional Decomposition. *Proceedings of the ASME 2022 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference IDETC-CIE2022*. https://doi.org/10.1115/DETC2022-89940

[7] Veljak, F. & Bojčetić, N. (2023). Functional modelling through Function Class Method: A case from DfAM domain. *Alexandria Engineering Journal, 66*, 191-209. https://doi.org/10.1016/j.aej.2022.12.001

[8] Nagel, R. L., Stone, R. B., Hutcheson, R. S., McAdams, D. A. & Donndelinger J. A. (2009). Function Design Framework (FDF): Integrated Process and Function Modeling for Complex Systems. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 273-286. https://doi.org/10.1115/DETC2008-49369

[9] Wu, J. C., Poppa, K., Leu, M. C. & Liu, X. F. (2012). Integrated function structure and object-oriented design framework. *Computers in Industry, 63*(5), 458-470. https://doi.org/10.1016/j.compind.2012.01.011

[10] Bojčetić, N., Veljak, F., Flegarić, S. & Štorga, M. (2020). Application for Product Functional Model Creation. *Tehnicki Vjesnik, 27*(3), 883-890. https://doi.org/10.17559/TV-20190923203841

[11] Meng, Z., Yong, C., Linfeng, C. & Youbai, X. (2019). A state-behavior-function model for functional modeling of multi-state systems. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 233*(7). https://doi.org/10.1177/0954406218791640

[12] Zadnik, Ž. (2011). Matrix of function and functionality in preliminary product development process. *PhD Thesis*, Univerza v Ljubljani, Fakulteta za strojništvo.

[13] Karakašić, M., Kljajin, M., Duhovnik, J. & Glavaš, H. (2018). Application of MFF Method in Conceptual Design. *Proceedings of VIII International Conference Industrial Engineering and Environmental Protection*, 72-80, ISBN 978-86-7672-309-6

[14] George, D. (2012). Concept Generation Using Morphological and Options Matrices. *PhD Thesis*, Graduate School of Clemson University. https://doi.org/10.1007/978-81-322-1050-4_16

[15] Asión-Suner, L. & López-Forniés, I. (2021). Analysis of Modular Design Applicable in Prosumer Scope. Guideline in the Creation of a New Modular Design Model. *Applied Sciences, 11*(22), 10620, 1-14. https://doi.org/10.3390/app112210620

[16] Gao, H. & Zhang, Y. (2020). Application of Modular Design Method in Product Design. *2020 International Conference on Intelligent Design (ICID)*. https://doi.org/10.1109/ICID52250.2020.00068

[17] Schuh, G., Rudolf, S. & Vogels, T. (2014). Development of modular product architectures. Procedia CIRP, 20, 120-125. https://doi.org/10.1016/j.procir.2014.05.042

[18] Jiu Mei, Z., Jing, X. & Bin, T. (2013). Introduction of Modular Design in the Conceptual Design of Refrigerators. *Applied Mechanics and Materials, 456*, 96-99. https://doi.org/10.4028/www.scientific.net/AMM.456

[19] Salonitis, K. (2014). Modular design for increasing assembly automation. *CIRP Annals, 63*(1), 189-192. https://doi.org/10.1016/j.cirp.2014.03.100

[20] Arvidsson, J. & Penndal, J. (2023). Modularity in Chassis Design. *Master's Thesis*, Chalmers University of Technology, Gothenburg, Sweden.

[21] Wei, Y., Qian, C. & Li, j. (2019). Modular Design of Mobile APP Interface Based on the Visual Flow. *Automatic Control and Computer Sciences, 53*, 56-62. https://doi.org/10.3103/S0146411619010127

[22] Sohn, J. & Chae, J. (2018). Revisiting Modular Design in a Contemporary Sociotechnical Context. *Meeting 4th NZAAR International Event Series on Natural and Built Environment Cities Sustainability and Advanced Engineering*, 49-56.

[23] Huang, G., Ceccarelli, M., Zhang, W. & Huang, Q. (2019). Modular Design Solutions of BIT Wheelchair for Motion Assistance. *IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. https://doi.org/10.1109/ARSO46408.2019.8948788

[24] Gershenson, J. K., Prasad, G. J. & Allamneni, S. (1999). Modular Product Design: A Life – Cycle View. *Journal of Integrated Design and Process Science, 3*(4), 13-26.

[25] Hirose, K. & Mishima, N. (2019). Eco-efficiency Evaluation of Modular Design Smartphones. *Procedia CIRP, 84*, 1054-1058. https://doi.org/10.1016/j.procir.2019.04.189

[26] You, Z.-H-. & Smith, S. (2016). A multi-objective modular design method for creating highly distinct independent modules. *Research in Engineering Design, 27*(2), 179-191. https://doi.org/10.1007/s00163-016-0213-8

[27] Kihlander, I. (2009). Decision making in concept phases – Towards improving product development processes. Licentiate Thesis, Royal Institute of Technology, Stockholm, Sweden.

[28] Lindley, J., Adams, R. & Wynn, L. (2017). Decision making in product design-bridging the gap between inception and reality. *Proceedings of the 19th International Conference on Engineering and Product Design Education (E&PDE17)*, 2012-2017.

[29] Marković, G., Zdravković, N., Karakašić, M. & Kolarević, M. (2020). Modified PROMETHEE Approach for Solving Multi-Criteria Location Problems with Complex Criteria Functions. *Tehnicki Vjesnik, 27*(1), 12-19. https://doi.org/10.17559/TV-20190225151515

[30] VDI 2225 (1997). VDI-Richtlinie 2225: Technisch-wirtschaftliches Konstruieren. VDI-Verlag, Düsseldorf.
[31] Short, T. & Harvey, J. (2008). Lightbulbs and nappies: sustainable development and customer perceptions. *International Journal of Sustainable Design, 1*(1), 13-28. https://doi.org/10.1504/IJSDES.2008.017054
[32] Taherdoost, H. (2017). Decision Making Using the Analytic Hierarchy Process (AHP): A Step by Step Approach. *International Journal of Economics and Management Systems, 2*, 244-246.
[33] Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences, 1*(1), 83-98. https://doi.org/10.1504/IJSSCI.2008.017590
[34] Ginting, R., Ishak, A., Malik, A. F. & Satrio, M. R. (2020). Product Development with Quality Function Deployment (QFD): A Literature Review. *IOP Conference Series: Materials Science and Engineering, 1003*(1), 1-6. https://doi.org/10.1088/1757-899X/1003/1/012022
[35] Reuter M. (2013). Technischer und wirtschaftlicher Vergleich von Herstellungsverfahren bei der Entwicklung von Kunststoffhohlkörpern in Automobilanwendungen. *Doktorale dissertation*, Fakultät für Ingenieurwissenschaften, Abteilung Maschinenbau der Universität Duisburg-Essen. (in German)

Authors' contacts:

Mirko Pastović, PhD
Croatian Sugar Industry d. d.,
Šećerana 63, 32270 Županja, Croatia
E-mail: mpastovic41@gmail.com

Mirko Karakašić, Full professor, PhD
(Corresponding author)
University of Slavonski Brod, Mechanical Engineering Faculty in Slavonski Brod,
Trg Ivane Brlić-Mažuranić 2, 35000 Slavonski Brod, Croatia
E-mail: mirko.karakasic@unisb.hr

Željko Ivandić, Full professor, PhD
University of Slavonski Brod, Mechanical Engineering Faculty in Slavonski Brod,
Trg Ivane Brlić-Mažuranić 2, 35000 Slavonski Brod, Croatia
E-mail: zivandic@unisb.hr

Ivan Grgić, PhD
University of Slavonski Brod, Mechanical Engineering Faculty in Slavonski Brod,
Trg Ivane Brlić-Mažuranić 2, 35000 Slavonski Brod, Croatia
E-mail: igrgic@unisb.hr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240514145648

Original scientific paper
Received: 2024-05-14, Accepted: 2024-06-22

Development and Optimization of a Differential Signal-Based Fabry-Perot Interferometer for Nanopositioning

Syuan-Cheng Chang, Chung-Ping Chang*, Yung-Cheng Wang

Abstract: In this study, we present the optimization of a Fabry-Perot interferometer with a differential signal utilized as the laser encoder to meet the stringent demands of nanopositioning. The proposed system aims to enhance stability and accuracy in nanopositioning applications by leveraging the common path structure and coaxial characteristics of Fabry-Perot interferometers. To improve the resolution of this system, an interpolation module is employed to increase the laser encoder resolution to 15.82 nm. Compared to the simulated interference signal from traditional Fabry-Perot interferometers, the differential interference signal proposed in this study is more sinusoidal, thus reducing errors in resolution subdivision. To verify the correspondence between the actual interference signal and the simulated one, a signal testing experiment is implemented in this study. Eventually, the experimental signal results demonstrate that the actual light intensity signals match the simulated results, indicating that this signal can be significantly beneficial for use as a laser encoder.

Keywords: differential signal processing; Fabry-Perot interferometer; interpolation module; laser encoder; nanopositioning

1 INTRODUCTION

In recent years, nanotechnology has become a pivotal field with widespread implications across industries, driving the demand for precise and reliable nanopositioning techniques. Among these techniques, laser interferometry particularly as an encoder in nanopositioning systems plays a crucial role due to its ability to provide high-resolution and accurate measurements at the nanoscale. The integration of laser interferometry into nanopositioning technologies has greatly enhanced the capabilities and functionalities of nanopositioners, enabling precise manipulation and control of objects with exceptional precision and accuracy [1].

However, commercial interferometers such as the Michelson interferometers have been commonly used for linear positioning calibration, they are susceptible to disturbances between the measurement and reference arms [2]. According to recent research, Fabry-Perot interferometers offer outstanding measuring stability due to their common-path structure which minimizes the effects of disturbances in the reference arm. Therefore, in some novel research, the Fabry-Perot interferometer has also been applied in the displacement measuring system, yielding remarkable results [3]. In this study, we have improved upon previous research and continued the development of this measurement technique by optimizing interferometric systems based on Fabry-Perot interferometers for nanopositioning technologies [4-6].

This study focuses on developing an interferometric system based on this common path structure to eliminate DC offset, making it suitable for precise positioning control in normal environments. The goal is to provide wide-ranging positioning control applications in precision mechanical industries.

2 THEORY AND PRINCIPLE

In this chapter, a brief introduction to displacement Fabry-Perot interferometer technology is provided. Additionally, the design of the proposed differential Fabry-Perot interferometer is demonstrated as follows.

2.1 Conventional Fabry-Perot Interferometer

Fabry-Perot interferometer was invented by French physicists Charles Fabry and Alfred Perot in the late 19th century [7]. The optical structure of the well-known FPI is depicted in Fig. 1. The incident beam (I_0) travels forwards and backwards repeatedly, dividing into numerous transmitted beams which interfere with each other. The resulting fringe signals are acquired by the detector. The correlative equations of the intensity distribution can be described by Eq. (1), where R, d, and λ represent reflectance of the mirror, the length of the cavity, and the laser wavelength, respectively. Since the measurement and reference beams travel through the same environment, the displacement measured is precisely defined by the distance between two parallel mirrors [8]. Hence, such an optical structure is feasible for precision length measurement.

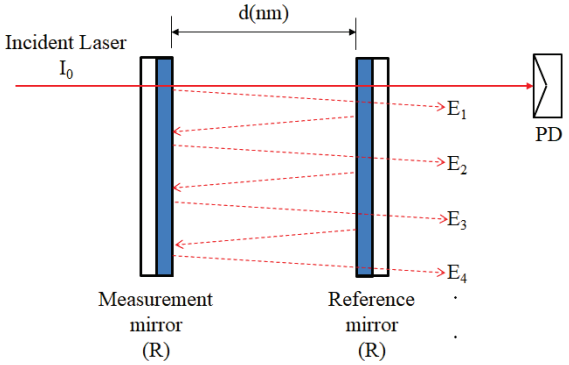


Figure 1 Conventional Fabry-Perot interferometer

$$I = I_0 \frac{(1-R)^2}{1+R^2-2\cdot R\cdot\cos\left(\frac{2\cdot\pi\cdot d}{\lambda\cdot\cos\alpha}\right)}. \quad (1)$$

Due to the conventional Fabry-Perot interferometer cannot determine the movement direction of the object. To address this limitation, a quadrature phase-shifted method based on polarizing technology has been proposed in previous research [9-11], as depicted in Fig. 2. In this polarizing phase-shifted method, a one-eighth waveplate is inserted into the optical cavity to generate quadrature phase-shifting. Consequently, the phase difference between the two interference signals is determined by the accuracy of the waveplate. The equations describing the intensity distribution of PD_S (I_S) and PD_P (I_P) are provided in Eq. (2) and (3).

The simulation result of the intensity distribution is shown in Fig. 3. Nevertheless, the signal offset of the polarizing phase-shifted method will be changed by the length of the cavity, the parallelism of the cavity, and the coherence of the laser source. It causes the interpolation error during the measurement procedure with the fine resolution.

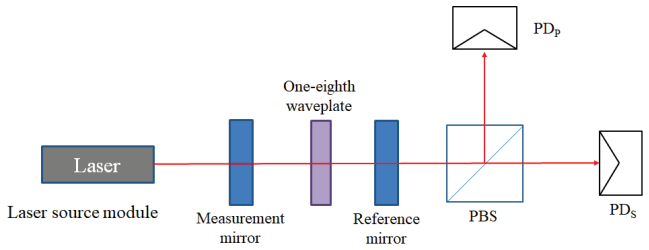


Figure 2 Quadrature phase-shift fiber FPI

$$I_s = E_s \cdot E_s^* = \frac{\frac{1}{2} \cdot A_0^2 \cdot T^2 \cdot T'}{1+R^2 \cdot T'^2 - 2 \cdot T' \cdot R \cdot \cos(2\delta)}, \quad (2)$$

$$I_p = E_p \cdot E_p^* = \frac{\frac{1}{2} \cdot A_0^2 \cdot T^2 \cdot T'}{1+R^2 \cdot T'^2 - 2 \cdot T' \cdot R \cdot \cos\left(2\delta + \frac{\pi}{2}\right)}. \quad (3)$$

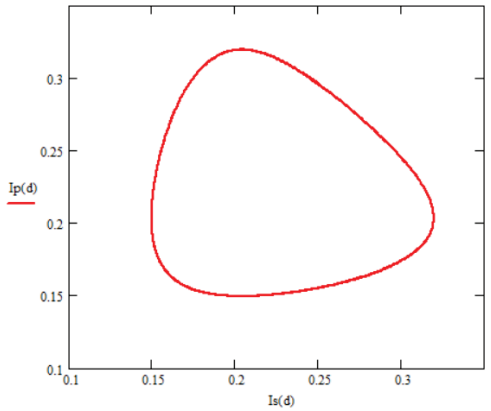


Figure 3 Intensity distribution of Quadrature phase-shift fiber FPI

2.2 Proposed Differential Fabry-Perot Interferometer

The optical structure of differential Fabry-Perot interferometer is shown in the Fig. 4. The incident laser passes through the optical cavity and the quarter-waveplate which is placed inside the cavity. The laser beams which include multiple right and left circular polarizing beams (E_1, E_2, E_3…) are divided into two circular polarizing beams by a beam splitter (BS). And each circular polarizing beam is divided into two multiple linear polarizing beams by a polarizing beam splitter (PBS). Those four interference signals can be detected by four photodetectors (PDs). The signals are divided by the same PBS and will have a phase shift of 180°. By rotating the half-waveplate, the phase shift of the interference signals can be adjusted into the proper position. Based on the previous research, the major purpose of the proposed structure can reduce the DC drift of the interference signal.

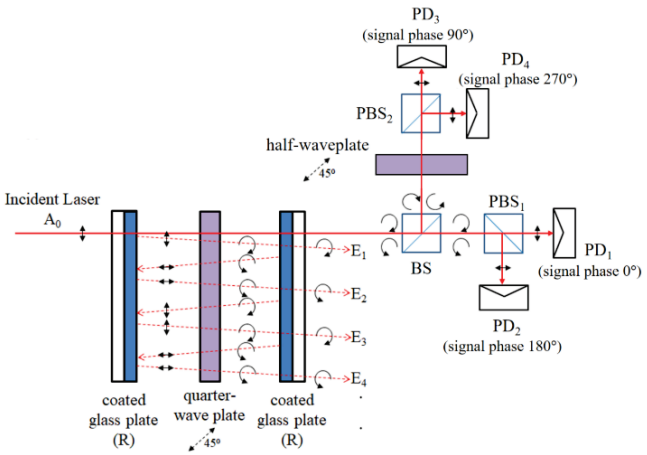


Figure 4 Differential Fabry-Perot interferometer

The equations of the transmitted electric field of the laser beams are expressed as E_1 to E_N (Eq. 4 to Eq. 7) and that of the whole interference beam is denoted with E_N (Eq. 7), where N is the order of the transmitted light beams. Here, $\boldsymbol{A}_0$, T, T_c, R, $\boldsymbol{WP}$, $OP(d)$ and $OPD(d)$ are the matrices and coefficients involved in the calculation of the electric field. $\boldsymbol{A}_0$ represents the matrix of the incident beam with linear polarization from the horizontal axis; T represents the transmittance of the coated glass plate; T_c represents the equivalent transmittance of the cavity; R represents the reflectance of the coated glass plate; WP represents the matrix of the quarter-waveplate with a 45-degree angle from the horizontal axis; $OP(d)$ represents the single-pass optical path within the optical cavity; $OPD(d)$ represents the optical path out of the optical cavity.

$$E_1 = \sqrt{T} \cdot \left(\sqrt{T_c} \cdot OPD(d) \cdot WP \cdot \sqrt{R}\right)^1 \cdot OP(d) \cdot \sqrt{T} \cdot \boldsymbol{A}_0, \quad (4)$$

$$E_2 = \sqrt{T} \cdot \left(\sqrt{T_c} \cdot OPD(d) \cdot WP \cdot \sqrt{R}\right)^3 \cdot OP(d) \cdot \sqrt{T} \cdot \boldsymbol{A}_0, \quad (5)$$

$$E_3 = \sqrt{T} \cdot \left(\sqrt{T_c} \cdot OPD(d) \cdot WP \cdot \sqrt{R}\right)^5 \cdot OP(d) \cdot \sqrt{T} \cdot \boldsymbol{A}_0, \quad (6)$$

$$E_N = \sqrt{T} \cdot \left(\sqrt{T_c} \cdot OPD(d) \cdot WP \cdot \sqrt{R}\right)^{2N-1} \cdot OP(d) \cdot \sqrt{T} \cdot \boldsymbol{A}_0. \quad (7)$$

In order to calculate the sum of the electric field (E_N), the odd term (E_{odd}) and the even term (E_{even}) of the electric field are separated for the calculation, as shown in Eq. (8) and (9).

$$E_N = E_{\text{odd}} + E_{\text{even}}, \tag{8}$$

$$E_N = \\ = \left(E_1 + E_3 + E_5 + \ldots + E_{2N-1}\right) + \left(E_2 + E_4 + E_6 + \ldots + E_{2N}\right). \tag{9}$$

Eventually, the total electric field obtained by summing the odd term and even term of the electric field is shown in Eq. (10).

$$E_N = \left\{ \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{i}{\sqrt{2}} \end{bmatrix} \cdot T \cdot e^{i(\omega t + kx)} \cdot \frac{\sqrt{T} \cdot \sqrt{R} \cdot e^{i\frac{2\pi d}{\lambda}}}{1 - \left(\sqrt{T} \cdot \sqrt{R} \cdot e^{i\frac{2\pi d}{\lambda}}\right)^4} \right\} + \\ + \left\{ \begin{bmatrix} \frac{-1}{\sqrt{2}} \\ \frac{i}{\sqrt{2}} \end{bmatrix} \cdot T \cdot e^{i(\omega t + kx)} \cdot \frac{\left(\sqrt{T} \cdot \sqrt{R} \cdot e^{i\frac{2\pi d}{\lambda}}\right)^3}{1 - \left(\sqrt{T} \cdot \sqrt{R} \cdot e^{i\frac{2\pi d}{\lambda}}\right)^4} \right\}. \tag{10}$$

When E_N has been obtained, the intensity of 0°, 45°, 90°, and 135° of the polarizer axis can be acquired by Eq. (11).

$$I = E \cdot E^*. \tag{11}$$

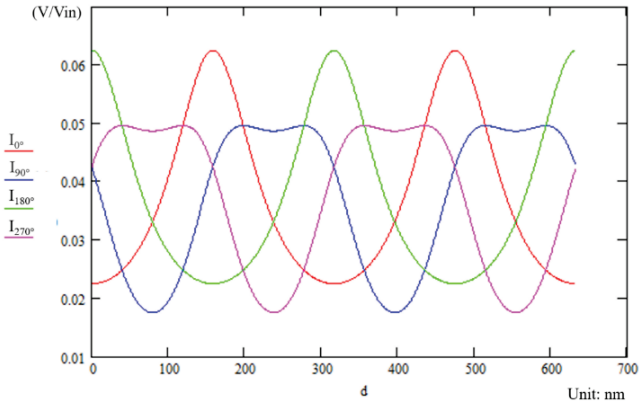


Figure 5 Simulation of the intensities in each phase

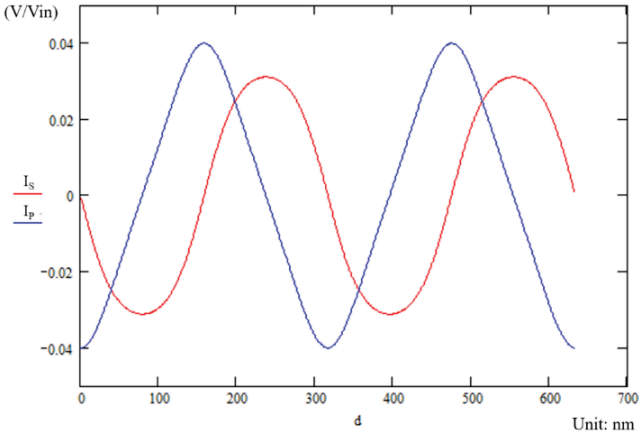
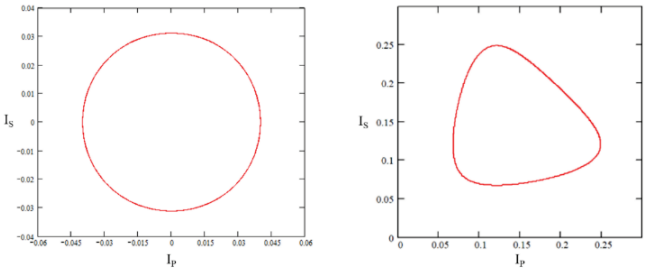


Figure 6 Simulation of differential signal

The simulation results of the intensity profiles for I (0°), I (45°), I (90°), and I (135°) are presented in Fig. 5. The processing principle of the differential signal involves subtracting one signal from the other, each having the same amplitude but opposite phase. Consequently, simulations of the differential signal and the Lissajous figure are depicted in Fig. 6 and Fig. 7(a), respectively. Comparing Fig. 7(a) and Fig. 7(b) can show that the simulation results of differential Fabry-Perot interferometer is a sinusoidal signal, whereas the simulation results of the QPSK FPI do not exhibit such sinusoidal signal.



(a) R=25 %, T'= 87 % (b) R=25 %, T'= 87 %

Figure 7 Simulation of Lissajous figure (a) Simulation results of Differential FPI (b) Simulation results of QPSK FPI

3 DESIGN OF PROPOSED POSITIOINING SYSTEM

The positioning system proposed in this article utilizes the interference signal of the differetial Fabry-Pert interferometer as a laser encoder. This system consists of three main components: the light source unit (laser source and polarizer), the measurement unit (reference mirror, quarter-waveplate, and measurement mirror), and the sensing unit (BS, two PBS, half-waveplate, and four photodiodes), as illustrated in Fig. 8. Since the measurement mirror is installed on a linear stage, any displacement generated can be determined by oberving the amplitude of the light intensity received by the photodetecors (PD_1 to PD_4).

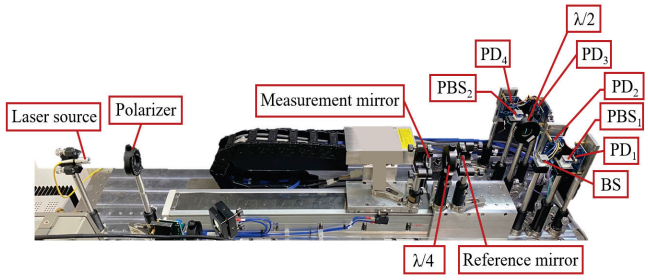


Figure 8 Design of proposed positioning system

The interpolation module is utilized in this study to convert the input orthogonal sine wave signal into a square wave signal. The resolution of the signal interpolation can be adjusted by modifying the circuit, and a interpolation reolusiton of 20 times is selected in this study. This means that within one sine wave cycle, five square waves are generated. Using the common fourfold frequency counting method in optical feedback signal processing, 20 counts can be obtained, as shown on Fig. 9. With an optical interference

cycle of 316.4 nm, the positioning resolution can be deduced as 15.82 nm. The interpolation module can configure through external voltage to set the reference baseline of the signal. The specific procedure involves inputting the two orthogonal interference signals into the interpolation module. This yields the differential signal A+, A–, B+, B– which are utilized as the laser encoder to achieve the positioning objective.

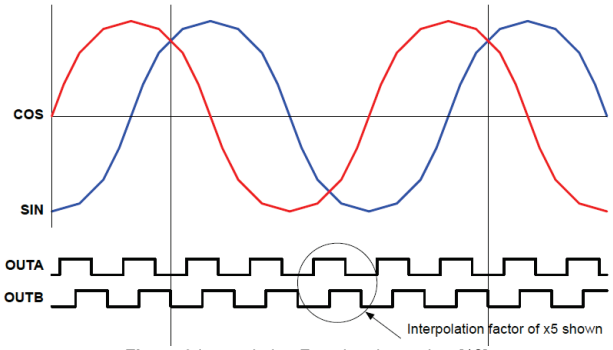


Figure 9 Interpolation Functional overview [12]

4 EXPERIMENTAL SIGNAL RESULTS

In order to verify the correspondence between the actual interference signal and the simulated one, experiments were conducted to measure the interference signal. According to the optical structure of the differential Fabry-Perot interferometer presented in Fig. 8, interference signal experiment was conducted. The light intensity signals detected by the photodetectors (PDs) are depicted in Fig. 9. It can be observed that the actual light intensity signals match the simulation shown in Fig. 5. The time-domain signal results from PD_1 to PD_4 are illustrated in Fig. 9(a) and 9(b), respectively.

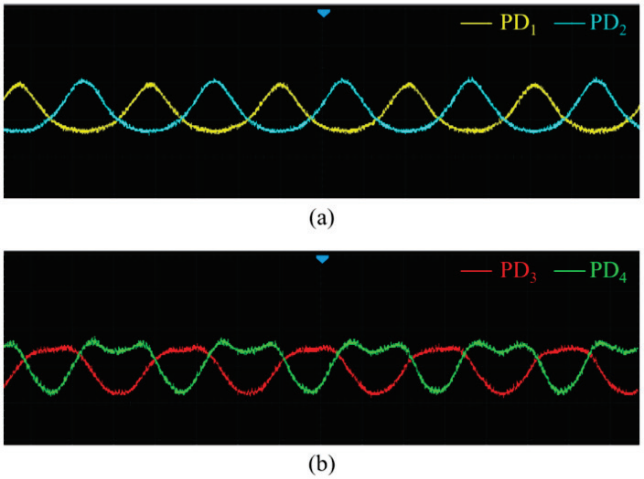


Figure 10 Interference signal from PD_1 to PD_4
(a) time-domain signal of PD_1 and PD_2, (b) time-domain signal of PD_3 and PD_4

By utilizing the principle of differential signal processing, the signals from PD_1 to PD_4 are subtracted to obtain the differential interference signal as shown in Figure 11. The light intensity signals from PD_1 and PD_2, with a phase difference of 180 degrees, are subtracted to obtain the sin signal. Similarly, subtracting the signals from PD_3 and PD_4, also with a phase difference of 180 degrees, yields the cos signal. The Lissajous figure can be represented by Fig. 12. These signals can represent the sinusoidal signal generated by the displacement of the linear stage.

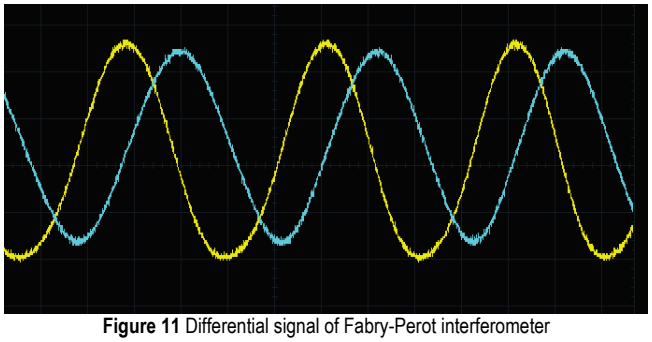


Figure 11 Differential signal of Fabry-Perot interferometer

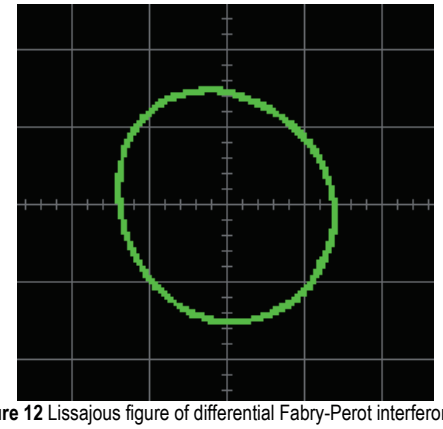


Figure 12 Lissajous figure of differential Fabry-Perot interferometer

After obtaining the differential signal through the principle of differential processing, signal processing can be conducted using the interpolation module proposed in this study. The differential signal is subjected to analog-to-digital conversion and further subdivided based on the selected subdivision ratio through this module. In this study, the resolution is subdivided by a factor of 20, thereby enhancing the resolution to 15.82 nm. If higher specifications are required, further subdivision can be performed according to the needs. The resulting signal after conversion can be represented by Fig. 13.

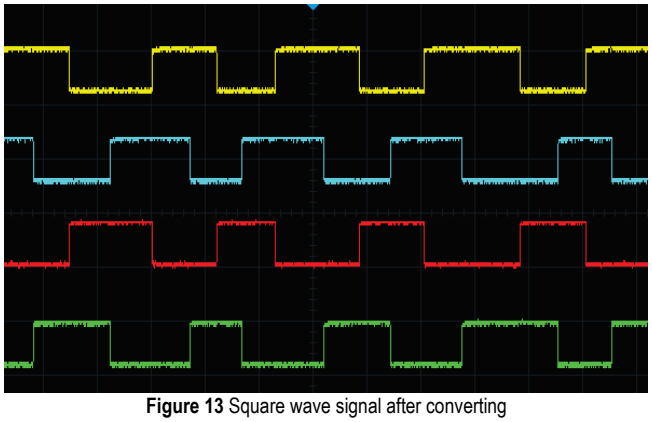


Figure 13 Square wave signal after converting

This square wave signal can be utilized for precision measurement purposes, allowing for the determination of the current displacement of the moving stage by the counting values. Additionally, this signal can serve as a laser encoder, providing feedback to the drive controller for feedback control. The integration of this signal as a laser encoder enhances the accuracy and reliability of feedback control systems, contributing to the overall advancement and competitiveness of industrial processes.

5 CONCLUSION

In this study, the optimization of interferometric system based on Fabry-Perot interferometer for advancing nanopositioning technologies in precision mechanical industries. This study has focused on developing a common-path structured interferometric system to eliminate DC offset, thereby enabling precise positioning control in normal environments. Furthermore, the utilization of polarizing phase-shifted methods and differential Fabry-Perot interferometers addresses limitations in determining movement direction and reduces signal drift, thereby enhancing measurement accuracy. Experimental validation demonstrates the correspondence between actual interference signals and simulations, thus validating the effectiveness of the proposed methods. Through signal processing and interpolation techniques, the resolution of the differential signal is enhanced, offering improved precision in displacement measurement and feedback control. Overall, the integration of interferometric systems as a laser encoder enhances the capabilities and reliability of nanopositioning systems, contributing to advancements in industrial processes and competitiveness.

6 REFERENCES

[1] Stuerzebecher, L., Fuchs, F., Zeitner, U. D. & Tuennermann, A. (2015). High-resolution proximity lithography for nano-optical components. *Microelectronic Engineering, 132*, 120-134. https://doi.org/10.1016/j.mee.2014.10.010

[2] Jaeger, G. (2010). Limitation of precision length measurements based on interfermeters. *Measurement*, *43*(5), 652-658. https://doi.org/10.1016/j.measurement.2009.12.030

[3] Chang, C. P., Tung, P. C., Shyu, L. H., Wang, Y. C., & Manske, E. (2013). Modified Fabry-Perot interferometer for displacement measurement in ultra large measuring range. *Review of Scientific Instruments*, *84*(5). https://doi.org/10.1063/1.4803672

[4] Chang, S. C., Chang, C. P., Wang, Y. C. & You, Z. F. (2022). Linear Displacement and Straightness Measurement by Fabry-Perot Interferometer Integrated with an Optoelectronic Module. *Tehnički glasnik, 16*(3), 420-425. https://doi.org/10.31803/tg-20220424124800

[5] Shyu, L. H., Chang, C. P. & Wang, Y. C. (2011). Influence of Intensity Loss in the Cavity of a Folded Fabry-Perot Interferometer on Interferometric Signals. *Review of Scientific instrument, 82*(6), 063103. https://doi.org/10.1063/1.3596451

[6] Chang, C. P., Tu, T. C., Huang, S. R., Wang, Y. C. & Chang, S. C. (2021). Development of the Heterodyne Laser Encoder System for the X-Y Positioning stage. *Sensors, 21*(17), 5775. https://doi.org/10.3390/s21175775

[7] Fabry, C. (1899). Theorie et applications d'une nouvelle methods de spectroscopie intereferentielle. *Ann. Chim. Ser. 7(16)*, 115-144.

[8] Lawall, J. R. (2005). Fabry–Perot metrology for displacements up to 50 mm. *JOSA A*, *22*(12), 2786-2798. https://doi.org/10.1364/JOSAA.22.002786

[9] Chang, C. P., Tung, P. C., Shyu, L. H., Wang, Y. C., & Manske, E. (2013). Multi-interferometric displacement measurement system with variable measurement mirrors. *Applied Optics*, *52*(17), 3902-3909. https://doi.org/10.1364/AO.52.003902

[10] Chang, C. P., Shih, Y. C., Chang, S. C. & Wang, Y. C. (2019). Laser encoder system for XY positioning stage. *Mechatronics*, *63*, 102274. https://doi.org/10.1016/j.mechatronics.2019.102274

[11] iC Haus Interpolation Products (2017, October 19, p. 15), online resource. Available from: https://www.ichaus.de/wp-content/uploads/TW8_datasheet_D3en.pdf

Authors' contacts:

Syuan-Cheng Chang, Assistant Professor
National Yunlin University of Science and Technology,
123 University Road, Section 3, Douliou, Yunlin 64002, Taiwan
tso1147279@gmail.com

Chung-Ping Chang, Associate Professor
(Corresponding author)
National Chiayi University,
300 Syuefu Road, Chiayi 600355, Taiwan
cpchang@mail.ncyu.edu.tw

Yung-Cheng Wang, Professor
National Yunlin University of Science and Technology,
123 University Road, Section 3, Douliou, Yunlin 64002, Taiwan
wangyc@yuntech.edu.tw

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240516154449

Original scientific paper
Received: 2024-05-16, Accepted: 2024-05-22

Bridging Technology and Healthcare: The Impact of AI in Surgical Instrument Classification

Matea Grdić, Sven Maričić*, Damjana Mihaljević, Lucia Labinjan

Abstract: Emerging technologies have sparked a surge in demand for artificial intelligence (AI), aimed at optimising various industries and simplifying daily tasks. In response to the rapid evolution of technology and the evolving requirements of surgical practices, we introduce an innovative application leveraging machine learning. Utilising Cloud Annotations, Collaboratory, and Node-RED, we developed a platform for accurate surgical instrument classification. By training our model on a diverse set of instruments and enabling user-friendly image capture, our application stands poised to revolutionise surgical workflows. This paper highlights the role of AI in healthcare and outlines the potential of our application to enhance surgical practices, improve instrument recognition, and contribute to patient care advancement. We also address challenges and opportunities in integrating AI into healthcare while proposing avenues for future research and development.

Keywords: artificial intelligence; healthcare innovation; image recognition; machine learning; Node-RED

1 INTRODUCTION

The rapid growth of innovative technologies reveals a wide range of breakthroughs that have the potential to achieve unprecedented levels of development and practical implementation [1]. Industries such as artificial intelligence (AI), machine learning, biotechnology, and cybernetics are leading the way in this emerging period, bringing about significant advances and reshaping conventional frameworks. The inherent potential of these technologies to enhance socio-economic standards worldwide is a key factor in their revolutionary impact [2, 3]. Significantly, within the domain of healthcare, they provide a potential opportunity for augmenting diagnostic precision and spearheading innovative approaches to illness prevention and treatment. The context of innovation provides the foundation for the creation of an application that aims to identify and classify surgical tools based on camera input. This program application utilizes artificial intelligence and machine learning to optimize surgical processes by assuring the prompt availability and precise identification of appropriate equipment [4]. This undertaking not only encompasses the fundamental concept of utilizing developing technologies to address certain obstacles but also showcases the capacity of these technologies to enhance medical methodologies and enhance patient care results.

This paper presents the development of an application capable of recognizing surgical instruments [4] through camera input and categorising them into respective classes. The machine learning model was trained using the Python programming language and TensorFlow library in the Colaboratory development environment. Subsequently, the model was converted into the TensorFlow.js format for seamless integration into web browsers. The application's graphical interface was built using the Node-RED development environment. The application successfully implements the classification of specific surgical instruments, with future enhancements aiming at detection and localization to enable recognition of multiple instruments within a single photograph. Furthermore, suggestions for future improvements include expanding the repertoire of recognized surgical instruments. The development tools employed include Cloud Annotations, Colaboratory, and Node-RED. Cloud Annotations facilitated the annotation of surgical instruments, while Colaboratory was used for model development and training. Node-RED enabled the creation of the graphical user interface, offering users the convenience of capturing images directly for recognition or uploading images from their computers. The model was trained on five different surgical instruments: curved surgical scissors, straight surgical scissors, scalpel, surgical forceps, and artery forceps. Additionally, the paper outlines the hardware setup consisting of a Raspberry Pi computer connected to a camera, providing a cost-effective solution for edge computing. The software environment involved Colaboratory for model development, Node-RED for interface design, and Cloud Annotations for dataset annotation and storage. The resulting application bridges the gap between machine learning algorithms and practical surgical instrument recognition, with potential applications in healthcare settings. Previous research efforts have delved into the specific area of computer vision and machine learning for surgical instrument recognition. These studies have highlighted the potential [1-3] of mentioned technologies in streamlining surgical workflows and reducing the risk of errors associated with human intervention.

2 HEALTHCARE AND ARTIFICIAL INTELLIGENCE

Artificial intelligence has emerged with a significant impact technology in modern healthcare system, revolutionizing diagnostic and treatment methodologies. With the proliferation of medical data and advancements in big data diagnostic techniques, AI has seamlessly integrated into healthcare systems, heralding a new era of efficiency and precision [5-9]. Through its adept analysis of information, medical records, and systems, AI augments digital automation, yielding swifter and more consistent outcomes while aiding physicians in achieving superior results [10].

However, despite its transformative potential, skepticism persists among practitioners regarding AI's future role in primary care. Many express concerns over its perceived lack

of empathy and ethical dilemmas, reflecting a nuanced dialogue surrounding the integration of AI into the healthcare landscape [11]. This scepticism underscores the importance of addressing not only the technical capabilities of AI but also its ethical and interpersonal implications as it continues to shape the future of healthcare delivery.

The primary objective of this research is to enhance the classification of surgical instruments, thereby streamlining the workflows of medical personnel and improving operational efficiency in clinical settings.

The incorporation of machine learning models for identifying and categorizing surgical instruments marks a significant advancement in the operational efficiency of surgical settings. By automating instrument recognition, these models considerably reduce the time required by healthcare personnel to locate necessary tools, thereby facilitating the optimization of surgical procedures. This is especially critical during urgent operations where time efficiency is paramount, enhancing both the safety and effectiveness of surgical interventions. Machine learning technologies are proving invaluable as educational tools within the medical field, particularly for training medical and nursing staff. These technologies enable quick and precise identification of various surgical instruments, enhancing the training process and ensuring medical trainees are better prepared for real-life clinical environments. These models in healthcare systems can reduce the cognitive load on medical staff, allowing them to concentrate more on patient care and less on ancillary tasks.

The integration of such sophisticated machine learning models not only streamlines surgical operations but also significantly improves the training and preparedness of medical personnel, thus elevating the standard of patient care across healthcare settings. There are key constraints that affect the implementation of new technologies in clinical settings. These include challenges in compatibility and integration with existing systems, the need for high computing resources, ensuring data security and privacy in accordance with regulations, and the need for regular maintenance and technical support. Understanding and addressing these limitations is critical to the successful integration of IT tools into medical practices. The use of this application in the identification and classification of surgical instruments represents a significant advance in the technological support of operative procedures. However, the role of nurses who assist surgeons during operations remains irreplaceable. Although the app can improve efficiency and accuracy in identifying and preparing surgical instruments, nurses have a wide range of responsibilities that include managing the operating room.

The surgical instrument classification application described in this article potentially improves surgical operations by enabling rapid identification of required tools, reducing operative time and risks of human error. It also may serve as an educational tool for medical staff, improving in a certain way instrument inventory management in healthcare facilities. Also, in situations where surgeons perform remote operations using robotic systems, the application can possibly provide additional support in the identification and selection of the necessary instruments in real time.

2.1 Materials and Methods

The study employed a combination of implemented additive manufactured housing with hardware and software system: machine learning techniques and edge computing to develop a surgical instrument recognition application. It involved data collection, model training, and application development phases, each of which is described below.

Hardware and 3D model:

- 3D printed special housing to implement easy access of surgical instruments.
- Raspberry Pi 3 Model B: This credit card-sized computer served as the main processing unit for edge computing.
- Camera: A camera was connected to the Raspberry Pi for capturing images of surgical instruments. Compatible with all models of Raspberry Pi using the CSI (Camera Serial Interface) port. Supports up to 3280 x 2464 pixels for stills and video resolutions of 1080p at 30 fps, 720p at 60 fps, and VGA at 90 fps. Uses a 15 cm ribbon cable for connection to the Raspberry Pi board, allowing for flexible integration into various setups.
- MicroSD Card: A 15 GB microSD card with Raspbian GNU Linux 10 operating system was used as storage.

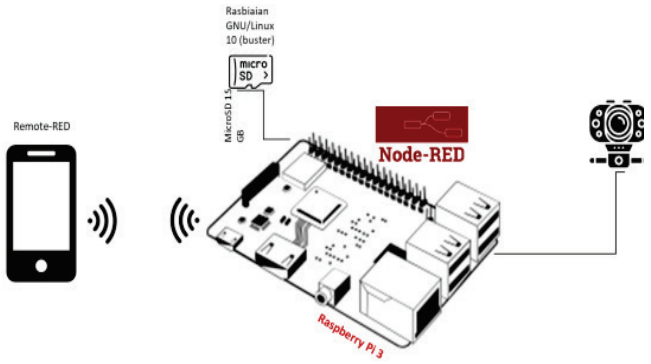


Figure 1 Schematic representation of the development environment [12]

Software:

- Python: The primary programming language used for model training.
- JavaScript: employed to develop the user interface, enabling interactive and dynamic interactions within application.
- TensorFlow: An open-source machine learning framework utilized for developing and training the model.
- Colaboratory: An integrated development environment (IDE) provided by Google for Python-based machine learning tasks, used for model training.
- Node-RED: A flow-based development tool used for creating the graphical user interface of the application.
- Cloud Annotations: A tool employed for annotating surgical instrument images to create a labelled dataset.

Procedures Followed:

Data collection: The dataset used for model training comprised 1500 original images, covering five distinct categories of surgical instruments: bandage and plaster scissors, surgical forceps, scalpel, straight surgical scissors, and artery forceps. Each category represents a different type of surgical tool commonly utilised in medical procedures, ensuring diversity in the dataset (Fig. 2). The overall accuracy is determined by adding the count of accurately identified values and dividing it by the entire count of values.

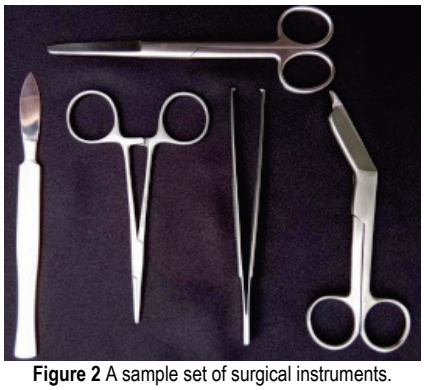


Figure 2 A sample set of surgical instruments.

Model training: To train the object detection model, a dataset with labelled bounding boxes was prepared using IBM Cloud Annotations tool. This tool facilitated systematic labelling of images, ensuring efficient data annotation and storage in IBM Cloud Object Storage. The Machine Learning model was created using Convolutional Neural Networks (CNN), TensorFlow and MobileNet CNN for analyzing visual images. Convolutional Neural Networks (CNNs) are a type of deep neural network primarily used for analyzing visual imagery. Also, it is worth to mentioned that the confusion matrix is important tool for enhancing Convolutional Neural Networks (CNNs) as it provides a detailed breakdown of accuracy and mistakes for each class, allowing for focused improvements. Additionally, it aids in the improvement of the network by fine-tuning thresholds, hence enhancing performance in practical applications. TensorFlow, a free, open-source machine learning library, excels in training and deploying neural networks, featuring tools for symbolic math and differentiable programming. MobileNet, a TensorFlow component, is an optimized CNN for mobile vision applications with low computational demands. Neural networks, central to TensorFlow, are algorithms that mimic the brain's function to detect data patterns and can be either organic or synthetic neuron systems.

The training process involved accessing the labelled dataset, installing necessary dependencies such as TensorFlow Object Detection API, and initiating training in Colaboratory. Once trained, the model was converted to TensorFlow.js format for web deployment and downloaded for integration into the application. Testing of the trained model confirmed its efficacy, with high accuracy in detecting surgical instruments (Fig 3). In order to further increase the accuracy, the model was trained several times with an increased number of images in the dataset. In addition, controlled environmental lighting was also identified as a key factor influencing the improvement of model performance. Adjusting these parameters enabled more precise identification and classification of surgical instruments, thereby minimizing potential room for error and increasing overall efficiency in the operating room.

Figure 3 Data training procedure

Application development: For the application development phase, the trained model underwent conversion into TensorFlow.js format, facilitating seamless integration into web browsers. Leveraging Node-RED, we crafted a user-friendly graphical interface enabling users to either capture images directly for recognition or upload images from their computers. The application was meticulously designed to process these images, utilizing the trained model, showcasing the recognized instrument along with its classification in an intuitive manner (Fig. 4).

Deployment: The developed application was deployed on the Raspberry Pi, enabling real-time instrument recognition at the edge. Users could access the application interface via smartphones using the Remote-RED application.

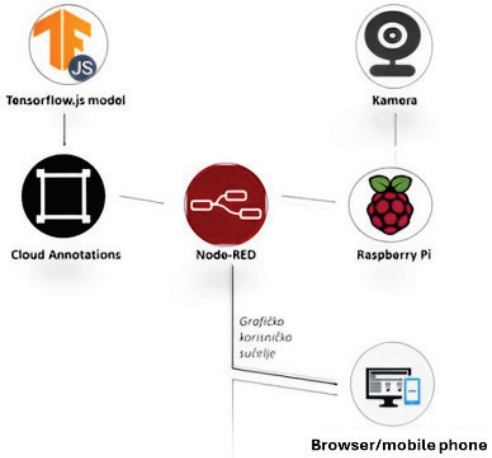


Figure 4 Schematic representation of the system

3 INTERFACE MODELLING

The development of the application for the classification of surgical instruments in the Node-RED environment is divided into several key functional subgroups that enable efficient image management and processing. These functionalities include:

Flow for uploading images from a computer: Allows users to upload images from a computer to the application, which is the first step in the processing process.

- **Direct image capture**: Enables direct image capture via a built-in or connected camera, allowing users to instantly capture and send images of surgical instruments for analysis.
- **Display of uploaded or captured images on the application interface**: After uploading or capturing, images are displayed on the application's user interface, providing visual feedback to users.
- **Saving captured images to the computer:** This functionality enables saving captured images to the local computer, thus ensuring that all data is safely archived.
- **Detection and display of detection results:** After image processing, the application identifies and classifies surgical instruments and displays the detection results on the interface.
- **Display of time and date:** The interface also displays the current time and date, which can be useful for records and documentation in clinical settings.
- **Remote access:** Allows users to remotely access the application, which is especially useful in situations where fast and efficient remote diagnostics are required.

These functionalities within the Node-RED environment form the basis for an efficient and flexible application that can be adapted to the specific needs of medical and research institutions.

4 RESULTS

The interface of the application is depicted in Fig. 5, enabling users to capture a picture or upload it from their computer. The results of the recognized class, along with their probabilities, are displayed in a table.

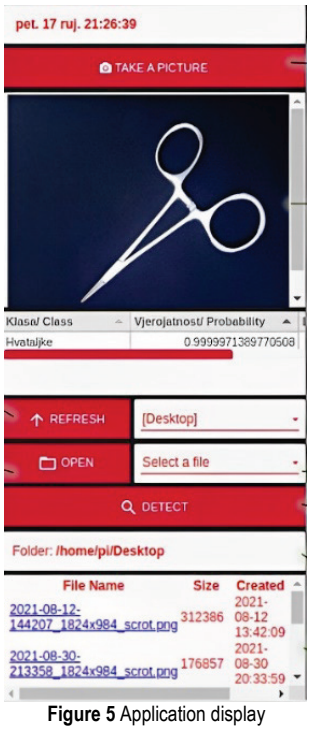


Figure 5 Application display

Additionally, remote access is facilitated through a gateway node, allowing users to access the website locally or remotely from a mobile application. By downloading the mobile application Remote-RED, available on Google Play, users can access the application from a remote location.

During the implementation of the project, we encountered significant challenges regarding the functionality of the software and hardware components. The main problem we faced was the limitation of our system to perform only the classification of surgical instruments, without the possibility of detection. This means that our current model can identify the type of instrument within a given pattern but cannot locate the presence or position of an instrument within an image. This limitation represents a significant obstacle for further application in real surgical environments, where precise detection and localization of instruments is crucial for successful operational support. In response to this problem, we plan to develop and integrate more advanced object detection algorithms into future versions of our system. These improvements will enable more precise tracking and identification of surgical instruments during operations, thereby increasing their operational efficiency and patient safety. These upgrades represent a key focus of our future research and development efforts.

5 CONCLUSION

In this study, we have presented an innovative application leveraging machine learning techniques for the classification of surgical instruments. By combining edge computing with image recognition algorithms, we developed a platform capable of accurately categorising surgical instruments in real-time. Our approach involved the utilisation of Python programming language, TensorFlow framework, and Node-RED development environment for model training, conversion, and interface design, respectively. Through systematic data collection, model training, and application development phases, we successfully created a user-friendly interface enabling both image capture and upload for instrument recognition. The deployment of the application on Raspberry Pi offers a cost-effective solution for edge computing, ensuring accessibility in healthcare settings. The results demonstrate the effectiveness of our approach in recognizing surgical instruments with high accuracy, paving the way for enhanced surgical workflows and improved patient care. Future enhancements will focus on expanding the repertoire of recognized instruments and integrating additional functionalities to address the evolving needs of surgical practices. Overall, our study contributes to the advancement of surgical practices through the integration of machine learning technologies, fostering innovation in healthcare delivery. The study on AI-based surgical instrument recognition holds several notable advantages for healthcare. By integrating machine learning with surgical practices, it offers improved training opportunities, minimises errors during procedures and enhances patient safety. Additionally, the application streamlines surgical workflows, leading to

increased efficiency and potentially shorter operation times. Its deployment on edge computing platforms ensures cost-effectiveness, making it accessible to a wide range of healthcare facilities. Moreover, the tool's capability for remote access enables surgeons to utilise it beyond the operating room, facilitating telemedicine and expert consultations. With the flexibility to customise and scale according to specific surgical needs, this study represents a significant stride towards advancing surgical practices and patient care. By combining machine learning with edge computing and web technologies, this study showcases a scalable and adaptable approach to deploying intelligent applications in medical environments. It sets the stage for further advancements in medical technology and opens doors to innovative solutions for healthcare challenges. In summary, integrating advanced machine learning models into healthcare practices not only streamlines operational procedures but also significantly enhances educational outcomes for medical personnel, thereby reinforcing the overall quality of patient care. This work establishes the scientific foundations for future research and enables potential patenting and practical application of observed innovations, which are currently not implemented in healthcare.

6 REFERENCES

[1] Harry, A. (2023). The future of medicine: harnessing the power of AI for revolutionizing healthcare. *International Journal of Multidisciplinary Sciences and Advanced Technology, 2*(1), 36-47. https://doi.org/10.47709/ijmdsa.v2i1.2395

[2] Rozario, M., Zainuddin, A. & Gamage, S. (2021). Artificial Intelligence and Machine learning in the Healthcare Sector: A Review. *Malaysian Journal of Science and Advanced Technology, 1*(3), 89-96. https://doi.org/10.56532/mjsat.v1i3.18

[3] Gupta, N. (2022). Machine Learning Applications in Healthcare. *The 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO2022)*, 1-6. https://doi.org/10.1109 /ICRITO56286.2022.9964865

[4] Grdić, M. (2021). Izrada 3D modela za strojno učenje kirurških instrumenata. *Diplomski rad*, Sveučilište Jurja Dobrile u Puli. https://urn.nsk.hr/urn:nbn:hr:137:571618 (in Croatian)

[5] Harry, A. (2023). Transforming Patient Care: The Role of Artificial Intelligence in Healthcare; A mini Review. *Bullet: Jurnal Multidisiplin Ilmu, 2*(3), 530-533. https://journal.mediapublikasi.id/index.php/bullet/article/view/2760

[6] Supriya, M. & Chattu, V. (2021). A Review of Artificial Intelligence, Big Data, and Blockchain Technology Applications in Medicine and Global Health. *Big Data Cogn. Comput., 5*, 41. https://doi.org/10.3390/bdcc5030041

[7] Datta, S., Barua, R. & Das, J. (2020). Application of Artificial Intelligence in Modern Healthcare System. IntechOpen. https://doi.org10.5772/intechopen.90454

[8] Ball, H. (2021). Improving Healthcare Cost, Quality, and Access through Artificial Intelligence and Machine Learning Applications. *Journal of Healthcare Management, 66*, 271-279. https://doi.org/10.1097/JHM-D-21-00149

[9] Rybin, S. & Ripka, D. (2023). Overview of Machine Learning Technologies in Medicine. *Seminar on Digital Medical and Environmental Systems and Tools (DMEST2023)*, 123-125. https://doi.org/10.1109/DMEST60476.2023.10339562

[10] Haleem, A., Javaid, M. & Khan, I. H. (2019). Current status and applications of Artificial Intelligence (AI) in medical field: An overview. *Current Medicine Research and Practice, 9*(6), 231-237. https://doi.org/10.1016/j.cmrp.2019.11.005

[11] Blease, C., Kaptchuk, T. J., Bernstein, M. H., Mandl, K. D., Halamka, J. & DesRoches, C. M. (2019). Artificial intelligence and the Future of Primary Care: Exploratory Qualitative study of UK general practitioners' views. *Journal of Medical Internet Research*, *21*(3). https://doi.org/10.2196/12802

[12] Node-RED. (n.d.). Node-RED. Retrieved February 20, 2024, from https://nodered.org/

Authors' contacts:

Matea Grdić, system integrator
Juraj Dobrila University of Pula,
Faculty of Engineering,
Zagrebačka ul. 30, 52100 Pula, Croatia
matea.grdic1@gmail.com

Sven Maričić, PhD, Associate Professor
(Corresponding author)
Juraj Dobrila University of Pula,
Faculty of Engineering,
Laboratory for Robotics and Artificial Intelligence,
Zagrebačka ul. 30, 52100 Pula, Croatia
smaricic@unipu.hr

Damjana Mihaljević, student
Juraj Dobrila University of Pula,
Faculty of Engineering,
Zagrebačka ul. 30, 52100 Pula, Croatia
dmihaljev@student.unipu.hr

Lucia Labinjan, student
Juraj Dobrila University of Pula,
Faculty of Engineering,
Zagrebačka ul. 30, 52100 Pula, Croatia
llabinjan@student.unipu.hr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240603185810

Original scientific paper
Received: 2024-06-03, Accepted: 2024-07-12

Computing the Deep Semantics of Visual Communications

Trpimir Jeronim Ježić, Marko Maričević, Ivana Pavlović, Miroslav Mikota*

Abstract: The research field of computational aesthetics gives crucial contributions to the development of mechanisms for filtering and/or generating value-laden informational content. This paper acknowledges a recognized escalating problem in the development of contemporary informational technologies and presents a practical solution for communicational quality management by employing an innovative approach to the computational aesthetic evaluation (CAE). After discussing the problem and attempted approaches to its alleviation, the paper offers a novel expert solution by presenting an original research approach and its resulting open-sourced model which outperforms its current state-of-the-art competition in semantic and stylistic classification, at the same time providing an idiomatic measure for objective aesthetic evaluation and demonstrating semantically rich and professionally recognized explanatory power which can serve as the solid basis for development of reliable and user friendly content retrieval, generative or auxiliary design applications. Presented model is resource- and privacy-wise utmost conservative. Its use evades all ethical, legal or security concerns that beset all currently prominent models. Its developmental and operational costs are practically nil.

Keywords: computational aesthetic evaluation; convolutional neural networks; feature engineering; graphic engineering; interpretable machine learning; semantic embeddings

1 INTRODUCTION

One of the first definitions for the term *computational aesthetics* as a name for the research field in informational studies appeared in Hoenig's paper in 2005. He defined it as "the research of computational methods that can make applicable aesthetic decisions in a similar fashion as humans can" [1, 2]. In that way, he emphasized its two major aspects: the use of computational methods (i.e. it provides measurable output) and the enhancement of applicability. Computational aesthetic evaluation, as any aesthetic judgement, is generally considered to be a highly idiosyncratic act [3]. As Galanter noted, notions of computational aesthetic evaluation led to "deep philosophical waters regarding phenomenology and consciousness" [4]. The area of aesthetic research thus calls for a strongly interdisciplinary approach [5-8]. Nevertheless, computational aesthetics, considered as a field at the intersection of science and art [9, 10], has seen significant progress in recent years. It became focused on aesthetic measurement, generative art, and design generation [7, 11, 12]. Among many applications of computational aesthetics established in last few decades, we focus our attention here on one of its most beneficial aspects: integrating contributions of computational aesthetic evaluation (CAE) and computer vision (CV) algorithms for the sake of advancement of human information accumulation and retrieval capacities. The research question posed in this paper is whether there is a way to computationally detect salient visual features which provide semantic context for decoding the meaning of visual communications.

2 RESEARCH METHODOLOGY

A necessary condition for success of any machine learning (ML) project is the appropriateness and quality of its training data. Our research question imposes on samples a few criteria that would make them suitable contributors to its answer. Firstly and obviously, the sample instances need to be recognized as successful communicators of their intended meaning. Secondly, to prove the universality of their status, they need to be acknowledged as role models for other successful communication attempts time and again since their first occurrence. Thirdly, there are a couple of bonus features of visual communication instances that would propound them as exceptionally suitable for enlightening our proposed research question. It would be preferable if the syntactic form of those instances were as arbitrary as possible, measurable by some standardized laboratory equipment, under the authors' control, not dictated by available resources or other means of production and not constrained by any predefined linguistic or other cultural mandate in adding or removing features of shapes to attune the entire composition for transmitting the intended semantic meaning, and lastly explainable by some design theory. In accordance with the research question and its qualifiers, we concluded that the narrowed category of pictorial high art would best serve as a representative of successful unhindered visual communications [13-15]. For that reason, we chose the community driven WikiArt site as the source of our sample.

2.1 Data Accumulation

First, we gathered data by scraping the whole of WikiArt website. In this way we gathered information on over 215,000 instances of artworks made by 5300 artists featured on the site at the time. The data featured information about exhibits such as artwork's title, author, style and genre and unique ID given to the artwork in the WikiArt site's database. We gathered additional data about artists such as their lists of names, birthdays, deathdays, and URLs of their Wikipedia pages. WikiArt organizes artworks into 220 styles and 68 genres. Artworks could be labelled with multiple style and genre categories, culminating in total 1552 unique style and 726 unique genre combinations. In the next step we cleaned up the dataset so it would contain only research relevant instances. Our focus was on achieving high precision in this selection without any pressure to achieve a high recall rate due to the abundant cardinality of the sample. Beside

pictorial art, WikiArt presents many exhibits of architectural, product design, VR and AR, calligraphic, ornamental, and other types of art. All instances belonging to those genre categories were not relevant to our research. We also applied an ensemble of custom filters on the scraped data to ensure that the remaining instances represent high-quality works, exemplars in communication design, authored by artists confident in their style, affirmed by their influence on culture and other artists and recognized by experts in the field. We based the construction of those filters on some proxy features. Some of these filters connote the requirements: that the instance's author has a substantial body of appreciated work; has a Wikipedia page dedicated to his influence as a visual artist; the style in which the image was categorized has substantial following and a body of authors who adhere to it; that the style has its dedicated Wikipedia page; that the work managed to stay relevant through at least couple of major cultural shifts (which eliminated a plethora of recent artworks); and so on.

Second, we needed to decide on the semantically relevant features for whose detection we will train our ML model. In the beginnings, CAE researchers attempted to construct expert systems for detecting hand-crafted and statistically generic aesthetic features, accomplishing modest but promising results [16, 17]. In 2014 [18], motivated by the advances in CV technologies [19], researchers turned their focus on convolutional deep neural network (CNN) architectures which immediately yielded better results in blind prediction of past ratings of images, but forced the researchers to abandon their search for aesthetic features that could clarify the problem of understanding the appreciation of visual communications [20]. Many black-box approaches have been tried over the next decade with increasing success rates due to the addition of more precise search criteria: by discriminating between portions of images that feature subjects from those of backgrounds [21]; or by discriminating portions by the compositional relevance of their positions in an image format [22, 23]; or by giving higher priority to features prominent in art and design theory [24]; or by adding descriptive "semantic" features that give context to aesthetic evaluation [25-31], which at times proved to be alone better predictors of aesthetic judgements than the pixel data.

Current research in CAE recognizes the need and benefits of focusing on semantically relevant features of images even when aiming solely to create a naive aesthetic approval prediction machine. However, all up to date and state-of-the-art research approaches try to add semantic features to their datasets by appending columns featuring linguistic labels gathered through some survey on dataset images or through semantic-web scraping. The idea is that these inputted features are universally relevant for instances of the dataset because they represent human judgement of those images, and that those features are automatically semantic because they are linguistically labelled. None of those stances are exculpatory. There are no guarantees that the surveys and their multiple choices present a valid spectrum or measure of semantically potent attributes of images, nor are there any guarantees that the surveys' subjects are representative of any model's future user. But crucially, there is no justification for claiming that the added feature columns store any semantic information at all. Those appended columns bring to the dataset only additional syntax, foreign to the original samples and absent during the original sample selection which qualified its instances as valid. Without guarantees for semantic stability of added syntax, additional columns only enhance the problem complexity of information extraction and cast doubt on the representative validity of altered samples. To avoid those common pitfalls, we abstained from conducting any *ad hoc* surveys on the collected dataset instances. For this we decided to base our decision mechanism exclusively on the well-established art-historical classification of styles already present in the given dataset. Those style labels served as a meaningful (semantically fixed) way of separating the instances into syntactically distinct buckets. Many of those buckets appeared to be uninformative for our purpose. For example, we found that sample sets labelled with style names prefixed with terms such as "post-", "neo-" or "new" more often than not presented aggregates of images with no internal syntactic or semantic coherence. Those labels rather functioned as negative signifiers, encompassing artworks associated only by the signal they are responding to, and not by the response communicated through the artworks. Similarly incoherent aggregations appeared in sets based on style labels that contained vague qualifying terms such as "classical", "analytical", "period" or "school". Using the criteria described in this and previous paragraphs alone we managed to reduce the sample size by 25%, purifying it down to 160,000 relevant instances.

2.2 Feature Selection

For the relevant features to help semantically categorize visual communications we decided to use two dimensions, well established and basic to any classification of possible thoughts and their expressions. The first dimension is the one on which the scientific method with its task of extracting generalizations from samples is founded; which is to say, the dimension relied upon when distinguishing concrete objects via abstractions. The second dimension we chose is the one on which all theories of applied communications, whether linguistic or visual, base their discussion on the structure of the relationships between the signifier and the signified; which is to say, when mapping the syntax to semantics of a message. Specifically, we base our second line of distinction on Charles S. Peirce's original theory of *semiotics* [32], which in practice underlies all design theories. Our first dimension determines the nature of the subjects in communication, and the second one determines the way the subjects are referred to. Clearly, these two dimensions describe the necessary basic semantic context which is a prerequisite for any further semantic decoding. We will refer to these dimensions as the *breadth* and the *depth* of a communication, in the same spirit in which the terms are used in common speech. It is likewise important to note that these two dimensions are by definition orthogonal to each other which makes them very convenient when mapping the

communicative space. Humans can talk about concrete objects or abstract concepts via examples or through symbolic means. There is no predetermined correlation at hand.

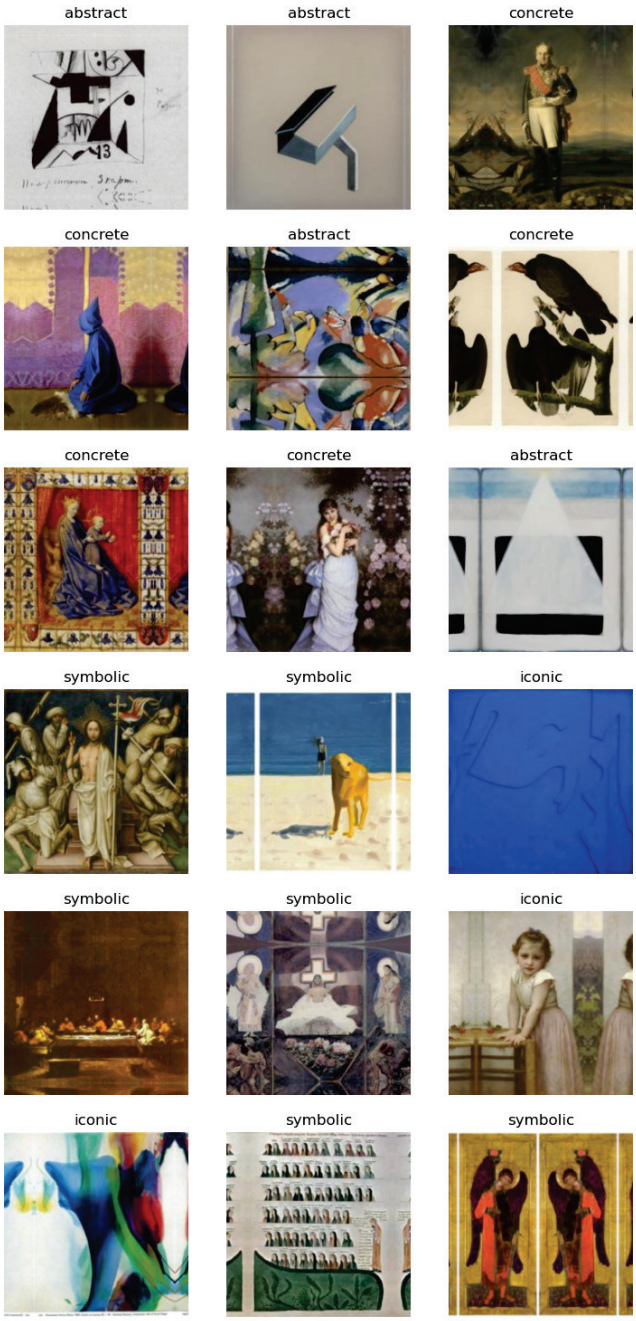


Figure 1 Random subsets of representative instances collected from WikiArt

To select the representative instances for the extrema of these dimensions, we created buckets of styles that are manifestly and universally recognized as belonging by impressions and communicative intentions to one of our four extrema. The concrete-abstract (breadth) dimension is represented on one end by figurative art styles such as Renaissance, Academicism, Naturalism, Realism and Hyper-Realism; and on the other end by abstract styles such as Concretism, Suprematism, Abstract Expressionism and Action painting. The iconic-symbolic (depth) dimension is represented on one end with naturalistic, visceral art styles such as Rococo, Academicism, Pointillism, Lyrical Abstraction and Color Field Painting; and on the other end by the intellectual styles, rich in symbolism, such as Byzantine, International Gothic, Classicism, Romanesque and Pop Art. While the distinction between the styles on the breadth dimension is visually obvious, there is no such obviousness in the depth dimension. The styles were handpicked based on consensus among the art historians and critics regarding the tone and intended meaning of the artworks belonging to these styles [33-35]. When it has been decided which styles are representative, further selection of specific artworks belonging to those stylistic categories was left to random choice. An image dataset containing about 500 images per extrema was obtained by downloading digital replicas of a stylistically curated, but otherwise randomly chosen samples from the WikiArt's website (Fig. 1).

2.3 Architecture Design and Training

To guarantee objectivity, training was conducted on pixel-data alone, supervised only by meta-stylistic labels for four extrema of two semantic dimensions defined above. By recognizing the fact that the two chosen dimensions are mutually orthogonal and that there could be very little or none instances that could serve as good representatives of extrema on both dimensions simultaneously, and that forcing the neural networks to distinguish between intermediate representations would motivate the model to overfit to any given dataset, we decided to train two separate neural networks, each for implementing one semantic dimension, and thereafter to pair the two unidimensional models into one proficient parallelized system. Model development was conducted with the PyTorch library for ML in Python, more specifically, using Jeremy Howards fastai wrapper-framework. By following the advice of Yosinski et al. [36], Dong et al. [37], and Howard [38], we base our training approach on fine-tuning an existing state-of-the-art CV model. This approach is conventional in CAE [18], [39], [28], [40] for its noted benefits. Fine-tuning on a trained CNN has been proven to be a resource effective initialization approach, but the benefits of fine-tuning do not stop at conservation. As has been shown in Deng et al. [20], fine-tuning for aesthetic evaluation from the vanilla AlexNet yields better performance than simply training the base net from scratch [18]. One possible explanation is the claim that multitask model training, or in case of fine-tuning, task accumulation, forces ML models to construct more realistic embedding space for inputted data, and thus build a better understanding of real-world relations between data points [41, 42, 30]. Another, more reasonable explanation is that the task overload serves as a realistic regularization mechanism which prevents models from overfitting to sampled data.

Papers referenced in previous paragraph used the AlexNet [19], a 13-layered convolutional neural network (CNN) that brought neural networks into the spotlight back in 2012 when it outperformed all other ML architectures

competing on ILSVRC, and by doing so, kick-started the ongoing AI revolution. We found its successor, the winner of ILSVRC 2014, a 22-layered CNN named GoogleNet [43], most efficient for our task. All foundational models were accessed through PyTorch's Timm library of open-sourced legacy CV models. In our training, the GoogleNet model outperformed AlexNet and all other up to date and moderately sized pretrained CNNs. It was deemed appropriate because of its minimalist design architecture and extremely small number of layers, considering today's trends. The interpolation between semantic extrema is expected in reality to be somewhat linear and we were cautious in preventing the possibility of overfitting a model to the data by giving it too many degrees of freedom. The only significant alteration we conducted on the model was that of severing GoogleNet's original classification head, for it was modeled for ImageNet's one thousand classes, and replacing it with a binary classification head.

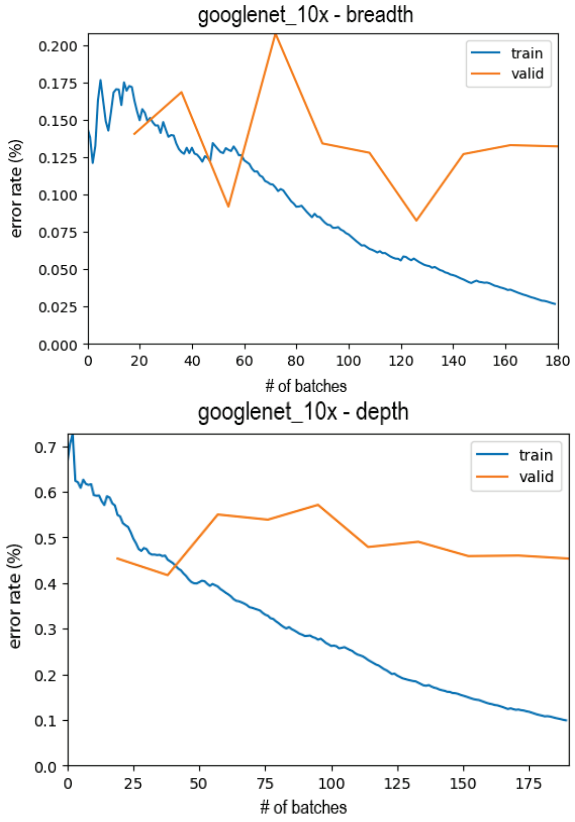


Figure 2 Plot of training and validation losses over 10 epochs

Training was conducted on two twin models in parallel, each on its respective sample-set of about 1000 images, driving Nvidia RTX 3070Ti graphics processor. The models were trained for 10 epochs each using *error rate* as the dominant metric. Each epoch took 4 seconds to compute. Results were similar for all tried models, but here we focus on the training of GoogleNet. For the breadth dimension, best results were achieved around seventh epoch, while for the depth dimension error rate kept declining through all ten epochs, although with diminishing returns. The results of best epochs were 2% error rate on the breadth dimension, and 14% error rate on the depth dimension. A surprising insight during training was, as visible in Fig. 2, the fact that the initial states of the CV models were inversely well fitted for the two tasks of semantic classification. Models were initially guessing classes on the breadth dimension correctly well beyond the binary chance probability, and on the depth dimension, missing the mark in the same manner. The other interesting aspect of those results, as visible on plots below, is the grade to which the weights of a CV model pretrained for object recognition needed to be altered to accommodate for a new task of semantic feature detection, especially for the breadth (abstract-concrete) dimension - the one that was initially good at blind guessing. Less surprising was the fact that the training for the depth (iconic-symbolic) dimension was more confounding to the algorithms, but the resulting model is still surprisingly efficient, given that there is no obvious visual cue that distinguishes those two semantic extrema. The training pipeline and detailed results have been made available online [44].

3 RESULTS AND DISCUSSION

To ensure the robustness of results, models were additionally tested on a completely new and unseen stylistically conditioned random set of images downloaded from the WikiArt's website. The respective *F1* scores are 0.978 for the breadth dimension and 0.882 for the depth dimension (Fig. 3). Those results are impressive for any practical feature detection, but state-of-the-art in general and universal semantic feature detection.

To further evaluate potential applicability of models, we downloaded entirely new dataset of 1000 images from the WikiArt's website, unconditioned on style classification, and plotted the models' predictions on a joint scatter plot. This shows how the models handle hundreds of styles unseen during training. The plot shown in Fig. 4a provides us with a couple of insights. Firstly, we can see that the breadth and depth dimensions are not strongly correlated, but there seems to be a slight imbalance in randomly downloaded sample which suggests weak artistic tendency to produce more figurative than abstract artworks when meaning to communicate a symbolic message. Likewise, there is a greater aspiration of visual artists for production of iconic rather than symbolic images (which explains model's iconic bias visible in Fig. 3).

The same plot displays how the CNN models tend to sharply distinct between classes, especially on the breadth dimension. All of those observations make real-world sense. High art is expected to be more iconic in contrast with applied art's need for heavy symbolic transmission. It is reasonable to expect that the artworks recognized as highly successful would feature clear indications of their intended subject matter and wouldn't be visually ambiguous about position of their content along the breadth dimension. Greater ambiguity on the depth dimension can be explained in couple of ways. There could be a greater tendency in visual communications to mix symbolic and iconic aspects when communicating layered messages common to high art. Moreover, Peirce's semiotic theory itself foresees that there should be a

discernible space for indexical messages amid the iconic-symbolic extrema. On the other hand, the moderate spread of distribution of artworks along the depth dimension could be the consequence of the loose method of sample collection, where the style labels by themselves could not demarcate the boundaries between the two limits of the category or that the WikiArt's labelling method wasn't stringent enough in this respect to begin with. Nevertheless, both dimensions provide a significant leverage for predicting the cultural period from which the artworks originated. By using the year of origin as a semi-reliable proxy indicator of cultural background, both dimensions show a noticeable measure of correlation with their semantic context (see Fig. 4b). Since the boundaries of visual styles are lax and cyclic, predictive power is measured through degree of mutual information.

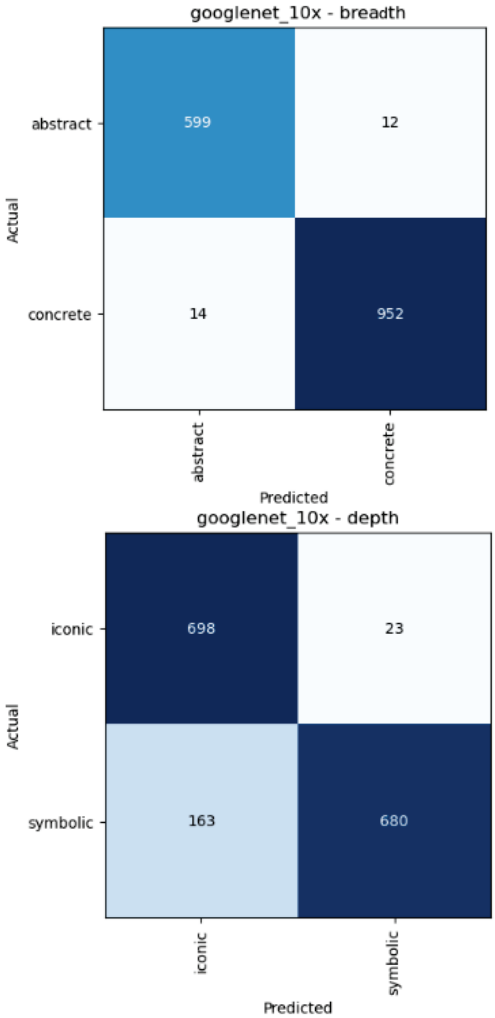


Figure 3 Confusion matrices of two fine-tuned GoogleNets, evaluated on a completely new validation set.

To get a clearer understanding of art styles distribution in this semantic space, Fig. 5 shows a scatter plot of joint distribution of random collection of artworks accompanied with style labels for random subset of 50 artworks. The plot clearly shows how the networks were able to cluster together artworks from diverse cultural periods, featuring distinct styles but associated by common semantic themes. Predominantly, symbolic and concrete artworks belong to periods of great narratives; concrete and iconic artworks convey subjective impressions and concerns; iconic and abstract works deal with avant-garde and conceptual expressions; while the abstract and symbolic art focuses on imaginative, intellectual and decorative trends. All art styles that were not presented in the training dataset found their reasonable placement within the proposed semantic space placing their artworks in a clear and explicatory relation to other, more unequivocal and less ambiguous works of art.

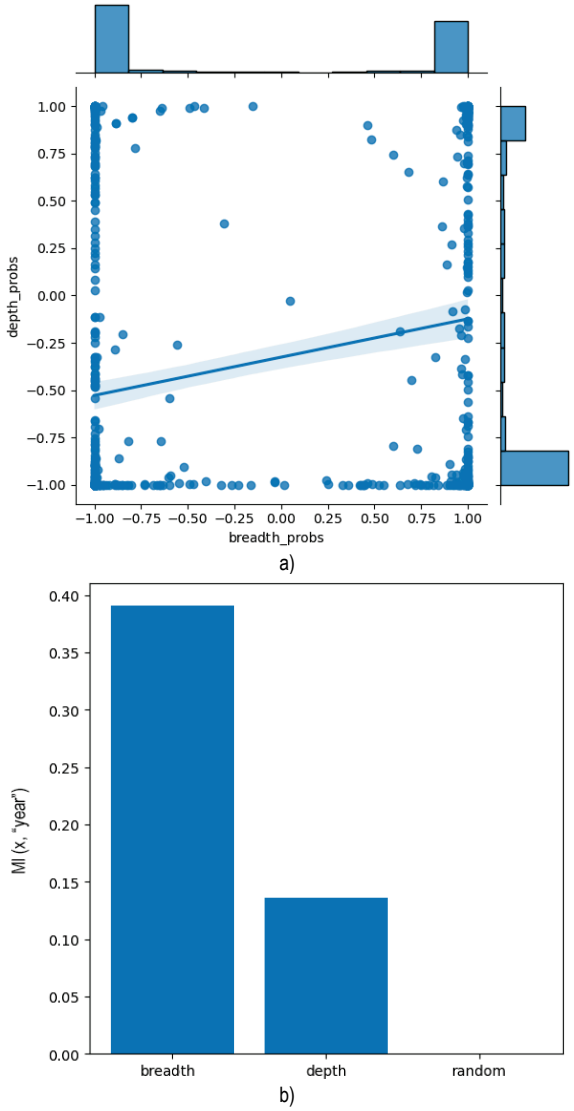


Figure 4 Diagrams of (a) joint plot of semantic feature predictions, (b) graph of mutual information between semantic features and the year of artwork's creation

Alongside aiding semantic interpretation, the model provides an idiomatic aesthetic evaluation criterion. Since all of the instances used in the sample present successful visual communications, and since the sample was picked at random from the entire body of historically and globally collected artworks, then the emptiness of large areas in the constructed space suggests the impossibility of creating successful visual communications that would feature compositional attributes congruent to those areas. If the model places the evaluated

artwork in a stranded area of semantic space, it is highly likely that the artwork wouldn't satisfy many peoples' aesthetic tastes. Likewise, visible in the same manner, if the model places the evaluated artwork in an area distant from its thematic kin, it is highly unlikely that it will efficiently communicate its intended meaning or be regarded as aesthetically significant.

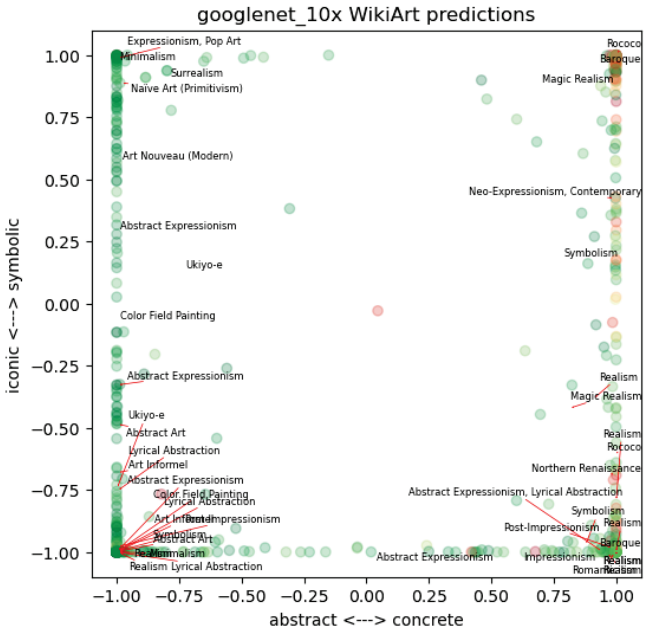


Figure 5 Scatter plot of semantic embeddings with style labels shown for a subset of sample's instances.

4 CONCLUSION

In conclusion, this study presents a novel approach to semantic feature detection using convolutional neural networks fine-tuned on stylistically conditioned images of artworks from WikiArt. The models achieved high F1 scores and provided robust results in distinguishing between abstract-concrete (breadth) and iconic-symbolic (depth) dimensions of artistic expression. The findings suggest that the proposed method can be used for universal semantic feature detection, aiding in the interpretation of visual communications across different, past and future cultural periods. Additionally, the models provide an idiomatic aesthetic evaluation criterion, allowing for the assessment of artworks' success in visual communication based on their positioning within the semantic space. Further research could explore the model's applicability expanding the scope of its application to include more diverse and historically representative samples of artworks, as well as investigating the potential applications of this method in other domains such as graphic or multimedia design.

5 REFERENCES

[1] Hoenig, F. (2005). Defining Computational Aesthetics. *Computational Aesthetics in Graphics, Visualization and Imaging*, p. 6.

[2] Greenfield, G. (2005). On the origins of the term "Computational aesthetics". *Proceedings of the First Eurographics conference on Computational Aesthetics in Graphics, Visualization and Imaging*, 9-12.

[3] Debnath, S. & Changder, S. (2020). Computational Approaches to Aesthetic Quality Assessment of Digital Photographs: State of the Art and Future Research Directives. *Pattern Recognit. Image Anal., 30*(4), 593-606. https:/doi.org/10.1134/S1054661820040082

[4] Galanter, P. (2012). Computational Aesthetic Evaluation: Past and Future. *Computers and Creativity*, McCormack, J. & d'Inverno, M., Eds. Springer Berlin Heidelberg, 255-293. https://doi.org/10.1007/978-3-642-31727-9_10

[5] Graham, D. J. & Redies, C. (2010). Statistical regularities in art: Relations with visual coding and perception. *Vision Research, 50*(16), 1503-1509. https:/doi.org/10.1016/j.visres.2010.05.002

[6] Shimamura, A. P. & Palmer, S. E. (Eds.) (2011). Aesthetic science: Connecting minds, brains, and experience. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199732142.001.0001

[7] Brachmann, A. & Redies, C. (2017). Computational and Experimental Approaches to Visual Aesthetics. *Front. Comput. Neurosci., 11*, p. 102. https:/doi.org/10.3389/fncom.2017.00102

[8] Li, R. & Zhang, J. (2020). Review of computational neuroaesthetics: Bridging the gap between neuroaesthetics and computer science. *Brain Inform, 7*(1), p. 16, https:/doi.org/10.1186/s40708-020-00118-w

[9] Spratt, E. L. & Elgammal, A. (2014). Computational Beauty: Aesthetic Judgment at the Intersection of Art and Science. https://arxiv.org/abs/1410.2488. https://doi.org/10.48550/arXiv.1410.2488

[10] Bo, Y., Yu, J. & Zhang, K. (2018). Computational aesthetics and applications. *Vis Comput Ind Biomed Art, 1*, p. 6. https:/doi.org/10.1186/s42492-018-0006-1

[11] Zhang, J., Miao, Y. & Yu, J. (2021). A Comprehensive Survey on Computational Aesthetic Evaluation of Visual Art Images: Metrics and Challenges. *IEEE Access, 9*, 77164-77187. https:/doi.org/10.1109/ACCESS.2021.3083075

[12] Valenzise, G., Kang, C. & Dufaux, F. (2022). Advances and challenges in computational image aesthetics. *Human Perception of Visual Information: Psychological and Computational Perspectives*, Springer, 133-181. https://doi.org/10.1007/978-3-030-81465-6_6

[13] Elgammal, A., Mazzone, M., Liu, B., Kim, D. & Elhoseiny, M. (2018). The Shape of Art History in the Eyes of the Machine. http://arxiv.org/abs/1801.07729. https://doi.org/10.48550/arXiv.1801.07729

[14] Mohammad, S. & Kiritchenko, S. (2018). WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art,. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[15] Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M. & Guibas, L. (2021). ArtEmis: Affective Language for Visual Art. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 11564-11574. https://doi.org/10.1109/CVPR46437.2021.01140

[16] Datta, R., Joshi, D., Li, J. & Wang, J. Z. (2006). Studying Aesthetics in Photographic Images Using a Computational Approach. *Computer Vision – ECCV 2006, vol. 3953*, Leonardis, A., Bischof, H. & Pinz, A., Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 288-301. https://doi.org/10.1007/11744078_23

[17] Ke, Y., Tang, X. & Jing, F. (2006). The Design of High-Level Features for Photo Quality Assessment. *The IEEE Computer*

Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06), 419-426. https://doi.org/10.1109/CVPR.2006.303

[18] Lu, X., Lin, Z., Jin, H., Yang, J. & Wang, J. Z. (2014). RAPID: Rating Pictorial Aesthetics using Deep Learning. *Proceedings of the 22nd ACM international conference on Multimedia*, 457-466. https://doi.org/10.1145/2647868.2654927

[19] Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM, 60*(6), 84-90. https://doi.org/10.1145/3065386

[20] Deng, Y., Loy, C. C. & Tang, X. (2017). Image Aesthetic Assessment: An experimental survey. *The IEEE Signal Processing Magazine, 34*(4), 80-106. https://doi.org/10.1109/MSP.2017.2696576

[21] Luo, Y. & Tang, X. (2008). Photo and Video Quality Evaluation: Focusing on the Subject. *Computer Vision – ECCV 2008, vol. 5304*, Forsyth, D., Torr, P. & Zisserman, A. Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 386-399. https://doi.org/10.1007/978-3-540-88690-7_29

[22] Mai, L., Jin, H. & Liu, F. (2016). Composition-Preserving Deep Photo Aesthetics Assessment. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016)*, 497-506. https://doi.org/10.1109/CVPR.2016.60

[23] Ma, S., Liu, J. & Chen, C. W. (2017). A-Lamp: Adaptive Layout-Aware Multi-patch Deep Convolutional Neural Network for Photo Aesthetic Assessment. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR2017)*, 722-731. https://doi.org/10.1109/CVPR.2017.84

[24] Aydın, T. O., Smolic, A. & Gross, M. (2015). Automated Aesthetic Analysis of Photographic Images. *The IEEE Transactions on Visualization and Computer Graphics, 21*(1), 31-42. https://doi.org/10.1109/TVCG.2014.2325047

[25] Dhar, S., Ordonez, V. & Berg, T. L. (2011). High level describable attributes for predicting aesthetics and interestingness. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, 1657-1664. https://doi.org/10.1109/CVPR.2011.5995467

[26] San Pedro, J., Yeh, T. & Oliver, N. (2012). Leveraging user comments for aesthetic aware image search reranking. *Proceedings of the 21st international conference on World Wide Web*, 439-448. https://doi.org/10.1145/2187836.2187896

[27] Tang, X., Luo, W. & Wang, X. (2013). Content-Based Photo Quality Assessment. *The IEEE Transactions on Multimedia, 15*(8), 1930-1943. https://doi.org/10.1109/TMM.2013.2269899

[28] Kao, Y., He, R. & Huang, K. (2017). Deep Aesthetic Quality Assessment with Semantic Information. *The IEEE Trans. on Image Process., 26*(3), 1482-1495. https://doi.org/10.1109/TIP.2017.2651399

[29] Liu, X., Li, N. & Xia, Y. (2018). Affective Image Classification by Jointly Using Interpretable Art Features and Semantic Annotations. *Journal of Visual Communication and Image Representation, 58*. https://doi.org/10.1016/j.jvcir.2018.12.032

[30] Tian, X. (2021). Using multi-task residual network to evaluate image aesthetic quality. *The IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC2021)*, 171-174. https://doi.org/10.1109/IAEAC50856.2021.9391005.

[31] Duan, J., Chen, P., Li, L., Wu, J. & Shi, G. (2022). Semantic Attribute Guided Image Aesthetics Assessment. *The IEEE International Conference on Visual Communications and Image Processing (VCIP2022)*, 1-5. https://doi.org/10.1109/VCIP56404.2022.10008896

[32] Jappy, T. (2013). *Introduction to Peircean visual semiotics*. London: Bloomsbury.

[33] Clark, K. (1969). Civilisation: *A personal view*, 1st publ. London: British Broadcasting Corp.

[34] Hauser, A. (1999). *The social history of art. 1: From prehistoric times to the Middle Ages*. 3rd ed. London: Routledge.

[35] Davies, P. J. E., Denny, W. B., Hofrichter, F. F., Jacobs, J., Roberts, A. M. & Simon, D. L. (2016). *Janson's History of art: The Western tradition*. Reissued eighth edition. Boston: Pearson.

[36] Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. (2014). How transferable are features in deep neural networks? https://arxiv.org/abs/1411.1792. https://doi.org/10.48550/arXiv.1411.1792

[37] Dong, C., Deng, Y., Loy, C. C. & Tang, X. (2015). Compression Artifacts Reduction by a Deep Convolutional Network. *The IEEE International Conference on Computer Vision (ICCV2015)*, 576-584. https://doi.org/10.1109/ICCV.2015.73

[38] Howard, J., Gugger, S. & Chintala, S. (2020). *Deep learning for coders with fastai and PyTorch: AI applications without a PhD*. First edition. Sebastopol, California: O'Reilly Media, Inc.

[39] Denzler, J., Rodner, E. & Simon, M. (2016). Convolutional Neural Networks as a Computational Model for the Underlying Processes of Aesthetics Perception. *Computer Vision – ECCV 2016 Workshops*, 871-887. https://doi.org/10.1007/978-3-319-46604-0_60

[40] Anwar, A., Kanwal, S., Tahir, M., Saqib, M., Uzair, M., Rahmani, M. K. I. & Ullah, H. (2022). A Survey on Image Aesthetic Assessment. https://arxiv.org/abs/2103.11616. https://doi.org/10.48550/arXiv.2103.11616

[41] Cavanagh, P. (2005). The artist as neuroscientist. *Nature, 434*(7031), 301-307. https://doi.org/10.1038/434301a

[42] Peng, K.-C. & Chen, T. (2016). Toward correlating and solving abstract tasks using convolutional neural networks. *The IEEE Winter Conference on Applications of Computer Vision (WACV2016)*, 1-9. https://doi.org/10.1109/WACV.2016.7477616

[43] Szegedy C. et al. (2015). Going deeper with convolutions. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015), Boston, MA, USA, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[44] Ježić, T. J. (2024). Trpquo/art_critic: Repository for WikiArt model training. https://github.com/Trpquo/art_critic. (Accessed: 26-Apr-2024).

Authors' contacts:

Trpimir Jeronim Ježić, B.Sc. Eng.
University of Zagreb, Faculty of Graphic Arts,
Getaldićeva 2, 10 000 Zagreb, Croatia
trpimir.jeronim.jezic@grf.unizg.hr

Marko Maričević, Assist. Prof., PhD
University of Zagreb, Faculty of Graphic Arts,
Getaldićeva 2, 10 000 Zagreb, Croatia
marko.maricevic@grf.unizg.hr

Ivana Pavlović, Teach. Assist., PhD
University of Zagreb, Faculty of Graphic Arts,
Getaldićeva 2, 10 000 Zagreb, Croatia
ivana.pavlovic@grf.unizg.hr

Miroslav Mikota, Assoc. Prof., PhD
(Corresponding author)
University of Zagreb, Faculty of Graphic Arts,
Getaldićeva 2, 10 000 Zagreb, Croatia
miroslav.mikota@grf.unizg.hr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20250115224222

Original scientific paper
Received: 2025-01-15, Accepted: 2025-02-21

Time Dependent Load Capacity of the Press Fit

Vinko Močilnik, Nenad Gubeljak*, Jožef Predan

Abstract: This study investigates the loading capacity of a press fit using experimental, numerical and theoretical methods. Tests on specimens with different interferences showed that the loading capacity increases over time as long as the plasticity remains at the micro level. At larger interferences, the plasticity extends to the macro level, which in the long term means a reduction in the loading capacity of the press fit due to creep. Numerical simulations using finite element modelling showed the influence of surface roughness and time-dependent effect on contact pressure and friction. Models in text books do not take in account plasticity and creep of the material in press fit. The phenomenon can lead to a weakening of the press fit over time. The results highlight the importance of optimizing the interference and surface preparation to improve the loading capacity and joint performance. The article presents an approach for calculating the press fit taking in to account the Bowden Tabor friction model and the visco-plasticity of the material used.

Keywords: FEM analysis; interface oversize; hub; shaft; loading creep; press fit; roughness

1 INTRODUCTION

A press fit is commonly used in engineering structures to connect a shaft and a hub. The parts fit together due to a radial pressure that depends on the amount of interference at the diameter where both parts touch. The two parts (shaft and hub) are permanently joined as a solid unit. Although interference is an effective way to transmit large torques, its disadvantage is that in some cases disassembly is very difficult or impossible, Fig. 1.

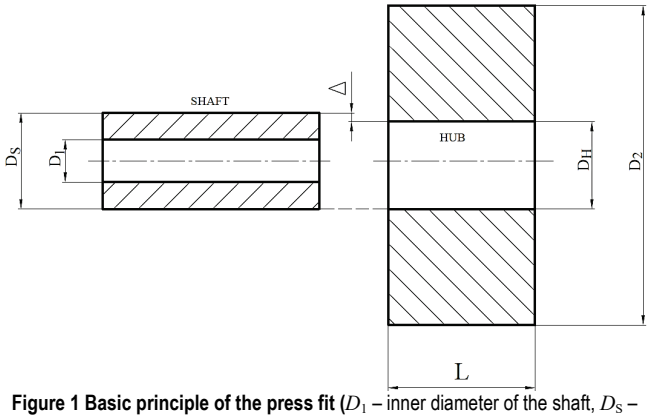


Figure 1 Basic principle of the press fit (D_1 – inner diameter of the shaft, D_S – nominal diameter of the shaft, D_H – nominal diameter of the hub, D_2 – external diameter of the hub, L – length of the press fit, Δ – interference of the press fit)

The composition of the such produced parts is possible in two ways:

a) By pressing the shaft into the hub.
b) By heating the hub or/and cooling the shaft to create a clearance between both parts.

The assembly of them is very simple and does not require a large assembly force. When the temperature, when the shaft be inserted, is gradually equalized, both parts want to return to their original dimensions and thus crash into each other; a pressure is created between them.

Due to the contact pressure at the nominal diameter, friction is established between the both parts, which resists against external forces in the axial and/or tangential direction. The greater the friction force, the greater the load capacity of the press fit. The average contact pressure along the joint and thus also the friction force depends on the size of the interference and the stiffness of the shaft and hub. The relationship between pressure and interference is linear if we assume that the material remains in the elastic region.

Lewis and co-workers in [1] measured the contact pressure in the varicose vein by using an ultrasound method to measure it. He showed that there is a region of uniform pressure within the contact zone and that on each side of the contact zone the contact pressure increases sharply. The average pressure along the length of the kink fits well with the results of the Lamé analysis, which is based on the plane stress state with considerable simplifications.

He concluded that the pressure distribution along the contact zone is comparable whether it is press fit.

Madej in [2] carried out the calculation of the interference fit with FEM analysis and compared the results with the experiment. He proved that the use of Lame's analysis gives a deficient estimation in some cases. The difference in the loading capacity of the interface fit can be up to 20 %. He showed that for accurate modelling it is necessary to use the size of the finite element less than 1% of the diameter of the axle. Autor's assessment is that in a more accurate analysis it is also necessary to model the friction. The Coulomb friction model is not sufficient to describe high-pressure sliding contact. The author suggests using the Bay-Wanheim (1976) model, which relates the frictional stress to the contact pressure.

In [3], the authors state that the force of static friction determines the strength of many different types of joints that are exposed to high loads, such as press-fits. The coefficient of static friction depends on many parameters, such as the mechanical properties of both materials, surface roughness, lubrication, impurities, hardness of contact surfaces, duration of contact and so on. In engineering calculations, the mean value of the coefficient of static friction, which is determined experimentally, is used. In order to determine the appropriate coefficient of static friction in a given case, a lot of experimental work is required. In order to simplify the process of determining the coefficient of static friction, the authors proposed a new calculation procedure based on the

mechanical-molecular theory of friction. The results were compared with experimentally determined values and a sufficiently accurate match was found.

In [4], the author discusses the influence of surface roughness on the design of interference fits. It gives a detailed study of the influence of the surface roughness on the loading capacity of the interference fits. Here, the aspect of plastic deformation is neglected.

The authors in [5] and [6] measured the deformation of the bushing circumference using strain gauges on the outer diameter of the bushing. This then allowed them to estimate the contact pressure using analytical equations.

Croccolo et al. in [7] found that at interference fits between aluminium and steel, the coefficient of friction for dry aluminium-steel surfaces is 0.46. They developed an analytical model and reported that if the stiffness of both parts is close, the actual interference is likely to be less than the estimated interference. Mori et al. [8] analysis plastic deformation in press fit too. They recognized that the joining strength may be increased by forming beads and dimples. An example of the second type occurs when workpieces are mechanically interlocked by plastic deformation. An example of the second type occurs when the workpieces are mechanically connected by plastic deformation.

Yang et all in [9] point out that the optimum pressure between the two parts is reached only before the ring begins to plastically yield and that any further increase in the deformation pressure results in a negligible increase in the interference pressure and joint strength. That it is possible to increase joint strength by increasing the surface roughness or cleaning the contact surfaces.

The authors in [10] emphasize the importance of surface roughness in interference fit and demonstrate experimentally that the extraction load varies by up to 300 % for *Ra* values of 0.24 to 6.82 microns. They used an elastic two-dimensional model to model surface asperities on both the pin and bushing and report that despite high stresses that tend to crush the asperities, they tend to persist under high pressure. The authors in [11] analysed the interference fit of ring gear and stepped shaft using numerical and analytical methods and concluded that the Lame's equations underestimate the contact pressure by up to 78 %. The author in [12] used an analytical model to show that contact pressure varies with temperature; an increase in temperature leads to a decrease in the yield stress, which causes plastic deformation in the interference fit, thus reducing its loading capacity. Numerical methods were used to predict the quality of the joint. The author in [13] compared the load transfer at different values of interference and concluded that the use of a carefully selected interference, increases the durability of the joint.

2 MATERIAL PROPERTIES AND TESTING SPECIMEN

Eight test specimens of the press fit with different interferences were made eighteen years ago, with the aim of establishing the agreement between the results of the load capacity calculation of the press fit and the realistically measured values. The dimensions of the specimen are shown in Fig. 2. The inside of the hub was finely turned, the shaft was grinded. Assembly was carried out using the hub heating from 150 up to 450 °C in furnace, while shaft was at room temperature +20 °C. The smallest interference was 0.015 mm and increased up to 0.15 mm. The test specimens were loaded in the axial direction until the shaft began to slide in the hub. The maximum force was measured; the load capacity of the press fit. As the interference increased, it was to be expected that the load capacity of the press fit would also increase.

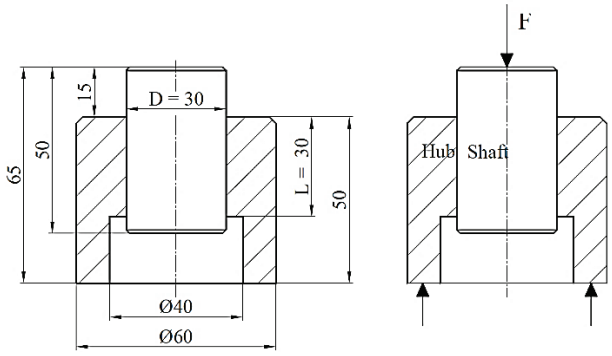


Figure 2 Specimen of press fit

S355 material with the following mechanical properties was used:

- Young's modulus E = 206 GPa
- Poisson's ratio v = 0.3
- elastic limit Re = 357 MPa
- tensile strength Rm = 512 MPa.

2.1 Rate-dependent Plasticity Creep Behaviour

A creep test of the material at constant stress was carried out. Fig. 3 shows setup for creep tension testing with elongation measurement at the INSTRON 1255 servo-hydraulic testing machine. Fig. 4 shows the white points at which the relaxation was measured. The three tensile test pieces were loaded to a white stress point and then keeping under constant plastic strain. The plastic strain was measured using an extensometer, and the moment when the stress was reached was considered as the time zero. Fig. 5 shows the dependences of the creep strain rate on time at constant stress. Three points 327 MPa, 398 MPa and 452 MPa were considered. Parameters A, n, and m were experimentally determined for ABAQUS FEM calculations.

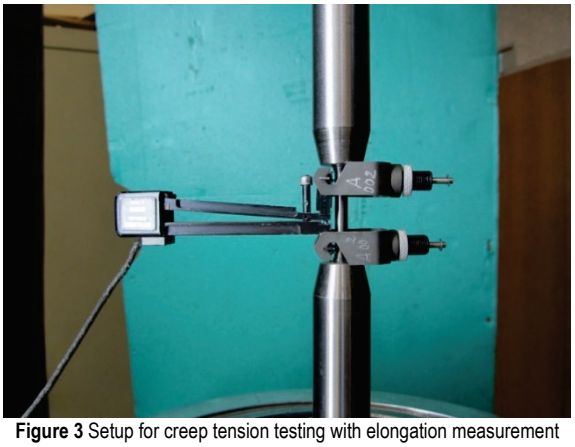


Figure 3 Setup for creep tension testing with elongation measurement

Fig. 4 shows the true stress-strain curve for this material in plastic range.

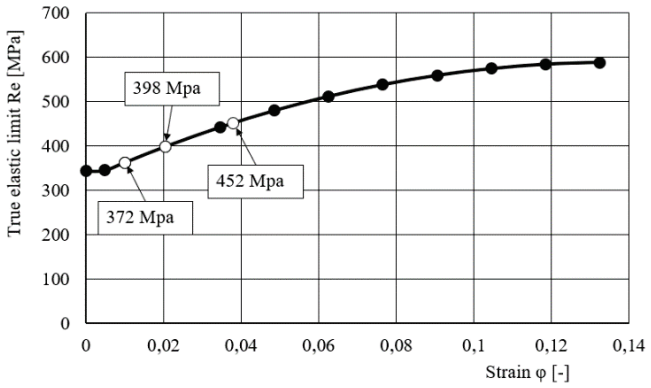


Figure 4 Tensile characteristic of the material S355 in plastic range

$$\dot{\varepsilon} = A \cdot \sigma^n \cdot t^m, \tag{1}$$

where: $\dot{\varepsilon}$ – uniaxial equivalent creep strain rate; σ – uniaxial equivalent deviatoric stress; A, n, m – constants determined experimentally at room temperature.

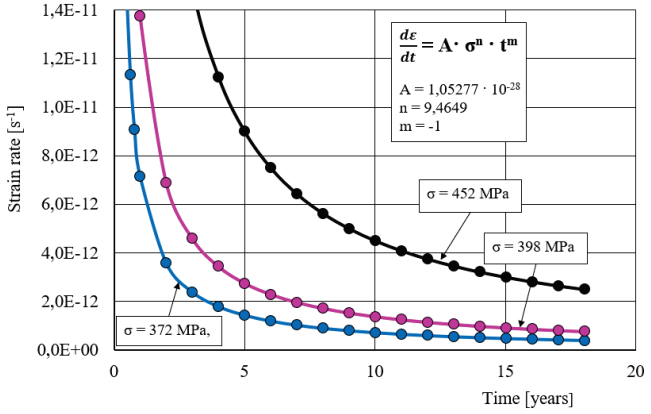


Figure 5 Time dependent properties of the used material

3 THEORETICAL BACKGROUND

3.1 Engineering Calculations

In literature for mechanical engineers, such as Decker [18] and Nieman [19], the procedure for calculating the press fit load capacity is described. The calculation is based on a theory that deals with stresses in a two-layer ring with the assumption of axial symmetry and a plane stress state, in elastic state of material.

The problem is solved using the first Lamé equation, written in the polar coordinate system, without considering volume forces, at a homogeneous temperature field, and has a form as follows:

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \cdot \frac{\partial u}{\partial r} - \frac{u}{r^2} = 0. \tag{2}$$

The variable u is the displacement in the radial direction $u = u(r)$. Using $u = r^k$, where k is parameter, gives the general solution of the differential equation:

$$u(r) = Ar + Br^{-1}, \tag{3}$$

where A and B are integration constants. Using Hooke's Law:

$$\begin{aligned} \sigma_r &= \frac{E}{1-\nu^2}\left(\varepsilon_r + \nu\varepsilon_\varphi\right) = \frac{E}{1-\nu^2}\left(\frac{\mathrm{d}u}{\mathrm{d}r} + \nu\frac{u}{r}\right), \\ \sigma_\varphi &= \frac{E}{1-\nu^2}\left(\varepsilon_\varphi + \nu\varepsilon_r\right) = \frac{E}{1-\nu^2}\left(\frac{u}{r} + \nu\frac{\mathrm{d}u}{\mathrm{d}r}\right), \end{aligned} \tag{4}$$

can be written:

$$\begin{aligned} \sigma_r &= \frac{E}{1-\nu^2}\left[A(1+\nu) - \frac{B}{r^2}(1-\nu)\right], \\ \sigma_\varphi &= \frac{E}{1-\nu^2}\left[A(1+\nu) + \frac{B}{r^2}(1-\nu)\right], \end{aligned} \tag{5}$$

where σ_r in σ_φ are principal stress components in polar coordinate system.

In Eqs. (4) and (5), the designations mean: σ_r – stress in radial direction; σ_φ – stress in tangential direction (hoop stress); E – modulus of elasticity and ν – Poisson's ratio.

Eqs. (5) are written separately for the inner ring (shaft) denoted by I and for the outer ring (hub) denoted by II. This gives us four integration constants, which are calculated from the following boundary conditions:

a) $r = D_1/2 \rightarrow \sigma_r^{\mathrm{I}} = 0,$

b) $r = D/2 \rightarrow \sigma_r^{\mathrm{I}} = \sigma_r^{\mathrm{II}} = -p,$

c) $r = D/2 \rightarrow u^{\mathrm{II}} - u^{\mathrm{I}} = \Delta,$

d) $r = D_2/2 \rightarrow \sigma_r^{\mathrm{II}} = 0.$

where p - interference pressure and Δ – interference.

For the case, when $D_1 = 0$, can be used $\sigma_r(r) = \text{const.} = \sigma_r(r = D/2)$.

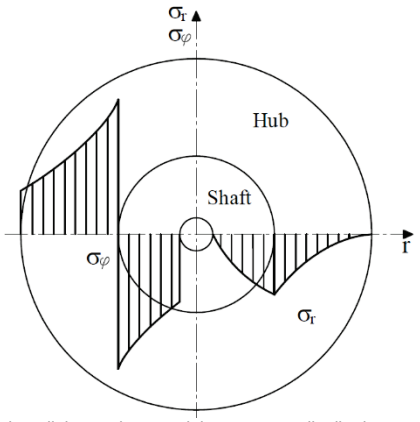


Figure 6 Elastic radial σ_r and tangential σ_φ stresses distribution on macro level of the press fit

The hub presses on the shaft due to elastic action as a spring. At larger interference, plastic deformation of the hub partially occurs, which could reduce the pressure between hub and shaft. This can happen if the plastic strain penetrates too deep into the material. Fig. 6 shows the elastic stress distribution on macro level of the press fit.

3.2 Coefficient of Static Friction at Large Pressure According to the Bowden-Tabor Model

The contact area, which is limited by the circumference at the nominal diameter and the length of the press fit, is called the apparent contact area. Due to the rough surface of the shaft and the hub, only some of the highest roughness peaks are in contact, which form the real contact area, which is one of the contact properties of the press fit at the micro level. The real contact area of the press fit increases with increasing normal force and is independent of the apparent contact area, [14]. This influence is not covered by the classic Coulomb law of sliding friction. The static coefficient of sliding friction can be significantly different on finely polished surfaces than on a realistic surface processed by turning and/or grinding. It should be emphasized that the roughness, both technologically and in terms of friction, must be in the optimal range, as described in [15] and [16].

The theory of the kinetic friction between two pure metal surfaces according to Bowden-Tabor [17] is based on the cold-press junction. At the points, at micro-level, where the two bodies are in touch, the surface of both, due to the great pressure, come so close together that a solid metallic bond is formed. Due to the pressure, the peaks on the rough surface deform plastically, new points come into contact and the contact surface increases. The oxide layer breaks down, impurities are squeezed out and a pure metal contact is formed. Cold pressed points at micro-contact are called contact area bridges. Under high pressure material plasticization, stress relaxation over time occurs, so the contact area bridges are enlarged and press fit load-capacity increases over time. As shown in this study, the load capacity increases with time only as long as the interference is small enough. However, with larger interferences, plasticization penetrates deeper into the shaft and bushing material and the opposite effect occurs; load capacity decreases. For pure metals without lubrication, strong adhesion and high normal force, the coefficient of static friction, according to the Bowden-Tabor, is approximated as:

$$\mu_0 \approx \frac{1}{\sqrt{3}} \cdot \left(\frac{2}{3-\xi} \right), \tag{6}$$

where ξ is so called Bowden-Tabor coefficient; in generally $\xi < 3$. In the case of $\xi = 0$ to 2.35, the coefficient of static friction is $\mu_0 = 0.4$ up to 1.8 [14].

4 SURFACE ROUGHNESS MEASUREMENT

With the help of a microscope KEYENCE VHX-7000, the roughness of the contact surface of the shaft and the hub was measured. The contact surface of the hub was turned to *Ra* 2.38 µm, and the contact surface of the shaft was ground to *Ra* 0.43 µm. Fig. 7 shows the image of both contact surfaces, and Fig. 8 shows the surface roughness profile for the shaft and the hub at reference length of 800 µm in initial contact.

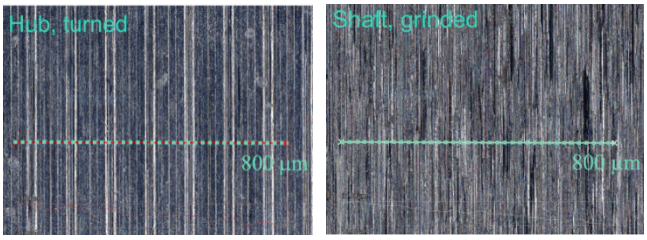


Figure 7 Surface image at reference length of 800 µm for shaft and hub

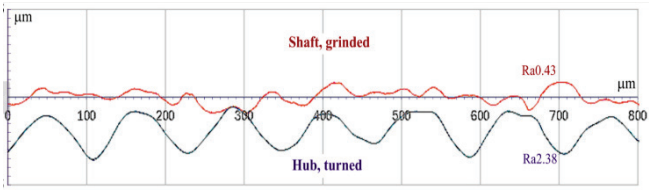


Figure 8 Surface roughness profile for the shaft and the hub at reference length of 800 µm in initial contact

At initial contact the roughness profile of the contact surface on the hub is deeper and periodically repeated, while the roughness profile of the ground shaft is shallower and random, Fig. 8.

5 FEM ANALYSIS

Fig. 9 shows the FEM analysis of the modelled contact roughness under a specified initial pressure. Contact bridges at micro-level and plasticization are visible.

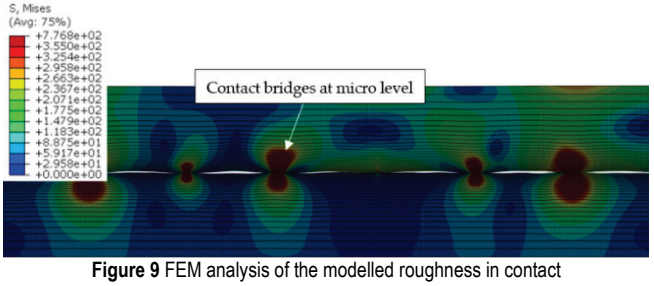


Figure 9 FEM analysis of the modelled roughness in contact

The numerical simulation of the stress-strain state in contact was performed using the software package SIMULIA Abaqus 2024. An elastic, plastic, and visco-plastic material behaviour model was employed. A hard contact model with a friction coefficient of 0.4 was used for modelling the contact. This approach allowed us to simulate material yielding around the contact area and stress relaxation over time due to material creep. In the linear elastic region, the Young's modulus was 206 GPa and Poisson's ratio was 0.3. The material has a yield strength of 357 MPa and strain-hardens to 591 MPa. A simple model accounting for time-dependent hardening was used for creep, with physical constants for this model experimentally determined as A = 1.05277e-28, n = 9.4649, and m = –1. The geometry was represented as a plane axisymmetric model. In the contact

region, finite elements were refined to capture the relatively complex profile of the measured surface; nevertheless, the simulation required 556800 CAX4 type finite elements to describe the shaft and hub. The model was constrained in a way that allowed it to deform freely. Due to micro-plasticization in contact bridges, additional time-dependent plasticization occurs.

6 RESULTS AND DISCUSSIONS

The load capacity of the press fit was measured as a function of the interference, at the beginning immediately after assembly and after 18 years. The test results are collected in Tab. 1 and shown on the graph in Fig. 10. In addition to the measurement of the load capacity of the press fit, a calculation based on the linear theory of elasticity was performed for comparison, as presented in chapter 3. At the calculation, the coefficient of static friction $\mu_0 = 0.2$ has been taken in account, in accordance with the recommendations from the literature Decker [18] and Nieman [19].

The results of the test of the load capacity of the press fit, depending on the interference, show that the actual load capacity is much higher than the calculated one. With time, the load capacity of the press fit increases to an interference of 0.05 mm, after which a noticeable drop in the load capacity is detected with respect to the initial state.

Table 1 Measured and calculated load capacity of the press fit

Δ (mm)	Measured load capacity F_M (kN)	Load capacity after 18 years F_M (kN)	Calculated load capacity F_C (kN), $\mu_0 = 0.2$	Integration constant A	Integration constant B
0.015	66.1	181.12	22.2	4.10e-5	0.075
0.030	124.5	177.16	44.5	8.38e-5	0.150
0.045	167.8	182.37	66.8	1.25e-4	0,224
0.060	233.4	190.0	89.1	1.67e-4	0,299
0.075	288.1	211.26	111.4	2.09e-4	0.374
0.085	326.1	220.0	126.2	2.37e-4	0.424
0.100	384.1	198.0	148.5	2.79e-4	0.498
0.115	410.5	145.97	170.7	3.21e-4	0.573

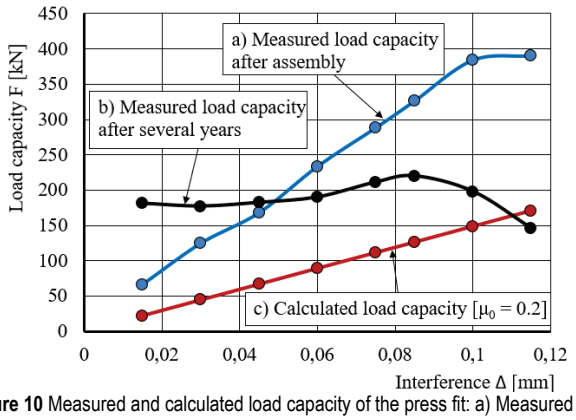


Figure 10 Measured and calculated load capacity of the press fit: a) Measured load capacity after assembly; b) Measured load capacity after 18 years; c) Calculated load capacity, using linear elastic theory of Lamé, and $\mu_0 = 0.2$

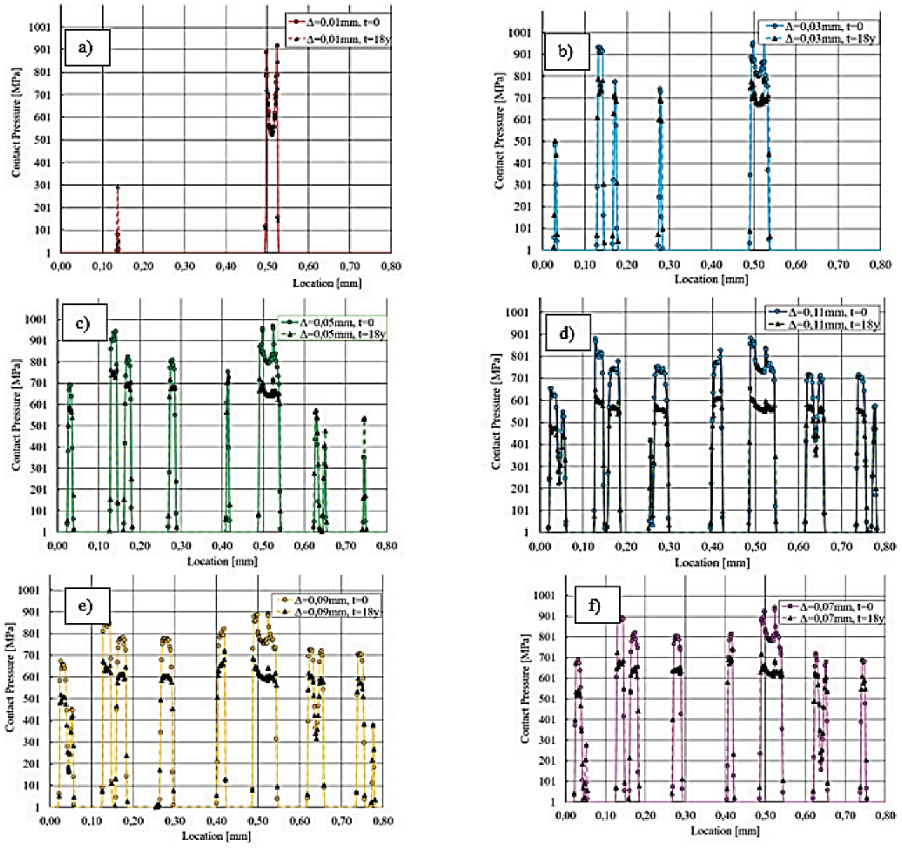


Figure 11 Numerical calculated contact pressure at reference length of 0,8 mm of contact surface at assembly, $t = 0$, and after 18 years, $t = 18y$ in dependence on interference Δ: a) Δ = 0.01 mm; b) 0.03; c) 0.05; d) 0.07; e) 0,09 and f) 0.11 mm.

Fig. 11 from a) to f) shows the results of the numerical analysis of the contact pressure for interferences from 0.01 to 0.11 mm at a reference contact length of 0.8 mm, immediately after assembly and after the passage of 18 years. The numerical analysis includes an elastic-visco-plastic material model. With increasing interference, the number of contact bridges increases, and also real contact area on the micro level. As the interference enlarges, the contact pressure on the real contact area also increases, the material plasticizes there, and with time stress relaxation occurs, which can be seen in Fig. 12 as a drop in the contact pressure during time.

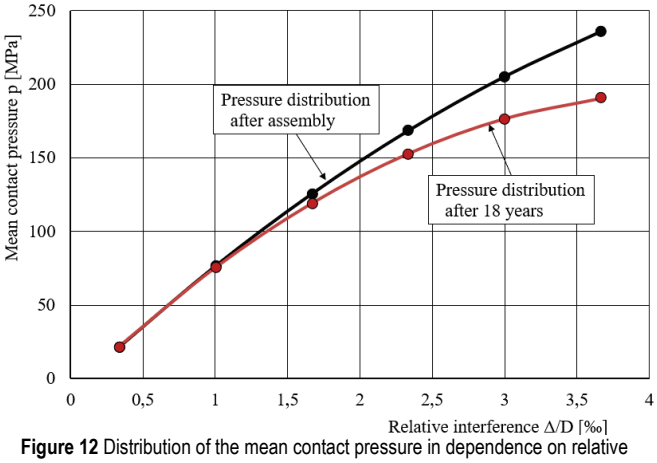


Figure 12 Distribution of the mean contact pressure in dependence on relative interference *Δ/D*

Fig. 12 shows the distribution of the mean contact pressure as a function of the relative interference, immediately after the mechanical joining of the press fit and after 18 years. The curves are a summary of the FEM analysis results from Fig. 11. Up to a relative interference of 1.6 ‰, there is no significant drop in contact pressure, while at larger relative interferences the drop in the contact pressure is greater due to stress relaxation of the plasticized material.

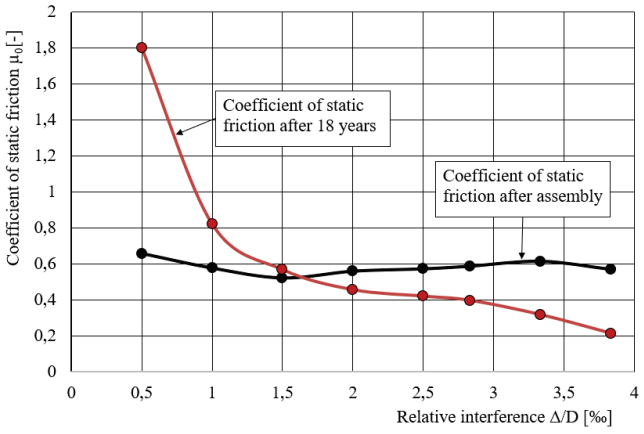


Figure 13 Distribution of the coefficient of static friction in dependence on relative interference *Δ/D* at the apparent contact area

Based on the numerically calculated mean contact pressure and the measured load capacity of the press fit, the coefficient of static friction on the apparent contact area was calculated. Fig. 13 shows the distribution of the coefficient of static friction μ_0 in dependence on relative interference of the press fit. After assembly, the mean coefficient of static friction was around 0.6, while after 18 years the coefficient of static friction is very variable and amounts to $\mu_0 = 1.8$ for a relative interference of 0.5 ‰ and 0.2 for a relative interference of 3.83 ‰.

Fig. 14 shows the distribution of the Bowden-Tabor coefficient ξ in dependence on relative interference of the press fit. The coefficient ξ was calculated based on Eq. (6). At the beginning, the value of the coefficient was around 1, while after 18 years, the value of this coefficient changes from $\xi = 2.36$ for a relative interference of 0.5 ‰ to $\xi = -2.42$ for a relative interference of 3.83 ‰.

In any case, regardless of the size of the interface, plasticization of the material occurs due to the high pressure and the small actual contact area. With a relative interference of up to 1.6 ‰, the pacification of the material is at the micro level, which means that only the peaks of the roughness of the actual contact area are plastically deformed, which increases as a result. Due to the cohesive forces, the two materials in contact are cold pressed and very strong contact bridges are formed. Over time, creep of the material occurs and the contact bridges become even stronger. The contact pressure does not decrease, because creep occurs only at the micro level. The outer layers of the material remain in the elastic range and act as a spring at all times, maintaining the contact pressure.

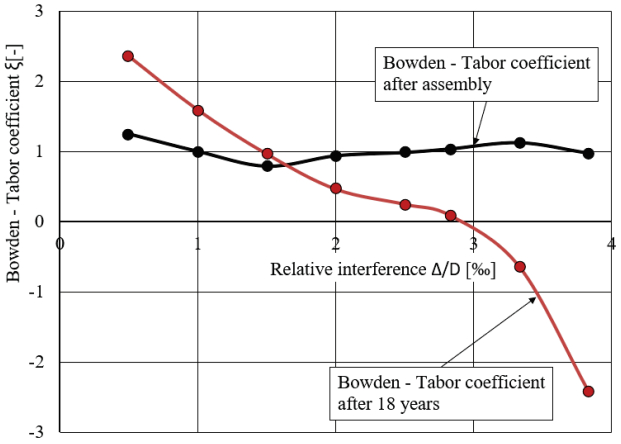


Figure 14 Distribution of the Bowden – Tabor coefficient ξ in dependence on relative interference *Δ/D* at the apparent contact area

Fig. 15 shows an equivalent creep strain, and Figure 16 shows an equivalent plastic strain in dependence on interference. At interference above 0.05 mm, plasticization extends to the macro level, which causes more intensive creep of the material in the hub. Contact pressure, the coefficient of sliding friction, and consequently the load capacity of the press fit decrease.

At an interference above 1.6 ‰, plasticization penetrates deeper into the material at the macro level, Figs. 15 and 16. In the initial state, when the material still has its initial strength, the load capacity of the press fit is relatively high. Over time, creep (relaxation) also occurs at the macro level, which reduces the spring action and contact pressure. As a

result, the friction coefficient and the load capacity of the entire press fit are strongly reduced.

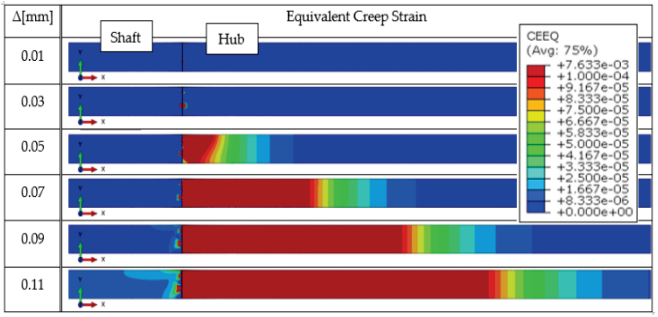


Figure 15 An equivalent creep strain in dependence on interference

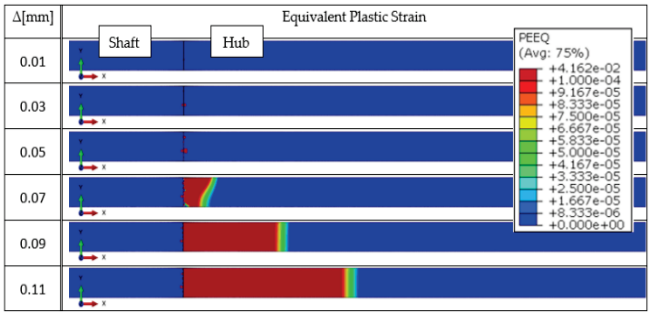


Figure 16 An equivalent plastic strain in dependence on interference

7 CONCLUSIONS

The study demonstrates that the loading capacity of a press fit is significantly affected by time-changing material and interference at used material and surface roughness. Experimental results have shown that the loading capacity increases with increasing interference up to a critical point, above which excessive plastic deformation at the macro level, over 1.6 ‰, reduces the strength of the joint. At a smaller interference, when plasticization is located only at the peaks of the rough surface at the micro level, over time, creep of the material causes the formation of a larger actual contact surface, increasing the coefficient of friction and thus the loading capacity of the press fit. At a larger interference, over 1.6 ‰, plasticization penetrates deeper into the material at the macro level. This causes stronger creep, which causes the contact pressure and coefficient of friction to drop; the press fit loses its original loading capacity.

Numerical simulation at a model of a rough contact surface confirms that the use of computational models based on the material elasticity and simpl Coulon's friction law, is not sufficient. For a deeper understanding of the process, it is necessary to use the visco-plastic behaviour of the material on a real rough contact surface. It turns out that it is more appropriate to use a friction model that takes in to account material cohesion and the formation of contact bridges. The results and findings show that at used material and surface roughness, it is also necessary to choose an appropriate oversize interference, if we want to produce a time-reliable and stiff joint.

The usual approach to calculating press fit in engineering practice uses the traditional Coulomb sliding friction model, where the sliding friction coefficient is more or less constant $\mu_0 \approx 0.2$. A solution approach is given based on a visco-plastic material and a more complex friction model; $\mu_0 \approx 0.6$. The results presented are valid for the used material and surface roughness. If we want to give a comprehensive recommendation for engineering use, it would be necessary to carry out a series of studies using different engineering materials and different combinations of surface roughness.

Acknowledgments

The authors acknowledge the Slovenian Research Agency ARIS for funding the Research Program P2-0137 "Numerical and experimental analysis of nonlinear mechanical systems".

8 REFERENCES

[1] Lewis, R., Marshall, M. B. & Dwyer-Joyce, R. S. (2005). Measurement of interface pressure in interference fits. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 219*(2), 127-139. https://doi.org/10.1243/095440605X8432

[2] Madej, J. (2018). A strength analysis of the interference-fit joints. *Mechanik*, **91**(11), 1032-1034. https://doi.org/10.17814/mechanik.2018.11.185

[3] Stamenković, D., Milošević, M., Mijajlović, M. & Banic, M. (2011). Estimation of the static friction coefficient for press fit joints. *Journal of the Balkan Tribological Association, 17*(3), 341-355. https://www.researchgate.net/publication/241688933

[4] Persson, B.N.J. (2023). Influence of Surface Roughness on Press Fits. *Tribology Letters, 71*(19). https://doi.org/10.1007/s11249-022-01688-y

[5] Croccolo, D., De Agostinis, M. & Vincenzi, N. (2011). How to improve static and fatigue strength in press-fitted joints using anaerobic adhesive. *Proceedings of the Institution of Mechanical Engineers, Part C. Journal of Mechanical Engineering Science, 225*(12), 2792-2803. https://doi.org/10.1177/0954406211411402

[6] Kim, S. S. & Lee, D. G. (2006). Design of the hybrid composite bearing assembled by interference fit. *Composite Structures, 75*(1-4), 222-230. https://doi.org/10.1016/j.compstruct.2006.04.026

[7] Croccolo, D., Agostinis, M. D. & Vincenzi, N. (2012). Design of hybrid steel-composite interference fitted and adhesively bonded connections. *International Journal of Adhesion and Adhesives, 37*, 19-25. https://doi.org/10.1016/j.ijadhadh.2012.01.011

[8] Mori, K., Bay, N., Fratini, L., Micari, F. & Tekkaya, A. E. (2013). Joining by plastic deformation. *CIRP Annals – Manufacturing Technology, 62*(2), 673-694. https://doi.org/10.1016/j.cirp.2013.05.004

[9] Yang, G. M., Coquille, J. C., Fontaine, J. F. & Lambertin, M. (2001). Influence of roughness on characteristics of tight interference fit of a shaft and a hub. *International Journal of Solids and Structures, 38*(42-43), 7691-7701. https://doi.org/10.1016/S0020-7683(01)00035-X

[10] Zhang, Y., McClain, B. & Fang, X. (2000). Design of interference fits via finite element method. *International Journal of Mechanical Sciences, 42*(9), 1835-1850. https://doi.org/10.1016/S0020-7403(99)00072-7

[11] Lippmann, H. (1992). The effect of a temperature cycle on the stress distribution in a shrink fit. *International Journal of Plasticity*, (1), 567-582. https://doi.org/10.1016/0749-6419(92)90031-7
[12] Wang, G. S. (1994). Stress analysis for a lug under various conditions. *The Journal of Strain Analysis for Engineering Design, 29*(1), 7-16. https://doi.org/10.1243/03093247V291007
[13] Rudnytskyj, A. (2018). Simulations of contact mechanics and wear of linearly reciprocating block-on-flat sliding test. *Mechanical Engineering, master level, Lulea University of Technology, department of Engineering Sciences and Mathematics*. https://www.divaportal.org/smash/record.jsf?pid=diva2%3A1242768&dswid=7465
[14] Popov, V. L. (2010). Contact Mechanics and Friction. *Springer-Verlag Berlin Heidelberg*. https://doi.org/10.1007/978-3-642-10803-7
[15] Vukelic D., Prica M., Ivanov V., Jovicic G., Budak I. & Luzanin O. (2022). Optimization of Surface Roughness Based on Turning Parameters and Insert Geometry. *Int. Journal of Simulation Modelling, 21*(3), 417-428. https://doi.org/10.2507/IJSIMM21-3-607
[16] Milošević, A., Simunovic, G., Kanović, Ž., Simunovic, K., Santosi, Z., Šokac, M. & Vukelić, D. (2024). Comprehensive Evaluation of Dimensional Deviation, Flank Wear, SurfaceRoughness and Material Remuval Rate in Dry Turningof C45. *Steel Facta Universitatis, Series: Mechanical Engineering, 22*(4), 547-566. https://doi.org/10.22190/FUME240403024M
[17] Bowden, F. P. & Tabor, D. (2001). The Friction and Lubrication of Solids. *Clarendon Press, Oxford*.
[18] Decker, K.-H. (2024). Maschinenelemente. *Carl Hanser Verlag München Wien.*
[19] Nieman, G. (1981). Maschinen-elemente Band I Konstruktion und Berechnung von Verbindungen, Lagern, Wellen, *Springer-Verlag Berlin Heidleberg GmbH*.

Authors' contacts:

Vinko Močilnik, Assist. Prof. Dr.
University of Maribor, Faculty of Mechanical Engineering
Smetanova 17, 2000 Maribor, Slovenia
+386 31-331-664, vinko.mocilnik@gmail.com

Nenad Gubeljak, Full Prof. Dr.
(Corresponding author)
University of Maribor, Faculty of Mechanical Engineering
Smetanova 17, 2000 Maribor, Slovenia
+386 31 659 279, nenad.gubeljak@um.si

Jožef Predan, Assoc. Prof. Dr.
University of Maribor, Faculty of Mechanical Engineering
Smetanova 17, 2000 Maribor, Slovenia
+386 41 333 701, jozef.predan@um.si

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20250221185213

Original scientific paper
Received: 2025-02-21, Accepted: 2025-03-11

Advancing PFMEA Decision-Making: FRADAR Based Prioritization of Failure Modes Using AP, RPN, and Multi-Attribute Assessment in the Automotive Industry

Nikola Komatina, Dragan Marinković*, Danijela Tadić, Dragan Pamučar

Abstract: This research proposes a novel way to improve Process Failure Modes and Effects Analysis (PFMEA) by using the Fuzzy RAnking based on the Distances And Range (FRADAR) method to prioritize activities for mitigating or eliminating failure modes in the automotive industry. The suggested approach seeks to improve classic PFMEA by using fuzzy sets to better assess risk-related criteria and their inherent uncertainty. The criteria used to prioritize actions for mitigating failure modes include the Action Priority (AP) and Risk Priority Number (RPN) approach, as well as the cost-effectiveness of actions, the time required to resolve issues, and their impact on production, all of which are assessed by a PFMEA team using predefined linguistic terms and suggestions. Applied to a case study of a Tier-1 automotive supplier, the FRADAR method effectively ranks failure modes, providing a structured and precise approach for action prioritization. The results highlight the model's potential to enhance decision-making processes, offering a robust framework for implementing PFMEA recommendations in the automotive industry.

Keywords: Action Priority; Automotive industry; FRADAR; PFMEA; RPN

1 INTRODUCTION

One of the important tools for improving the reliability of the manufacturing process in the automotive industry is Failure Mode and Effect Analysis (FMEA), specifically its version related to the manufacturing process, Process Failure Mode and Effect Analysis (PFMEA). By applying this method, a specific methodological procedure is used to determine the priority of the considered failure modes, or those factors that may potentially disrupt the realization or affect the final outcome of the manufacturing process.

The importance of FMEA analysis in the automotive industry is also reflected in the fact that its application is mandatory and prescribed by the IATF 16949:2016 standard [1]. Therefore, the latest version of the handbook [2] provides joint guidelines for applying different types of FMEA analysis in the automotive industry. In addition, the new manual modifies the FMEA methodology for risk assessment/priority determination of failure modes, shifting from the traditional Risk Priority Number (RPN) approach to the Action Priority (AP) approach. Although the new approach definitely addresses some of the methodological issues and ambiguities of the old approach, there is still space for improvement, as explained in previous research [3, 4].

The traditional FMEA analysis determines the priority of failure modes based on three risk factors: Severity of failure mode consequence (S), occurrence or frequency of failure mode (O), and the possibility of detecting failure modes (D). Like RPN, the AP approach is based on the analysis of these three risk factors. In both cases, the FMEA team determines or evaluates the value of these risk factors, which are stated on a scale of (1-10). These three risk factors are multiplied to determine RPN, which has a value between 1 and 1000. The AP approach has predefined scenarios for each combination of the values of these three risk factors, which means that no calculation is required. The priority of failure modes in the RPN approach is determined based on its value, and the threshold values (which may vary) primarily depend on the type of product, while in the AP approach, priority can be defined as Low (L), Medium (M), or High (H) (see [2]).

Although the priority of failure modes, as determined by the RPN or AP approaches, gives useful information about the importance of each failure mode, neither approach provides recommendations on the sequence in which activities should be done to address the causes of their occurrence. Therefore, the problem discussed in this paper is to determine the sequence of actions at the failure mode level by applying the RPN parameter and considering additional criteria. The AP approach is used to perform the primary selection of failure modes (failure modes of L priority are not considered).

In this case, the traditional RPN approach is used rather than the AP approach, as the application of the RPN approach provides a numerical value, which corresponds to the problem being addressed. This type of problem has been scarcely discussed in the literature. A similar problem was only considered in the paper [5], where the authors extended the PFMEA analysis by applying the Interval Type-2 Fuzzy Analytic Hierarchy Process (IT2FAHP) to determine the importance of the risk factors S, O, and D. The priority of actions was determined using two heuristic methods, namely the Genetic Algorithm (GA) and Variable Neighborhood Search (VNS). Additional criteria for determining the sequence of actions included downtime costs and maintenance costs caused by the occurrence of the failure mode.

This study investigates the application of the RAnking Based on Distances And Range (RADAR) method to improve FMEA analysis for defining the sequence of actions during failure modes. This is a novel method developed in the studies [6, 7] and is part of the Multi-Attribute Decision-Making (MADM) methods based on distance. The author applied the method to determine the priority of failure modes [6]. Through a comparison with the RPN approach, the weighted RPN approach, and two MADM methods, it was demonstrated that the RADAR method produces results most similar to conventional FMEA analysis. This indicates that the RADAR method is suitable for this type of problem. Furthermore, in the study [7], the author applied the RADAR method, as well as its modification RADAR II, for equipment

selection in the automotive industry. By comparing the results obtained using multiple MADM methods, it was found that the method provides reliable and robust solutions. For this reason, this research employs the RADAR method extended with the use of fuzzy sets theory [8, 9].

In this paper, triangular fuzzy numbers (TFNs) were used to describe uncertain values. Although various approaches for representing uncertainty have been used in combination with different MADM methods in the literature, such as type-2 fuzzy sets [10], intuitionistic fuzzy sets [11] or rough sets [12], the authors believe that triangular fuzzy numbers are the most suitable for this type of problem due to their simplicity. They do not require high computational complexity and user friendly for practical application.

The main goal of this paper is to enhance the traditional FMEA analysis by applying a methodology for prioritizing actions, specifically for determining the sequence of treatment or intervention on the causes of failure modes. For this purpose, the fuzzy RADAR (FRADAR) approach has been used.

This paper is structured as follows: Section 1 introduces the research issue and discusses the study's objectives and importance. Section 2 provides a literature review that summarizes relevant studies and approaches in the topic. Section 3 outlines the methodology, including the FRADAR approach used in the study. Section 4 includes a case study demonstrating the suggested model's use in the automotive industry. Finally, Section 5 presents conclusions, discusses the findings, and suggests future study areas.

2 LITERATURE REVIEW

As previously stated, the subject of prioritizing failure mode activities, that is, selecting the sequence in which to treat the causes or mitigate the impacts of failure modes, has not received much attention in the literature. However, multiple studies have been conducted in which FMEA analysis or another production issue has been improved through the use of various MADM techniques [13-15].

This study primarily focuses on using a fuzzy MADM methodology to improve FMEA. The relevant literature identifies many methodologies and domains of application for fuzzy and fuzzy MADM concepts [16-18]. Some authors [19, 20] expand the FMEA analysis by employing fuzzy sets theory, specifically TFNs, to characterize risk factor values and/or weights. On the other hand, in paper [21] authors employ a combination of triangular and trapezoidal fuzzy numbers to represent uncertain values. In the study [22] the authors applied a similar principle; however, they used the VIKOR method for ranking failure modes.

FMEA analysis combined with the VIKOR method (Serb. VIšekriterijumsko KOmpromisno Rangiranje) [23] also can be found in the papers of [24, 25]. In both studies, the authors use triangular fuzzy numbers to describe uncertainties.

Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) [26], being one of the most widely used methods in various MADM domains [27, 28], is also combined with FMEA analysis in several studies [29-31]. All of these studies utilize triangular fuzzy numbers. It is worth noting that [24] applied the TOPSIS method in combination with fuzzy sets theory to test the results obtained by the VIKOR method.

Besides these two methods, the Decision-Making Trial and Evaluation Laboratory (DEMATEL) [32, 33] method was used in combination with triangular fuzzy numbers in the study by Liu et al. [34], while it was applied with the Evaluation based on Distance from Average Solution (EDAS) [35] method in the paper of [36]. In studies by [31, 36], the Analytic Hierarchy Process (AHP) [37] method was employed to determine the weights of risk factors. The criteria weights (or risk factors, in this case), as in this study, can be directly assessed by decision-makers and aggregated using an appropriate operator. However, numerous other methods can also be used for this purpose, as has been done in various types of optimization problems [7, 38-40].

When it comes to the application domain, it is quite broad regarding fuzzy FMEA analysis. In addition to the automotive industry [21], it is applied in the food industry [20], electronics [34], agriculture [31], transportation [24], enterprise architecture [25], well drilling [29], renewable energy investments [36], industrial processes [30], project risk management [19], floating production storage and offloading [22] and more.

As previously mentioned in the earlier sections of this paper, each of these applications focused on prioritization or risk assessment. However, no one in the relevant literature has addressed the problem of determining the sequence in which failure modes should be addressed. In this study, this issue is tackled through the application of a combined PFMEA-FRADAR approach.

3 METHODOLOGY

In this study, the traditional PFMEA analysis was extended by the application of the FRADAR approach to determine the specific sequence of addressing failure modes i, $i = 1, \ldots, I$. This approach streamlines the implementation of proposed activities for PFMEA team members. Given that PFMEA recommends focusing on removing or minimizing the causes of failure modes with high (H) and medium (M) priority, only these failure modes were investigated in this study.

In Fig. 1, the proposed methodology is presented, while the explanation of the phases shown in the figure is provided later in this chapter.

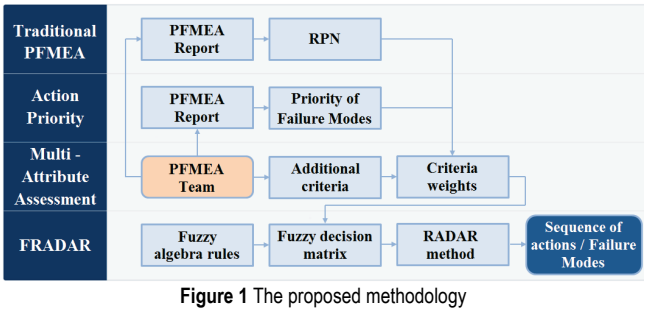


Figure 1 The proposed methodology

The failures' priority was identified using the AP technique. The RPN parameter value was then calculated for each failure mode and used to determine the failure mode value using the RPN criterion.

The criteria used to determine the sequence of actions are k, k = 1, …, K: RPN value (k = 1), cost-effectiveness of mitigation actions (k = 2), time required to resolve the issue (k = 3), and impact on production process realization (k = 4).

The criteria were defined in collaboration with the PFMEA team of an automotive industry company. The company is a Tier-1 supplier in the automotive supply chain, with its production facility located in the Republic of Serbia.

The first criterion pertains to the RPN parameter value, traditionally calculated by multiplying the risk factor values S (Severity), O (Occurrence), and D (Detection). The second criterion represents the cost-effectiveness of mitigation actions for the considered failure mode. The third criterion refers to the time required to resolve the issue, as estimated by the PFMEA team members. Finally, the fourth criterion considers the impact of the proposed actions on potentially halting certain production phases or even the entire production process.

3.1 Determination of Criteria Weights

A total of 5 members of the PFMEA team participated in the research, where e, e = 1, …, E. Each of them assessed the importance of the considered criteria based on pre-defined linguistic statements. The linguistic expressions used were:

- Absolutely unimportant criterion (L1): (0, 0, 0.3),
- Slightly important criterion (L2): (0.1, 0.3, 0.5),
- Moderately important criterion (L3): (0.25, 0.5, 0.75),
- Very important criterion (L4): (0.5, 0.7, 0.9),
- Absolutely important criterion (L5): (0.7, 1, 1).

The domain of fuzzy numbers is defined on the measurement scale (0-1). The evaluations of the PFMEA team members were aggregated using the fuzzy arithmetic mean operator [41]. In this way, the weights of the considered criteria were determined.

3.2 Modelling of Uncertain Values of Criteria

Unlike determining the criteria weights, where the assessments of the PFMEA team members are aggregated using the fuzzy arithmetic mean operator, in this case, the assessments are made by consensus. Using predefined linguistic expressions (Tab. 1), the PFMEA team members evaluate the values for each considered failure mode i, i = 1, …, I, according to each criterion k, k = 1, …, K. The domain of fuzzy numbers is defined on the measurement scale (1-10).

In this context, the RPN value (k = 1) is considered as a benefit-type criterion, as higher RPN values suggest a higher priority for addressing failure modes, implying that steps should be taken to limit the risks associated with these modes. This guarantees that failure modes with higher RPN values receive priority for corrective measures. Similarly, the cost-effectiveness of mitigation actions (k =2) is a benefit-type criterion since it assesses the efficiency of risk reduction in terms of cost, with the goal of maximizing benefits while minimizing expenses. On the other side, both the time necessary to remedy the issue (k = 3) and the impact on production process realization (k = 3) are cost-type criteria, as they relate to the possible rise in costs due to longer resolution timeframes or larger impact on production process.

Table 1 Linguistic expressions modelled using triangular fuzzy numbers for evaluating the values of failure modes according to the considered criteria

Fuzzy number notation	TFN	Cost-effectiveness of mitigation actions (k = 2)	Time required to resolve the issue (k = 3)	Impact on production process realization (k = 4)
		Description		
Very low value (V1)	(1, 1, 3.5)	Actions are extremely costly, with little to no benefit in terms of risk reduction.	Very quick actions that require minimal time and effort to implement.	The issue has a negligible impact on the production process.
Low value (V2)	(2, 3.5, 5)	Actions provide some cost-effectiveness but with limited benefit.	Actions that require a short amount of time to implement but are still relatively simple.	The issue may slightly disrupt the production process but does not halt it.
Medium value (V3)	(4, 5.5, 7)	Actions provide balanced cost-effectiveness with noticeable risk reduction.	Actions requiring a moderate amount of time and effort to implement.	The issue may cause moderate disruption but can be managed without halting the process.
High value (V4)	(6, 7.5, 9)	Actions are very cost-effective, significantly reducing risks and costs.	Actions that require considerable time and resources to implement.	The issue can cause significant disruption to the production process, requiring attention to avoid delays.
Very high value (V5)	(7.5, 10, 10)	Actions are extremely cost-effective, almost eliminating risks and costs.	Actions that require a substantial amount of time and effort to fully implement.	The issue has a severe impact, potentially halting the production process and requiring immediate resolution.

3.3 Proposed Algorithm

In this research, a model for determining the priority sequence of actions to address failure modes in an automobile industry company was developed and tested. The proposed model can be given using the following algorithm:

Step 1. Data collection on failure modes from the existing PFMEA report. Based on the application of the AP methodology, failure modes of H and M priority categories were selected, i, i = 1, …, I.

Step 2. In collaboration with the members of the PFMEA team from a Tier-1 company in the automotive supply chain, a set of criteria was defined, k, k = 1, …, K.

Step 3. The importance of the considered criteria, $\tilde{w}_k^e$, was assessed by the members of the PFMEA team, e, e = 1, …, E. Their assessments, were aggregated using the fuzzy arithmetic mean operator to determine criteria weights, $\tilde{\omega}_k$.

Step 4. Based on the assessments made by the PFMEA team members, which were reached through consensus, a fuzzy decision matrix is formed:

$$\left[\tilde{M}_{ik}\right]_{I\times K} \tag{1}$$

The next steps of the proposed algorithm represent an extension of the RADAR method [6, 7] through the

application of fuzzy set theory and fuzzy algebra rules [9, 42]:

Step 5. Construct the fuzzy maximum proportion matrix, $\tilde{\alpha}$:

$$\left[\tilde{\alpha}_{ik}\right]_{I\times K} \tag{2}$$

For the benefit type of criteria:

$$\tilde{\alpha}_{ik} = \frac{\frac{\max\limits_{i}\tilde{M}_{ik}}{\tilde{M}_{ik}}}{\left[\left(\frac{\max\limits_{i}\tilde{M}_{ik}}{\tilde{M}_{ik}}\right)+\left(\frac{\tilde{M}_{ik}}{\min\limits_{i}\tilde{M}_{ik}}\right)\right]} \tag{3}$$

For the cost type of criteria:

$$\tilde{\alpha}_{ik} = \frac{\frac{\tilde{M}_{ik}}{\min\limits_{i}\tilde{M}_{ik}}}{\left[\left(\frac{\max\limits_{i}\tilde{M}_{ik}}{\tilde{M}_{ik}}\right)+\left(\frac{\tilde{M}_{ik}}{\min\limits_{i}\tilde{M}_{ik}}\right)\right]} \tag{4}$$

Step 6. Construct the fuzzy minimum proportion matrix, $\tilde{\beta}$:

$$\left[\tilde{\beta}_{ik}\right]_{I\times K} \tag{5}$$

For the benefit type of criteria:

$$\tilde{\beta}_{ik} = \frac{\frac{\tilde{M}_{ik}}{\min\limits_{i}\tilde{M}_{ik}}}{\left[\left(\frac{\max\limits_{i}\tilde{M}_{ik}}{\tilde{M}_{ik}}\right)+\left(\frac{\tilde{M}_{ik}}{\min\limits_{i}\tilde{M}_{ik}}\right)\right]} \tag{6}$$

For the cost type of criteria:

$$\tilde{\beta}_{ik} = \frac{\frac{\max\limits_{i}\tilde{M}_{ik}}{\tilde{M}_{ik}}}{\left[\left(\frac{\max\limits_{i}\tilde{M}_{ik}}{\tilde{M}_{ik}}\right)+\left(\frac{\tilde{M}_{ik}}{\min\limits_{i}\tilde{M}_{ik}}\right)\right]} \tag{7}$$

Step 7. Construct the empty range matrix:

$$\left[E_{ik}\right]_{I\times K} \tag{8}$$

where:

$$E_{ik} = \left|\alpha_{ik} - \beta_{ik}\right| \tag{9}$$

where: α_{ik} is defuzzified value of $\tilde{\alpha}_{ik}$, and β_{ik} is defuzzified value of $\tilde{\beta}_{ik}$. Defuzzification of this values is performed by procedure defined in [43].

Step 8. Construct the fuzzy relative relationship matrix:

$$\left[\widetilde{RR}_{ik}\right]_{I\times K} \tag{10}$$

where:

$$\widetilde{RR}_{ik} = \frac{\tilde{\alpha}_{ik}}{\tilde{\beta}_{ik} + E_{ik}} \tag{11}$$

Step 9. Construct the fuzzy weighted relative relationship matrix:

$$\left[\widetilde{WRR}_{ik}\right]_{I\times K} \tag{12}$$

where:

$$\widetilde{WRR}_{ik} = \widetilde{RR}_{ik} \cdot \tilde{\omega}_{k} \tag{13}$$

Step 10. Aggregated ranking index, RI_i:

$$RI_i = \frac{\min\sum_{k=1}^{K} defuzz\,\widetilde{WRR}_i}{\sum_{k=1}^{K} defuzz\,\widetilde{WRR}_i} \tag{14}$$

Defuzzification of $\widetilde{WRR}_i$ is performed using the defuzzification procedure defined in the [43].

Afterward, the ranking of the considered failure modes is determined, where the highest value of RI_i indicates the failure mode that should be addressed first. The reverse is also true. In the next section, the proposed algorithm is tested on a case study from the automotive industry.

4 CASE STUDY

The case study used to test the developed model was conducted in a company that is a Tier-1 supplier in the automotive supply chain. The company specializes in manufacturing leather upholstery for automobile interiors. Data from the PFMEA report were collected from the sewing production process phase. The considered failure modes, related to the sewing machine, (with H and M priority) are presented in Tab. 2, along with their S, O, and D values, AP approach category, and RPN values (step 1 of the proposed algorithm).

Out of a total of 27 identified failure modes for considered production phase, 16 with H or M priority are shown in Tab. 2. The remaining failure modes have L priority and were therefore not considered in this analysis. Of course, if financial resources and other resources are available, the PFMEA team may choose to include them for consideration.

In this case, the limitation is the available resources, which is why failure modes with L priority were not considered.

According to the new AP approach, the priority of failure modes is determined. It can be seen in Tab. 2 that this priority is not always compatible with the RPN value. Some failure modes, such as (i = 5), have a higher RPN value compared to, for example, (i = 10), even though (i = 10) has a H priority and (i = 5) has a M priority. This is because the AP approach prioritizes Severity over the other two risk factors. In this study, the set of failure modes under consideration was selected according to the AP approach, but the RPN value was used in the decision matrix, as it allows for the quantitative expression of priority/risk.

Table 2 Identified failure modes and their priority according to AP and RPN approaches

Number of failure mode	Failure mode	S	O	D	AP	RPN
i = 1	Misaligned stitching	8	6	6	H	288
i = 2	Thread breakage	9	6	7	H	378
i = 3	Skipped stitches	8	6	5	H	240
i = 4	Needle damage	9	4	4	H	144
i = 5	Incorrect thread tension	7	5	6	M	210
i = 6	Material slippage during sewing	8	5	5	M	200
i = 7	Uneven stitch length	7	4	6	M	168
i = 8	Puckering of fabric	8	5	6	M	240
i = 9	Loose stitches	7	4	6	M	168
i = 10	Sewing machine's feed mechanism failure	9	4	4	H	144
i = 11	Incorrect alignment of panels	7	5	5	M	175
i = 12	Fabric tearing during stitching	9	5	7	H	315
i = 13	Color mismatch in stitching	6	4	7	M	168
i = 14	Excessive thread wastage	7	5	4	M	140
i = 15	Needle overheating	8	4	5	M	160
i = 16	Damage to embroidery patterns	6	6	5	M	180

As already mentioned, and as presented in Step 2 of the proposed algorithm, a set of criteria has been defined: RPN value (k = 1), cost-effectiveness of mitigation actions (k = 2), time required to resolve the issue (k = 3), and impact on production process realization (k = 4).

The criteria were defined in collaboration with the members of the PFMEA team during a panel discussion. During the same panel discussion, the PFMEA team members expressed their assessments regarding the importance of each criterion, as well as the value of each failure mode according to each considered criterion.

According to Step 3 of the proposed algorithm, the importance of the considered criteria is evaluated by the members of the PFMEA team. Their assessments are:

$$\tilde{w}_1^1 = L4 \quad \tilde{w}_2^1 = L3 \quad \tilde{w}_3^1 = L2 \quad \tilde{w}_4^1 = L3$$
$$\tilde{w}_1^2 = L4 \quad \tilde{w}_2^2 = L2 \quad \tilde{w}_3^2 = L1 \quad \tilde{w}_4^2 = L3$$
$$\tilde{w}_1^3 = L3 \quad \tilde{w}_2^3 = L3 \quad \tilde{w}_3^3 = L2 \quad \tilde{w}_4^3 = L3$$
$$\tilde{w}_1^4 = L4 \quad \tilde{w}_2^4 = L4 \quad \tilde{w}_3^4 = L3 \quad \tilde{w}_4^4 = L3$$
$$\tilde{w}_1^5 = L5 \quad \tilde{w}_2^5 = L3 \quad \tilde{w}_3^5 = L2 \quad \tilde{w}_4^5 = L2$$

The aggregated and normalized weight values of the criteria are:

$$\tilde{\omega}_1 = (0.17,\ 0.37,\ 0.82) \qquad \tilde{\omega}_2 = (0.10,\ 0.26,\ 0.67)$$
$$\tilde{\omega}_3 = (0.04,\ 0.14,\ 0.47) \qquad \tilde{\omega}_4 = (0.08,\ 0.23,\ 0.64)$$

According to Step 4 of the proposed algorithm, the PFMEA team members provided consensus-based evaluations for the value of each failure mode against each considered criterion. The only exception is the RPN value, which was directly taken from the PFMEA report previously developed by the same PFMEA team. The fuzzy decision matrix is presented in Tab. 3.

Each of the five PFMEA team members recorded their assessments on paper for later discussion and deliberation. A compromise solution was reached through agreement, or, in cases where consensus could not be achieved, an average score (assessment) was calculated. Such cases were very rare. In most instances, their assessments largely coincided and were consistent.

Table 3 The fuzzy decision matrix

	k = 1	k = 2	k = 2	k = 2
i = 1	288	V4	V3	V4
i = 2	378	V3	V3	V4
i = 3	240	V4	V3	V5
i = 4	144	V2	V2	V3
i = 5	210	V3	V3	V4
i = 6	200	V3	V3	V4
i = 7	168	V3	V2	V3
i = 8	240	V4	V3	V4
i = 9	168	V3	V3	V4
i = 10	144	V4	V4	V5
i = 11	175	V4	V3	V4
i = 12	315	V5	V3	V5
i = 13	168	V3	V2	V4
i = 14	140	V2	V2	V2
i = 15	160	V2	V2	V3
i = 16	180	V4	V3	V4

By applying step 5 of the proposed algorithm, the fuzzy maximum proportion matrix, $\tilde{\alpha}$, is formed. Given that the minimum and maximum values are clearly identifiable due to the linguistic terms being arranged in a gradient, there is no need for comparing fuzzy numbers in this step.

The first element of the fuzzy maximum proportion matrix, $\tilde{\alpha}$, is calculated as follows:

$$\tilde{\alpha}_{11} = \frac{\dfrac{\max\limits_i \tilde{M}_{i1}}{\tilde{M}_{11}}}{\left[\left(\dfrac{\max\limits_i \tilde{M}_{i1}}{\tilde{M}_{11}}\right)+\left(\dfrac{\tilde{M}_{11}}{\min\limits_i \tilde{M}_{i1}}\right)\right]} = \frac{\dfrac{378}{288}}{\dfrac{378}{288}+\dfrac{288}{140}} = 0.39$$

A crisp value can also be represented as a fuzzy number:

$$\tilde{\alpha}_{11} = (0.39,\ 0.39,\ 039)$$

Tab. 4 presents the remaining values of the fuzzy maximum proportion matrix, $\tilde{\alpha}$. Fuzzy algebra rules were applied for operations with fuzzy numbers.

Tab. 5 presents values of the fuzzy minimum proportion matrix, $\tilde{\beta}$ (Step 6 of the proposed algorithm). Also, fuzzy algebra rules were applied for operations with fuzzy numbers.

By applying the procedure shown in step 7, the empty range matrix, E_{ik} (Tab. 6), was determined. In order to determine the values of the elements of this matrix, the values of the matrices $\tilde{\alpha}$ and $\tilde{\beta}$ were defuzzified. The subsequent steps of the proposed algorithm continue using fuzzy numbers and fuzzy algebra rules.

Table 4 The fuzzy maximum proportion matrix, $\tilde{\alpha}$

i	$k=1$	$k=2$	$k=2$	$k=2$
$i=1$	(0.39, 0.39, 0.39)	(0.14, 0.38, 0.82)	(0.14, 0.54, 7.19)	(0.19, 0.62, 5.14)
$i=2$	(0.27, 0.27, 0.27)	(0.18, 0.54, 1.34)	(0.14, 0.54, 7.19)	(0.19, 0.62, 5.14)
$i=3$	(0.48, 0.48, 0.48)	(0.14, 0.38, 0.82)	(0.14, 0.54, 7.19)	(0.24, 0.74, 4.22)
$i=4$	(0.72, 0.72, 0.72)	(0.20, 0.74, 2.63)	(0.06, 0.32, 17.50)	(0.13, 0.46, 7.50)
$i=5$	(0.55, 0.55, 0.55)	(0.18, 0.54, 1.34)	(0.14, 0.54, 7.19)	(0.19, 0.62, 5.14)
$i=6$	(0.57, 0.57, 0.57)	(0.18, 0.54, 1.34)	(0.14, 0.54, 7.19)	(0.19, 0.62, 5.14)
$i=7$	(0.65, 0.65, 0.65)	(0.18, 0.54, 1.34)	(0.06, 0.32, 17.50)	(0.13, 0.46, 7.50)
$i=8$	(0.48, 0.48, 0.48)	(0.14, 0.38, 0.82)	(0.14, 0.54, 7.19)	(0.19, 0.62, 5.14)
$i=9$	(0.65, 0.65, 0.65)	(0.18, 0.54, 1.34)	(0.14, 0.54, 7.19)	(0.19, 0.62, 5.14)
$i=10$	(0.72, 0.72, 0.72)	(0.14, 0.38, 0.82)	(0.20, 0.68, 5.00)	(0.24, 0.74, 4.22)
$i=11$	(0.63, 0.63, 0.63)	(0.14, 0.38, 0.82)	(0.14, 0.54, 7.19)	(0.19, 0.62, 5.14)
$i=12$	(0.35, 0.35, 0.35)	(0.12, 0.26, 0.59)	(0.14, 0.54, 7.19)	(0.24, 0.74, 4.22)
$i=13$	(0.65, 0.65, 0.65)	(0.18, 0.54, 1.34)	(0.06, 0.32, 17.50)	(0.19, 0.62, 5.14)
$i=14$	(0.73, 0.73, 0.73)	(0.20, 0.74, 2.63)	(0.06, 0.32, 17.50)	(0.05, 0.26, 18.75)
$i=15$	(0.67, 0.67, 0.67)	(0.20, 0.74, 2.63)	(0.06, 0.32, 17.50)	(0.13, 0.46, 7.50)
$i=16$	(0.62, 0.62, 0.62)	(0.18, 0.54, 1.34)	(0.14, 0.54, 7.19)	(0.19, 0.62, 5.14)

Table 5 The fuzzy minimum proportion matrix, $\tilde{\beta}$

i	$k=1$	$k=2$	$k=2$	$k=2$
$i=1$	(0.61, 0.61, 0.61)	(0.19, 0.62, 5.14)	(0.15, 0.46, 1.36)	(0.14,0.38,0.82)
$i=2$	(0.73, 0.73, 0.73)	(0.13, 0.46, 7.50)	(0.15, 0.46, 1.36)	(0.14,0.38,0.82)
$i=3$	(0.52, 0.52, 0.52)	(0.19, 0.62, 5.14)	(0.15, 0.46, 1.36)	(0.12,0.26,0.59)
$i=4$	(0.28, 0.28, 0.28)	(0.05, 0.26, 18.75)	(0.17, 0.68, 2.81)	(0.18,0.54,1.34)
$i=5$	(0.45, 0.45, 0.45)	(0.13, 0.46, 7.50)	(0.15, 0.46, 1.36)	(0.14,0.38,0.82)
$i=6$	(0.43, 0.43, 0.43)	(0.13, 0.46, 7.50)	(0.15, 0.46, 1.36)	(0.14,0.38,0.82)
$i=7$	(0.35, 0.35, 0.35)	(0.13, 0.46, 7.50)	(0.17, 0.68, 2.81)	(0.18,0.54,1.34)
$i=8$	(0.52, 0.52, 0.52)	(0.19, 0.62, 5.14)	(0.15, 0.46, 1.36)	(0.14,0.38,0.82)
$i=9$	(0.35, 0.35, 0.35)	(0.13, 0.46, 7.50)	(0.15, 0.46, 1.36)	(0.14,0.38,0.82)
$i=10$	(0.28, 0.28, 0.28)	(0.19, 0.62, 5.14)	(0.11, 0.32, 0.80)	(0.12,0.26,0.59)
$i=11$	(0.37, 0.37, 0.37)	(0.19, 0.62, 5.14)	(0.15, 0.46, 1.36)	(0.14,0.38,0.82)
$i=12$	(0.65, 0.65, 0.65)	(0.24, 0.74, 4.22)	(0.15, 0.46, 1.36)	(0.12,0.26,0.59)
$i=13$	(0.35, 0.35, 0.35)	(0.13, 0.46, 7.50)	(0.17, 0.68, 2.81)	(0.14,0.38,0.82)
$i=14$	(0.27, 0.27, 0.27)	(0.05, 0.26, 18.75)	(0.17, 0.68, 2.81)	(0.20,0.74,2.63)
$i=15$	(0.33, 0.33, 0.33)	(0.05, 0.26, 18.75)	(0.17, 0.68, 2.81)	(0.18,0.54,1.34)
$i=16$	(0.38, 0.38, 0.38)	(0.13, 0.46, 7.50)	(0.15, 0.46, 1.36)	(0.14,0.38,0.82)

Table 6 The empty range matrix, E_{ik}

i	$k=1$	$k=2$	$k=2$	$k=2$
$i=1$	0.22	1.54	1.96	1.54
$i=2$	0.46	2.02	1.96	1.54
$i=3$	0.04	1.54	1.96	1.41
$i=4$	0.44	5.16	4.74	2.02
$i=5$	0.09	2.02	1.96	1.54
$i=6$	0.14	2.02	1.96	1.54
$i=7$	0.30	2.02	4.74	2.02
$i=8$	0.04	1.54	1.96	1.54
$i=9$	0.30	2.02	1.96	1.54
$i=10$	0.44	1.54	1.55	1.41
$i=11$	0.27	1.54	1.96	1.54
$i=12$	0.30	1.41	1.96	1.41
$i=13$	0.30	2.02	4.74	1.54
$i=14$	0.46	5.16	4.74	5.16
$i=15$	0.35	5.16	4.74	2.02
$i=16$	0.24	2.02	1.96	1.54

By applying the procedure presented in step 8 of the proposed algorithm, the values of the elements of the fuzzy relative relationship matrix (Tab. 7) are calculated.

Example of calculating the element of the fuzzy relative relationship matrix:

$$\widetilde{RR}_{12} = \frac{\tilde{\alpha}_{12}}{\tilde{\beta}_{12} + E_{12}} =$$

$$= \frac{(0.14,\ 0.38,\ 0.82)}{(0.19,\ 0.62,\ 5.14) + (1.54,\ 1.54,\ 1.54)} = (0.02,\ 0.18,\ 0.47)$$

In step 9 of the proposed algorithm, the values are made more difficult, as shown in Tab. 8.

Example of calculating the element of the fuzzy weighted relative relationship matrix:

$$\widetilde{WRR}_{11} = \widetilde{RR}_{11} \cdot \tilde{\omega}_1 =$$
$$= (0.47,\ 0.47,\ 0.47) \cdot (0.17,\ 0.37,\ 0.82) = (0.08,\ 0.17,\ 0.38)$$

According to step 10 of the proposed algorithm, the aggregated ranking index, RI_i, is calculated. An example of the calculation for the first failure mode:

$$RI_1 = \frac{\min \sum_{k=1}^{K} defuzz\, \widetilde{WRR}_i}{\sum_{k=1}^{K} defuzz\, \widetilde{WRR}_i} = \frac{1.39}{1.56} = 0.89$$

The value of the aggregated ranking index, RI_i, for the remaining failure modes, as well as the ranking of failure modes, is shown in Tab. 9.

Table 7 The fuzzy relative relationship matrix, $\left[\widetilde{RR}_{ik}\right]_{I\times K}$

i	$k = 1$	$k = 2$	$k = 2$	$k = 2$
$i = 1$	(0.47, 0.47, 0.47)	(0.02, 0.18, 0.47)	(0.04, 0.22, 3.40)	(0.08, 0.32, 3.07)
$i = 2$	(0.28, 0.28, 0.28)	(0.02, 0.27, 0.80)	(0.04, 0.22, 3.40)	(0.08, 0.32, 3.07)
$i = 3$	(0.65, 0.65, 0.65)	(0.02, 0.18, 0.47)	(0.04, 0.22, 3.40)	(0.11, 0.41, 2.55)
$i = 4$	(1.43, 1.43, 1.43)	(0.01, 0.41, 1.65)	(0.01, 0.12, 8.20)	(0.05, 0.22, 4.37)
$i = 5$	(0.81, 0.81, 0.81)	(0.02, 0.27, 0.80)	(0.04, 0.22, 3.40)	(0.08, 0.32, 3.07)
$i = 6$	(0.87, 0.87, 0.87)	(0.02, 0.27, 0.80)	(0.04, 0.22, 3.40)	(0.08, 0.32, 3.07)
$i = 7$	(1.15, 1.15, 1.15)	(0.02, 0.27, 0.80)	(0.01, 0.12, 8.20)	(0.05, 0.22, 4.37)
$i = 8$	(0.65, 0.65, 0.65)	(0.02, 0.18, 0.47)	(0.04, 0.22, 3.40)	(0.08, 0.32, 3.07)
$i = 9$	(1.15, 1.15, 1.15)	(0.02, 0.27, 0.80)	(0.04, 0.22, 3.40)	(0.08, 0.32, 3.07)
$i = 10$	(1.43, 1.43, 1.43)	(0.02, 0.18, 0.47)	(0.07, 0.30, 2.41)	(0.11, 0.41, 2.55)
$i = 11$	(1.08, 1.08, 1.08)	(0.02, 0.18, 0.47)	(0.04, 0.22, 3.40)	(0.08, 0.32, 3.07)
$i = 12$	(0.40, 0.40, 0.40)	(0.02, 0.11, 0.33)	(0.04, 0.22, 3.40)	(0.11, 0.41, 2.55)
$i = 13$	(1.15, 1.15, 1.15)	(0.02, 0.27, 0.80)	(0.01, 0.12, 8.20)	(0.08, 0.32, 3.07)
$i = 14$	(1.49, 1.49, 1.49)	(0.01, 0.41, 1.65)	(0.01, 0.12, 8.20)	(0.01, 0.11, 10.79)
$i = 15$	(1.23, 1.23, 1.23)	(0.01, 0.41, 1.65)	(0.01, 0.12, 8.20)	(0.05, 0.22, 4.37)
$i = 16$	(1.03, 1.03, 1.03)	(0.02, 0.27, 0.80)	(0.04, 0.22, 3.40)	(0.08, 0.32, 3.07)

Table 8 The fuzzy weighted relative relationship matrix, $\left[\widetilde{WRR}_{ik}\right]_{I\times K}$

i	$k = 1$	$k = 2$	$k = 2$	$k = 2$
$i = 1$	(0.08, 0.17, 0.38)	(0.002, 0.05, 0.32)	(0.002, 0.03, 1.60)	(0.01, 0.07, 1.97)
$i = 2$	(0.05, 0.11, 0.23)	(0.002, 0.07, 0.54)	(0.002, 0.03, 1.60)	(0.01, 0.07, 1.97)
$i = 3$	(0.11, 0.24, 0.53)	(0.002, 0.05, 0.32)	(0.002, 0.03, 1.60)	(0.01, 0.09, 1.63)
$i = 4$	(0.24, 0.53, 1.17)	(0.001, 0.11, 1.11)	(0.000, 0.02, 3.85)	(0.004, 0.05, 2.80)
$i = 5$	(0.14, 0.30, 0.66)	(0.002, 0.07, 0.54)	(0.002, 0.03, 1.60)	(0.01, 0.07, 1.97)
$i = 6$	(0.15, 0.32, 0.72)	(0.002, 0.07, 0.54)	(0.002, 0.03, 1.60)	(0.01, 0.07, 1.97)
$i = 7$	(0.19, 0.42, 0.94)	(0.002, 0.07, 0.54)	(0.000, 0.02, 3.85)	(0.004, 0.05, 2.80)
$i = 8$	(0.11, 0.24, 0.53)	(0.002, 0.05, 0.32)	(0.002, 0.03, 1.60)	(0.01, 0.07, 1.97)
$i = 9$	(0.19, 0.42, 0.94)	(0.002, 0.07, 0.54)	(0.002, 0.03, 1.60)	(0.01, 0.07, 1.97)
$i = 10$	(0.24, 0.53, 1.17)	(0.002, 0.05, 0.32)	(0.003, 0.04, 1.13)	(0.01, 0.09, 1.63)
$i = 11$	(0.18, 0.40, 0.88)	(0.002, 0.05, 0.32)	(0.002, 0.03, 1.60)	(0.01, 0.07, 1.97)
$i = 12$	(0.07, 0.15, 0.33)	(0.002, 0.03, 0.22)	(0.002, 0.03, 1.60)	(0.01, 0.09, 1.63)
$i = 13$	(0.19, 0.42, 0.94)	(0.002, 0.07, 0.54)	(0.000, 0.02, 3.85)	(0.01, 0.07, 1.97)
$i = 14$	(0.25, 0.55, 1.22)	(0.001, 0.11, 1.11)	(0.000, 0.02, 3.85)	(0.001, 0.03, 6.91)
$i = 15$	(0.21, 0.46, 1.01)	(0.001, 0.11, 1.11)	(0.000, 0.02, 3.85)	(0.004, 0.05, 2.80)
$i = 16$	(0.18, 0.38, 0.85)	(0.002, 0.07, 0.54)	(0.002, 0.03, 1.60)	(0.01, 0.07, 1.97)

Table 9 Order of taking actions for the considered failure modes (ranking)

i	RI_i	*Rank*
$i = 1$	0.89	3-4
$i = 2$	0.89	3-4
$i = 3$	0.90	2
$i = 4$	0.42	15
$i = 5$	0.77	7
$i = 6$	0.76	8-9
$i = 7$	0.47	13
$i = 8$	0.85	5
$i = 9$	0.71	11
$i = 10$	0.80	6
$i = 11$	0.76	8-9
$i = 12$	1.00	1
$i = 13$	0.52	12
$i = 14$	0.30	16
$i = 15$	0.43	14
$i = 16$	0.73	10

The presented rankings show the priority order in which the identified failure modes should be handled. Fabric tearing during stitching (i = 12) is given the highest priority, indicating its importance to the production process and overall quality. This is followed by skipped stitches (i = 3), with misaligned stitching (i = 1) and thread breakage (i = 2) tied for third place. These failure modes reveal severe concerns with sewing quality and machine functionality that need to be addressed immediately.

The results are reasonable and consistent with practical expectations, prioritizing failure modes with the greatest potential to interrupt production and damage product quality. Lower-priority failure modes, such as excessive thread wastage (i = 14) and needle damage (i = 4), are placed 15th and 16th, indicating that they have a less immediate impact than other issues. However, they should be addressed as part

of a larger reform strategy. The suggested technique effectively distinguishes between failure modes, directing decision-makers to more targeted and economical mitigation actions.

If the obtained results are compared with the traditional RPN approach, as well as the new AP approach, certain similarities and differences can be observed in the considered case. According to the proposed approach, as well as the RPN parameter, the same top five failure modes are highlighted, but their rankings vary. The key reason for this occurrence is that the RPN criterion carries the greatest weight in this case study. However, changing the weights of the criteria can significantly influence the final ranking.

Although there are similarities, there are also significant differences. For example, failure mode (i = 10), which ranks sixth according to the proposed approach, is ranked fifteenth (second to last) according to RPN. Similarly, (i = 7), which is ranked thirteenth according to the proposed model, holds the tenth place according to RPN. There are also some minor variations (a change of one or two positions in the ranking), but these are not particularly significant.

When compared to the AP approach, certain differences can also be observed. For example, (i = 4), which has a priority rating of H, is ranked fifteenth according to the proposed model. The reason for this is the low cost-effectiveness of the proposed measures. Simply put, it is not economically viable for the company to allocate resources to eliminate the impact of this failure mode

5 CONCLUSION

The primary objective of this research was to improve traditional PFMEA by incorporating the FRADAR approach to properly prioritize failure modes during the manufacturing process. The suggested model addresses the shortcomings of the traditional RPN-based technique by introducing additional decision criteria and utilizing fuzzy set theory to handle uncertainty.

The proposed methodology, which was based on the FRADAR approach, included four essential criteria: RPN value, cost-effectiveness of mitigation actions, time required to resolve the issue, and impact on the production process. Fuzzy numbers represented the language assessments provided by PFMEA team members, ensuring accurate evaluations. The methodology was validated through a case study conducted in a Tier-1 automotive supplier, with a focus on leather seat cover production. The results showed rational prioritization, with higher-ranked failure types indicating their disruptive potential.

The proposed model has multiple advantages, including more flexibility in dealing with complex failure modes, higher precision in prioritization via fuzzy logic, and adaptability to different industrial environments. However, the dependence on expert judgment adds the possibility of subjectivity, and the fuzzy methodology may increase computational complexity.

Future study could concentrate on improving the model by incorporating criteria weight modifications, sophisticated computational methodologies, and expanding its application to other automobile manufacturing processes. This would improve the robustness of risk management systems and enable more effective decision-making in the automobile industry and beyond.

6 REFERENCES

[1] IATF 16949:2016. (2017). *Quality management system requirements for automotive production and relevant service parts organizations, 1st edition*. International Automotive Task Force.

[2] AIAG&VDA. (2019). *Failure Mode and Effects Analysis—FMEA Handbook: Design FMEA, process FMEA, supplemental FMEA for monitoring & system response*. Southfild, Michigan: Automotive Industry Action Group.

[3] Komatina, N., Tadić, D., Aleksić, A., & Banduka, N. (2022). The integrated PFMEA approach with interval type-2 fuzzy sets and FBWM: A case study in the automotive industry. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, *236*(6), 1201-1212. https://doi.org/10.1177/09544070211034799

[4] Liu, H.-C., Liu, L., & Liu, N. (2013). Risk evaluation approaches in failure mode and effects analysis: A literature review. *Expert Systems with Applications*, *40*(2), 828-838. https://doi.org/10.1016/j.eswa.2012.08.010

[5] Komatina, N., Tadić, D., Đurić, G., & Aleksić, A. (2023). Determination of manufacturing process failures priority under type 2 fuzzy environment: Application of genetic algorithm and Variable neighborhood search. *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering*, 095440892311605. https://doi.org/10.1177/09544089231160510

[6] Komatina, N. (2024). A compromise-based MADM approach for prioritizing failures: Integrating the RADAR method within the FMEA framework. *Jurnal Sistem Dan Manajemen Industri*, *8*(2), 73-88. https://doi.org/10.30656/jsmi.v8i2.9283

[7] Komatina, N. (2025). A Novel BWM-RADAR Approach for Multi-Attribute Selection of Equipment in the Automotive Industry. *Spectrum of Mechanical Engineering and Operational Research*, *2*(1), 104-120. https://doi.org/10.31181/smeor21202531

[8] Eti, S., Dinçer, H., Yüksel, S., & Gökalp, Y. (2025). A New Fuzzy Decision-Making Model for Enhancing Electric Vehicle Charging Infrastructure. *Spectrum of Decision Making and Applications*, *2*(1), 94-99. https://doi.org/10.31181/sdmap21202513

[9] Mendel, J. M. (2024). Type-1 Fuzzy Sets and Fuzzy Logic. In J. M. Mendel, *Explainable Uncertain Rule-Based Fuzzy Systems* (pp. 17-73). Springer International Publishing. https://doi.org/10.1007/978-3-031-35378-9_2

[10] Aleksić, A., Milanović, D. D., Komatina, N., & Tadić, D. (2023). Evaluation and ranking of failures in manufacturing process by combining best-worst method and VIKOR under type-2 fuzzy environment. *Expert Systems*, *40*(2), e13148. https://doi.org/10.1111/exsy.13148

[11] Sudžum, R., Nestić, S., Komatina, N., & Kraišnik, M. (2024). An Intuitionistic Fuzzy Multi-Criteria Approach for Prioritizing Failures That Cause Overproduction: A Case Study in Process Manufacturing. *Axioms*, *13*(6), 357. https://doi.org/10.3390/axioms13060357

[12] Božanić, D., Epler, I., Puška, A., Biswas, S., Marinković, D., & Koprivica, S. (2024). Application of the DIBR II – rough MABAC decision-making model for ranking methods and techniques of lean organization systems management in the

process of technical maintenance. *Facta Universitatis, Series: Mechanical Engineering*, *22*(1), 101. https://doi.org/10.22190/FUME230614026B

[13] Komatina, N., Tadić, D., Aleksić, A., & Jovanović, A. D. (2023). The assessment and selection of suppliers using AHP and MABAC with type-2 fuzzy numbers in automotive industry. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, *237*(4), 836-852. https://doi.org/10.1177/1748006X221095359

[14] Marković, A., Stojanović, B., Komatina, N., & Ivanović, L. (2024). Multi-attribute approach for selection of polymeric materials for manufacturing gears: A case study in the automotive industry. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *238*(22), 10744-10755. https://doi.org/10.1177/09544062241271690

[15] Sahoo, S. K., Choudhury, B. B., & Dhal, P. R. (2024). A Bibliometric Analysis of Material Selection Using MCDM Methods: Trends and Insights. *Spectrum of Mechanical Engineering and Operational Research*, *1*(1), 189-205. https://doi.org/10.31181/smeor11202417

[16] Mehdiabadi, A., Sadeghi, A., Yazdi, A. K., & Tan, Y. (2025). Sustainability Service Chain Capabilities in the Oil and Gas Industry: A Fuzzy Hybrid Approach SWARA-MABAC. *Spectrum of Operational Research*, *2*(1), 92-112. https://doi.org/10.31181/sor21202512

[17] Mishra, A. R., Rani, P., Cavallaro, F., & Alrasheedi, A. F. (2023). Assessment of sustainable wastewater treatment technologies using interval-valued intuitionistic fuzzy distance measure-based MAIRCA method. *Facta Universitatis, Series: Mechanical Engineering*, *21*(3), 359. https://doi.org/10.22190/FUME230901034M

[18] Mishra, A. R., & Rani, P. (2025). Evaluating and Prioritizing Blockchain Networks using Intuitionistic Fuzzy Multi-Criteria Decision-Making Method. *Spectrum of Mechanical Engineering and Operational Research*, *2*(1), 78-92. https://doi.org/10.31181/smeor21202527

[19] Roghanian, E., & Mojibian, F. (2015). Using fuzzy FMEA and fuzzy logic in project risk management. *Interdisciplinary Journal of Management Studies*, *8*(3), 373-395.

[20] Wessiani, N. A., & Sarwoko, S. O. (2015). Risk Analysis of Poultry Feed Production Using Fuzzy FMEA. *Procedia Manufacturing*, *4*, 270-281. https://doi.org/10.1016/j.promfg.2015.11.041

[21] Godina, R., Silva, B. G. R., & Espadinha-Cruz, P. (2021). A DMAIC Integrated Fuzzy FMEA Model: A Case Study in the Automotive Industry. *Applied Sciences*, *11*(8), 3726. https://doi.org/10.3390/app11083726

[22] Wang, L., Yu, Y., Liu, Z., Liu, Z., & Liu, X. (2024). An Enhanced Failure Mode and Effects Analysis Risk Identification Method Based on Uncertainty and Fuzziness. *Journal of Engineering Management and Systems Engineering*, *3*(3), 116-131. https://doi.org/10.56578/jemse030301

[23] Opricovic, S., & Tzeng, G.-H. (2007). Extended VIKOR method in comparison with outranking methods. *European Journal of Operational Research*, *178*(2), 514-529. https://doi.org/10.1016/j.ejor.2006.01.020

[24] Hajiagha, S. H. R., Hashemi, S. S., Mohammadi, Y., & Zavadskas, K. (2016). Fuzzy belief structure-based VIKOR method: An application for ranking delay causes of Tehran metro system by FMEA criteria. *Transport*, *31*(1), 108-118. https://doi.org/10.3846/16484142.2016.1133454

[25] Safari, H., Faraji, Z., & Majidian, S. (2016). Identifying and evaluating enterprise architecture risks using FMEA and fuzzy VIKOR. *Journal of Intelligent Manufacturing*, *27*(2), 475-486. https://doi.org/10.1007/s10845-014-0880-0

[26] Hwang, C. L., & Yoon, K. (1981). *Methods for multiple attribute decision making. Multiple attribute decision making: Methods and applications a state-of-the-art survey* (pp. 58-191).

[27] Biswas, A., Gazi, K. H., Sankar, P. M., & Ghosh, A. (2025). A Decision-Making Framework for Sustainable Highway Restaurant Site Selection: AHP-TOPSIS Approach based on the Fuzzy Numbers. *Spectrum of Operational Research*, *2*(1), 1-26. https://doi.org/10.31181/sor2120256

[28] Kizielewicz, B., & Sałabun, W. (2025). Benchmark study of re-identification methods based on stochastic fuzzy normalization and their application to decision-making problems in engineering. *Facta Universitatis, Series: Mechanical Engineering*, *Online first*, 1-22. https://doi.org/10.22190/FUME240916004K

[29] Khodadadi-Karimvand, M., & Shirouyehzad, H. (2021). Well drilling fuzzy risk assessment using fuzzy FMEA and fuzzy TOPSIS. *Journal of Fuzzy Extension and Applications*, *2*(2). https://doi.org/10.22105/jfea.2021.275955.1086

[30] Magalhães, W. R. D., & Lima Junior, F. R. (2021). A model based on FMEA and Fuzzy TOPSIS for risk prioritization in industrial processes. *Gestão & Produção*, *28*(4), e5535. https://doi.org/10.1590/1806-9649-2020v28e5535

[31] Zandi, P., Rahmani, M., Khanian, M., & Mosavi, A. (2020). Agricultural Risk Management Using Fuzzy TOPSIS Analytical Hierarchy Process (AHP) and Failure Mode and Effects Analysis (FMEA). *Agriculture*, *10*(11), 504. https://doi.org/10.3390/agriculture10110504

[32] Falatoonitoosi, E., Leman, Z., Sorooshian, S., & Salimi, M. (2013). Decision-Making Trial and Evaluation Laboratory. *Research Journal of Applied Sciences, Engineering and Technology*, *5*(13), 3476-3480. https://doi.org/10.19026/rjaset.5.4475

[33] Gazi, K. H., Raisa, N., Biswas, A., Azizzadeh, F., & Mondal, S. P. (2025). Finding Most Important Criteria in Women's Empowerment for Sports Sector by Pentagonal Fuzzy DEMATEL Methodology. *Spectrum of Decision Making and Applications*, *2*(1), 28-52. https://doi.org/10.31181/sdmap21202510

[34] Liu, H.-C., You, J.-X., Lin, Q.-L., & Li, H. (2015). Risk assessment in system FMEA combining fuzzy weighted average with fuzzy decision-making trial and evaluation laboratory. *International Journal of Computer Integrated Manufacturing*, *28*(7), 701-714. https://doi.org/10.1080/0951192X.2014.900865

[35] Alinezhad, A., & Khalili, J. (2019). EDAS Method. In A. Alinezhad & J. Khalili, *New Methods and Applications in Multiple Attribute Decision Making (MADM)* (Vol. 277, pp. 149-155). Springer International Publishing. https://doi.org/10.1007/978-3-030-15009-9_21

[36] Karatop, B., Taşkan, B., Adar, E., & Kubat, C. (2021). Decision analysis related to the renewable energy investments in Turkey based on a Fuzzy AHP-EDAS-Fuzzy FMEA approach. *Computers & Industrial Engineering*, *151*, 106958. https://doi.org/10.1016/j.cie.2020.106958

[37] Saaty, T. L. (2013). The Modern Science of Multicriteria Decision Making and Its Practical Applications: The AHP/ANP Approach. *Operations Research*, *61*(5), 1101-1118. https://doi.org/10.1287/opre.2013.1197

[38] Božanić, D., Pamučar, D., Milić, A., Marinković, D., & Komazec, N. (2022). Modification of the Logarithm Methodology of Additive Weights (LMAW) by a Triangular Fuzzy Number and Its Application in Multi-Criteria Decision Making. *Axioms*, *11*(3), 89.

https://doi.org/10.3390/axioms11030089
[39] Rakić, M., Žižović, M. M., Miljković, B., Njeguš, A., Žižović, M. R., & Đorđević, I. (2023). Multi-criteria selection of standards for system analyst activities in organizations. *Facta Universitatis, Series: Mechanical Engineering*, *21*(3), 433. https://doi.org/10.22190/FUME230521023R
[40] Žižović, M., Miljković, B., & Marinković, D. (2020). Objective methods for determining criteria weight coefficients: A modification of the CRITIC method. *Decision Making: Applications in Management and Engineering*, *3*(2), 149–161. https://doi.org/10.31181/dmame2003149z
[41] Gao, L. S. (1999). The fuzzy arithmetic mean. *Fuzzy Sets and Systems*, *107*(3), 335-348. https://doi.org/10.1016/S0165-0114(98)00050-5
[42] Jibril Alkali, A., & Kupongoh Samaila, S. (2021). A Study of Operators on Fuzzy Sets. *Mathematics Letters*, *7*(2), 30. https://doi.org/10.11648/j.ml.20210702.13
[43] Kahraman, C., Öztayşi, B., Uçal Sarı, İ., & Turanoğlu, E. (2014). Fuzzy analytic hierarchy process with interval type-2 fuzzy sets. *Knowledge-Based Systems*, *59*, 48-57. https://doi.org/10.1016/j.knosys.2014.02.001

Authors' contacts:

Nikola Komatina, PhD, Scientific Associate
University of Kragujevac, Faculty of Engineering,
Sestre Janjić 6, 34000 Kragujevac, Serbia
nkomatina@kg.ac.rs

Dragan Marinković, PhD, Senior Lecturer and Researcher
(Corresponding author)
1) Faculty of Mechanical Engineering and Transport Systems, TU Berlin,
Straße des 17. Juni 135, 10623 Berlin, Germany
2) Institute of Mechanical Science, Vilnius Gediminas Technical University,
LT-10105 Vilnius, Lithuania
dragan.marinkovic@tu-berlin.de

Danijela Tadić, PhD, Full professor
University of Kragujevac, Faculty of Engineering,
Sestre Janjić 6, 34000 Kragujevac, Serbia
galovic@kg.ac.rs

Dragan Pamučar, PhD, Full professor
1) Széchenyi István University,
Győr, Hungary
2) Department of Operations Research and Statistics, Faculty of Organizational Sciences, University of Belgrade,
Jove Ilića 154, 11000 Belgrade, Serbia
3) Department of Applied Mathematical Science, College of Science and Technology, Korea University
Sejong 30019, Republic of Korea
draganpamucar@gmail.com

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20250225100638

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

Replacing Backpropagation with the Forward-Forward (FF) Algorithm in Transformer Models: A Theoretical and Empirical Study on Scalable and Efficient Gradient-Free Training

Hyun Jung Kim, Sang Hyun Yoo*

Abstract: This study proposes a novel integration of the Forward-Forward (FF) algorithm into Transformer architectures as an efficient and gradient-free alternative to Backpropagation (BP). Motivated by the computational limitations of BP-such as high memory usage and gradient instability-we aim to examine whether FF can maintain comparable model performance while improving training efficiency. We present both theoretical justifications and empirical evaluations on the IMDB sentiment analysis dataset. Our experiments show that FF reduces training time by approximately 20% and memory usage by 30%, with only a marginal decrease in BLEU score (27.8 vs. 28.3) and slight increase in Perplexity (13.2 vs. 12.5). Furthermore, we extend our evaluation across varying model depths and hardware platforms (desktop GPU, cloud GPU, SoC-based laptop), and perform statistical testing and ablation studies to investigate FF's behavior within Transformer components. These results highlight the viability of FF for scalable, rethereforeurce-efficient Transformer training and provide a foundation for future research in hybrid and distributed deep learning frameworks.

Keywords: Backpropagation Alternative; Computational Efficiency; Efficient AI Training; Forward-Forward (FF) Algorithm; Training Stability; Transformer Models

1 INTRODUCTION

Recent advancements in deep learning have transformed fields like Natural Language Processing (NLP) and Computer Vision (CV), with Transformer models playing a crucial role. Leveraging self-attention mechanisms, these models have demonstrated remarkable performance in tasks such as machine translation, text generation, and image analysis [1-5]. However, Backpropagation (BP), the standard training algorithm, presents several challenges: high computational costs due to full gradient traversal through all layers [6], memory inefficiency from gradient storage during backward passes [7, 8], and susceptibility to vanishing/exploding gradients in deep architectures [9].

In 2022, Geoffrey Hinton introduced the Forward-Forward (FF) algorithm as an alternative to BP [10]. This method removes the need for backpropagation by using two forward passes to independently optimize effectiveness values for both positive and negative data. Compared to BP, FF reduces memory consumption and computational overhead while mitigating gradient vanishing and exploding issues, which commonly affect deep learning models. Additionally, FF enhances training efficiency in rethereforeurce-constrained environments, making it a promising alternative to BP in practical implementations.

Preliminary studies have demonstrated the computational efficiency and training stability of FF compared to BP [10, 11]. Furthermore, prior research has explored its advantages over non-gradient descent methods, such as genetic algorithms, across various neural network architectures, highlighting its potential as a transformative learning approach [12].

Despite the widespread use of Backpropagation (BP), a growing body of research has explored gradient-free alternatives for training deep networks. These alternatives include evolutionary strategies (EvoGrad) [22], neuroevolution frameworks [23], and, more recently, distributed FF implementations [24, 25] designed for federated environments. However, these approaches either require expensive population-based evaluations (as in GA), fail to scale to large architectures like Transformers, or lack direct layer-wise control. To date, no prior studies have demonstrated a complete FF-based training pipeline for Transformer models, which are inherently deep and resource-demanding. Our study addresses this critical gap by theoretically analyzing FF's applicability to Transformer components (e.g., self-attention and feedforward networks), empirically benchmarking FF versus BP across multiple training conditions, and providing a reproducible implementation plan with statistical validation.

2 RELATED WORKS & BACKGROUND

2.1 BP-Based Transformer Training and Emerging Alternatives

BP-based Transformer training presents several challenges, including high computational and memory costs [6, 7] and gradient instability [9]. Alternative techniques such as Sparse Attention [16] and Gradient Checkpointing [8, 17] have been proposed to mitigate these issues. However, these methods still rely on BP, limiting their ability to address its inherent inefficiencies fully. Tab. 1 compares different training methods, including BP, Sparse Attention, Gradient Checkpointing, and the FF algorithm [6, 9, 10, 12, 14, 15].

Table 1 Comparison of neural network training methods including Backpropagation (BP), Sparse Attention, Gradient Checkpointing, and Forward-Forward (FF)

Feature	BP	Sparse Attention	Gradient Checkpointing	FF Algorithm
Memory Usage	High	Moderate	Moderate	Low
Computational Complexity	High	Moderate	Moderate	Low
Gradient Stability	Susceptible	Limited Improvement	Limited Improvement	Stable

Unlike existing methods that offer only partial improvements, the FF algorithm eliminates backpropagation entirely by leveraging two forward passes. This approach reduces memory and computational costs and ensures stable gradient calculations, positioning FF as a promising and

potentially groundbreaking alternative for Transformer training [10, 12].

2.2 The Forward-Forward Algorithm

The Forward-Forward (FF) algorithm replaces Backpropagation (BP) with two forward passes, optimizing a goodness function that increases scores for positive samples while decreasing them for negative samples [10]. By eliminating backpropagation, FF removes the need to store gradients, significantly reducing memory usage [6]. Additionally, it improves training stability by mitigating gradient explosion and vanishing issues [9, 10]. Its architecture-agnostic nature allows it to be applied to a variety of neural network structures, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

Early studies demonstrated FF's effectiveness in tasks such as semantic segmentation, where disentangled attention mechanisms improved edge detection and small object recognition [18]. Hinton's research further highlighted FF's ability to enhance training stability and reduce memory usage while maintaining competitive accuracy [10-12]. Expanding on this, the author's 2024 study, "Analyzing the Characteristics of Gradient Descent and Non-Gradient Descent-Based Algorithms in Neural Network Learning," compared FF with BP and non-gradient descent methods such as Genetic Algorithms (GA). Across multiple architectures, including Competitive Memory Networks (CMM) and Multilayer Perceptrons (MLPs), FF consistently outperformed these methods in terms of computational efficiency and training stability [12].

Moreover, FF's flexibility and architecture-agnostic nature make it a promising alternative for future AI training methodologies, enabling efficient adaptation to diverse neural network architectures.

2.3 Justification for Applying FF to Transformers

Despite its promising attributes, integrating the Forward-Forward (FF) algorithm with Transformers remains largely unexplored. Given that Transformers heavily rely on Backpropagation (BP) for gradient updates, adopting FF could lead to significant improvements in computational efficiency and training stability, particularly for ultra-large models used in natural language processing (NLP) and computer vision (CV) applications.

Comprehensive testing on large-scale datasets such as GLUE and WMT is necessary to further validate its scalability and effectiveness. These evaluations will determine whether FF can maintain competitive accuracy while reducing computational overhead, making it a viable alternative for large-scale real-world applications.

2.4 Potential of FF as a BP Replacement

In addition to comparing FF with Backpropagation (BP), we conducted a literature-based comparative analysis with other gradient-free algorithms, notably Genetic Algorithms (GA) and EvoGrad. EvoGrad, as proposed in [19], has shown success in specific optimization tasks but exhibits limited generalizability in high-dimensional neural networks. Similarly, although GA offers flexibility through population-based optimization, it typically converges 2–5 times slower and requires higher memory usage compared to gradient-based methods.

In contrast, the Forward-Forward (FF) algorithm not only achieves lower computational costs and faster convergence but also maintains robust performance in large-scale models.

Table 2 Qualitative comparison of BP, Genetic Algorithms (GA), EvoGrad, and Forward-Forward (FF) across various performance dimensions

Aspect	Backpropagation (BP)	Genetic Algorithm (GA)	EvoGrad	Forward-Forward (FF)
Gradient Usage	Yes	No	No	No
Memory Usage	High	High	Moderate	Low
Convergence Speed	Fast	Slow	Moderate	Fast
Stability	Medium	High Variance	Low Variance	High
Applicability to Large Models	Yes	Limited	Limited	Yes

Unlike GA or EvoGrad, FF is architecture-agnostic and demonstrates high stability under parameter shifts. Additionally, it achieves competitive BLEU and perplexity scores in Transformer-based NLP tasks, further supporting its potential as a scalable, gradient-free training approach.

Further evaluation against hybrid methods (e.g., Sparse Attention combined with Checkpointing) is proposed in Section 4.2 Future Research Directions to more broadly validate FF's position within the landscape of training algorithms.

2.5 Novelty of the Study

Although previous research has highlighted the potential of the FF algorithm in neural network training [11, 12], no prior studies have explored its application to large-scale architectures such as Transformer models. The attention mechanisms and multilayered structures of Transformers significantly increase the computational complexity of BP, making them an ideal testbed for evaluating the efficiency and scalability of the FF algorithm [20].

This study represents the first attempt to integrate the FF algorithm into Transformer training and assess its potential as a BP replacement. By extending the application scope of the FF algorithm, this research aims to enhance the efficiency of Transformer model training while introducing a new framework for deep learning methodologies.

3 PROPOSED METHODOLOGY: FF-BASED TRANSFORMER DESIGN

This study proposes applying the Forward-Forward (FF) algorithm as an alternative to Backpropagation (BP) in training Transformer models. By eliminating backpropagation and conducting two forward passes, the FF

algorithm introduces a novel approach in which each layer learns "goodness" values for positive and negative data. This section details the design of integrating the FF algorithm into the Transformer training process.

3.1 Theoretical Analysis: FF vs. BP

The Forward-Forward (FF) algorithm offers several theoretical advantages over Backpropagation (BP), particularly in terms of computational cost, memory usage, and training stability.

Table 3 Comparison of Backpropagation (BP) and Forward-Forward (FF) training algorithms in Transformer models

Feature	Backpropagation (BP)	Forward-Forward (FF)
Memory Usage	High ($O(n2)$)	Low ($O(n)$)
Computational Cost	High ($O(n \cdot m)$)	Moderate ($O(n \cdot m/2)$)
Gradient Stability	Susceptible to issues	Stable

By addressing the bottlenecks of BP, the FF algorithm simplifies training, particularly for large-scale models, such as Transformers.

3.2 Existing Transformer Training Structure

Transformer models follow a structured training process consisting of the following key components.

1) Input Embedding and Self-Attention

The input data is first transformed into embeddings, serving as numerical representations of textual information. The self-attention mechanism then computes relationships between tokens by analyzing contextual interactions, enabling the model to capture long-range dependencies effectively [1, 8, 19, 22].

2) Feedforward Network

The attention-processed outputs are passed through a feedforward network incorporating non-linear activation functions, enhancing the model's learning capabilities and representation power [1, 7].

3) Loss Calculation and BP-Based Weight Update

The loss is computed by measuring the difference between the predicted and ground-truth outputs [1, 6]. This process involves propagating gradients through multiple layers, requiring the storage of intermediate activations. Consequently, Backpropagation (BP) incurs significant computational and memory costs [6, 8].

3.3 Applying the FF Algorithm to Transformers

The Forward-Forward (FF) algorithm introduces an alternative training approach that eliminates BP by employing two forward passes. The proposed process consists of the following steps.

Step 1. Generating Positive and Negative Data

- Positive Data: Original sequences from the dataset (e.g., "The cat sits on the mat.").
- Negative Data: Noisy or perturbed sequences generated by modifying word positions or introducing random noise (e.g., "The mat sits on the cat.")

Step 2. Calculating Goodness Values

For each layer in the Transformer model, a goodness score is computed by aggregating the activation values, which serves as an indicator of the layer's overall effectiveness in processing and propagating information through the network.

$$Goodness = \sigma(W \cdot X + b) \quad (1)$$

In our model, each layer computes its goodness score based on the weighted inputs. Specifically, let W represent the layer weights, X the input data, b the bias term, and σ an activation function (e.g., ReLU or Sigmoid). The model is trained to produce higher goodness values for positive data and lower goodness values for negative data.

For each layer, the goodness score is computed by applying the activation function to the weighted inputs plus the bias. Then, to aggregate the contributions from all layers, the overall positive and negative goodness scores are computed as follows.

$$G_+ = \sum_{l=1}^{L} \left\| \sigma(W_{lx_+} + b_l) \right\|_2 \quad \text{(Positive Goodness)}$$

$$G_- = \sum_{l=1}^{L} \left\| \sigma(W_{lx_-} + b_l) \right\|_2 \quad \text{(Negative Goodness)} \quad (2)$$

Here, x_+ and x_- denote the positive and negative input samples, respectively. These equations succinctly capture how the model aggregates the layer-wise goodness scores to assess its overall performance in distinguishing between positive and negative data.

Step 3. Loss Function and Local Weight Update

In our approach, rather than propagating gradients through the entire network as in traditional backpropagation, we adjust the weights of each layer independently based on a margin-based loss function. This loss function is defined as.

$$\mathcal{L} = max(0, Goodness_{negative} - Goodness_{positive} + \delta) \quad (3)$$

where G_{positive} and G_{negative} denote the aggregated goodness scores for positive and negative input samples, respectively, and δ is a stability margin that prevents trivial solutions.

$$W_l \leftarrow W_l + \eta \cdot \nabla_{w_l} \left(\left\| \sigma(W_{lx_+} + b_l) \right\|_2 - \left\| \sigma(W_{lx_-} + b_l) \right\|_2 \right) \quad (4)$$

In this formulation, W_l represents the weight matrix of the l-th layer, while η denotes the learning rate. The symbol σ refers to a non-linear activation function, such as ReLU or Sigmoid. The input vectors x_+ and x_- correspond to the positive and negative samples, respectively, and b_l indicates the bias term of the l-th layer. The term ∇W_l denotes the gradient computed with respect to W_l. This equation derives the gradient updates based on the goodness scores calculated for each layer from both the positive and negative inputs. The weights are then updated directly using this gradient, which illustrates a key feature of the Forward-Forward (FF) algorithm: it enables learning without relying on traditional backpropagation. Unlike backpropagation, which requires gradient propagation through multiple layers, the FF algorithm updates weights locally based on goodness values, significantly reducing memory usage and computational overhead.

Fig. 1 illustrates the FF training process, consisting of:

1) Generating positive and negative data samples.
2) Calculate goodness values for both positive and negative samples.
3) Updating weights based on the margin between these goodness values.

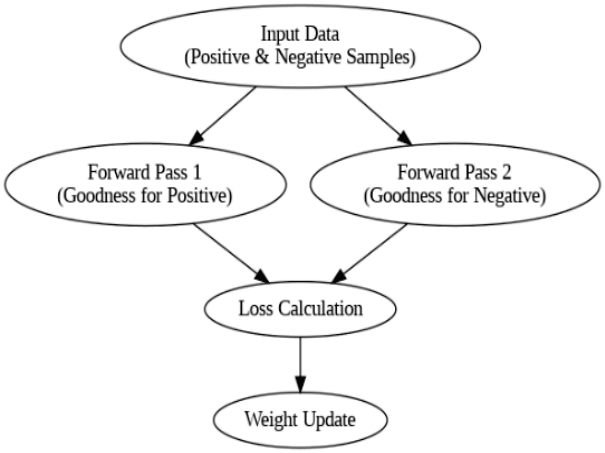


Figure 1 Flowchart illustrating the FF-based Transformer training process including goodness evaluation and weight update using positive and negative samples

3.4 Algorithm Design

The FF-based Transformer training process modifies the conventional Transformer training pipeline by replacing BP with goodness-based learning. The modified training steps are as follows.

1) Input Embedding and Self-Attention [1, 8, 19]

Positive and negative data samples are fed into the self-attention mechanism, which computes attention scores based on contextual relationships. Positive samples yield higher attention scores, while negative samples produce lower scores due to noise [23].

2) Feedforward Network

Goodness values are computed using the outputs from the attention mechanism, and the loss function is applied independently at each layer to optimize these values. Unlike BP-based training, FF does not require storing gradients, improving computational efficiency.

3) Algorithm for Generating Negative Data

To effectively train a Transformer using FF, negative samples must introduce slight perturbations while maintaining semantic similarity. This ensures the model learns to differentiate between coherent and perturbed sequences.

The following Python function generates negative samples by adding Gaussian noise to the original data.

Table 4 Algorithm for generating negative data in FF-based transformer training

```
import numpy as np

def generate_negative_data(positive_data):
    noise = np.random.normal(0, 0.1,
positive_data.shape)
    return positive_data + noise
```

This method perturbs the original data slightly, ensuring that negative samples remain contextually similar while challenging the model to distinguish correct from incorrect patterns.

4) Pseudo-Code for FF-Based Transformer Training

The following pseudo-code outlines the FF-based Transformer training process, clearly illustrating how goodness values and local updates are computed without backpropagation.

Table 5 Step-by-step pseudocode outlining the FF-based Transformer training process.

```
# FF-Based Transformer Training
for epoch in range(num_epochs):
  for x_pos in dataset:
    # Step 1: Generate a negative sample
    x_neg = generate_negative_sample(x_pos)

    for layer in transformer_layers:
      # Step 2: Forward pass for positive and
negative inputs
      g_pos = layer.forward(x_pos)
      g_neg = layer.forward(x_neg)

      # Step 3: Compute Goodness scores
      G_pos = torch.sum(g_pos ** 2)
      G_neg = torch.sum(g_neg ** 2)

      # Step 4: Compute hinge margin loss
      loss=torch.maximum(torch.tensor(0.0), delta-
G_pos + G_neg)

      # Step 5: Local weight update without
backpropagation
      layer.update_weights(g_pos, g_neg, loss)
```

This training routine reinforces positive samples by increasing their Goodness score while penalizing negative samples. By avoiding global gradient propagation, FF achieves significant efficiency gains in training deep models such as Transformers. This pseudocode aligns with the mathematical framework described in Sections 3.3 and 3.4.

3.5 Reproducibility Considerations

To ensure the reproducibility of the proposed FF-based Transformer training approach, all experiments were originally conducted on an NVIDIA RTX 3080 GPU, which provided sufficient computational power for training large-scale models. The Transformer model was trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 64, ensuring a stable learning process and convergence. Standard tokenization and normalization techniques were applied for dataset preprocessing to maintain consistency across input sequences, thereby enhancing the model's ability to generalize across different data distributions.

To evaluate real-world applicability across different hardware platforms, we benchmarked the training time of the same Transformer model on three configurations: an

NVIDIA RTX 3080 (a desktop GPU with 10GB VRAM), an NVIDIA Tesla V100 (a high-performance GPU accessible via cloud services such as Google Cloud and AWS, equipped with 16GB VRAM), and an Apple M2 Pro (a laptop-grade SoC featuring a 16-core neural engine).

Table 6 Training time comparison of FF and BP across different hardware platforms

Device	Method	Training Time (s)
RTX 3080	BP	360
RTX 3080	FF	290
Tesla V100	BP	312
Tesla V100	FF	248
Apple M2 Pro	BP	510
Apple M2 Pro	FF	418

These results consistently demonstrate the computational efficiency of the FF algorithm across different hardware, with time savings ranging from 17% (Tesla V100) to 20% (Apple M2 Pro). This confirms that FF is practical for deployment on both high-performance and consumer-grade devices.

3.6 Comparison with Existing Methods

The FF algorithm offers several advantages over the existing methods, including BP, Sparse Attention, and Gradient Checkpointing.

Table 7 Evaluation of Transformer training algorithms using BLEU, Perplexity, Time, and Memory metrics

Aspect	BP	Sparse Attention	Gradient Checkpointing	FF Algorithm
Computational Complexity	High	Moderate	Moderate	Low
Memory Usage	Very High	High	Moderate	Low
Training Stability	Gradient Issues	Limited Improvement	Limited Improvement	Stable (with margin)
Additional Complexity	Moderate	High	Very High	Moderate

3.7 Expanding Application Scenarios

The efficiency and stability of the Forward-Forward (FF) algorithm make it suitable for various real-world applications across different domains, including Natural Language Processing (NLP), Computer Vision (CV), and resource-constrained environments.

In NLP, FF can enhance the efficiency of machine translation by reducing training time while maintaining competitive BLEU scores. Additionally, its ability to improve perplexity in text generation makes it a promising approach for developing more natural and coherent language models. By replacing Backpropagation (BP) with FF, NLP models can achieve faster convergence while preserving overall translation quality and fluency.

In Computer Vision (CV), FF demonstrates significant advantages in image synthesis and classification. The algorithm improves computational efficiency without sacrificing visual quality, making it an effective alternative for generative models. Furthermore, FF-based models can achieve high accuracy in classification tasks while reducing computational requirements, making them suitable for large-scale vision applications such as medical imaging, object detection, and facial recognition.

Beyond traditional AI applications, FF is particularly beneficial for resource-constrained environments, including mobile devices, embedded systems, and edge computing platforms.

In real-world applications such as on-device voice assistants (e.g., Google Assistant, Siri), the FF algorithm can reduce training memory requirements, enabling incremental learning directly on the device without offloading computation to the cloud. Similarly, autonomous drones or robots with limited GPU resources can benefit from its reduced computational footprint.

Despite these advantages, certain deployment challenges persist. For example, current deep learning frameworks are not yet optimized for FF-style dual-pass training, which may necessitate the development of customized training backends or lightweight runtime modules. Furthermore, because FF does not leverage gradient information, integrating it into existing mixed-precision or quantization pipelines could require additional engineering efforts.

Overall, these factors highlight both the opportunities and the technical hurdles associated with deploying FF in real-world systems. Continued support for toolchain development and empirical validation on embedded AI hardware will be critical to fully exploit FF's potential as a scalable, efficient, and sustainable alternative to traditional backpropagation.

4 EXPERIMENTAL RESULTS

4.1 Preliminary Experimental Results

To validate the theoretical advantages of the Forward-Forward (FF) algorithm, a preliminary experiment was conducted comparing FF with Backpropagation (BP) using a small-scale Transformer model on the IMDB Sentiment Analysis dataset. Our experimental setup involved a stratified sampling method to maintain class balance, resulting in 2,500 training samples (1,250 positive and 1,250 negative) and 500 test samples. All texts were lowercased, tokenized using the standard NLTK tokenizer, and transformed into fixed-length sequences of 256 tokens with uniform padding and truncation.

The Transformer model architecture includes 2 layers, 4 attention heads, and a hidden dimension of 128, with layer normalization and a dropout rate of 0.1 applied after both the attention and feedforward sublayers. Training was performed using the Adam optimizer (*learning rate* = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with a batch size of 64 over 10 epochs. Evaluation was based on BLEU score, perplexity, training time, and memory usage, with experiments conducted under Windows 11 (CUDA 12) and Python 3.9 on an NVIDIA RTX 3080 GPU.

In addition to the baseline experiments, we conducted further studies to examine the influence of hyperparameters, including model depth (2-layer vs. 4-layer Transformer encoders), batch sizes (32 vs. 64), and learning rates (0.001

vs. 0.0005). For each configuration, a standard BP-based Transformer was used as a control.

Table 8 Detailed experimental configurations and performance outcomes under varying model parameters and training conditions

Model Depth	Batch Size	Learning Rate	Training Method	BLEU	Perplexity	Time (s)	Memory (GB)
2 layers	64	0.001	BP	28.3	12.5	360	4.5
2 layers	64	0.001	FF	27.8	13.2	290	3.1
4 layers	64	0.001	FF	27.2	13.5	410	3.9
4 layers	32	0.0005	FF	26.9	13.8	330	3.2

The results indicate that the FF algorithm significantly improves computational efficiency by reducing training time by approximately 20% compared to BP. Moreover, memory consumption decreased by 30%, making FF particularly beneficial for deployment in resource-constrained environments.

To ensure statistical validity, each experimental configuration was run five times with different random seeds. The results were averaged, and standard deviations (*Std*) were calculated. Furthermore, a two-tailed independent samples t-test was performed to compare the BLEU and Perplexity scores of the FF and BP-based models. Tab. 9 presents the updated results including standard deviations and p-values.

Table 9 Statistical analysis of BLEU and Perplexity scores for BP and FF including standard deviations and p-values

Metric	*BP* (*Mean* ± *Std*)	*FF* (*Mean* ± *Std*)	*p*-value
BLEU Score	28.3 ± 0.6	27.8 ± 0.7	0.072
Perplexity	12.5 ± 0.4	13.2 ± 0.5	0.049
Training Time (s)	360 ± 10	290 ± 12	–
Memory (GB)	4.5 ± 0.2	3.1 ± 0.1	–

These updated results indicate that while the BLEU score difference was not statistically significant at the 5% level ($p = 0.072$), the increase in Perplexity for FF was marginally significant ($p = 0.049$). Nonetheless, performance metrics remain within acceptable margins, supporting the claim that FF offers substantial computational benefits without significantly compromising model accuracy.

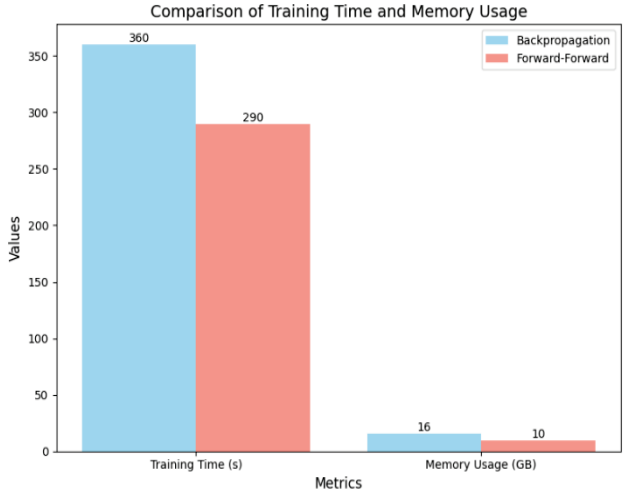


Figure 2 Comparison of training time and memory usage between Backpropagation (BP) and Forward-Forward (FF) algorithms

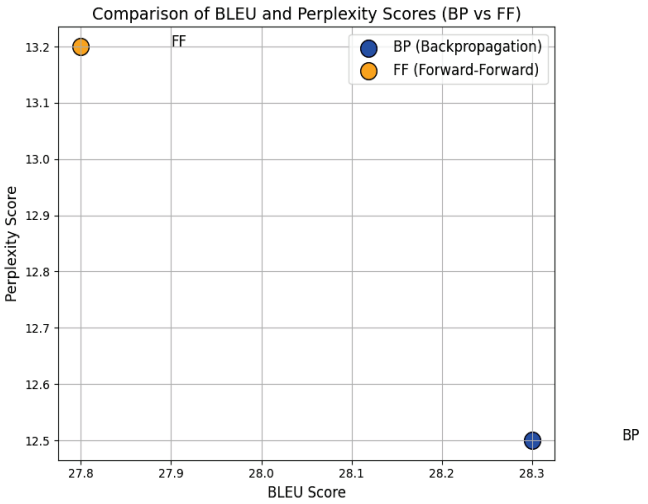


Figure 3 Comparison of BLEU and Perplexity scores between BP and FF algorithms

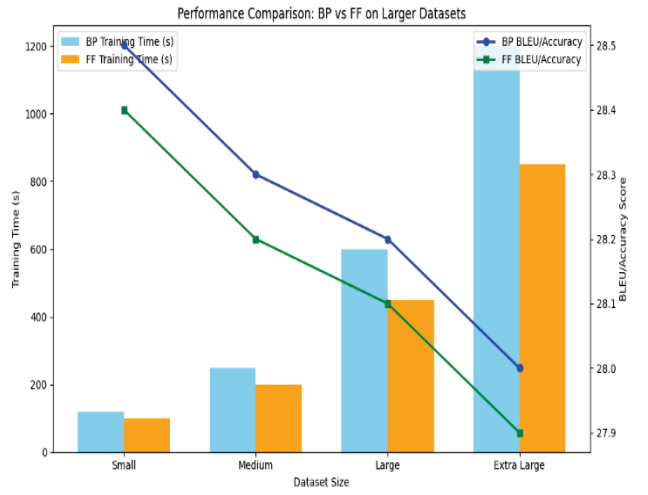


Figure 4 Performance comparison of BP and FF on larger datasets showing scalability of FF

Figs. 2, 3, and 4 provide visual comparisons of training time, memory usage, BLEU scores, and perplexity between BP and FF across different configurations and hardware types, further emphasizing the robustness and efficiency of the FF algorithm.

The Fig. 3 illustrates that while BP slightly outperforms FF in both BLEU and Perplexity, the performance gap is minimal. FF maintains competitive quality in language generation despite its gradient-free nature.

By incorporating these multi-hardware evaluations alongside statistical analyses, we demonstrate both the reproducibility and practical advantages of the FF-based Transformer training approach across various real-world deployment scenarios.

4.2 Future Research Directions

While the FF algorithm has demonstrated notable improvements in computational efficiency and training stability, further research is required to validate its scalability and applicability in large-scale AI models. Future studies should evaluate FF on larger datasets, such as WMT for

machine translation, GLUE for NLP classification, and ImageNet for vision tasks. Additionally, testing FF on ultra-large models like GPT-4 and PaLM would provide valuable insights into its feasibility for high-parameter architectures.

Optimizing the loss function is critical to enhance FF's convergence and generalization. Techniques such as adaptive margin loss (which dynamically adjusts the margin δ and regularization strategies (e.g., L1/L2 norms) can be explored to improve stability and prevent overfitting.

Beyond Transformers, FF's applicability to other AI architectures presents an exciting research avenue. Hybrid approaches-for example, combining FF with Sparse Attention for memory efficiency or integrating it with Gradient Checkpointing to selectively apply backpropagation-could further enhance performance. Future evaluations should also include comparative studies of FF against hybrid paradigms that blend gradient-free and gradient-based techniques, such as Sparse Attention plus Checkpointing, to clarify its unique contributions and trade-offs.

Investigating FF's potential in reinforcement learning models may further improve training stability. Scaling FF for distributed training is another crucial research area; future work should explore multi-GPU/TPU parallelism and its integration into federated learning, enabling decentralized AI systems to leverage FF efficiently. Finally, expanding FF's applications in specialized architectures—such as Generative Adversarial Networks (GANs), Spiking Neural Networks (SNNs), and biomedical AI models (e.g., medical image analysis and drug discovery)—could unlock new frontiers in AI model training.

Future studies will also include component-wise ablation analysis to evaluate the sensitivity of FF when applied selectively to different submodules of the Transformer. Specifically, we propose evaluating the following three variants.

1) FF-Attn: Apply FF only to self-attention layers, while training the feedforward layers using backpropagation.
2) FF-FFN: Apply FF solely to feedforward networks (FFN), with the self-attention mechanisms trained via backpropagation.
3) FF-All: Apply FF to all layers, as implemented in the current study.

This ablation analysis will help determine whether FF impacts the self-attention and feedforward components differently and identify optimal hybrid training strategies. Such insights could lead to a fine-grained integration of FF and BP, optimizing both performance and resource efficiency. By addressing these research directions, FF has the potential to evolve into a fundamental deep-learning training paradigm, eventually replacing backpropagation in future AI methodologies.

5 DISCUSSION AND FUTURE DIRECTIONS

The Forward-Forward (FF) algorithm introduces an alternative training approach for Transformer models, aiming to address the limitations of traditional Backpropagation (BP) while improving computational efficiency, training stability, and energy consumption. This section discusses the trade-offs, challenges, and future research directions necessary for the broader adoption of FF-based models.

5.1 Trade-offs in Computational Efficiency and Performance

The primary advantage of FF lies in its ability to eliminate backpropagation, reducing computational overhead and memory consumption. Experimental results demonstrate a 20% decrease in training time and 30% lower memory usage than BP. However, this efficiency gain comes at a slight cost—an increase in Perplexity. This trade-off suggests that, while FF improves training efficiency, additional refinements are needed to maintain or surpass BP-based models' performance. Future studies must investigate techniques to optimize FF's performance while preserving its computational benefits.

5.2 Limitations and Challenges

Despite its potential, FF faces several challenges that must be addressed before it can be widely adopted. These challenges include its scalability to large datasets, the optimization of its loss function, and its integration with self-attention mechanisms in Transformer models.

First, scalability remains a major concern, as current evaluations have primarily focused on smaller datasets. FF must be tested on large-scale datasets such as GLUE and WMT to more rigorously assess its real-world applicability, where model performance can be systematically analyzed.

Second, optimizing the loss function is essential for improving accuracy across NLP and computer vision tasks. The current goodness function requires further refinement to effectively minimize the difference in goodness values while preserving training stability. Future work should focus on developing adaptive loss functions that enhance FF's robustness.

Finally, integrating FF with the self-attention mechanism in Transformer models presents a technical challenge. Self-attention involves complex token interactions, and FF must be carefully adapted to function seamlessly within this framework. Ensuring that FF-based models achieve comparable performance without disrupting the fundamental operations of Transformers will be critical for their broader adoption.

Addressing these challenges is crucial to enhancing FF's reliability, scalability, and applicability across diverse AI domains. Moreover, in real-world deployments where privacy is critical (e.g., healthcare or federated learning), FF's gradient-free nature may reduce the risk of gradient-based leakage attacks, offering an added layer of model security.

5.3 Future Research Directions

Future research should focus on the following key areas to further explore the applicability and effectiveness of the Forward-Forward (FF) algorithm.

1) Large-Scale Model Evaluation

FF must be tested on ultra-large models such as GPT-4 and PaLM to assess its feasibility in high-parameter architectures.

Comparative evaluations with BP-based models will be necessary to analyze scalability and efficiency in large-scale neural networks.

2) Resource-Constrained AI Systems

Given FF's reduced memory and computational demands, it could be highly advantageous in IoT devices, edge computing, and real-time AI applications.

Future studies should examine FF's energy efficiency and latency in resource-constrained environments such as autonomous vehicles, embedded AI, and federated learning systems.

3) Hybrid Training Approaches

It can be integrated with Sparse Attention for memory efficiency or Gradient Checkpointing for selective backpropagation to enhance FF's effectiveness.

Additionally, FF's role in reinforcement learning could be explored to investigate its impact on training stability and sample efficiency.

4) Parallelization and Distributed Learning

Scaling FF for multi-GPU/TPU distributed training is crucial for its adoption in large-scale AI applications.

Further research should investigate FF's compatibility with federated learning and decentralized AI models, ensuring its feasibility in privacy-preserving AI training.

5) Expanding Applications beyond Transformers

Beyond Transformers, FF has potential applications in Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs).

Applying FF to biomedical AI models, such as medical image analysis and drug discovery, could unlock new possibilities in AI-driven healthcare.

By addressing these research directions, FF could evolve into a scalable, efficient, and broadly applicable training paradigm, potentially replacing backpropagation in future AI methodologies.

5.4 Evaluation Metrics for FF Performance Validation

Key performance metrics must be clearly defined to objectively assess the effectiveness of the Forward-Forward (FF) algorithm.

These metrics enable a structured comparison of FF-based training with Backpropagation (BP)-based methods in terms of computational efficiency and model performance.

Tab. 10 presents the evaluation criteria used for experimental validation.

Table 10 Evaluation metrics for experimental validation

Metric	Purpose	Application
BLEU	Measures translation quality	Machine translation (e.g., WMT)
Perplexity	Evaluates predictive accuracy of language models	Text generation (e.g., GLUE, IMDB)
Latency	Analyzes training and inference speed	All tasks
Accuracy	Benchmarks classification performance	Image classification (e.g., CIFAR-10, ImageNet)

These metrics provide a quantitative basis for evaluating FF's performance across different machine learning tasks.

By analyzing these indicators, researchers can determine whether FF maintains model accuracy while improving computational efficiency, ensuring its viability as an alternative to BP.

6 CONCLUSIONS

This study introduced a Forward-Forward (FF)-based training framework for Transformer models as a resource-efficient, gradient-free alternative to Backpropagation (BP). We addressed the growing need for scalable and memory-efficient training methods, particularly for edge and privacy-sensitive applications.

Our contributions include a theoretical formulation of the FF algorithm within Transformer layers, the development of a structured training pipeline complete with pseudocode and a margin-based loss optimization strategy, empirical evaluations across different hardware platforms and model depths, statistical validation of our results along with plans for component-wise ablation studies, and a comparative analysis with other gradient-free methods, such as Genetic Algorithms (GA) and EvoGrad.

While FF does not yet outperform BP in every metric, our results demonstrate that it achieves comparable accuracy with significantly lower computational costs. This validates its feasibility for real-world systems and highlights its potential to redefine deep learning methodologies by reducing training time, memory usage, and energy consumption.

Future research directions include exploring hybrid FF-BP integrations to further enhance performance, establishing formal convergence proofs and theoretical bounds, and implementing FF in privacy-preserving and federated learning scenarios. These advancements mark FF as a promising candidate for next-generation training frameworks in deep learning, paving the way for more accessible and resource-efficient AI solutions.

7 REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://doi.org/10.5555/3295222.3295349

[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1810.04805

[3] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI*. Retrieved from https://openai.com/research/language-understanding

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale [Preprint]. *arXiv*.

https://doi.org/10.48550/arXiv.2010.11929
[5] Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1409.0473
[6] Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1412.6980
[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90
[8] Chen, T., Rubanova, Y., Bettencourt, J., & Duvenaud, D. (2016). Training deep nets with sublinear memory cost [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1604.06174
[9] Chen, Q., Sun, C., Lu, Z., & Gao, C. (2022). Enabling energy-efficient inference for self-attention mechanisms in neural networks. In *IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (pp. 25–28). IEEE. https://doi.org/10.1109/AICAS54282.2022.9869924
[10] Hinton, G. (2022). The forward-forward algorithm: Some preliminary investigations [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2212.13345
[11] Noh, H., Kim, H., & Yoo, S. (2023). Research on forward-forward algorithm. In *Annual Conference of KIPS Proceedings* (pp. 469–470). Korean Information Processing Society.
[12] Kim, H. (2024). Analyzing the characteristics of gradient descent and non-gradient descent-based algorithms in neural network learning. In *Proceedings of the IEEE International Conference on Electrical, Computer and Energy Technologies (ICECET)*. IEEE. https://doi.org/10.1109/ICECET61485.2024.10698114
[13] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://doi.org/10.5555/3455716.3455856
[14] Real, E., Aggarwal, A., Huang, Y., & Le, Q. V. (2019). Regularized evolution for image classifier architecture search. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 4780–4789. https://doi.org/10.1609/aaai.v33i01.33014780
[15] Zoph, B., & Le, Q. V. (2017). Neural architecture search with reinforcement learning [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1611.01578
[16] Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1904.10509
[17] Ham, T. J., Jung, S., Kim, S., Oh, Y. H., Park, Y., Song, Y., Park, J.-H., Lee, S., Park, K., & Lee, J. W. (2020). A³: Accelerating attention mechanisms in neural networks with approximation. In *IEEE International Symposium on High Performance Computer Architecture (HPCA)* (pp. 328–341). IEEE. https://doi.org/10.1109/HPCA47549.2020.00035
[18] Xie, Y. (2023). Efficient disentangled attention network for semantic segmentation. In *IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE)* (pp. 213–217). IEEE. https://doi.org/10.1109/ICSECE58870.2023.10263368
[19] Pham, H., Guan, M., Zoph, B., Le, Q. V., & Dean, J. (2018). Efficient neural architecture search via parameter sharing [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1802.03268
[20] Salimans, T., & Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1602.07868
[21] Yao, M., Zhao, G., Zhang, H., Hu, Y., Deng, L., Tian, Y., Xu, B., & Li, G. (2022). Attention spiking neural networks [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2209.13929
[22] Wang, Q., Wu, B., Zhu, P. F., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11531–11539). IEEE. https://doi.org/10.1109/CVPR42600.2020.01155
[23] Gandhi, S., Gala, R., Kornberg, J., & Sridhar, A. (2023). Extending the forward-forward algorithm [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2307.04205
[24] Aktemur, E., Zorlutuna, E., Bilgili, K., Bok, T. E., Yanikoglu, B., & Mutluergil, S. O. (2024). Going forward-forward in distributed deep learning [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2404.08573
[25] Reyes-Angulo, A., & Paheding, S. (2024). Forward-forward algorithm for hyperspectral image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE. https://doi.org/10.1109/CVPRW63382.2024.00321
[26] Gao, Y., & Zhang, Y. (2023). The predictive forward-forward algorithm [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2301.01452

Authors' contacts:

Hyun Jung Kim, Assistant Professor
Sang-Huh College and the Graduate School of Information & Communication,
Department of Convergence Information Technology (Artificial Intelligence Major),
Konkuk University, Republic of Korea
nygirl@konkuk.ac.kr

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

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240112104444

Preliminary communication
Received: 2024-01-12, Accepted: 2024-05-21

Efficient Deep Learning Job Allocation in Cloud Systems by Predicting Resource Consumptions including GPU and CPU

Abuda Chad Ferrino, Tae Young Choe*

Abstract: One objective of GPU scheduling in cloud systems is to minimize the completion times of given deep learning models. This is important for deep learning in cloud environments because deep learning workloads require a lot of time to finish, and misallocation of these workloads can create a huge increase in job completion time. Difficulties of GPU scheduling come from a diverse type of parameters including model architectures and GPU types. Some of these model architectures are CPU-intensive rather than GPU-intensive which creates a different hardware requirement when training different models. The previous GPU scheduling research had used a small set of parameters, which did not include CPU parameters, which made it difficult to reduce the job completion time (JCT). This paper introduces an improved GPU scheduling approach that reduces job completion time by predicting execution time and various resource consumption parameters including GPU Utilization%, GPU Memory Utilization%, GPU Memory, and CPU Utilization%. The experimental results show that the proposed model improves JCT by up to 40.9% on GPU Allocation based on Computing Efficiency compared to Driple.

Keywords: cloud computing; convolutional neural network; deep learning; GPU job scheduling; performance estimation

1 INTRODUCTION

Deep learning is used in various fields such as Chat-GPT [1] that generates responses to user inquiries. It is also used for structure design to predict equipment failure [2], automatic vehicle incident model [3, 26], and video anomaly detection [4]. With the success of deep learning used in various fields of study, the complexity of these deep learning models also increases as a higher number of parameters and layers yield better accuracy.

As proposed in the Convolutional Neural Network (CNN) used in [4], the stacked 3×3 convolution kernels from VGG16 have been increased up to 7×7 convolution kernels. The increase in parameters and layers of a neural network model also increases the training time needed to train a given model to obtain a desirable accuracy. The trend in the increased complexity of deep learning models not only increases training times but also increases the hardware requirements of training such models. Scheduling in deep learning is to allocate computational resources and jobs that will optimize and accelerate model training. One of the difficulties of scheduling jobs for deep learning models is that there is a diverse amount of model architectures, resulting in different hardware requirements to train such models. Some of these models require high GPU performance and some other models require higher CPU performance.

As a result, scheduling mechanisms must be able to adapt to these different model architectures and allocate the jobs accordingly based on the specific needs of a given model. Previous work [5] claimed that operations used in a model have different impacts on execution time, but in this work, they only calculated the execution time of each operation and scaled it into an entire iteration. Meanwhile [6] focused on the 'features' derived from the network being trained, the data being trained, and the hardware that the model is being trained on, for evaluating resource consumption to estimate execution time. However, this work only utilized a VGG16 model to test their model. On the other hand, others focused only on the Central Processing Unit (CPU) [7] to estimate the execution time of Convolutional Neural Networks (CNN) on a single CNN model with three different architectural sizes. Lastly, [8] introduced Graph Neural Network (GNN) and utilized Graphics Processing Unit (GPU) and network parameters for resource consumption prediction to estimate the resource consumption based on three prediction targets, for each resource consumption parameter totaling up to 12 total prediction targets, while it considers the entire graph as an input, which involves all the operations found within the graph, they only focused on GPU parameters such as GPU Utilization% and GPU Memory Utilization%.

In this paper, resource consumption and execution time prediction is explored by evaluating the different Resource Consumption parameters that are obtained from the GPU, and CPU, during training, to create a model that will be utilized to predict resource consumption and execution time for predicting the GPU allocation of a given Deep Learning task, and lastly, applying a scheduling algorithm, such as First-In First-Out (FIFO) and Shortest Job First (SJF) to test the effectiveness of this approach in real-world applications. This paper also introduces a new approach for predicting GPU Job allocation by using computing efficiency, which is derived from the resource consumption prediction parameters. To achieve these objectives this study aimed to:

- Create a dataset with four resource consumption parameters (GPU Utilization, GPU Memory Utilization, CPU Utilization, and GPU Memory) and four prediction targets for each parameter (Average Burst Time, Average Idle Time, Average Peak Consumption, Execution Time) which amounts to 16 prediction metrics.
- Develop a model that will not only predict resource consumption but also the execution time of a given input which can be used for scheduling jobs.
- Evaluate the performance of prediction targets as parameters for GPU allocation and its effects on job scheduling using FIFO and SJF scheduling policies.

2 BACKGROUND AND RELATED WORKS

2.1 Job Profiling

Previous works such as [6] dissected a given model architecture and utilized what they called Layer Features, which includes batch size, optimizer, and activation function. In addition, they also included layer-specific features, such as Convolutional features, pooling features, and so on, which are mostly found on a CNN model, in this example they only used VGG16. They also included some GPU hardware specification as part of their training features to create a profiler that will predict the execution time of each layer, to predict the full model execution time. However, this is very limited to models that are like VGG16. This approach requires dissecting a target model and figure out all the Layer Features, Convolutional features, and GPU hardware specifications, which will be difficult to implement on different neural network models due to the varying model architectures and operations used in each model.

Another work involves the use of a CPU [7] instead of a GPU for profiling CNNs of small, medium, and large CNN architectures, with varying amount of convolutional layers. They mainly utilized forward propagation and backward propagation per image for their measurements to predict the execution time, which is also found in [9]. Contrary to [7], in [9], they used a multi-GPU approach and were limited by small-scale CNN model architectures. While these works predicted execution time, it only covered very specific neural network models

Lastly, in [8] they profiled an input task, as shown in Fig. 1, and utilized TensorFlow to convert it to a graph, which produces an adjacency matrix and feature matrix. This job profiler produces three outputs that characterize resource consumption: active time, idle time, and average peak consumption. With these, they were able to estimate the resource consumption of various models with different hyperparameter settings.

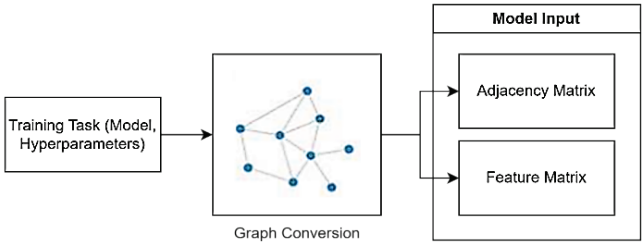


Figure 1 Input task conversion to graph

2.2 Resource Consumption Parameters

The system used in this paper uses a job profiler based on Driple [10] to create a model that will predict resource consumption by the usage of resource consumption parameters on GPU, and CPU along with execution time, which will then be used for scheduling. While other works only used GPU, CPU, or multi-GPU hardware settings. This paper proposes to utilize both GPU and CPU parameters, namely GPU Utilization%, GPU Memory, GPU Memory Utilization%, and CPU Utilization%. The GPU parameters are obtained via the usage of the NVIDIA System Management Interface (nvidia-smi) library [11] and CPU parameters were obtained using psutil library which calculates the CPU Utilization% for all cores. These parameters are profiled in an interval of 1/6 per second. Previous work [8], utilized only GPU Utilization% and GPU Memory Utilization for GPU parameters.

However, in this approach the prediction output of these input parameters will be used for GPU allocation of jobs, hence the addition of GPU Memory and CPU Utilization% on the resource consumption parameters are proposed. As these parameters can also affect the performance of training a model as larger architectures would prefer bigger amounts of GPU Memory for optimal performance, in some cases, it will not be able to train the model at all [12]. Moreover, based on initial experiments using the same hyperparameter settings of a VGG model on 3 machines, as seen in Table 1, the RTX2070 machine simulated as a low-performance system shows that the resource usage between the other two machines is significant. This was identified as a CPU bottleneck that can occur on lower-performance systems, which makes the training time longer and makes it difficult to perform job profiling during training as shown on the vgg19 model.

Table 1 Profiling of average GPU utilization (%), GPU memory (MB) for VGG models

Dataset	GTX1080	RTX2070	Titan X	Model
ImageNet	98.84%, 7840.87	26.59%, 5738.25	98.27%, 11926.75	vgg11
	99.18%, 7856.85	16.16%, 5622.33	98.74%, 11929.67	vgg16
	89.95%, 7712.80	-	98.88%, 11930.06	vgg19

2.3 Evaluation of Resource Consumption

In this work, similar output parameters were utilized. However, in the context of a scheduler, the amount of resources consumed by a job is of importance. Hence, the output parameters are proposed to be defined as follows: burst time is the upper half of the median for each resource type, while idle time is defined by the lower half of the median for each resource type.

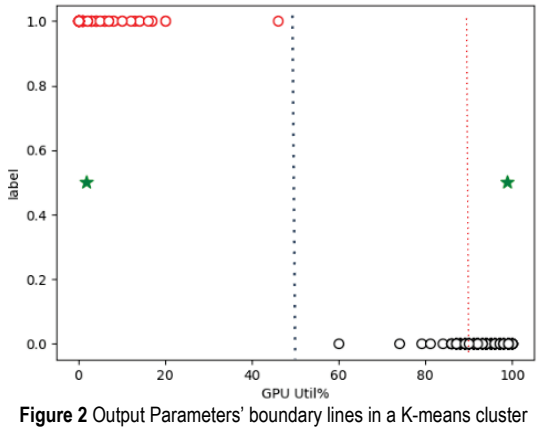


Figure 2 Output Parameters' boundary lines in a K-means cluster

For example, for a workload having a GPU utilization between 0-60%, the burst time is 31-60%, while idle time is 0-30% utilization. For the peak consumption, it considers the

maximum capacity of each resource type, which is > 90% of its maximum value. In terms of GPU utilization, it will be data that's > 90% utilization and for memory, it will be > 90% of the total available memory in the GPU. These are represented by the black dotted line representing the median and red dotted lines representing the boundary for peak consumption data points on the right side, as seen on an example benchmark in Fig. 2.

By measuring GPU Memory Utilization%, GPU Utilization%, GPU Memory, and CPU Utilization% for each workload, they are converted into these three parameters using these criteria. These are then grouped together by their respective hardware setup to create a dataset, which will be used to train a model to produce the predicted resource consumption on a given GPU or hardware setup. These predicted values are then used for GPU allocation of succeeding deep learning workloads.

For scheduling, the execution time of a workload also helps with decision-making for job allocation. Hence, in addition to the mentioned parameters, the execution time will also be included as part of the input and output parameters. The input execution time will be measured during the job profiling phase, and it will be normalized based on the maximum and minimum execution time values found on the dataset. The output execution time will be part of the prediction targets along with the three mentioned parameters, burst time, idle time, and average peak consumption.

Therefore, this paper profiles five different resource consumption parameters, GPU Utilization%, GPU Memory Utilization%, GPU Memory, CPU Utilization%, and Wall-Clock Execution Time. Among the hardware resource consumption parameters, each will have four prediction output targets (burst time, idle time, average peak consumption, and execution time), having a total of 16 prediction output parameters, as shown in Tab. 2.

Table 2 Resource consumption parameters for profiling

Proposed Parameters	Prediction Targets
GPU Utilization%	Burst Time, Idle Time, Average Peak Consumption, Execution Time
GPU Memory Utilization%	Burst Time, Idle Time, Average Peak Consumption, Execution Time
GPU Memory	Burst Time, Idle Time, Average Peak Consumption, Execution Time
CPU Utilization%	Burst Time, Idle Time, Average Peak Consumption, Execution Time
Wall Clock Execution Time	*-Part of the prediction targets above-*

2.4 Job Scheduling

Previous works on job scheduling implemented a generalized wide range of scheduling policies [13], such as First-In-First-Out (FIFO) and Shortest Job First (SJF) policies, to measure the makespan, which is the amount of time it takes to complete all the jobs scheduled, and job completion time for deep learning training workloads. While Gavel focused on job allocation based on policies, and to improve makespan and JCT, they did not use a prediction model and they did not consider hardware resource consumption parameters.

Another scheduler focused on a directed acyclic graph (DAG) [14], which involves adding idle time in between jobs to decrease the average job completion time of the workloads. While they used graphs as inputs, it did not involve deep learning workloads. The proposed model has additional input parameters and prediction targets compared to the previous work [8]. Specifically, the proposed model intends to profile jobs based on GPU, and CPU parameters, such as GPU Utilization%, GPU Memory Utilization%, GPU Memory, and CPU Utilization%, while also including Execution Time.

This is different compared to the previous model which only uses GPU Resources such as, GPU Utilization%, and GPU Memory Utilization%, and network parameters. However, in this research, network parameters were not considered since deep learning workloads are allocated on a single machine after GPU Allocation and network parameters does not impact the performance of training. This research introduces computing efficiency that is derived from the resource consumption prediction to schedule jobs. The performance of computing efficiency is evaluated by measuring the makespan and average job completion time using FIFO and SJF policies. A similar approach is also applied to job allocation based on predicted execution time.

2.5 Challenges of GPU Scheduling

GPU Scheduling is the process of allocating and managing GPU resources to various computational tasks or processes. Ref. [17] addressed various difficulties in GPU scheduling especially in larger scale systems where the workload has various computational requirements, which was divided into two types which were identified as High-GPU Tasks and Low-GPU Tasks.

High-GPU Tasks involves Natural Language Processing (NLP) [18] with advanced language models, Image classification with a large output such as a modified ResNet model [19]. High GPU tasks could potentially consume the entire GPU resource usage while having low CPU resource utilization.

For Low-GPU Tasks it was discovered that these tasks spend a considerable amount of time on CPUs for data processing. Some of these tasks involve click-through rate (CTR) [26] prediction models, Graph Neural Network (GNN) [20] Training models and Reinforcement learning. Among these tasks the GPU is underutilized on Low-GPU tasks mainly because of the CPU bottleneck where the CPU limits the amount of GPU resource that can be used by the model, leaving the GPU usage underutilized [28].

As such, it was mentioned that considering a multi-resource scheduler would be preferable to alleviate the scheduling difficulties. Hence, the proposed model intends to profile jobs based on GPU, and CPU parameters, such as GPU Utilization%, GPU Memory Utilization%, GPU Memory, and CPU Utilization%, while also including Execution Time. This is different compared to the previous model [8] which only uses GPU Resources such as, GPU Utilization%, and GPU Memory Utilization%, and network parameters. However, in this research, network parameters were not considered since deep learning workloads are allocated on a single machine after GPU Allocation and network parameters does not impact the performance of training. This research introduces computing efficiency that

is derived from the resource consumption prediction to schedule jobs. The performance of computing efficiency is evaluated by measuring the makespan and average job completion time using FIFO and SJF policies. A similar method is also applied to job allocation based on predicted execution time. This approach introduces a way of scheduling deep learning workloads by only specifying the model architecture.

3 METHODOLOGY

3.1 Job Profiler

The system workflow consisted of two major parts, namely the job profiler and the job scheduler. The job profiler is responsible for the creation of the proposed prediction model that will be utilized to help the schedule decide which GPU to allocate for a given job. First, the job profiler, as seen in Fig. 3, takes in a training workload, which is represented by a given neural network model, the dataset to be used, and its hyperparameters such as batch size, epochs, and optimizer.

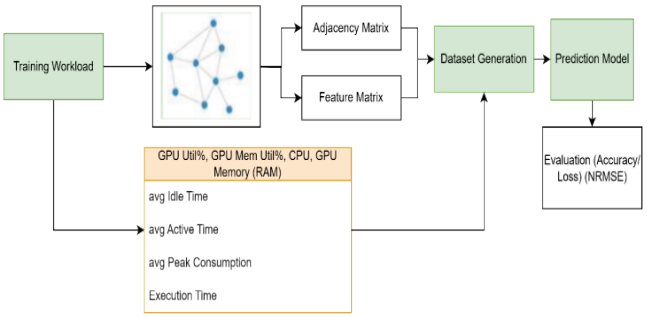


Figure 3 Job profiler workflow

The training workload consists of all the combinations of these settings, as shown in Tab. 3, separating generative neural networks with convolutional neural networks, which totals up to 432 training workloads for a given GPU.

Table 3 Hyperparameter settings

Datasets	Optimizer	Batch Size	Epoch	Models
ImageNet, Cifar10	SGD, Adam	8, 16, 32, 64	10, 20	vgg11, vgg16, vgg19, lenet, googlenet, densenet40-k12, densenet100-k12, inception3,inception4, resnet20, resnet50, resnet101
MNIST	SGD, Adam	8, 16, 32, 64	10, 20	dcgan, conditional_gan, cvae

When starting a training workload, the model being trained is converted into a graph, producing an adjacency matrix that describes the connectivity between nodes and edges, with a value of 1 if there is a connection and 0 if there is none, while the feature matrix consists of tensor size and node type. The node type is the operations found in each model, such as Conv2D, that is converted into a float number by implementing frequency encoding, which estimates the number of occurrences of a given operation found in a dataset. This implementation is similar to the one used in Driple [10].

At the same time, the resource consumption of the system was profiled based on four parameters: GPU Memory Utilization%, GPU Utilization%, GPU Memory, and CPU Utilization% at a rate of 1/6 per second, including the total Execution Time of a given workload. These parameters are segregated into Burst data points and Idle data points using Kmeans, shown previously in Fig. 2, to label them into two clusters according to their resource consumption. These labeled data points are then used to obtain the data for average burst time, average idle time, and average peak consumption rate. Once the data for all parameters are converted, it is then combined with the adjacency matrix, feature matrix, and execution time to create a single entry into the dataset which also includes, the GPU type, the dataset used for training, and the hyperparameters used for the training. Given the settings in Tab. 3, a dataset will contain 384 training workloads for each GPU that will be used for training the proposed model. In this study, three GPUs were used for job profiling which means that there were three datasets that were created.

Once the necessary datasets have been created, with the help of the Driple model [10], these datasets were trained and evaluated based on the loss function Mean Square Error (*MSE*) where N is the number of data, y is the ground truth and $\hat{y}$ is the predicted value.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2. \quad (1)$$

3.2 Job Scheduler

Once the prediction model is trained, the best model will be used by the job scheduler, as shown in Fig. 4. The job scheduler takes in a deep learning training workload as an input, containing the model architecture, and hyperparameter settings. Similarly, since the prediction model takes adjacency matrices and feature matrices as input, the input workload must be converted to a graph to obtain these matrices.

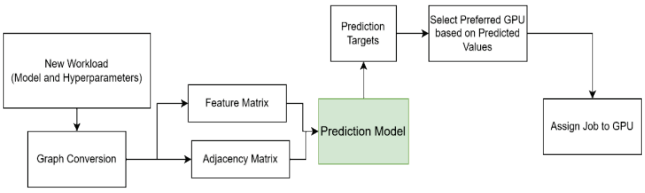


Figure 4 Scheduler workflow

After the conversion, the matrices will then be fed as an input to the prediction model, and the prediction model will produce 16 prediction targets, one for each resource consumption type, namely Average Burst Time, Average Idle Time, Average Peak Consumption, and Execution Time. Among these prediction targets, the scheduler will decide which GPU should a given job be allocated to by measuring the execution time for each dataset and measuring the

computing efficiency ($\hat{y}_{\text{eff}}$) by obtaining a summation of the quotient of the predicted average peak consumption rate ($\hat{y}_{\text{peak}}$) divided by the predicted average burst time ($\hat{y}_{\text{burst}}$) for all resource consumption parameters.

$$\hat{y}_{\text{eff}} = \sum_{i=1}^{N} \frac{\hat{y}_{\text{peak}}}{\hat{y}_{\text{burst}}}. \quad (2)$$

For the execution time, the lower the value, the better the performance of a given workload, hence, GPUs with the lowest predicted execution time will be preferred when scheduling this task. On the other hand, for computing efficiency, the higher the average peak consumption-average burst time ratio translates to the computing resources being fully utilized by the workload while having minimal idle time. Therefore, when considering computing efficiency, the GPU with the highest ratio will be preferred when scheduling a given task. This will be done for all available jobs in the scheduler until all the jobs are assigned.

The performance of the scheduler will be measured by calculating the average job completion time, and the makespan of the schedule. The scheduler workload that will be used for the experiments will be using the combinations of the hyperparameter settings in Tab. 4. Since the models trivial, alexnet, resnet32, and gan are not a part of the dataset in which the model was trained on, these models will be considered as unknown inputs. The scheduler workload will involve up to 128 jobs for GPU allocation.

Table 4 Scheduler workload

Datasets	Optimizer	Batch Size	Epoch	Models
ImageNet, Cifar10	SGD, Adam	8, 16, 32, 64	10, 20	trivial, alexnet, resnet32
MNIST	SGD, Adam	8, 16, 32, 64	10, 20	gan

The algorithm for GPU allocation based on execution time as shown in Algorithm 1, jobs are loaded into the queue, and for each job in the queue, it is converted to a graph. Once the conversion is finished it is then loaded into the prediction model along with the GPU dataset to obtain the execution time prediction for a given dataset.

```
queue = get.Jobs()

for job in queue:
        inputs = convert_to_graph(job)

        for gpu in datasets:
                execution_time_pred += prediction_model(inputs,gpu)
job.GPUAllocation(argmin(execution_time_pred))

if SJF:
        for each GPU:
                queue = sort.by(execution_time_pred,ascending)
```

Algorithm 1 Pseudocode for scheduling based on execution time predictions.

Since there are four resource consumption types, GPU Utilization%, GPU Memory Utilization%, GPU Memory, and CPU Utilization%, the prediction model produces an execution time for each resource consumption type. Therefore, the predicted execution time among the four resource consumption types will be added together. From the predictions obtained from all the datasets, the GPU with the lowest predicted execution time will be the designated allocation for the given job. If the scheduling policy is SJF, then the jobs are sorted after allocation on the GPU based on the predicted execution time.

```
queue = get.Jobs()

for job in queue:
        inputs = convert_to_graph(jobs)

        for gpu in datasets:
                for each resource_consumption_type:
                        resource_consumption_pred = prediction_model(inputs,gpu)
                        comp_eff += resource_consumption_pred[peak] / resource_consumption_pred[burst]
job.GPUAllocation(argmax(comp_eff))

if SJF:
        for each GPU:
                queue = sort.by(comp_eff,ascending)
```

Algorithm 2 Pseudocode for scheduling based on computing efficiency.

On the other hand, GPU allocation based on computing efficiency as shown in Algorithm 2, similarly, jobs are loaded into the queue, then the jobs are converted into graphs, and loaded into the prediction model with the GPU dataset. However, this time, there are four resource consumption types, GPU Utilization%, GPU Memory Utilization%, GPU Memory, and CPU Utilization%, which produces four prediction targets each, the average burst time, average idle time, average peak consumption, and execution time.

For the computation of computing efficiency, using Eq. (2), the ratio of the predicted peak consumption and predicted burst time for each resource type will be combined to a single dataset. After going through all datasets, the highest computing efficiency score will be the designated GPU allocation of a given job. If the scheduling policy is SJF, the jobs will be sorted after allocation based on their computing efficiency scores.

4 EXPERIMENTS AND RESULTS

4.1 Experiment Setup

The experiment was applied on three different servers, which are GTX 1080 8GB, RTX 2070 8GB, and Titan X 12GB, as shown in Tab. 5. The operating system used was Ubuntu 18.04, while using Tensorflow 2.5, and CUDA 11.2 to train a workload on three datasets, which are ImageNet, Cifar10, and MNIST dataset.

Table 5 Machine specifications

GPU Type	CPU Model	GPU Memory Capacity	Memory (RAM)
GTX1080	Intel Core i7-4790K	8 GB	32GB
RTX2070	AMD Ryzen 5 3600	8 GB	32GB
TitanX	Intel Core i7-6900k	12 GB	64GB

Training setting is specified by using a combination of different optimizers, batch sizes, epochs, and models as shown in Tab. 3, with the help of TensorFlow benchmark library [15], Deep Convolutional Generative Adversarial Network (DCGAN) model [23][27], Conditional Generative Adversarial Network (Conditional GAN) [24], Convolutional Variational AutoEncoder (VCAE) [25] were

also included. For training the prediction model, recommended setup by [8] was used and executed on PyTorch 1.8.1

4.2 Hyperparameter Analysis

To better understand the effects of each hyperparameter to the execution time, along with the hardware specifications, Batch Size, Epochs, Optimizers, and GPU Execution times were analyzed as shown in Figs. 5, 6, and 7. It can be inferred from the epoch comparison that for the settings used in this experiment, when the number of epochs is doubled, the execution time is also doubled. On the other hand, as the batch size increases, the execution time decreases, and the improvement is approximately 20% from batch size 8 to batch size 16. However, as the batch size continues to increase, the difference in execution time becomes minimal as seen on batch size 32 and batch size 64.

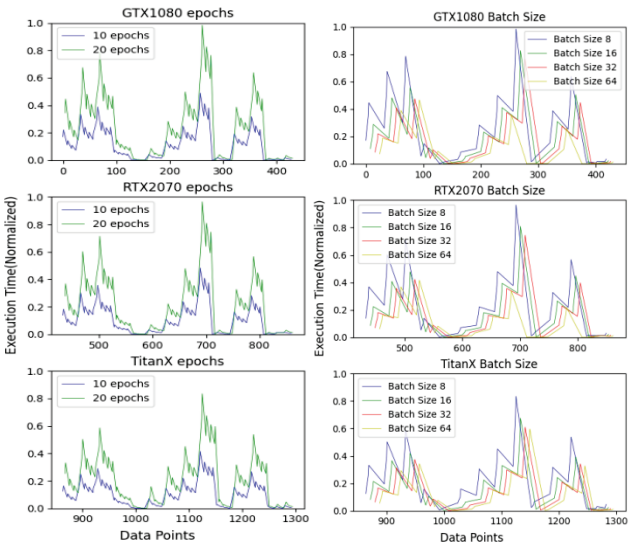


Figure 5 Execution times based on epochs and batch sizes among the 3 GPU machines

For the optimizers, the difference in execution time is very minimal as shown in Fig. 6. Hence, when considering the execution time of a model based on hyperparameters, epochs and batch size has a higher priority compared to optimizers. Lastly, in Fig. 7, the execution time of these models on three machines for these hyperparameter settings shows that GTX1080 has the highest average execution time, and TitanX has the lowest average execution time. For the graphs shown in Figs. 5, 6, and 7, there are some points where the execution time is 0, this is due to the GPU Memory being insufficient, hence the training model cannot be loaded onto the machine and was unable to train the model. It can be observed in Fig. 7, that for GTX1080, and RTX2070 it was unable to train the models that were configured to be Adam optimizer and batch size 64, while TitanX, having a bigger memory capacity, was able to load and train the model.

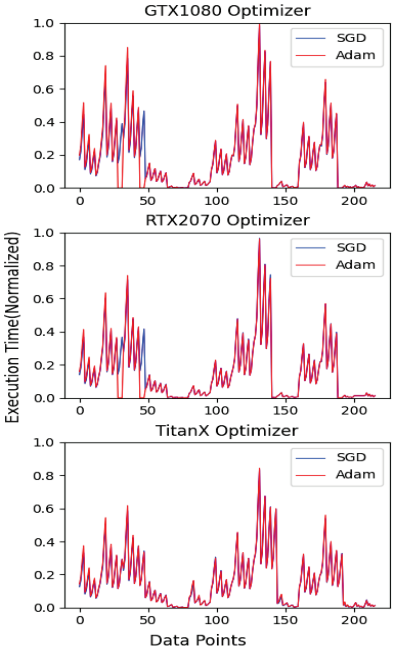


Figure 6 Execution Times based on optimizers among the 3 GPU machines

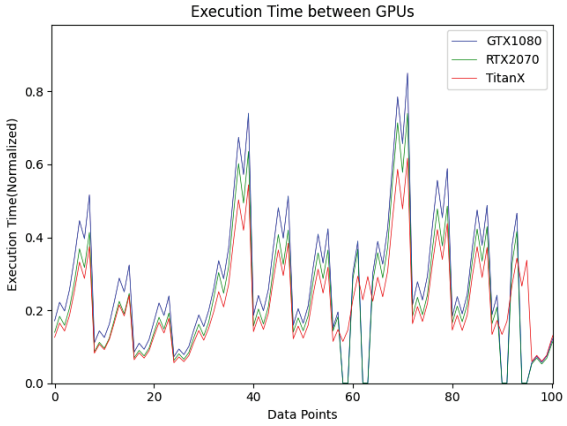


Figure 7 Execution time comparison between 3 GPU machines

4.3 Evaluating the Proposed Model

To test the accuracy of the prediction model, the proposed model was trained on the three GPU datasets generated using the data obtained from the previous section, to evaluate the validation loss using MSE. Driple [8] uses GPU Utilization%, and GPU Memory Utilization%, Network Transmission (Tx), Network Received (Rx) as inputs.

The proposed model changes the input parameters into GPU Utilization%, GPU Memory Utilization%, CPU Utilization%, and GPU Memory, with also the inclusion of Execution Time. Due to the nature of a training workload not being distributed on multiple machines unlike Driple, this approach only considers training the entire model on a single machine. Hence, Network parameters are not considered.

In Tab. 6, from the validation loss perspective, Driple has the best accuracy on the RTX2070 dataset, while the Proposed Model has the best accuracy on the other two datasets, the GTX1080 dataset and TitanX dataset. While the

proposed model also has a higher loss compared to the original model, this is expected due to the increase in parameter inputs, with an increase in total loss of about 6.84%.

Table 6 Validation loss comparison with driple

Model	Validation Loss			Evaluation	
	GTX1080	RTX2070	TitanX	Total Loss	Total Loss vs Original Model
Driple	0.022	0.0072	0.028	0.057	0
Proposed Model	0.017	0.024	0.0199	0.061	+6.84

Considering Validation Loss, it shows that the increase in input parameters increases the complexity of the model, which leads into higher validation loss. Hence, in line with the previous work [8], Transfer Learning is also explored in this research to determine the effects of transfer learning on the validation loss and how it is constructed.

4.4 Transfer Learning

Transfer learning involves using a pre-trained model and using its features as a reference when starting to train a new model. Using these features as a starting point, the training time and epochs required to train a new model will be decreased while also maintaining a similar amount of validation loss. It is also advantageous for models trained on smaller datasets [16], since the datasets used in the experiments only consist of 320 samples for each GPU.

To select which model to use as a pre-trained model for transfer learning, based on the validation loss in Table 6 for the proposed model, the GTX1080 dataset has the lowest validation loss, based on this, it has the best accuracy among the other two datasets. The original model, despite having the best accuracy between all test cases, was not considered because it uses a different set of inputs compared to the proposed model and it would be lacking the features that are necessary to be transferred to the new model that will be trained. Therefore, GTX1080 for the proposed model is selected to be the pre-trained model used for transfer learning.

In Tab. 7, the validation loss is evaluated between the proposed model with Transfer Learning and the model without Transfer Learning. For the validation loss, there is a 5% improvement in the accuracy when Transfer Learning is applied, as the total loss goes down to 0.0578. This is almost comparable to the previous model, Driple, which has a validation loss of 0.057.

Table 7 Validation loss

Model	Validation Loss			Evaluation	
	GTX1080	RTX2070	TitanX	Total Loss	Total Loss vs Proposed Model w/o TL
Proposed Model w/o TL	0.0171	0.0239	0.0199	0.0609	0
Proposed Model w/ TL	0.0174	0.0172	0.0232	0.0578	–5.090311987

Due to the improvement in the validation loss, the proposed model with transfer learning was used as the prediction model that was used for scheduling jobs using FIFO and SJF.

4.5 GPU Allocation for Job Scheduling

For the GPU Allocation of the proposed model, it will be evaluated based on two parameters, Computing Efficiency and Execution Time. Both parameters are outputs of the prediction model and are computed as the sum of these prediction values across all resource consumption parameters. In the case of Computing Efficiency, it is the sum of the ratio of the Average Peak Time and Average Burst Time per resource consumption parameter, while Execution Time is the sum of the predicted Execution Time per resource parameter.

4.5.1 GPU Allocation based on Computing Efficiency

To demonstrate how the GPU Allocating based on Computing Efficiency is determined, 6 jobs found on the dataset as specified in Tab. 8 are used as input to the prediction model. Each job arrives one minute apart from each other.

Table 8 Jobs for FIFO simulation (seconds)

	Optimizer	Batch Size	Epoch	Arrival (minute)
resnet20_cifar10	Adam	8	20	0
densenet40-k12_cifar10	SGD	64	10	1
inception3_imagenet	Adam	16	20	2
vgg11_imagenet	SGD	32	10	3
DCGAN_mnist	Adam	32	20	4
googlenet_imagenet	SGD	64	10	5

To compute the computing efficiency, there will be four different resource consumption parameters per GPU Dataset. Each of these resource consumption parameters has outputs for Average Burst Time, Average Idle Time, and Average Peak Consumption Rate. By utilizing Eq. (2) to obtain the computing efficiency of these output parameters and getting the summation of the ratio of values for four different resource consumption parameters, values shown in Tab. 9 are obtained for each GPU Dataset.

Table 9 Scheduling jobs based on computing efficiency

	GTX1080	RTX2070	TitanX	Allocation	JCT (s)
resnet20_cifar10	1.12	3.12	3.55	TitanX	849.21
densenet40-k12_cifar10	1.91	4.73	3.39	RTX2070	411.09
inception3_imagenet	2.01	4.99	4.51	RTX2070	11,179.09
vgg11_imagenet	2.42	4.76	4.79	TitanX	3,397.21
DCGAN_mnist	1.88	–485.58	–1861.76	GTX1080	248.48
googlenet_imagenet	1.91	–471.49	5.23	TitanX	4358.18

As observed in Tab. 9, the prediction model selects the highest computing efficiency amongst the dataset for GPU Allocation. In the case of DCGAN_mnist, it has the highest

computing efficiency compared to RTX2070 and TitanX, therefore this job is allocated to GTX1080. By following this criteria, the other workloads are allocated to their respective GPUs, and the JCT for each job is measured.

4.5.2 Performance Comparison based on Computing Efficiency

Since Driple [8] uses different input parameters compared to the proposed model, which includes CPU Util%, GPU Memory, and Execution Time, the proposed model and the Driple model were compared based on scheduling based on FIFO and SJF (sorting after job allocation), with the average JCT and makespan as the evaluation parameters. Furthermore, since the previous work does not have the Execution Time as its parameter, the comparison was based on the computing efficiency by using Eq. (2) to calculate the computing efficiency for each prediction target and obtaining the maximum total value for each resource consumption parameter as basis for GPU allocation.

In the case of Driple, the GPU allocation mostly ignored the GTX1080 GPU for this setup, having 0 jobs allocated on this machine, and the rest of the jobs allocated on the other two machines for both FIFO and SJF, as observed in Tabs. 10 and 12. It can be observed that while the makespan is higher on FIFO, the JCT is about 35.44% lower than Driple. However, looking at the makespan for each scheduling algorithm, it can be considered that Driple has the marginal advantage on makespan, even though the proposed model is marginally higher by approximately 0.035% on SJF. In terms of Average JCT, improvements on both models are observed when SJF is applied, with the proposed model 40.9% lower on average JCT on the SJF scheduling algorithm compared to Driple.

Table 10 FIFO based on computing efficiency (seconds)

	GTX1080	RTX2070	TitanX	Average JCT	Makespan
Driple	0	13814.65	20354.29	18719.38	**35120.41**
Proposed Model	3627.14	11135.98	19748.27	**12084.68**	35929.98

Table 11 FIFO Job Distribution based on computing efficiency

	GTX1080	RTX2070	TitanX
Driple	0 (0%)	10 (7.81%)	118 (92.19%)
Proposed Model	32 (25%)	54 (42.19%)	42 (32.81%)

Table 12 SJF based on computing efficiency (seconds)

	GTX1080	RTX2070	TitanX	Average JCT	Makespan
Driple	0	13134.74	13791.36	13627.21	34773.36
Proposed Model	3137.301	5941.055	14526.52	8057.22	34785.50

It is also observed that the job distribution does not change from FIFO to SJF since SJF is only reordering the jobs after allocating it to a designated GPU as shown in Tab. 13. Furthermore, since the input consists of training workloads, it can be seen in Tab. 10 for Driple that despite having only 10 jobs allocated to RTX2070, it has way higher JCT Compared to the Proposed Model having 54 jobs allocated to the same GPU. This shows that the training time of training workloads varies a lot depending on the parameters and architecture and reordering them and allocating them correctly improves the JCT of the given jobs.

Table 13 SJF Job Distribution based on computing efficiency

	GTX1080	RTX2070	TitanX
Driple	0 (0%)	10 (7.81%)	118 (92.19%)
Proposed Model	32 (25%)	54 (42.19%)	42 (32.81%)

4.5.3 GPU Allocation based on Execution Time

As Driple does not contain Execution Time, the GPU allocation based on Execution Time are only observed using the proposed model based on FIFO and SJF scheduling algorithms. Using the same set of jobs found in Table 8, the prediction model will produce a predicted execution time value for each resource consumption parameter, up to a total of four predicted values, for each GPU dataset. The predicted values will be summed up together per GPU Dataset as shown in Tab. 14. However, for allocating jobs based on execution time, the lowest predicted execution time for each dataset is considered. Hence, the GPU allocation are decided based on this criterion.

Table 14 FIFO based on predicted execution time (seconds)

	GTX1080	RTX2070	TitanX	Allocation	JCT (s)
resnet20_cifar10	–0.017	–0.036	–0.045	TitanX	819.21
densenet40-k12_cifar10	0.158	0.256	0.265	GTX1080	485.49
inception3_imagenet	0.82	1.52	2.51	GTX1080	13979.41
vgg11_imagenet	0.89	1.64	2.46	GTX1080	17452.4
DCGAN_mnist	1105.15	8929.84	17820.21	GTX1080	17700.88
googlenet_imagenet	1.21	2.22	3.11	GTX1080	19228.15

Contrary to the FIFO Allocation based on computing efficiency, the allocation based on Execution Time ended up assigning most of the selected jobs on GTX1080, which led to a very high JCT for most jobs, while also the RTX2070 not being selected for any of the given jobs for allocation. Based on the results from the allocation based on computing efficiency, shown in Tab. 9 and allocation based on execution time, shown in Tab. 14, it can be inferred that the Allocation performs way better when utilizing computing efficiency for the given jobs.

4.5.4 Performance Evaluation based on Execution Time

This time, the proposed model based on execution time prediction was observed for FIFO and SJF algorithm. As shown in Tab. 15, the average JCT of the proposed model increases by approximately 8.01% when using SJF scheduling compared to FIFO. However, the makespan decreases by approximately 1.09% when using SJF scheduling. Comparing this with the prediction results based on computing efficiency, the average JCT and makespan is increased by up to 124.87% an 23.13%, respectively.

Table 15 Scheduling based on execution time prediction (seconds)

	GTX1080	RTX2070	TitanX	Average JCT	Makespan
Proposed Model (FIFO)	11054.36	2361.27	23781.35	16665.91	45747.11
Proposed Model (SJF)	14445.05	1911.83	24226.93	18118.01	45252.55

Table 16 Job Distribution based on execution time

	GTX1080	RTX2070	TitanX
Proposed Model (FIFO)	48 (37.5%)	14 (10.94%)	66 (51.56%)
Proposed Model (SJF)	48 (37.5%)	14 (10.94%)	66 (51.56%)

As seen in the GPU allocation based on Execution Time and Computing Efficiency, the scheduler generally performed better on average JCT and Makespan, when based on computing efficiency, while using the proposed model. In contrast to Tabs. 11 and 13, it can be seen in Tab. 13 that jobs with lower JCT are allocated to RTX2070 and majority of the jobs with higher JCT were allocated to GTX1080 and TitanX.

4.6 Ablation Study

To understand the effects of the resource consumption parameters on the GPU allocation based on Average JCT and Makespan, two scenarios were considered by utilizing the same jobs as shown in Tab. 4. First, the resource consumption parameters are separated from each other and treated as an individual parameter for GPU Allocation. While the other scenario involves taking away one of each parameter and observing its effects on Average JCT and Makespan.

4.6.1 Resource Consumption Prediction of Individual Parameters

As shown in Tab. 17, utilizing only one of each resource consumption parameters for scheduling jobs with FIFO and SJF were observed. It shows that GPU Memory Utilization% does not consider GTX1080 for allocation while GPU Memory does not consider TitanX.

For each resource consumption parameter, the Average JCT and Makespan has seen improvements when SJF is applied, except for GPU Memory Utilization%, where it caused the Average JCT to increase by 5.20%. On the other hand, the CPU Utilization% has the lowest average JCT which is comparable to the output of the entire proposed model on Table 11. Furthermore, the makespan for both CPU Utilization% and GPU Memory is also lower than the entire proposed model.

The job distribution in Tab. 18 shows that the number of jobs allocated to a given GPU does not directly represent the JCT of a given job. This is due to the varying nature of the JCT of a given training workload which depends on the hyperparameter settings and model architecture. Therefore, despite having more jobs allocated to RTX2070 on GPU Memory Utilization%, and GTX1080 for CPU Utilization%, the other GPUs that have the workload with higher execution time will have the higher JCT for both cases.

Table 17 Scheduling based on computing efficiency using individual parameters (seconds)

	GTX1080	RTX2070	TitanX	Average JCT	Makespan
GPU Memory Utilization% (FIFO)	0	10625.55	24724.39	16973.80	46014.70
GPU Memory Utilization% (SJF)	0	12004.40	25492.16	17905.29	45740.25
GPU Utilization% (FIFO)	20498.89	5540.84	6413.51	14681.26	44198.14
GPU Utilization% (SJF)	13450.48	2776.40	5337.25	9874.39	42313.61
CPU Utilization% (FIFO)	3627.14	12201.24	15968.43	10646.34	31539.89
CPU Utilization% (SJF)	3362.36	8663.90	13568.19	8104.81	31469.66
GPU Memory (FIFO)	17076.69	11176.63	0	15786.05	30495.19
GPU Memory (SJF)	15543.98	12847.58	0	14954.14	29714.42

Table 18 Job distribution based on computing efficiency using individual parameters

	GTX1080	RTX2070	TitanX
GPU Memory Utilization% (FIFO)	0 (0%)	72 (56.25%)	56 (43.75%)
GPU Memory Utilization% (SJF)	0 (0%)	72 (56.25%)	56 (43.75%)
GPU Utilization% (FIFO)	76 (59.37%)	14 (10.94%)	38 (29.69%)
GPU Utilization% (SJF)	76 (59.37%)	14 (10.94%)	38 (29.69%)
CPU Utilization% (FIFO)	48 (37.5%)	14 (10.94%)	66 (51.56%)
CPU Utilization% (SJF)	48 (37.5%)	14 (10.94%)	66 (51.56%)
GPU Memory (FIFO)	100 (78.12%)	28 (21.88%)	0 (0%)
GPU Memory (SJF)	100 (78.12%)	28 (21.88%)	0 (0%)

4.6.2 Resource Consumption Prediction Removing One of Each Parameter

For Tab. 19, removing one of each resource consumption parameters from the model for scheduling jobs with FIFO and SJF were observed. It shows that when CPU Utilization% is removed, the jobs are only allocated on RTX2070 and TitanX. It also shows that removing GPU Memory Utilization% from the model increases the average JCT and Makespan of the entire proposed model by up to 2.25% and 5.99%, respectively.

Based on these results, it can be assumed that the GPU Memory Utilization% is not needed for the prediction model for scheduling, since removing one of the other parameters increases the JCT and Makespan of the scheduler. Also, it is notable that the scheduling based on GPU Memory Utilization% alone worsened the JCT and Makespan of the scheduler when using SJT.

In addition, it can be seen in Tab. 20 that CPU Utilization% highly influences the GPU allocation because

once it was removed there were 0 jobs allocated on GTX1080. On other cases, even if the other parameters were removed, the allocated jobs on GTX1080 remained consistent.

Table 19 Scheduling based on computing efficiency by removing one of each parameter (seconds)

	GTX1080	RTX2070	TitanX	Average JCT	Makespan
w/o GPU Memory Utilization% (FIFO)	3627.14	15216.02	10768.57	11901.85	33678.68
w/o GPU Memory Utilization% (SJF)	3285.68	9244.58	10534.79	7875.81	32699.8
w/o GPU Utilization% (FIFO)	3627.14	9212.53	22607.66	12630.06	41815.10
w/o GPU Utilization% (SJF)	3129.87	6839.89	15589.24	9056.68	41001.02
w/o CPU Utilization% (FIFO)	0	10933.61	27365.87	21460.53	47644.84
w/o CPU Utilization% (SJF)	0	6733.22	28965.25	20975.61	45447.84
w/o GPU Memory (FIFO)	3627.14	9586.68	21969.74	12546.96	41262.48
w/o GPU Memory (SJF)	3201.46	5358.39	15065.50	8307.65	39544.48

Table 20 Job Distribution based on computing efficiency by removing one of each parameter

	GTX1080	RTX2070	TitanX
w/o GPU Memory Utilization% (FIFO)	32 (25%)	84 (65.62%)	12 (9.38%)
w/o GPU Memory Utilization% (SJF)	32 (25%)	84 (65.62%)	12 (9.38%)
w/o GPU Utilization% (FIFO)	32 (25%)	50 (39.06%)	46 (35.94%)
w/o GPU Utilization% (SJF)	32 (25%)	50 (39.06%)	46 (35.94%)
w/o CPU Utilization% (FIFO)	0 (0%)	46 (35.94%)	82 (64.06%)
w/o CPU Utilization% (SJF)	0 (0%)	46 (35.94%)	82 (64.06%)
w/o GPU Memory (FIFO)	32 (25%)	50 (39.06%)	46 (35.94%)
w/o GPU Memory (SJF)	32 (25%)	50 (39.06%)	46 (35.94%)

Furthermore, it is also notable that by removing GPU Memory Utilization%, despite having only 12 jobs on TitanX, most of the jobs that has higher execution time was allocated to it and has comparable JCT compared to the RTX2070 that has 84 jobs allocated to it when using SJF Scheduling.

4.7 Scalability

Scalability of the GPU scheduling algorithm is tested to determine its behavior on larger scale systems. By increasing the number of available GPUs, the number of datasets that give out an efficiency score also increases. Furthermore, using non-neural network models as inputs and only having a simple graph as an input were analyzed to demonstrate that the algorithm does not specifically work only on neural network workloads, but it could also be designed to process non-neural network workloads.

The time complexity of the algorithm when using FIFO is shown in Fig. 8. It shows that the execution time grows linearly when GPU servers increase with the same number of jobs but when the jobs are also scaled up twice or four times the number of available GPU servers, the execution time of the algorithm grows based on the number of jobs x GPU servers. In addition, it is noticeable that there is a small bump in execution time on lower GPU servers, this is due to the overhead caused by loading up the Tensorflow library. Ideally, the algorithm only needs to boot up once and does not need to reload the Tensorflow library every time a new job appears in the queue.

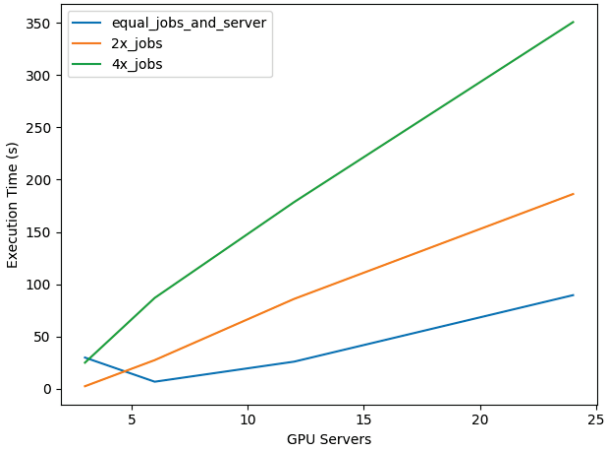


Figure 8 Execution time of the FIFO algorithm by scaling GPU servers and jobs

```
queue = get.Jobs()

for job in queue:                                                    #cost = O(n)
    inputs = convert_to_graph(jobs)

    for gpu in datasets:                                             #cost = O(m)
        resource_consumption_pred = prediction_model(inputs,gpu)
        comp_eff = sum(resource_consumption_pred[peak] / resource_consumption_pred[burst])
job.GPUAllocation(argmax(comp_eff))

if SJF:
    for each GPU:                                                    #cost = O(m)
        queue = sort.by(execution_time_pred,ascending)
```

Figure 9 Computational costs of the scheduling algorithm using big O notation

To further describe the behavior of the algorithm, Big O notation was utilized. As shown in Fig. 9, for O(n) where n stands for the number of jobs loaded into the queue, and O(m) where m stands for the number of GPU servers or datasets that are being used in the system. Based on the n and m parameters using the Big O notation, it is evaluated that the execution time for the FIFO Scheduling Algorithm will be O(n*m), which is determined by the number of jobs and number of GPU servers used in the system. For the SJF Scheduling Algorithm however, after allocation jobs are sorted for each GPU which adds another O(m) to the execution time.

This results in O(n*m) + O(m) for the SJF Algorithm. The expected average workload of the scheduling algorithm is 1 job and m number of GPU servers which leads into O(1*m) or O(m). Lastly, in the worst-case scenario there will be multiple jobs and multiple GPU servers to consider which

leads into a O(n*m) execution time. The summary of this data is shown in Tab. 21.

Table 21 Execution time for FIFO and SJF scheduling algorithms using big O notation

	Average	Worst-Case
(FIFO)	O(m)	O(n*m)
(SJF)	O(m) +O(m)	O(n*m) + O(m)

5 DISCUSSION AND CONCLUSION

5.1 Discussion

The experiments in this work used three GPU servers, specifically, TitanX, RTX2070, and GTX1080, solely due to availability. This work is not limited to these three GPUs only and can be implemented on other NVIDIA GPU types as well, since the framework of the system mostly runs on TensorFlow.

In scenarios where there are multiple GPU of the same type, a single dataset can be used to represent these GPU servers. For example, if there are three GTX1080 servers, a single dataset can be used for each of the GTX1080 servers since they all have similar capacity. However, since it is highly likely that due to them having a shared dataset, the scores will have the same or similar values, it is recommended that a priority system should be introduced when scheduling in systems with multiple GPU servers of the same type. Having the same GPU server also reduces the execution time of the algorithm since one dataset can represent multiple GPU servers in this scenario.

Limitations on the experiment setup used in this research is that the GPUs used in the work is limited to GTX1080, RTX2070 and TitanX and does not cover a wide range of GPU types which can help with analyzing the GPU performance for different training workloads that can contribute in creating a more generalized model used for scheduling jobs in cloud systems.

In addition, this work limited only with deep learning workloads. However, as long as a code can be represented into a graph, it can be trained and analyzed for a prediction model of different workloads not only limited to deep learning workloads. Adjusting the parameters and introducing new variables to incorporate a change in input.

There is also the lack of a general dataset for this type of application. Unlike other applications for images, videos, sentences, and the like, there is no fundamental dataset that can be used for resource consumption of deep learning workloads on different types of GPU. At the moment, this is very difficult since there's always new deep learning models that are being developed. However, the complexity of these models continues to increase and along with it the computing requirements to train such models in a reasonable amount of time also increases.

It is also worth noting that a lot of these deep learning models are also derived from a simpler form of its application. There are various deep learning models such as VGG and ResNet that fall under the CNN category where we can also infer that the architecture of these models have some similarity between them and when there is enough common application regarding the analysis of such architectures, a general dataset can be created that can be used as a reference data for future research regarding performance analysis and also job allocation of deep learning applications.

5.2 Using Non-Neural Networks as Input

When a different type of input is not a neural network, the scheduler will still be able to determine the GPU allocation given that the input is converted into a graph format using Tensorflow. By utilizing the tf.get_concrete_function() [21] to obtain the graph representation of a given function and the use of tf.io.write_graph() [22] to write the graph representation into a file for the scheduler to be able to process and convert it to its own adjacency and feature matrix that can be used against the prediction model to determine GPU allocation as shown in Figs. 10 and 11. The limitation for this approach is that the inputs must be in graph format, mainly using Tensorflow.

```
import tensorflow as tf

def simple_relu(x):
  if tf.greater(x, 0):
    return x
  else:
    return 0

tf_simple_relu = tf.function(simple_relu)
graph_def = tf_simple_relu.get_concrete_function(tf.constant(1)).graph.as_graph_def()
tf.io.write_graph(graph_def, '.', 'sample_graph' +'.pbtxt')
```

Figure 10 Obtaining the Graph representation of a function

```
node {
  name: "x"
  op: "Placeholder"
  attr {
    key: "_user_specified_name"
    value {
      s: "x"
    }
  }
  attr {
    key: "dtype"
    value {
      type: DT_INT32
    }
  }...
```

Figure 11 The output of the graph representation

5.3 Conclusion

The research was able to create multiple datasets, one for each machine used in this research, containing four resource consumption parameters (GPU Utilization, GPU Memory Utilization, CPU Utilization, and GPU Memory) and four prediction targets for each parameter (Average Burst Time, Average Idle Time, Average Peak Consumption, Execution Time) which totals up to 16 prediction metrics. By utilizing the datasets created, a prediction model that predicts both resource consumption parameters and the execution time of a given input was developed. To evaluate the performance of the proposed model against the previous work, both models are run on a FIFO and SJF scheduling mechanism.

The results show that GPU allocation based on computing efficiency, compared with the previous work, the average JCT is improved up to 40.9%. However, the makespan of the proposed model were still higher compared to the previous work. It has been discovered that removing one of the resource consumption parameters included in the proposed model, improves the overall JCT and Makespan by

an additional 2.25% and 5.99%, respectively. Due to these observations, future work is considered which involves removing GPU Memory Utilization% from the resource consumption parameters, and further analysis on Execution Time predictions, to seek improvements on both JCT and Makespan.

This work is also limited by using only Convolutional Neural Networks and some Generative models. Including more models from different applications and having a bigger dataset available for this application would help create a more generalized model for scheduling in the future.

Acknowledgements

This research was supported by Kumoh National Institute of Technology (2022-2023).

6 REFERENCES

[1] OpenAI (2023). GPT-4 Technical Report. *ArXiv*, abs/2303.08774. https://doi.org/10.48550/arXiv.2303.08774

[2] Jang, S. B. (2021). Deep neural network structure design for equipment failure prediction in smart factory. *Asia-pacific Journal of Convergent Research Interchange, 7*(12), 1-10. https://doi.org/10.47116/apjcri.2021.12.01

[3] Kim, D. (2018). Deep learning neural networks for automatic vehicle incident detection. *Asia-pacific Journal of Convergent Research Interchange, 4*(3), 107-117 https://doi.org/10.14257/apjcri.2018.09.11

[4] Kim, T. H. (2022).Video anomaly detection based on convolutional neural network. *Asia-pacific Journal of Convergent Research Interchange, 8*(11), 73-87. https://doi.org/10.47116/apjcri.2022.11.06

[5] Yu, G., Gao. Y, Golikov P. & Pekhimenko G. (2021). Habitat: A runtime-based computational performance predictor for deep neural network training. *Proceedings of the 2021 USENIX Annual Technical Conference*, USENIX ATC'21.

[6] Justus, D., Brennan, J., Bonner, S. & McGough, A. S. (2018). Predicting the computational cost of deep learning models. *2018 IEEE International Conference on Big Data (Big Data)*, 3873-3882. https://doi.org/10.1109/BigData.2018.8622396

[7] Viebke, A., Pllana, S., Memeti, S. & Kolodziej, J. (2019). Performance modelling of deep learning on Intel many integrated core architectures. *International Conference on High Performance Computing & Simulation (HPCS2019)*, 724-731. https://doi.org/10.1109/HPCS48598.2019.9188090

[8] Yang, G., Shin, C, Lee, J., Yoo, Y. & Yoo, C. (2022). Prediction of the resource consumption of distributed deep learning systems. *Proceedings of the ACM on Measurement and Analysis of Computing Systems, 6*, 1-25. https://doi.org/10.1145/3489048.3530962

[9] Pei, Z., Li, C., Qin, X., Chen, X. & Wei, G. (2019). Iteration time prediction for CNN in multi-GPU platform: Modeling and analysis. *IEEE Access* 1(1). https://doi.org/10.1109/ACCESS.2019.2916550

[10] Yang, G. (2022). Driple, https://github.com/gsyang33/Driple

[11] https://developer.nvidia.com/nvidia-system-management-interface

[12] Shin C., Yang G., Yoo Y., Lee J. & Yoo C. (2022). Xonar: Profiling-based job orderer for distributed deep learning. *IEEE 15th International Conference on Cloud Computing (CLOUD2022)*, 112-114. https://doi.org/10.1109/CLOUD55607.2022.00030

[13] Narayanan, D., Santhanam, K., Kazhamiaka, F., Phanishayee, A. & Zaharia, M. (2020). Heterogeneity-aware cluster scheduling policies for deep learning workloads. *Proceedings of the 14th USENIX Symposium on Operating Systems Design and Implementation,* 481-498.

[14] Duan, Y. &. Wu, J. (2021). Improving learning-based DAG scheduling by inserting deliberate idle slots. *IEEE Network, 35*(6), 133-139. https://doi.org/10.1109/MNET.001.2100231

[15] https://github.com/tensorflow/benchmarks/tree/master/scripts/tf_cnn_benchmarks, (2022)

[16] Shu, M. (2019). Deep learning for image classification on very small datasets using transfer learning.

[17] Weng, Q., Xiao, W., Yu, Y., Wang, W., Wang, C., He, J., Li, Y., Zhang, L., Lin, W. & Ding, Y. (2022). MLaaS in the wild: Workload analysis and scheduling in large-scale heterogeneous GPU clusters. *Symposium on Networked Systems Design and Implementation*. https://doi.org/10.21203/rs.3.rs-2266264/v1

[18] Lan, Z. Chen, M., Goodman, S., Gimpel, K., Sharma, P. & Soricut, R. (2019) ALBERT: A lite BERT for self-supervised learning of language representations. *ArXiv.* abs/1909.11942

[19] He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016)*, Las Vegas, NV, USA, 770-778. https://doi.org/10.1109/CVPR.2016.90.

[20] Zhu, R., Zhao, K., Yang, H., Lin, W., Zhou, C., Ai, B., Li, Y. & Zhou, J. (2019). AliGraph: A comprehensive graph neural network platform. *ArXiv*. abs/1902.08730 https://doi.org/10.48550/arXiv.1902.08730

[21] https://www.tensorflow.org/api_docs/python/tf/function

[22] https://www.tensorflow.org/api_docs/python/tf/io/write_graph

[23] Radford, A., Metz, L. & Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *ArXiv*. abs/1511.06434 https://doi.org/10.48550/arXiv.1511.06434

[24] Mirza, M. & Osindero, S. (2014). Conditional Generative Adversarial Nets. *ArXiv*. abs/1411.1784. https://doi.org/10.48550/arXiv.1411.1784

[25] Chollet, F. (2020). https://github.com/keras-team/keras-io/blob/master/examples/generative/vae.py

[26] Poon, S. M. Y. (2022). Simulation of vehicle target detection based on embedded system. *Journal of Science and Engineering Research, 1*(2), 1-20. https://doi.org/10.56828/jser.2022.1.2.1

[27] Tang, J. C. K. (2022). Deep learning-based analysis of voiceprint data mining. *Journal of Science and Engineering Research*, 1(1), 1-14. https://doi.org/10.56828/jser.2022.1.1.1

[28] Liu, X., Xu, H. & He, L. (2015). A Formal Method of CPU Resources Scheduling in the Cloud Computing Environment. *International Journal of Grid and Distributed Computing, 8*(1), 133-144. https://doi.org/10.14257/ijgdc.2015.8.1.13

Authors' contacts:

Abuda Chad Ferrino, M.S. Student
Department of Computer AI Convergence Engineering,
Kumoh National Institute of Technology,
61 Daehak-ro, Gumi-si, Gyeongsangbuk-do, 39177, Republic of Korea
cfpabuda@gmail.com

Tae Young Choe, Professor
(Corresponding author)
Department of Computer AI Convergence Engineering,
Kumoh National Institute of Technology,
61 Daehak-ro, Gumi-si, Gyeongsangbuk-do, 39177, Republic of Korea
choety@gmail.com

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240514152419

Preliminary communication
Received: 2024-05-14, Accepted: 2024-06-22

Method to Assess Computerised Systems Supporting Maintenance Services

Michał Pająk*, Bogdan Landowski, Łukasz Muślewski, Dragutin Lisjak

Abstract: There are many CMMS (Computerized Maintenance Management Systems) systems on the market in a wide range of prices and capabilities. An important problem for enterprise decision-makers is the selection of a system supporting maintenance services, appropriate for the specificity of the enterprise. The study presents an analysis of methods for assessing and selecting this type of systems and proposes a subjective-point method for analysing this type of issues. The issues presented in the study are part of the work on the development of materials and tools supporting enterprise decision-makers in the process of analysing existing solutions and selecting IT systems supporting maintenance services. Additionally, survey research was carried out, which enabled the generation of a database of basic values for the features of selected products, evaluation criteria, and in particular the determination of weight values for individual criteria, which contributed to the automatic execution of calculations for the adopted assumptions.

Keywords: CMMS; Computer Aided Systems; Computerised Maintenance Management; Maintenance; Selection Method

1 INTRODUCTION

Computerized Maintenance Management Systems (CMMS) are intended to support maintenance subsystem (maintenance engineering) of all types of enterprises where technical objects are operated. They enable the collection of information on failures of objects and operating processes carried out in the enterprise, along with their detailed descriptions for machines, devices and vehicles, as well as the development of periodic and preventive inspection schedules and their queuing [1]. Formal definition can be found in [2]. According to it a computerized maintenance management system (CMMS) is any software package that maintains a computer database of information about an organization's maintenance operations.

The pursuit of a continuous increase in the efficiency of enterprise operations with the constant increase in requirements regarding product quality and the growing complexity of manufacturing processes, as well as the increase in the number of regulations and formal and legal requirements mean that one of the key elements determining the development of an enterprise is the ability to effectively use IT and telecommunications techniques.

Methods for assessing IT systems supporting maintenance engineering and optimizing the selection of these systems for a specific enterprise can be divided into two main groups: objective methods based on mathematical optimization tasks and subjective methods based on the analysis of selected criteria with assignment appropriate weights of them. However, in industrial practice, in vast majority cases the evaluation methods belonging to the second group mentioned above are used.

The decision-maker selects a quasi-optimal CMMS system from many possible solutions, taking into account, often contradictory, selection criteria. In practice, the decision-maker builds a summary table of all the data "for" and "against" affecting the choice of a given solution and hierarchizes the systems from the used set of assessment criteria point of view.

In order to support the decision-making process of the CMMS systems selection and reduce the degree of its subjectivity, research was carried out, the result of which is the CMMS software evaluation system described in this paper.

The remaining part of the paper is organized as follows: Chapter 2 describes the basic goals and functions of CMMS systems, and Chapter 3 presents an analysis of existing multi-criteria analysis methods. On this basis, an evaluation method is proposed in Chapter 4 and a system for evaluating CMMS programs is defined in Chapter 5. Chapter 6 describes the survey research conducted, presents and analyses its results. The paper is summarized in Chapter 7.

2 GOALS AND BASIC FUNCTIONS OF CMMS SYSTEMS

The increasing degree of automation of production and service processes as well as technical and technological progress with the simultaneous constant increase in requirements regarding the safety of use of technological machines and products, as well as the pursuit of reducing production costs result in an increase in the importance of maintenance subsystems in the structure of the company. This also involves the need to use IT systems supporting maintenance engineering.

It is possible to distinguish general principles, which are widely accepted and must always be taken into account when selecting devices, including computer devices and IT systems. These are functionality, reliability and durability, efficiency, purchase and implementation costs, operating costs, availability of consumables, ease of use, ergonomics and compliance with applicable standards and regulations.

The purchase and implementation of an appropriate CMMS system depends on the requirements formulated for this type of systems by decision-makers, the type of production or services provided, the type of technological devices used (including their complexity and level of automation), the size of the enterprise and organizational structure of the company.

It should be noted that software systems supporting maintenance processes are only a tool and the benefits of their implementation depend not only on their proper selection from the point of view of the specific nature of the

enterprise and the goals set formulated for them by decision-makers, but also on their rational use and knowledge as well as the ability to use their capabilities and functions. The correct formulation of the goals and requirements for CMMS systems is a necessary condition for their proper selection and obtaining the expected effects from their implementation.

The general objectives that can be achieved by the implementation and rational use of software which supports maintenance subsystems are the following: reducing costs while ensuring the required level of readiness and reliability of the technical objects in operation, reducing machine downtime and increasing the efficiency of service processes. These objectives can be expressed by the following list of particulars goals [3, 4]:

- creating comprehensive documentation regarding the machines, tools and other technical objects used and ensuring easy and quick access to this data by authorized persons,
- standardization of the terminology used,
- supporting planning processes for handling and purchasing consumables and spare parts,
- recording and processing of data regarding operational events,
- reducing machine downtime,
- quick analysis of data regarding failures of technical objects,
- automatic generation of developed report templates for selected time intervals (working times, service times, downtime, etc.)
- identification of costs related to service processes, including the possibility of recording and analysing costs by type,
- optimization of the management of consumables and spare parts (including automation of the processes of making orders),
- control, analysis and optimization of material resources,
- identification of machine nodes particularly susceptible to damage,
- identification of repetitive failures,
- increasing the efficiency and quality of service processes thanks to increased employee involvement achieved by identifying their activities.

The basic functions of CMMS systems are:

- recording and processing data regarding technical objects in use (process lines, machines included in them, standalone devices) and means of operation,
- mapping the enterprise structure (decomposition of the organizational and technical structure to the adopted level of detail) and its visualization,
- recording and processing data regarding service processes (to the extent determined by the decision-makers),
- supporting and planning the service processes
- recording the scope of service activities and their automatic generation for a given service,
- automating the generation of orders for the execution of service processes, defining scope of activities, employees performing the service, list of necessary tools and consumables, forwarding orders and documentation for specific positions in the company
- management of authorizations of employees providing services,
- settlement of completed works,
- automatic generation of required reports and periodic and current reports,
- recording and settlement of costs including types and centres,
- monitoring (alerting) about scheduled service dates.

The scope of functions offered by the CMMS systems available on the market are significantly expanded in relation to the basic functions listed above. The diversity of functions of this type of systems, on the one hand, enables the appropriate selection of the system to suit the specificity and needs of the enterprise, but on the other hand, it makes it difficult (due to the lack of unification) to analyse these systems and their proper selection.

3 ANALYSIS OF EXISTING SELECTION METHODS

In fact, the choice of a CMMS system can be reduced to formulating a system of n inequalities with m variables, where n is the number of criteria and m is the number of arguments of these criteria. In the case of one-dimensional criteria, there is a relationship $m \leq n$. Since in a multi-criteria evaluation system there are in most cases contradictory or partially mutually exclusive criteria, the formulated system of inequalities is often contradictory (rarely undefined). This means that there is often a situation when none of the considered variants of the solution (CMMS systems) satisfies the system of inequalities, or many of the assessed systems meet such a system. It is therefore necessary to use such an evaluation method that in each case (contradictory, definite and undefined system of inequalities) it is possible to rank the assessed solutions and clearly select the best one. It is proposed to accomplished this task through the use of multi-objective analysis method [5].

The main goal of multi-objective analysis is to rank the assessed variants in a specific order and, if possible, to determine the overall quality of individual variants in the form of assigning appropriate grades or scores [6]. There are many methods of multi-criteria analysis, but they are all based on a common scheme. The first phase of the analysis is to determine the criteria that make up the multi-criteria evaluation system. Given that, each criterion can and usually concerns a completely different field and there are usually both quantitative and qualitative criteria. An evaluation method is introduced that allows assigning a grade to each variant according to each criterion. In addition, a system of weights for the criteria is introduced, allowing for differentiating the importance of individual assessments. It is also possible to introduce weights ranking the impact of the assessments of individual decision-makers on the final assessment, if there are many of them. The final stage is the process of calculating the total score of a given variant using a specific algorithm. This process includes the evaluation of the variant according to individual criteria, the weight of individual criteria and the weight of decision-makers, if they are introduced. Because of the final process, the variants are

ranked in terms of their total rating value. It should be noted that the possibility of simple interpretation of the obtained total assessment value is very helpful in the practical application of the analysis.

Analysing scientific databases as Web of Science and Scopus in the field of assessment methods it can be noticed that commonly used multi-criteria evaluation methods are linear and relational methods. These are the main assessment techniques groups. Therefore in the paper the comparison will be limited to them. The first group consists of hierarchical linear methods. Representative method of the first group is the point method. It occupies a special place in quality assessment because it combines all separately assessed features into one number that comprehensively expresses the overall quality of the examined object [7]. It is based on the hierarchy of system elements and their distance from the achievable maximum value. The assessment is based on the adoption of a specific scale. The hierarchy is characterized by an ascending or descending order indicating the degree of fulfilment of the global criterion, including all sub-criteria. The result of the point method is a ranking of the tested variants from the point of view of the degree to which the requirements are met. The significant advantages of this method are the resistance to high differences in observations, the ability to compare quantitative and qualitative features simultaneously, simple and understandable structure, the relative ease of interpretation, short implementation time and low testing costs.

The precision of the obtained results depends on the proper definition of individual quality levels and this is the first condition to obtain meaningful results. The second condition is the training of the assessment team, allowing clear understanding of the definition of individual features of the object. Definitions cannot contain concepts that are emotional or too general [8].

The second group are relational methods. The example of relational methods is the AHP (Analytic Hierarchy Process) method. There are two preparatory stages in it. First stage consists in a hierarchically superior determination of the relative dominance of criteria. It is obtained from pairwise comparison of them. The second one is the calculation of the relative dominance of individual solution in terms of subsequent criteria. The final step is the determination of synthetic scores organizing the assed solutions and the analysis and interpretation of the method results [9, 10].

While in point methods the assessment is given taking into account one's own feelings as to the fulfilment degree of a given criterion by the object separately from other objects, in relational methods the differences in fulfilment degree by individual objects are mainly taken into account. In point methods, relations between criteria can be externally imposed in the form of a preference vector. In contrast, in relational methods they are calculated as a preference vector from the dominance matrix. There are many similarities between the analysed methods, but the interpretation of the results of the AHP method is much more difficult. None of them takes into account the correlations between the criteria [8]. To summarize the differences, it should be emphasized that in the point method, points are awarded on the basis of predetermined criteria, so it is an absolute assessment, while in relational methods, elements are compared and ranked relative to each other. Therefore, it is a relative assessment.

4 ASSUMPTIONS OF CMMS ASSESSMENT METHOD

Based on the analysis performed, it is proposed to use the point method to evaluate IT systems supporting maintenance engineering.

The key element of the assessment method are the assessment criteria formulated on the basis of the objectives that should be achieved through the implementation and rational use of electronic systems supporting maintenance subsystems. Analysing the objectives identified above, it was stated that the evaluation criteria are both quantitative (reduction of machine downtime expressed in hours) and qualitative (standardization of the terminology used expressed on a subjective evaluation scale). Therefore, the proposed assessment method must meet the assumptions of a multi-criteria assessment combining both types of criteria into a coherent assessment system. It was decided to use a fuzzy extension of the SMART multi-objective assessment method.

It should be noticed that the use of elements of fuzzy logic allows for modelling ambiguously specified assessment criteria and taking into account the subjective nature of the assessment according to qualitative criteria [11].

In the SMART method [6], the criterion domain is determined based on the upper and lower range of variability of the criterion argument. It is assumed that the range of variability is divided into six intervals, the sizes of which increase with the distance from the optimum according to a geometric series with a quotient equal to two. Depending on the type of criterion, they are described by different functions. For the criterion where the most desirable value is the smallest value or the largest possible value, the function takes the form Eq. (1):

$$v = \log_2\left(\frac{P_v - P_{\min}}{P_{\max} - P_{\min}} \cdot 64\right), \tag{1}$$

For the criterion where the most desired value is the largest value, the function takes the form Eq. (2):

$$v = \log_2\left(\frac{P_{\max} - P_v}{P_{\max} - P_{\min}} \cdot 64\right), \tag{2}$$

For the function values corresponding to the arguments defined by Eq. (3) for criterion (1) and Eq. (4) for criterion (2), a scale from 4 to 10 is introduced according to Eq. (5).

$$P_v = P_{\min} + (P_{\max} - P_{\min}) \cdot \frac{2^v}{64}, \ v = 0, 1, ..., 6 \tag{3}$$

$$P_v = P_{\max} + (P_{\max} - P_{\min}) \cdot \frac{2^v}{64}, \ v = 0, 1, ..., 6 \tag{4}$$

$$g = 4 + v, \tag{5}$$

where: v - the value of the criterion function, P_{max} – the upper range of variability of the criterion argument, P_{min} – the lower range of variability of the criterion argument, P_v - the argument of the criterion function, g – the degree of the criterion fulfilment.

For criterion (1), with the desired value as high as possible, a scale from 4 to 10 is introduced according to the formula (6):

$$g = 10 - v. \tag{6}$$

A normalized weight c is determined for each criterion. The total rating of variant s is calculated according to formula (7)

$$s = \sum_{i=1}^{k} c_i \cdot g_i. \tag{7}$$

The value of the final grade ranges from 4 to 10. This allows for verbal interpretation of the value (Tab. 1).

Table 1 Interpretation of the variant's total rating value

Total rating values	Interpretation
10	Ideal
9	Very good
8	Good
7	Poor
6	Very poor
5	Bad
4	Very bad

The lack of precision in determining the degree of fulfilment of individual criteria and the possibility of distinguishing individual variants from the considered criteria point of view should be modelled by describing each criterion with a fuzzy set with a membership function consisting of straight lines. The shape of the broken membership function curves depends on the type of criterion determined using the theory of multi-objective analysis.

For the criterion for which the smallest value is most desirable (MINSIMP), a membership function with the shape shown in the Fig. 1, calculated based on Eqs. (1), (3) and (5). The criterion in which the biggest value is most desirable (MAXSIMP) is modelled as a membership function calculated using Eqs. (2), (4) and (5) (Fig. 2). While, in the case of the criterion for which the biggest possible value is the most desirable (MAXINV), the membership function presented in the Fig. 3 is used. It is calculated using Eqs. (1), (3) and (6).

The presented functions are scaled to the appropriate size of the fuzzy set domain by dividing the set fuzzy set support into 64 equal parts. The criterion domain is assumed equal to the support of the fuzzy set. The function values for the support elements range from 4 to 10. In order to describe the criteria by normal fuzzy sets, these values are divided by 10.

Developing an evaluation system according to the described above multi-objective analysis method involves defining a set of evaluation criteria, determining for each of them the type of criterion and the range of argument variability. Additionally, for each of them the weights determining the importance of individual criteria should be assign.

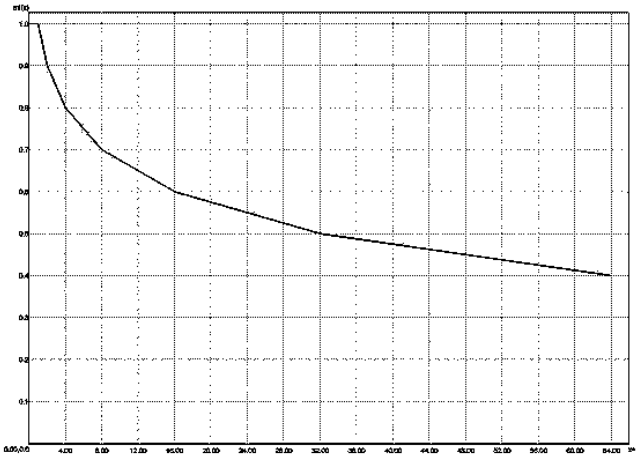


Figure 1 Membership function of the MINSIMP criterion

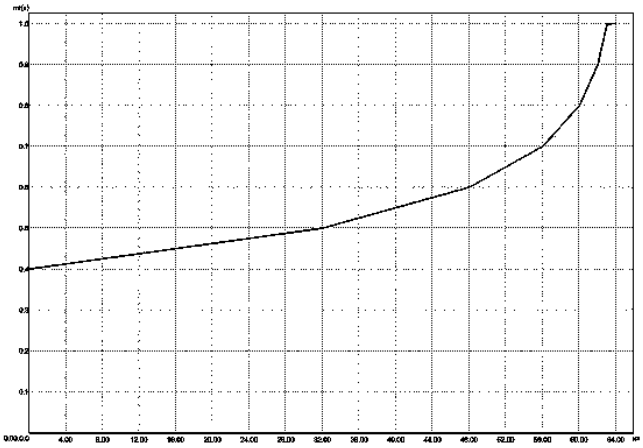


Figure 2 Membership function of the MAXSIMP criterion

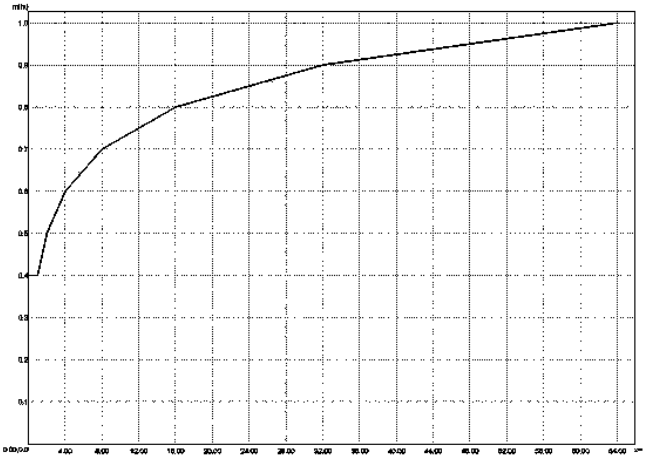


Figure 3 Membership function of the MAXINV criterion

For quantitative criteria, the degree of fulfilment of the criterion is clearly defined, while for qualitative criteria it is a subjective assessment dependent on an expert. In such cases, the use of fuzzy inference was proposed.

For the argument, uniform partitioning was introduced using seven fuzzy sets, the first of which is of the L type, the

last of the Γ type, and all intermediate ones of the Λ type [12]. The sets denote the linguistic terms very small (VL), small (L), medium small (ML), medium (M), medium high (MH), high (H) and very high (VH) (Fig. 4).

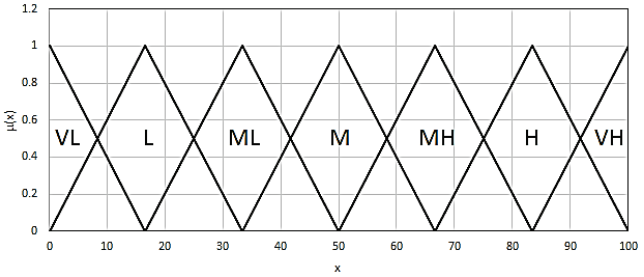


Figure 4 Partitioning of the argument range

Experts describe the argument value using the introduced scale. Then, similarly to the fuzzy controller [13], the degree of activation of individual fuzzy sets is determined based on the percentage of expert responses consistent with a given fuzzy set. On this basis, a numerical value is determined by using the height operator (8) [14] . Based on the determined value, the degree of fulfilment of the criterion is found.

$$x_{\mathrm{H}} = \frac{\sum_{i=1}^{k} h(FS_i) \cdot x(h(FS_i))}{\sum_{i=1}^{k} h(FS_i)}, \tag{8}$$

where: x_{H} – the sharp value of the criterion argument, k – the amount of the criteria, FS_i – i^{th} fuzzy set, $h(FS_i)$ – the height of i^{th} fuzzy set, $x(h(FS_i))$ – the position of height of i^{th} fuzzy set.

For the used selection criteria system, the team of experts will assess the degree to which individual criteria are met by the analysed CMMS systems. Additionally, each expert will determine the weight of individual criteria on a scale from 1 to 10. The evaluators (experts) may be specialists in the implementation and operation of this type of systems and managers of maintenance departments in enterprises of the analysed industries.

5 EVALUATION SYSTEM OF THE CMMS SOFTWARE AND PRODUCT EVALUATION SURVEY

Computer systems supporting maintenance engineering facilitate rational and effective management of enterprise resources. Existing solutions in most cases provide the basic functions required for this category of software. However, their degree of complexity, ease of implementation and use, range of additional functions offered, susceptibility to user intervention and adaptation to their specific needs make it very difficult and sometimes impossible to directly compare the offered solutions. This makes the process of selecting the best solution suited to the specificity of the company and its needs more difficult.

Producers and distributors of this type of software, for obvious reasons, do not indicate possible problems and costs related to the implementation and use of the offered systems, only enumerating their advantages, which most often include better use of company resources, which is supposed to enable: reduction of production costs, reduction of stocks and reducing the number of damages to machines and devices used in processes carried out therein.

In order to be able to compare and evaluate software packages to support maintenance engineering, a uniform evaluation system was developed. It was formulated as a result of literature research [15-18], analysis of offers from global [19] and local [20-22] CMMS system manufacturers, and industrial research conducted in selected SME transport companies. The assessment system were formulated based on functions and features of CMMS systems mentioned earlier in the paper. The resulting evaluating system consists of 37 criteria. For each of them, the type of criterion (MINSIMP, MAXSIMP, MAXINV), the type of argument (numerical value - NV, linguistic value - LV) and the range of argument variability were determined (Tab. 2). On this basis, a survey research card was developed in which experts in the field of the problem determine the weights of individual criteria as well as the values of arguments for each of the assessed systems.

Table 2 List of criteria of the formulated evaluation system

No	Criterion	Criterion type	Argument type	Argument range
1	Possibility to modify the form of data and reports presentation	MAXINV	LV	VL-VH
2	Data protection against unauthorized access	MAXSIMP	NV	0-7
3	Implementation time	MINSIMP	NV (days)	30-700
4	Availability of technical support	MAXSIMP	NV (%)	50-100
5	Functionality	MAXINV	LV	VL-VH
6	Comfort of work	MAXINV	LV	VL-VH
7	Annual system operating costs	MINSIMP	NV (KEUR)	0-15
8	Implementation costs	MINSIMP	NV (KEUR)	2-40
9	Purchase costs	MINSIMP	NV (KEUR)	0-400
10	Ease of analysis of operational events	MAXINV	LV	VL-VH
11	Guaranteed system availability	MAXSIMP	NV (%)	80-100
12	Possibility of adaptation to individual needs	MAXINV	LV	VL-VH
13	Possibility of configuration and modification by users	MAXINV	LV	VL-VH
14	Impact on increasing the efficiency of the service system	MAXINV	LV	VL-VH
15	Possibility to reduce stocks	MAXINV	LV	VL-VH
16	Possibility to reduce machine downtime	MAXINV	LV	VL-VH
17	Influence on the optimization of the exploitation strategy	MAXINV	LV	VL-VH
18	Possibility to predict the durability and reliability of machines	MAXINV	LV	VL-VH

Table 3 List of criteria of the formulated evaluation system (continuation)

No	Criterion	Criterion type	Argument type	Argument range
19	Possibility of localization of the system	MAXSIMP	LV	VL-VH
20	Possibility of expanding the system	MAXINV	LV	VL-VH
21	Possibility of cooperation with other systems	MAXINV	LV	VL-VH
22	Reliable operation	MAXSIMP	NV (%)	80-100
23	Resistance of data to damage of IT devices	MAXSIMP	NV	0-7
24	Optimization of the use of technical infrastructure	MAXSIMP	LV	VL-VH
25	Optimization of the use of human resources	MAXSIMP	LV	VL-VH
26	Position of the system manufacturer on the market	MAXSIMP	NV	0-7
27	Complication of use	MINSIMP	LV	VL-VH
28	Complexity of the implementation process	MINSIMP	LV	VL-VH
29	The intensity of system development by the manufacturer	MAXINV	LV	VL-VH
30	Completeness of documentation regarding operational events	MAXSIMP	NV	0-7
31	The degree of usefulness of the functions offered by the system	MAXSIMP	NV (%)	50-100
32	Speed of operation	MAXSIMP	LV	VL-VH
33	Versatility	MAXINV	LV	VL-VH
34	Support for the introduction of ISO performance quality standards	MAXINV	LV	VL-VH
35	Supporting of service processes planning	MAXSIMP	LV	VL-VH
36	Required qualifications of system users	MINSIMP	NV	0-7
37	Compliance to applicable standards and regulations	MAXINV	LV	VL-VH

In the evaluation system, in addition to the criteria formulated for numerical and linguistic values, there are also criteria whose arguments take values from the defined lists. The meaning of these values of the lists for particular criteria is described in the Tab. 3.

Table 4 The meaning of list values for particular criteria

	Criterion				
	2	23	26	30	36
0	Lack	Lack	Single installation	Lack	Lack
1	OS security	Local media backup	Local market participant	Event log	Primary education
2	User / password security	Disk array systems	A niche position on a national market	Log of events divided into groups	Primary with training

Table 5 The meaning of list values for particular criteria (continuation)

	Criterion				
	2	23	26	30	36
3	User / password security with user profiles	Local server backup	A well-known national market participant	Report	Second. education
4	2FA software security	Distributed server backup	Dominant position on a national market	Report divided into events groups	Second. education with training
5	2FA software security with user profiles	Local data mirroring	A niche position on an international market	Report divided into events groups and basic descript.	Higher
6	Hardware-based 2FA security	Remote data mirroring	A well-known international market participant	Report divided into events groups and extended descript.	Higher education with training
7	Hardware 2FA security with user profiles	Cloud solution	A dominant position on an international market	Report of events with full descript.	Expert level

6 SELECTED RESULTS OF ASSESSING CMMS SYSTEMS

In order to verify the applicability of the developed evaluation system, a survey was conducted. The research was carried out by surveying respondents. A set of 23 employees of maintenance departments of transport SME was randomly selected. All respondents had higher technical education. The group of respondents included both the employees supervising maintenance services and the employees directly performing maintenance tasks. A set of criteria for assessing CMMS systems was adopted as a survey. Respondents were to evaluate, on a scale from 1 to 10, the importance of individual criteria (the higher the rating value, the more important the criterion).

In Tab. 4 the results of the conducted surveys regarding the assessment of the significance of the criteria for assessing CMMS can be found. Additionally, the assessment of an exemplary CMMS system is also presented.

A set of 37 evaluation criteria was analysed. The high importance ratings for the analysed criteria are noteworthy. Only 3 out of 37 analysed criteria received an importance rating below 5.

Table 6 Assessment results (scale from 0 to 10) of the importance of the criteria and the grades of the exemplary CMMS system

Criterion no.	Mean weight	Standard deviation	Variation coefficient	CMMS
1	9,130	0,869	0,095	0,712
2	6,565	1,950	0,297	0,52857
3	7,217	1,808	0,251	0,55672
4	8,957	0,825	0,092	0,64
5	6,087	1,649	0,271	0,94
6	8,957	0,928	0,104	0,9334
7	9,348	0,714	0,076	0,59333
8	7,348	1,799	0,245	0,63158
9	4,130	1,180	0,286	0,58

Table 7 Assessment results (scale from 0 to 10) of the importance of the criteria and the grades of the exemplary CMMS system (continuation)

Criterion no.	Mean weight	Standard deviation	Variation coefficient	CMMS
10	9,261	0,689	0,074	0,9574
11	6,217	2,255	0,363	0,74
12	8,130	1,424	0,175	0,7072
13	5,565	1,502	0,270	0,7328
14	9,391	0,656	0,070	0,9026
15	5,261	2,281	0,434	0,9028
16	9,304	0,703	0,076	0,904
17	9,696	0,470	0,049	0,924
18	7,348	1,584	0,216	0,9214
19	6,870	2,117	0,308	0,5
20	6,174	2,037	0,330	0,9334
21	6,522	3,189	0,489	0,884
22	9,783	0,422	0,043	0,84
23	9,739	0,449	0,046	0,52857
24	8,783	0,951	0,108	0,4868
25	9,304	0,703	0,076	0,4974
26	3,043	1,637	0,538	0,48571
27	9,565	0,590	0,062	0,4532
28	5,913	1,505	0,255	0,448
29	8,217	0,850	0,103	0,9668
30	6,696	1,820	0,272	0,68571
31	9,435	0,662	0,070	0,58
32	9,217	0,850	0,092	0,5668
33	3,870	1,486	0,384	0,7328
34	8,696	0,926	0,107	0,9014
35	9,826	0,388	0,039	0,4948
36	8,739	1,137	0,130	0,48571
37	9,913	0,288	0,029	0,9334

Analysing the obtained results, it can be concluded that in the set of 37 analysed criteria, three subsets of criteria can be distinguished, based on the value of the coefficient of variation of the obtained assessments of their importance.

The first subset of 15 criteria is characterized by a low coefficient of variation ($v < 0.1$). This means that respondents were very consistent in their assessments of the importance of the analysed criteria. At the same time, this subset of criteria includes criteria that received the highest importance rating (above 9). There is a correlation between the average value of the obtained weights and the dispersion of individual respondents' ratings around the average value.

The second subset, consisting of 15 criteria, is characterized by a coefficient of variation value ranging from 0.1 to 0.3 ($0.1 < v < 0.3$). It seems that such a dispersion of the obtained ratings may result from the respondents' different preferences, goals and expectations for CMMS systems.

The third subset of 7 criteria is characterized by a relatively high value of the variation coefficient ($v > 0.3$). This means a large discrepancy among respondents as to their assessments of the importance of the analysed criteria. At the same time, this subset of criteria includes criteria that received a relatively low importance rating.

In order to obtain an assessment of the exemplary CMMS system, the obtained criteria weights were normalized by dividing each of the weights by their sum. Thanks to this, the weights sums to 1. Then, 10 multiplied the calculated ratings of the considered CMMS system in order to obtain a rating consistent with the SMART method. Finally, using Eq. (7), the total score of the system was calculated. The score was equal to 7.1, which can be interpreted in accordance with the table (Tab. 1) as poor quality of the analysed solution.

The obtained results confirm the validity of the adopted assumptions of the method for assessing CMMS systems, both in terms of the need to use the possibility of individual criteria weights selection and the suitability of the developed set of criteria for the assessment of this type of systems. The analysis of the obtained results may be useful in the process of selecting from among existing CMMS systems the solutions that best meets the expectations of enterprise decision-makers (hierarchization of criteria). It can be also useful for producers and enterprises dealing with the distribution and implementation of this type of software packages.

7 SUMMARY

An important problem for enterprise decision-makers is the selection of a system supporting maintenance services, appropriate for the specificity of the enterprise. The considerations presented in this work constitute the assumptions of the method for assessing IT systems supporting maintenance services. The elements of the method are a defined set of criteria, the method of determining the assessment in the case of criteria formulated for numerical and linguistic arguments, and the procedure for calculating the weights of the criteria, the degree of fulfilment of individual criteria, as well as the final assessment of the analysed CMMS system.

The study presents an analysis of methods for assessing and selecting this type of systems and proposes a subjective-point method to use. The issues presented in the study are part of the method of supporting enterprise decision-makers in the process of analysing existing solutions and selecting IT systems supporting maintenance services.

The conducted surveys proved the accuracy of the selection of evaluation criteria and the usefulness of the proposed method for assessing CMMS systems. There are many CMMS systems on the market in a wide range of prices and capabilities. Information about available packages of this type, compiled in a uniform and coherent form, will enable enterprise decision-makers to conduct a preliminary analysis of existing solutions and facilitate the process of selecting software tailored to the needs of a given enterprise.

The main limitations of the research are the comparison of only two most popular assessment methods. Therefore, it is planned to extend the analysis in the future. It is also planned to use the developed assessment method to create a computer system supporting the CMMS system selection process. Such a system will support the formulation of criteria, determining their types, the range of variability of arguments, and the shape and location of fuzzy sets used to partition the range. It will also automatically perform the calculations necessary to obtain the final score of the assessed solution.

8 REFERENCES

[1] Ogbo, A. I., Eneh, N. C. J., Mbah, C. N. & Isijola, D. O. (2018). Effects of comp manufacturing companies in Enugu state, Nigeria. *Sci Technol, 4*, 38-51.

[2] Cato, W. W. & Mobley, R. K. (2002). *Computer-Managed Maintenance Systems* (Second Edition). Chapter 2 - Definition of a CMMS. Woburn: Butterworth-Heinemann, 13-55. https://doi.org/10.1016/B978-075067473-7/50002-4
[3] Aliyu, A., Baglee, D. & Dixon, D. (2023). Computerised maintenance management system (CMMS) role in small and medium enterprise (SME). *Proc Int Conf Cond Monit Asset Manag 2023*, 1-7. https://doi.org/10.1784/cm2023.2f5
[4] Lopes, I., Senra, P., Vilarinho, S., Sá, V., Teixeira, C., Lopes, J., et al. (2016). Requirements specification of a computerized maintenance management system–a case study. *Procedia Cirp, 52*, 268-273. https://doi.org/10.1016/j.procir.2016.07.047
[5] Mansouri, S. A., Lee, H. & Aluko, O. (2015). Multi-objective decision support to enhance environmental sustainability in maritime shipping: A review and future directions. *Transp Res Part E Logist Transp Rev, 78*, 3-18. https://doi.org/10.1016/j.tre.2015.01.012
[6] Muślewski, Ł., Pająk, M., Migawa, K. & Landowski, B. (2022). An expert system for optimizing the operation of a technical system. *J Qual Maint Eng, 28*,131-153. https://doi.org/10.1108/JQME-05-2020-0033
[7] Silverman, D. (2020). *Qualitative Research*. London: SAGE Publications Ltd.
[8] Denzin, N. K. & Lincoln, Y. S. (2017). *The SAGE Handbook of Qualitative Research*. SAGE Publications, Inc.
[9] Bhatt, N., Guru, S., Thanki, S. & Sood, G. (2021). Analysing the factors affecting the selection of ERP package: A fuzzy AHP approach. *Inf Syst E-Bus Manag, 19*, 641-682. https://doi.org/10.1007/s10257-021-00521-8
[10] Kahraman, C., Onar, S. C., Öztayşi, B., Şeker, Ş. & Karaşan, A. (2020). Integration of fuzzy AHP with other fuzzy multicriteria methods: a state of the art survey. *J Mult Log Soft Comput, 35*.
[11] Ali, A., Ullah, K. & Hussain, A. (2023). An approach to multi-attribute decision-making based on intuitionistic fuzzy soft information and Aczel-Alsina operational laws. *J Decis Anal Intell Comput, 3*, 80-89. https://doi.org/10.31181/jdaic10006062023a
[12] Pedrycz, W. (2021). *An Introduction to Computing with Fuzzy Sets Analysis, Design, and Applications*. Cham: Springer Ltd. https://doi.org/10.1007/978-3-030-52800-3
[13] Mohammadzadeh, A., Sabzalian, M. H., Zhang, C., Castillo, O., Sakthivel, R. & El-Sousy, F. F. M. (2023). *Modern Adaptive Fuzzy Control Systems*. vol. 421. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-17393-6
[14] Veerraju, N. & Prasannam, V. L. (2021). Solving Fuzzy Linear Programming Problem Using Defuzzification Method. *Commun Math Appl, 12*, 335.
[15] Abu Bakar, Z. & Kamaruzzaman, S. N. (2023). Facility Manager's Acceptance of CMMS in Malaysia: an Exploratory Study Using PLS-SEM Approach. *J Surv Constr Prop, 14*, 83-93. https://doi.org/10.22452/jscp.vol14no1.7
[16] Mahabir R, Punb KF. Improving CMMS-integrated Equipment Reliability in Compliance with the IS0 41001: 2018 Standard: A Case Study n.d.
[17] Meira, D., Lopes, I. & Pires, C. (2020). Selection of computerized maintenance management systems to meet organizations' needs using AHP. *Procedia Manuf, 51*, 1573-1580. https://doi.org/10.1016/j.promfg.2020.10.219
[18] Hamodi, H. & Aljumaili, M. (2017). Data Quality of Maintenance Data: A Case Study in MAXIMO CMMS. *Maint. Perform. Meas. Manag. (MPMM 2016)*. Novemb. 28, Luleå, Sweden, Luleå tekniska universitet, 105-110.
[19] IBM Maximo Application Suite n.d. https://www.ibm.com/products/maximo (Accessed on March 2, 2024)
[20] System CMMS dla działów utrzymania ruchu n.d. https://qrmaint.pl/system-cmms/ (Accessed on March 2, 2024)
[21] System CMMS do utrzymania ruchu n.d. https://www.astor.com.pl/oferta/oprogramowanie-przemyslowe/systemy-cmms.html (Accessed on March 2, 2024)
[22] CMMS Maszyna SMART - centrum pomocy i dokumentacji n.d. https://www.neuron.com.pl/cmms-help/r1.html (Accessed on March 2, 2024)

Authors' contacts:

Michał Pająk, PhD D.Sc. Eng. Associate Professor
(Corresponding author)
Faculty of Mechanical Engineering, Casimir Pulaski Radom University,
ul. Stasieckiego 54, 26-600 Radom, Poland
+48483617149, m.pajak@uthrad.pl

Bogdan Landowski, PhD Eng. Assistant Professor
Faculty of Mechanical Engineering, University of Technology and Life Sciences,
al. Prof. S. Kaliskiego 7, 85-789 Bydgoszcz, Poland
+48523408211, lbogdan@pbs.edu.pl

Łukasz Muślewski, PhD D.Sc. Eng. Associate Professor
Faculty of Mechanical Engineering, University of Technology and Life Sciences,
al. Prof. S. Kaliskiego 7, 85-789 Bydgoszcz, Poland
+48523408298, lukasz.muslewski@pbs.edu.pl

Dragutin Lisjak, PhD Eng. Full Professor
Faculty of Mechanical Engineering and Naval Architecture,
University of Zagreb,
Ivana Lučića 5, 10000 Zagreb, Croatia
+ 385 6168 377, dragutin.lisjak@fsb.hr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240517165802

Preliminary communication
Received: 2024-05-17, Accepted: 2024-06-22

Integrating Robotic Systems into a Plasma Cutting Workstation - New Workstation Design Approach Using Techno-Economic Evaluation

Jakub Müller*, Tomáš Broum, Miroslav Malaga, Monika Milatová

Abstract: This paper proposes a procedure for implementing a robotic system in a plasma cutting workplace. The goal is to create an efficient and safe process for integrating a robot into the existing workspace, with an emphasis on increasing productivity and minimising the risk of human factors. The paper includes an analysis of the requirements for the robotic system, including compatibility with the plasma cutting equipment. The proposed procedure involves an analysis of the current state of the workplace, followed by recommended implementation steps. The procedure includes a comparison of the current state of the workplace with the proposed state and the establishment of criteria for evaluation, which are subsequently used in the techno-economic evaluation. The paper also presents an industrial case study to validate the proposed procedure.

Keywords: automation; industrial case study; plasma cutting; robotics; techno-economic evaluation

1 INTRODUCTION

The integration of robotic systems into industrial processes is a key step towards improving the competitiveness and efficiency of the production environment. When deciding on such an investment, a thorough economic and technical evaluation is essential. This paper focuses on just such an evaluation in the context of a robotic plasma cutting workstation. The research analyses not only the technical aspects, such as the reduction of the need for operators and quality control using laser technology, but also the economic implications, including savings and payback time of the investment, which according to the research conducted is often overlooked and not addressed in the available publications. Our findings suggest that automation through robotic systems is not only a suitable solution for existing manufacturing processes, but also a key to future innovative growth and competitiveness.

2 LITERATURE REVIEW

In today's industrial environment, there is a strong emphasis on innovative approaches and technologies to optimise production processes and increase their efficiency. Two of the key tools in this endeavour are robotic systems and simulations, which play an indispensable role in the design and improvement of production lines and processes [1].

Some analyses are devoted to the design of automated assembly stations to improve working conditions and efficiency, focusing on the integration of palletising stations with industrial robots and cobots to increase competitiveness [2], or they explore the use of multi-robot cells for production lines [2] e.g., with an extension to optimise reconfiguration capacities [3].

The studies also reflect the economic aspects of robotization of production processes [4] and at the same time deal with the practical problems associated with the implementation of Industry 4.0, especially in small and medium-sized enterprises [5]. In the context of the use of industrial robots, their impact on the qualitative development of the manufacturing industry [6] and their impact on employment and labour costs [7] are investigated. The contribution of industrial robots to labour productivity growth and economic convergence is also an important issue [8]. The issue of robot/cobot grippers is also very often discussed [9, 10].

Furthermore, specific aspects of the implementation of robotic systems and the optimisation of their number for the reconfiguration of modular manufacturing systems [11], or e.g. the modelling and evaluation of the use of workers in production [12] are investigated. Other authors focus on specific methods and tools for optimising manufacturing processes [13, 14].

The use of these innovative approaches can then be presented in scientific articles focusing on practical examples of the use of simulations for the optimisation of manufacturing processes often provided by robots [14, 15] [16, 17]. What is missing here is the economic aspect and directly targeting the area of plasma cutting.

In general, the authors do not deal with a more comprehensive techno-economic evaluation of the implementation/acquisition of industrial robots with a focus on plasma, with the exception of [18, 19], who touch lightly on these aspects.

3 METHOD

The proposed workflow describing the complete design and evaluation of an automated plasma workstation is shown in Fig. 1. A detailed description of the individual steps is given in the following subsections.

3.1 Analysis of the Current State

The first step of the workplace design is to analyse the current state of the workplace. In most cases, a comparison of the current state with the proposed state is made. But even in the case of a completely new workplace, it is advantageous to identify the criteria against which it will be evaluated. The criteria are then used for the later techno-economic

evaluation. They are determined based on the findings and results of the analysis of the current state, as well as the identified deficiencies at the workplace. The established criteria are divided into measurable criteria, the expression of which is explicit, and the remaining criteria which are non-measurable and thus need a qualitative assessment. They are also divided according to their nature into maximising, where the aim is to achieve the highest possible value of the criterion to ensure a good result, and minimising, where the aim is to achieve the lowest possible value to obtain the best result. To accurately analyse the movement of operators over time, it is necessary to create a Spaghetti diagram, which is a tool used to reduce waste in the form of unnecessary transportation, movement and waiting or downtime. This tool is used to track the movement of materials and people, where a visual representation of the physical movement on paper occurs along with details of direction and distance over a predetermined period of time. It is a simple method of analysis used to capture the progress of the work. Everything is recorded in a pencil drawing of the workplace layout. Different colours are used to differentiate between persons or materials. [20]

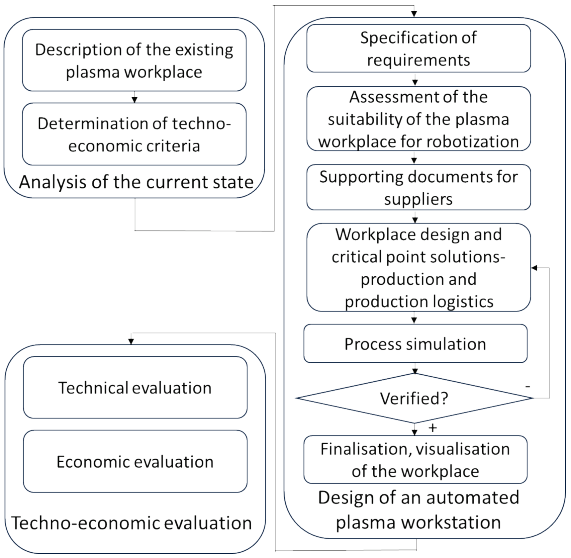


Figure 1 Complete design and evaluation of an automated plasma workstation

3.2 Design of an Automated Plasma Workstation

In this chapter, we outline the steps involved in the design of a new automated workplace for plasma cutting using robotic equipment. The design should be based on the requirements outlined by the company management and takes into account the analysis of the current state and the identified shortcomings of the existing workplace.

3.2.1 Specification of Requirements

An important part of the design process is the specification of the requirements. The basis of the entire design is the creation of a 2D spatial arrangement of the workstation, commonly known as a layout, detailing the placement of various components and elements. These elements include handling robots, plasma console, input and output material containers, scrap tray and other essential items that are key to ensuring the functionality of the workstation. This design should then be complemented by a 3D visualisation of the workplace and robots. It is important to have the information supported by data so that the economic evaluation and payback period calculation can be checked at the end of the design.

3.2.2 Assessment of the Suitability of the Plasma Workplace for Robotization

When deciding whether a plasma workplace is suitable for the introduction of robotics, consideration should be given to an analysis of the current situation and the individual activities performed by the operators at the workplace. In the current situation, it is assumed that all these activities are performed by humans, but the aim is to replace most of them with robots. It is assumed that some of the activities that are performed by humans will be performed in a different way in the new workplace, even though their purpose and output will be the same. The aim is that these activities should take no more than the same amount of time as it takes a human to do them, so that the introduction of robotics is beneficial. If there is complex handling of parts and complicated insertion of small parts into jigs, it would be worth considering whether such activities are strictly necessary to replace by robots. Sometimes it is possible that the complexity of the part insertion makes human work faster than robot work.

This decision is usually determined by consulting with experts from the companies supplying the robots and also from information and knowledge gained in the field of automation and robotics of production and manufacturing processes.

3.2.3 Supporting Documents for Suppliers

An important step is to have the right documentation for the supplier. The first very important parameter to determine is the average volume. The average volume is the ratio of the total volume per year to the number of working days per year (Eq. (1)).

$$AV = \frac{Total\ volume\ per\ year}{Number\ of\ working\ days\ per\ year}. \quad (1)$$

Another important calculation is the available time, where the length of one shift is determined (all calculations should be divided into hours and seconds). Next, the number of shifts per day is determined and lastly the number of hours and seconds per year is determined. The cycle time (Eq. (2)) required to produce one piece is obtained as the ratio of the available seconds per day and the average number of pieces per day shown in Eq. (1).

$$Cycle\ Time\ (CT) = \frac{Seconds\ per\ day\ (T_D)}{AV}. \quad (2)$$

This time must not be exceeded, so there is no room for any downtime or errors, but this is unrealistic in practice as the required number of units would not be produced in a given period. Hence, this cycle time is considered as the time at *OEE* - Overall Equipment Efficiency, Eq. (3). Therefore, the time needs to be recalculated to find the length of time at *OEE* = 100%. For this, it is sufficient to multiply the calculated CT by the *OEE* value shown in Eq. (4).

$$OEE = Availability * Performance * Quality, \quad (3)$$

$$CT_{100\%} = CT * OEE. \quad (4)$$

3.2.4 Workplace Design and Critical Point Solutions

When designing a new plasma workstation, there are almost always important points to consider and the best possible option to choose from.

Number of robots. In terms of full automation and robotization of the workplace, it is necessary to consider the number of robots.

Types of robots. In most cases, the supplier proposes the use of robots based on experience and expertise.

Workplace safety. This depends on the types of robots. In the case of collaborative robots (cobots), no fencing is necessary as they are equipped with a sensor for human contact. In the case of industrial robots, security of the workplace is required in the form of fencing.

Robot tool for gripping parts. Choosing the right gripping tool is a very important part and there are many factors that go into choosing the best tool.

Checking for the presence of holes. Once the hole cutting process is complete, it is important to verify that the holes are indeed there. It is possible that the cycle has been performed but incompletely, or that the cut holes have a different shape than desired. Another factor is whether or not the workplace is close to other workplaces, since because of the heat generated by the plasma torch and the cutting of the holes, it is necessary to secure the workplace with a welding transparent red fence to protect other workers. It works on the same principle as the glass that welders have in their welding helmets to protect their eyesight.

3.3 Techno-Economic Evaluation

In this chapter, the focus will be on the techno-economic evaluation of the design of the workplace from several perspectives. First, the technical aspects of the workplace will be evaluated based on technical criteria. For this purpose, a multi-criteria evaluation method will be used with the determination of weights for each criterion. This will be followed by an economic evaluation, where the payback period (also called return on investment) will be determined.

3.3.1 Technical Evaluation

For the comparison of the new workplace with the current one, the technical criteria for the evaluation are established in the next chapter of the paper as part of the analysis of the current state. These criteria are listed in descending order of importance, with the most important criterion at the top. The ranking of the criteria is part of the pairwise comparison method used to determine the weights of each criterion. This method works on the principle of listing the criteria in the order in the table in both column and row so that each can be compared with the other. If a criterion is more important in the row than in the column, a number one is written in the corresponding cell, otherwise zero. The number of u_{er} is then calculated by simply summing the values in each row. This number tells how many times the criterion is more important than the others. [21] The resulting weight of each criterion p_r is determined according to the Eq. (5).

$$p_r = \frac{\sum_{e=1}^{q} u_{er}}{\sum_{r=1}^{s}\sum_{e=1}^{q} u_{er}}, \quad (5)$$

where: p_r – weight of importance of the criterion, u_{er} – number of criteria preferences, q – number of experts, s – number of criterion.

This is followed by a comparison of the variants according to the individual technical criteria, where the calculation and evaluation of the variants according to the technical criteria is then shown (Tab. 5). In the rank function method, the highest value, in this case 3, was assigned to the best variant, the remaining variants were then ranked 2 and 1. The final rating is calculated according to the Eq. (6).

$$w_t = \sum_{K=1}^{n} p_r * g_r, \quad (6)$$

where: w_t – ordinal function value, g_r – value of the assigned ordinal function.

3.3.2 Economic Evaluation

When deciding on a manufacturing investment, it is important to make investment calculations. These calculations use objective criteria that are determined on the basis of the company's objectives. Various methods are used for the calculations, which include the calculation of the payback period. A static method is used to determine this, which works with average annual values, and therefore average annual savings. It uses static values that the company itself uses. The criterion itself is a minimisation criterion and the aim is to achieve the shortest payback period. A general formula is used for this purpose (7).

$$Payback\ period\ \ (PP) = \frac{Investment}{Cost\ saving}. \quad (7)$$

4 INDUSTRIAL CASE STUDY

In the introduction of the Industrial Case chapter, following on from the previous Method chapter, we explore

the practical application of the methodology in an industrial setting and analyse a specific case study that illustrates the effective implementation of the proposed procedures.

4.1 Analysis of the Current State

As there will be a comparison between the current situation and the proposed situation, it is necessary to determine the criteria against which the assessment will be made. A total of 8 criteria have been selected:

- K_1 - Number of produced pieces
- K_2 - Number of operators
- K_3 - Length of material flow
- K_4 - Method of disposal of cuttings
- K_5 - Level of automation
- K_6 - Method of checking the presence of holes
- K_7 - Method of supply
- K_8 - Workplace area.

For an accurate analysis, Spaghetti diagrams must be created. In our case study there are diagrams for 1 operator (Fig. 2) and 2 operators.

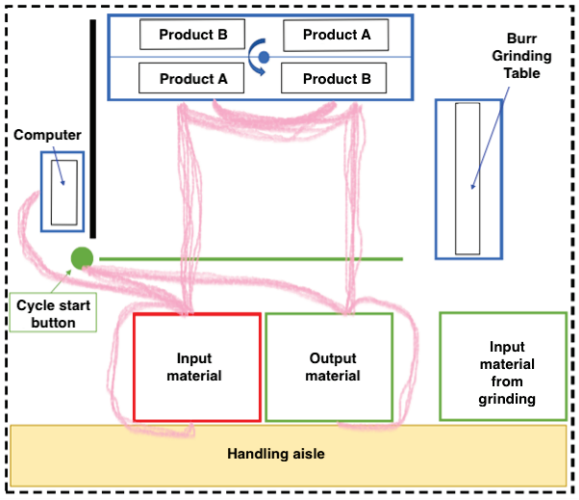


Figure 2 Spaghetti diagram – 1 operator

After their analysis, the routes and distances that operators have to travel during the part placement are visible. After measuring 15 cycles, the average times (cycle times) were calculated (Tab. 1).

Table 1 Operator times (Cycle times)

	One operator	Two operators
Average time (s)	41.99	24.45
Difference (s)	17.54	

The cycle time on a given project at *OEE*=100% should be 19.5 s. However, this value is difficult to achieve in practice. To achieve this value, there would have to be constant, and above all fast, loading of parts into the jigs, no downtime, no waiting and no machine failure. The average cycle time is 41.99 s for one operator and 24.45 s for two operators, a difference of 17.54 s. When the cycle times are converted to *OEE*, it is found that one operator achieves an *OEE* of only 46% and two operators 80%. The management of Shape Corp. decided to create a completely new design for a workplace using modern technologies - specifically robotics, which would eliminate the above-mentioned shortcomings.

4.2 Design of an Automated Plasma Workstation

The assessment of the suitability of the workplace for robotization includes the activities carried out at the plasma workstation and compares the current state with the proposed state. The activities to be assessed are as follows: 1. Opening the containers, 2. Removing the parts, 3. Transferring the part to the jig, 4. Loading the part into the jig, 5. Transferring the part from one jig to another, 6. Removing the finished part, 7. Placing the part in the final container, 8. Closing the containers, 9. Activities 2 to 7 are assumed to be performed by the new robot. The plan view of the new workplace shows the layout of all the items in Fig. 3. ("NOK" stands for "Not OK" = rejected parts.) It plays a key role in optimising the performance, efficiency and safety of the workplace. The right layout can greatly influence workflow, reduce production time and minimise the risk of operational errors.

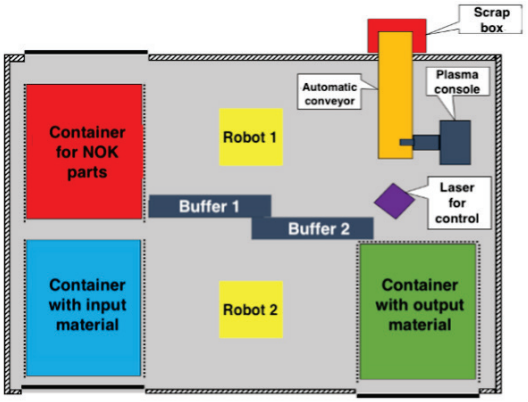


Figure 3 Location of workplaces

The objective was to cut holes at the workplace for two different projects, where each project contains one front and one rear bumper, for a total of four part types with different annual production volumes (Tab. 2).

Table 2 Production volume

Project	Name	Annual production volume (pcs)
P33C	FRT BEAM	250 000
P33C	RR BEAM	250 000
HHM	FRT BEAM	128 000
HHM	RR HS	25 000
Total		653 000

It is therefore necessary to produce a total of 653,000 bumpers in one year, which has 240 working days, which is the standard value considered by Shape Corp., including downtime, which is twice a year. The simple quotient of these two values gives the number of pieces that need to be produced in one day (Eq. (8)).

$$AV = \frac{653\ 000}{240} = 2721\ \frac{\text{pcs}}{\text{day}}. \quad (8)$$

For further calculations, it was agreed with the labour consultant that 2.5 shifts per day would be calculated rather than three, as this deducted time for machine maintenance, tool changes, container changes, emptying the scrap box, etc. The Tab. 3 shows a summary of the available time in hours and seconds required to produce the required number of pieces per shift, per day with 2.5 shifts and per year.

Table 3 Disposable time

	Number of hours (h)	Number of seconds (s)
Shift	7.5	27 000
Day	18.8	67 500
Year	4500	16 200 000

Cycle time for production of 1 piece is calculated afterwards (9). The company considers standard value of OEE 85%, it is considered in Eq. (10).

$$CT = \frac{67\ 500}{2721} = 24.81\ \text{s}, \quad (9)$$

$$CT_{100\%} = 24.8 * 0.85 = 21.09\ \text{s}. \quad (10)$$

In terms of the workplace design and solution, the following points had to be solved in terms of full automation and robotization of the workplace: number of robots, types of robots, types of effectors, removal of the part from the container, positioning of the part in the jig, cutting of holes, placement of the part in the final container, selection of active and passive elements, removal of cuttings, checking the presence of holes and workplace safety. The key points are briefly described below. Based on experience and expertise, the supplier proposed the use of two identical FANUC type M-710iC/50 handling robots due to their specific characteristics and parameters.

Due to the different activities performed by the two robots, it was proposed that each would have a different effector (Tab. 4).

Table 4 Selection of gripper

Robot Tool Concept (Effector Type and Type)		
Robot 1	Magnetic	SCHUNK GSW-M
Robot 2	Pneumatic	SCHUNK PGN+P 200-2

The variant which grips the part from the side was chosen. In this variant, the bumper (Fig. 4) is pulled upwards out of the container. When the part is taken out of the container by the first robot, it has to be moved and positioned in the alignment jig in a precise position, from which the second robot takes it and transports it to the plasma torch. Clamping the part from the side has the following advantages: fast handling, the possibility to remove all parts and no need to open the container door.

For the alignment of the parts in the handover point, the option of data holes or also RPS (reference point system) holes was chosen, which are holes on the back of the part that ensure the precise alignment of parts in jigs for example for welding, riveting and cutting. The build through the data holes is very precise and the emphasis was placed on this.

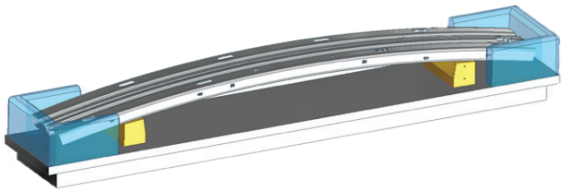


Figure 4 Bumper

To check for the presence of holes, the option of using a laser beam was chosen, which checks for the presence of holes by shining a beam on the spot and having it pass through. This is a fast and high-quality way of evaluating the presence of a hole although relatively more expensive than the other options. Another advantage is the large measurement distance and sufficient accuracy.

5 RESULTS

On the basis of specifications and consultations based on the defined items and processes in the previous chapter, a technical design of the new workplace was created by the supplier ARC-Robotics, s. r. o. Based on this information, the supplier also created a quotation for the delivery of the workstation. Several items were not included in the quotation, these are the robots that the company will use from the existing workplace, as well as the plasma source itself, the burner and the exhaust, which will also use the existing one. The quotation included additional information such as the delivery time of the workstation, which was approximately 18-22 weeks from ordering, defined payment terms, warranty, service coverage and a detailed description of the specification and design of the workstation.

Regarding the 3D design of the workplace, a simplified simulation of the workplace was received. This simulation was mainly used to test the speed of the robots, whether it is sufficient to ensure the process and the reach distances of the robots and thus determine the position of the different elements of the workplace.

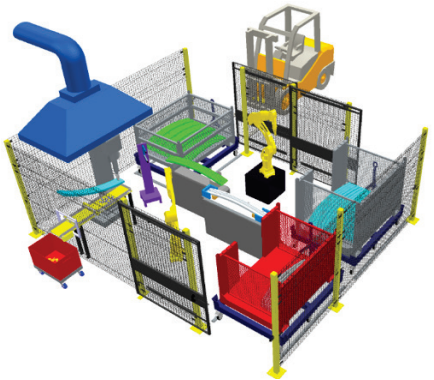


Figure 5 3D workplace design

The resulting detailed workstation design is shown in Fig. 5. The plasma process is carried out in several steps. The

starting point is the supply and placement of one input material container, one empty output container and one red container (NOK parts). The actual process is as follows. The first robot drives its arm to the input material container and grabs a workpiece. It proceeds with it towards the first setting jig, into which it inserts the material. In this exact position, the second robot grabs it using a shape effector to ensure its exact position. It then proceeds to the plasma burner where the holes are cut. The part drops out of the cut-off, which is conveyed by a conveyor to the red box. The holes on the finished part are checked with a laser. The robot inserts the matching part into the second setting jig. Here, the first robot takes over again and places it in the output container.

5.1 Technical Evaluation

The technical criteria set out in the previous chapter were used to compare the new workplace with the current one. For the criteria, their importance was determined in consultation with the responsible experts from the company. Subsequently, the weights of each criterion were determined using the pairwise comparison method. Using the ordinal function method, the degree of fulfilment of the criteria for the three variants was determined by comparing the current WELD003 site with two operators (V1), with one operator (V2) and the new site (V3). A comparison of the variants against each criterion is presented below. A list of all criteria was provided in the analysis of the current situation, and the comparison is shown on one representative criteria.

Workplace area K_8. The current WELD003 occupies an area of 66.8 m^2 with a table for burr grinding included, which is the same for V1 and V2. At the same time, there is a large area where workers only walk when they go to set up a part, otherwise they must not stand there when turning the jig for safety reasons. This area is not needed in the new workplace, so V3 only occupies an area of 30 m^2. The comparison of the variants and the selection of the best one is shown in Tab. 5.

Table 5 Comparison of variants

		V1		V2		V3	
K	p_r	g_r	w_t	g_r	w_t	g_r	w_t
K_1	0.25	3	0.75	1	0.25	2	0.50
K_2	0.21	3	0.64	1	0.21	2	0.43
K_3	0.18	2	0.36	1	0.18	3	0.54
K_4	0.14	1.5	0.21	1.5	0.21	3	0.43
K_5	0.11	1	0.11	2	0.21	3	0.32
K_6	0.07	1.5	0.11	1.5	0.11	3	0.21
K_7	0.04	1	0.04	1	0.04	1	0.04
K_8	0.00	1.5	0.00	1.5	0.00	3	0.00
$\sum w_t$			**2.21**		**1.21**		**2.46**

The pairwise comparison method [21] was used to calculate p_r. It is clear from the results that the best variant according to the technical criteria is V3, which is a new workplace.

5.2 Economic Evaluation

When deciding on a manufacturing investment, it is important to make investment calculations. Various methods are used to make the calculations, including the payback period calculation required by Shape Corp. and it is therefore the focus of this evaluation.

The size of the investment is based on the quotation from the supplier, and this may be added together with other purchased items. The cost savings of the workplace is, according to the company's decision, in the form of salary cost savings as there will be no operators at the new workplace. This item is considered essential. Overheads and material costs are considered to remain the same. The annual saving is therefore calculated as the difference between the total cost of the workplace and the total cost of the workplace excluding wages.

This is followed by a calculation of the payback period of the investment in years (Eq. (11)). Monetary units in the fraction are EUR.

$$PP = \frac{115\ 709,64}{99\ 583,72} = 1.16 \text{ years.} \tag{11}$$

This is a very short payback period, which may be an important factor for the company in deciding to go ahead with it, as it has requested a payback period of up to two years.

6 DISCUSSION

Focusing on the results achieved, it is first necessary to stress the need to finalise the design of the new workplace. This requires a specific techno-economic evaluation and comparison with the existing situation. In the technical evaluation, it was found that the number of units produced for the new workplace is almost comparable to the better original version. The reduction in the number of operators is the key factor, which in turn has a key impact on cost savings. Another key aspect is the minimisation of claims by means of laser inspection for the presence of cut-out holes, which is also fast. It is also worth mentioning the reduction of the workplace area, which can be used in other ways. The economic evaluation comes out well below two years. In summary, automation can be clearly recommended.

Based on previous research and analysis in robotic systems and their integration into industrial processes, our paper has moved towards greater specificity and innovation by focusing on the integration of robotic systems into a steel cutting workstation. Our innovative integration of robots into steel cutting workstations is shown in [1] in the design of robotic assembly stations and [22, 3] in the application of industrial and collaborative robots. The study also emphasises the selection of grippers, a topic that is often discussed in robot selection [10, 11]. With this approach, we extend the theoretical and practical knowledge in robotics and present new opportunities for improving competitiveness and innovation in manufacturing processes.

In terms of the economic impact of robotization, [3, 5, 6] provide examples of the economic aspects of robotizing manufacturing processes, while [18] and [20] focus on the factors affecting the cost side of implementing industrial

robots in the context of Industry 4.0. These data can support our claims about the possibilities of reducing costs and increasing efficiency in specific manufacturing sectors through the integration of robotic systems.

7 CONCLUSION

This paper presents a comprehensive approach to the design and evaluation of an automated plasma workstation. The steps to follow are detailed, including verification with a case study. The case study concludes that the integration of robotic systems into a plasma cutting workstation provides significant technical and economic benefits. The technical evaluation showed comparable production to existing methods, with the key benefits being a reduction in the need for operators and minimisation of errors due to laser control. The economic evaluation confirmed significant savings, making automation highly recommended for improving competitiveness and efficiency. To achieve these benefits, it is important to follow a systematic approach and evaluate the results afterwards, which this paper attempts to present.

Acknowledgments

This paper was created with the subsidy of the project SGS-2023-025 'Environmentally sustainable production' carried out with the support of the Internal Grant Agency of the University of West Bohemia.

8 REFERENCES

[1] Daneshjo, N., Sabadka, D., Malega, P., Dzuro, M. & Jankovič, M. (2022). Creation of More Efficient Work Environment through the New Design of the Automatic Robotic Assembly Station. *Adv. Sci. Technol. Res. J., 16*, 74-84. https://doi.org/10.12913/22998624/151547

[2] Baláž, V., Vagaš, M., Semjon, J. & Zubrzycki, J. (2014). Proposal of Multirobotic System with Two Robots. *AMM, 613*, 243-247. https://doi.org/10.4028/www.scientific.net/AMM.613.243

[3] Ulewicz, R. & Mazur, M. (2019). Economic Aspects of Robotization of Production Processes by Example of a Car Semi-trailers Manufacturer. *Manufacturing Technology, 19*, 1054-1059. https://doi.org/10.21062/ujep/417.2019/a/1213-2489/MT/19/6/1054

[4] Ingaldi, M. & Ulewicz, R. (2019). Problems with the Implementation of Industry 4.0 in Enterprises from the SME Sector. *Sustainability, 12*, 217. https://doi.org/10.3390/su12010217

[5] Guo, Q. & Su, Z. (2023). The Application of Industrial Robot and the High-Quality Development of Manufacturing Industry: From a Sustainability Perspective. *Sustainability, 15*(16), 12621. https://doi.org/10.3390/su151612621

[6] Jung, J. H. & Lim, D.-G. (2020). Industrial robots, employment growth, and labor cost: A simultaneous equation analysis. *Technological Forecasting and Social Change, 159*, 120202. https://doi.org/10.1016/j.techfore.2020.120202

[7] Eder, A., Koller, W. & Mahlberg, B. (2024). The contribution of industrial robots to labor productivity growth and economic convergence: a production frontier approach. *J Prod Anal., 61*, 157-181. https://doi.org/10.1007/s11123-023-00707-x

[8] Marschall, M., Gregor, M., Ďurica, L., Vavrík, V., Bielik, T., Grznár, P. & Mozol, Š. (2022). Defining the Number of Mobile Robotic Systems Needed for Reconfiguration of Modular Manufacturing Systems via Simulation. *Machines, 10*, 316. https://doi.org/10.3390/machines10050316

[9] Pollák, M. & Dobránsky, J. (2020). Structural Design and Material Cutting Using a Laser End Effector on a Robot Arm. TEM Journal, 9(4), 1455-1459. https://doi.org/10.18421/TEM94-17

[10] Xie, G., Holladay, R., Chin, L. & Rus, D. (2024). In-Hand Manipulation with a Simple Belted Parallel-Jaw Gripper. *IEEE Robot. Autom. Lett., 9*, 1334-1341. https://doi.org/10.1109/LRA.2023.3346750

[11] Krajčovič, M., Furmannová, B., Grznár, P., Furmann, R., Plinta, D., Svitek, R., Antoniuk, I. (2021). System of Parametric Modelling and Assessing the Production Staff Utilisation as a Basis for Aggregate Production Planning. *Applied Sciences, 11*, 9347. https://doi.org/10.3390/app11199347

[12] Vavrík, V., Gregor, M., Marschall, M., Grznár, P. & Mozol, Š. (2019). The design of manufacturing line configurations with multiagent logistics system. *Transportation Research Procedia, 40*, 1224-1230. https://doi.org/10.1016/j.trpro.2019.07.170

[13] Kováč, J., Malega, P., Rudy, V., Svetlík, J. (2023). Vumark's Method of Production Layout Designing. *Applied Sciences, 13*, 1496. https://doi.org/10.3390/app13031496

[14] Daneshjo, N., Mareš, A., Malega, P., Chlpek, S. & Baňas, T. (2023). Creating a Simulation of Assembly of a Selected Component in a Virtual Environment. *TEM Journal, 12*(1), 558-565. https://doi.org/10.18421/TEM121-66

[15] Daneshjo, N., Malega, P., Hlubeňová, J. & Štuller, P. (2022). Implementation of Simulation in the Design of Robotic Production Systems. TEM Journal, 11(1), 179-188. https://doi.org/10.18421/TEM111-22

[16] Daneshjo, N., Mareš, A., Malega, P. & Župčan, V. (2023). Proposal of Workplace Modification in the Assembly Line in Automotive Production. *Adv. Sci. Technol. Res. J., 17*, 88-100. https://doi.org/10.12913/22998624/163423

[17] Pollák, M. & Goryl, K. (2023). Simulation Design and Measurement of Welding Robot Repeatability Utilizing the Contact Measurement Method. *Machines, 11*, 734. https://doi.org/10.3390/machines11070734

[18] Zenisek, D. & Broum, T. (2020). Factors Affecting the Production Cost Function of Assembly Lines in Relation to the Level of Automation in Context of Industry 4.0. In: Soliman, K. (ed.) University of West Bohemia Pilsen, 18229-18239.

[19] Zenisek, D., Broum, T. & Simon, M. (2023). Payback Calculation Refinement of Industrial Robot Applications. *MM SJ*, 2023. https://doi.org/10.17973/MMSJ.2023_10_2023077

[20] LEAN SYSTEMS: applications and case studies in manufacturing, service. ROUTLEDGE, S.l. (2021)

[21] Kułakowski, K. (2020). *Understanding the analytic hierarchy process*. CRC Press, Boca Raton, FL https://doi.org/10.1201/9781315392226

[22] Kováč, J., Jenčík, R., Andrejko, P., Hajduk, M., Pilat, Z., Tomči, P., Varga, J. & Bezák, M. (2020). Integrated Palletizing Workstation with an Industrial Robot and a Cobot. In: Berns, K. & Görges, D. (eds.) *Advances in Service and Industrial Robotics*. Springer International Publishing, Cham, 202-209. https://doi.org/10.1007/978-3-030-19648-6_24

Authors' contacts:

Ing. **Jakub Müller**
(Corresponding author)
Department of Industrial Engineering and Management,
University of West Bohemia,
Univerzitní 22, 301 00 Pilsen, Czech Republic, Europe
+420 377 638 456, mullerja@fst.zcu.cz

Grupo Integrado de Ingeniería, Universidade da Coruña,
Campus de Esteiro, 15403, Ferrol, Spain, Europe
jakub.muller@udc.es

Ing. **Tomáš Broum**, PhD
Department of Industrial Engineering and Management,
University of West Bohemia,
Univerzitní 22, 301 00 Pilsen, Czech Republic, Europe
+420 377 638 431, broum@fst.zcu.cz

Ing. **Miroslav Malaga**, PhD
Department of Industrial Engineering and Management,
University of West Bohemia,
Univerzitní 22, 301 00 Pilsen, Czech Republic, Europe
+420 377 638 457, malaga@fst.zcu.cz

Ing. **Monika Milatová**
Shape Corp.,
Havířská 1388, 330 23, Nýřany, Czech Republic, Europe
+420 770 114 330, milatovam@shapecorp.com

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240522154747

Preliminary communication
Received: 2024-05-22, Accepted: 2024-06-28

The Mediating Role of Supply Chain Integration in the Relationship between TQM and Innovation Performance

Ehsan Masoudi*, Neda Rajabani, Arash Shahin

Abstract: The purpose of this study is to provide a theoretical framework that studies the mediating role of supply chain integration (SCI) in the relationship between total quality management (TQM) and innovation performance (IP). This Practical research was done by using the descriptive-correlation method. The statistical population of the study consisted of managers of SMEs in Golpayegan industrial town in Iran. The sampling method was simple random, and 137 companies were selected to determine the sample size by G*Power software. The validity of the questionnaires was confirmed using content and construct validity, and Cronbach's alpha was used to assess their reliability. The data was analyzed by structural equation modelling in the SMART_PLS software. The research findings showed that TQM has a significant effect on IP and SCI. The impact of SCI on IP is also significant and positive. In addition, the mediating role of SCI in the relationship between TQM and IP was confirmed.

Keywords: Innovation performance; SMEs; supply chain integration; Supply Chain Management; Total Quality Management

1 INTRODUCTION

The survival of companies in manufacturing organizations has been threatened due to the increase in competition both at the local and global levels. For this reason, these organizations have turned to improving innovation performance to achieve competitive advantage and competition [1]. In general, economic growth is related to innovation, and one of the key results of innovation is innovation performance [2]. Among manufacturing companies, small and medium-sized companies play an important role in creating innovation and economic progress, especially in developing countries [3]. SMEs are committed to taking risks to be the first to introduce new products, services, and operational technologies, which indicates that these companies are highly inclined to implement innovative strategies [4]. However, few SMEs embed good innovation processes in their companies, which can lead to the loss of their competitive advantage. Accordingly, SMEs should maintain innovation by keeping their employees' creativity active and identifying opportunities [5].

Organizations use different methods to improve innovation performance in providing new products, processes, and services, which can be referred to as total quality management and supply chain integration. Companies that implement total quality management and continue to improve product quality can improve their competitive position and business success and differentiate their products [6]. Quality management (QM) improves teamwork among employees of organizations and causes employees to provide innovative ideas to improve products, services, and processes. This has finally made organizational structures flexible, which is one of the vital factors in creating innovation [7]. The key to business success is recognizing the importance of innovation and quality. Companies should pursue both innovation and quality in the implementation phase. Accordingly, both TQM and innovation appear to be critical to the success of organizations [8].

In addition, in today's chaotic business environment, manufacturing companies are trying to improve their performance by strengthening their supply chain management (SCM) processes. For this reason, the concept of supply chain integration (SCI) was developed to strengthen the SCM of companies. SCI can be defined as a coordinated collaboration between different functions within the organization on the one hand and between the organization itself and its external partners, suppliers, and customers on the other hand, to effectively manage materials, services, information, money, and decisions. Supply chain integration helps companies develop and launch innovative products and services and improves innovation performance in organizations [9]. Also, just as total quality management and supply chain integration help to improve the performance of companies, total quality management can help to integrate and improve the companies' supply chain [10].

Ahinful et al. (2023) [11], Mushtaq et al. (2022) [12], and Kulenović et al. (2022) [6], in their research, concluded that total quality management has an impact on innovation performance. Yani (2022) [13], Tarigan and Kristianto (2019) [14], and Thai and Jie (2018) [10], concluded in their studies that total quality management impact on supply chain integration. Also, according to the research done by Bwaliez (2021) [9], Kumar et al. (2020) [15], and Somjai and Girdwichai (2019) [16], supply chain integration affects innovation performance.

As can be seen in the review of the background of the research, so far no research has been conducted that has studied the mediating role of supply chain integration in the relationship between total quality management and innovation performance, and for this reason, the present research has been conducted.

In the continuation of this paper, the theoretical foundations are first introduced, then the proposed model and its hypotheses are presented, and the research method is explained. The desired model is tested in a practical study and finally, according to the obtained results, a discussion and conclusion are made.

2 LITERATURE REVIEW

2.1 The Influence of TQM on Innovation Performance

Quality management systems such as TQM have a significant impact on the reputation of organizations and the trust of customers by introducing new products and services, and for this reason, quality has become the most important decision-making factor [17]. The impact of TQM on innovation performance is explained by Barney's (1991) resource-based theory, which considers TQM as a capability that allows organizations to achieve innovation performance. Innovation performance is considered an intangible resource of the organization that is impossible to imitate [11]. TQM and innovation have the same importance and goals in improving the performance of organizations and enabling companies to increase competitive advantage by considering customer expectations. The amount of organizational and product development, which are the main goals of innovation, is different from the implementation of TQM [8].

2.2 The Influence of TQM on Supply Chain Integration

Integration is an important principle in SCM literature. SCI includes internal company procedures and external procedures with customers and suppliers. Manufacturers who have properly connected their internal processes to external suppliers and consumers in supply chains are the most successful [18]. Increased performance in companies is achieved through the strategic integration of SCI and TQM principles. TQM acts as a systematic framework for quality management, permeating organizational structures and emphasizing the integration of quality principles into all operational dimensions. Likewise, supply chain integration facilitates coordinated interactions between companies, suppliers, and customers in an integrated system [19]. The implementation of TQM affects the integration of internal cooperation between supply chain departments. Quality management is needed to improve supply chain integration for the proper implementation of SCM inside and outside organizations [13].

2.3 The Influence of SCI on Innovation Performance

From a dynamic capability perspective, a company may be able to modify its SCM capabilities to better align with supply chain objectives. This includes having an enabling SCI that can significantly contribute to innovation performance. For example, customer integration is an important dimension of SCI that improves innovation performance. It is argued that establishing long-term and close relationships with important suppliers reduces opportunistic behaviour and irregularities in transactions, improves product quality, and reduces monitoring, delivery, and performance costs, thereby improving innovation performance [5]. Improving SCI is one way to achieve innovation. SCI is a key driver in improving product innovation. SCI improves organizational functions such as innovation performance through the coordination of activities and information flow between a company and its suppliers and customers [20].

2.4 SCI Mediates the Relationship between TQM and IP

The philosophy of quality management is used in business, industry, and services to ensure maximum efficiency and effectiveness, which can increase stability, increase efficiency, and prevent mistakes in the decision-making process of managers. If the quality management system is implemented properly, it can be effective in improving the management and increasing the effectiveness, and improving the decision-making processes and the performance of the organization. In addition, the findings of studies have shown that TQM has a positive correlation with the innovative performance of companies [21]. One of the important dimensions of TQM is management commitment, which is a prerequisite for achieving an integrated supply chain. In addition, the impact of SCI on the performance of companies has been confirmed. Integration enhances company performance by allowing better coordination and collaboration within the company between different departments. In this regard, achieving high innovation performance is possible if companies establish a strong relationship with foreign partners and then penetrate deep into foreign companies to gain access to valuable information and resources for the company [5]. According to the mentioned reasons, SCI can play a mediating role in the relationship between TQM and innovation performance.

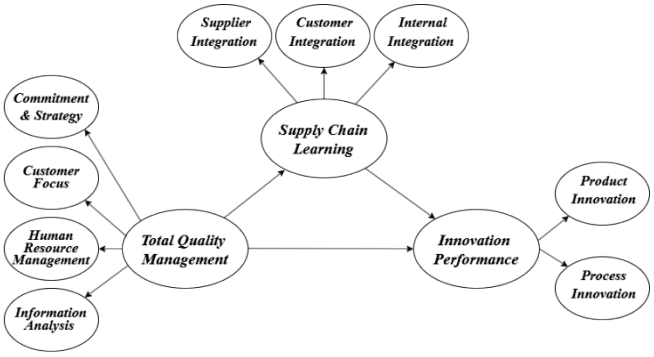


Figure 1 The conceptual model of research

Fig. 1 illustrates the proposed conceptual model. As it is known, the independent variable in this research is TQM, the dependent variable is IP, and the mediating variable is SCI. Accordingly, the hypotheses of the present research are as follows:

H1: TQM impacts innovation performance.
H2: TQM impacts supply chain integration.
H3: SCI impacts innovation performance.
H4: SCI mediates the relationship between TQM and IP.

3 RESEARCH METHODOLOGY

3.1 Measurement Instrument

A questionnaire tool was used to measure the research variables. The questionnaire consisted of two parts. The first part included the demographic information of the respondents and included 4 items (Industry sector, Respondent's work experience, Duration of the company and No. of employees). The second part was the items related to the research variables, which were measured based on a five-

point Likert scale from 1 (strongly disagree) to 5 (strongly agree). The following studies were used to measure the research variables.

TQM: For the total quality management variable, we used the Vanichchinchai and Igel (2011) questionnaire. This questionnaire measures TQM with 17 items and includes 4 dimensions: commitment and strategy (4 items), customer focus (3 items), human resource management, and information analysis (3 items) [22].

SCI: For the supply chain integration variable, we used the Thai and Jie (2018) questionnaire. This questionnaire measures SCI with 12 items and includes 3 dimensions: customer integration (4 items), supplier integration (4 items), and internal integration (4 items) [10].

IP: For the innovation performance variable from the questionnaire of Escrig-Tena et al. (2018), we used. This questionnaire measures IP with 9 items and includes 2 dimensions of product innovation (5 items) and process innovation (4 items) [23].

Also, to check the content validity of the survey, two expert academicians and two senior industry experts reviewed the survey items.

3.2 Sample Size and Data Analysis Approach

The statistical population of this research consists of quality managers of small and medium companies in Golpayegan Industrial Town in Iran. A simple random sampling method was used to collect data in the winter of 2024. To determine the required sample size in PLS-SEM, G*Power software was used to perform power analysis related to model settings [24]. In this research, 137 observations were needed to reach 80% statistical power and detect R^2 values of at least 0.1 with a 5% error probability by G*Power software.

Because the sample size of the study was small (n=137), and by performing the Kolmogorov-Smirnov (K-S) test, it was determined that the distribution of the structures was not normal at the level of $\rho = 0.05$; for this reason, PLS-SEM was used to analyze the data [25]. In this paper, SEM was used because it performed confirmatory factor analysis (CFA) to explore the theoretical and conceptual dimensions of observable and latent scales in complex models. The causal inference appears to be accurate due to the nature of access and correction of observed measurement errors in the analysis [26].

4 RESULTS

These results were obtained in the analysis of demographic variables. In the industrial sector, the Manufacturers of construction materials, with 24.82%, equivalent to 34 companies, were the most frequent, and the Manufacturers of chemical products were the least frequent with 6.57%, equivalent to 12 companies. Regarding the respondent's work experience, less than 10 years with 47.45%, equivalent to 65 people, had the most frequency, and more than 20 years with 16.05%, equivalent to 22 people, had the lowest frequency. Regarding the duration of the company, companies with 10 to 20 years of operation had the highest frequency, with 51.83%, equivalent to 71 companies, and companies with less than 10 years of operation had the lowest frequency, with 21.90%, equivalent to 30 companies. In addition, about the number of employees, companies with fewer than 50 employees were the most with 45.98%, equivalent to 63 companies, and companies with more than 200 employees were the least frequent, with 6.57%, equivalent to 9 companies.

Regarding research variables, according to Chin (2010), the analysis of studies by PLS consists of two stages of evaluation of the external model (measurement) and estimation of the internal model (structural), and the tests of the measurement model are used to check the validity and reliability of the structures [27]. As can be seen in Tab. 1, all factor loadings of the constructs are above 0.7, which indicates the validity of the proposed model [28].

Table 1 Measurement Model

Construct	Label	Factor	*C-α*	*CR*	*AVE*
Commitment and Strategy	CS1	0.815	0.810	0.876	0.640
	CS2	0.870			
	CS3	0.815			
	CS4	0.712			
Customer focus	CF1	0.858	0.752	0.858	0.669
	CF2	0.793			
	CF3	0.802			
Human resource management	HRM1	0.840	0.893	0.916	0.610
	HRM2	0.847			
	HRM3	0.758			
	HRM4	0.755			
	HRM5	0.744			
	HRM6	0.757			
	HRM7	0.756			
Information analysis	IA1	0.835	0.758	0.860	0.673
	IA2	0.819			
	IA3	0.807			
Customer integration	CI1	0.743	0.758	0.846	0.580
	CI2	0.811			
	CI3	0.731			
	CI4	0.757			
Supplier integration	SI1	0.844	0.819	0.881	0.651
	SI2	0.813			
	SI3	0.705			
	SI4	0.857			
Internal integration	II1	0.854	0.833	0.889	0.667
	II2	0.816			
	II3	0.783			
	II4	0.811			
Product Innovation	PrI1	0.783	0.880	0.913	0.677
	PrI2	0.811			
	PrI3	0.857			
	PrI4	0.825			
	PrI5	0.837			
Process Innovation	PsI1	0.843	0.861	0.906	0.706
	PsI2	0.881			
	PsI3	0.866			
	PsI4	0.768			

To check the reliability of the external model, Cronbach's alpha (*C-α*), composite reliability (*CR*), and average variance extracted (*AVE*) were used, which, as can be seen in Tab. 1, *C-α* and *CR* are more than 0.7, and the reliability of the variables has been confirmed. *AVE* was also higher than 0.5 for all constructs, which was confirmed [24].

In addition, the value of *CR* was higher than *AVE*, which indicates the confirmation of composite reliability [29]. Tabs. 2 and 3 show the discriminant validity results. It is shown in Tab. 2 that all the variables are consistent with the criteria proposed by Fornell and Larcker (1981) because all the AVE squares of the variables were higher than the correlation of that variable with other variables [30]. Tab. 3 also shows that the HTMT indices were less than 0.9, and discriminant validity is established [31].

Table 2 Fornell and Larcker coefficients

	CS	*CF*	*CI*	*HRM*	*IA*	*II*	*PsI*	*PtI*	*SI*
CS	**0.800**								
CF	0.580	**0.818**							
CI	0.303	0.351	**0.761**						
HRM	0.517	0.461	0.203	**0.781**					
IA	0.321	0.326	0.134	0.349	**0.820**				
II	0.370	0.446	0.217	0.457	0.642	**0.817**			
PsI	0.257	0.191	0.438	0.248	0.273	0.341	**0.840**		
PtI	0.348	0.308	0.174	0.385	0.539	0.521	0.251	**0.823**	
SI	0.358	0.278	0.532	0.287	0.163	0.329	0.605	0.271	**0.807**

Table 3 HTMT criterion

	CS	*CF*	*CI*	*HRM*	*IA*	*II*	*PsI*	*PtI*	*SI*
CS									
CF	0.737								
CI	0.388	0.464							
HRM	0.602	0.558	0.244						
IA	0.398	0.426	0.176	0.415					
II	0.451	0.563	0.269	0.528	0.820				
PsI	0.300	0.236	0.537	0.274	0.339	0.397			
PtI	0.407	0.378	0.212	0.435	0.660	0.609	0.284		
SI	0.440	0.355	0.669	0.332	0.198	0.398	0.728	0.315	

In the analysis of the structural model of the research, a three-step approach including R^2 value, Q^2 model quality, and the significance of the path coefficient of the structural model was used [32], and its results can be seen in Tabs. 4 and 5 and Figs. 2 and 3.

According to the three values of 0.25, 0.50, and 0.75 (low, medium, and high), the R^2 value for *SCI* and *IP* was between low and medium [33]. According to the three values of 0.02, 0.15, and 0.35 (low, medium, and high), the redundancy index structural model of *SCI* and *IP* was medium [24]. In addition, the commonality index for *SCI* and *IP* was medium to high. Finally, considering the three values of 0.1, 0.25, and 0.36 (low, medium, and high) suggested by Tenenhaus et al. (2005), goodness of fit test (*GOF*) was used to evaluate the overall research model, and the overall value of the research model was high [34].

Table 4 R^2, cross validity redundancy and communality

Variables	R^2	Redundancy index	Communality index
Supply chain integration	0.343	0.113	0.266
Innovation performance	0.478	0.192	0.316

$$GOF = \sqrt{\overline{AVE} \cdot \overline{R}^2} = 0.518$$

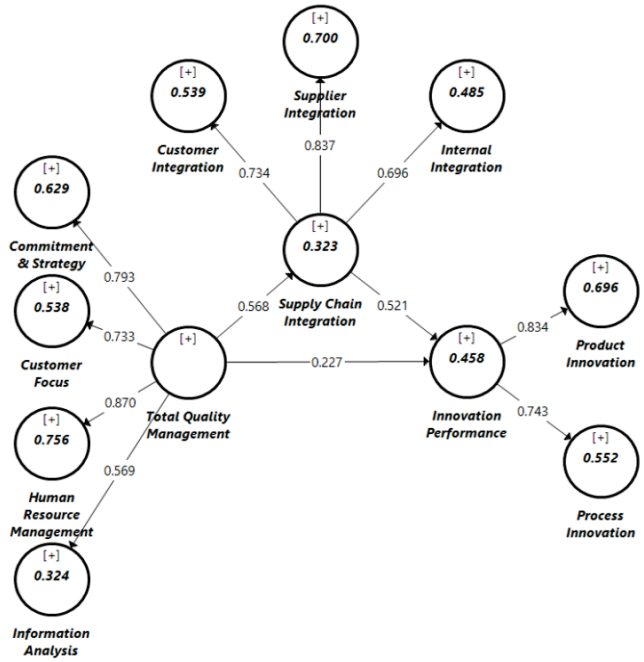


Figure 2 Path coefficient result

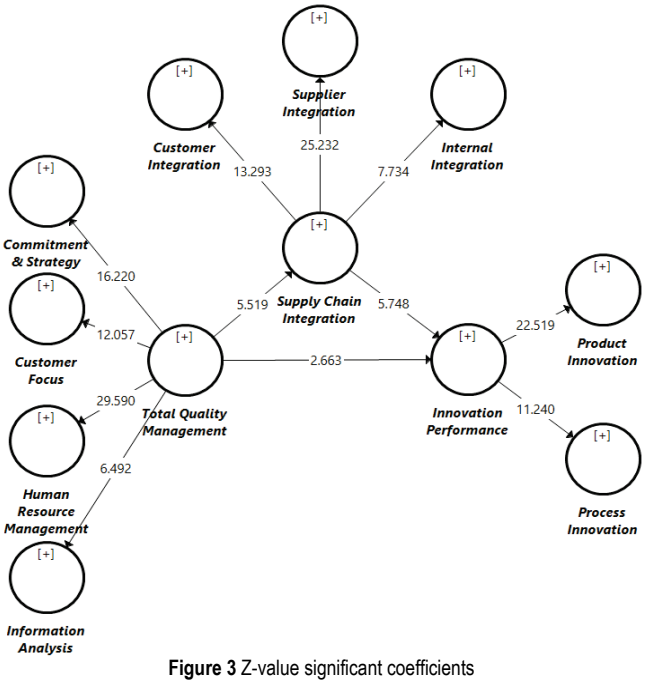


Figure 3 Z-value significant coefficients

Table 5 Hypotheses testing results

Hypothesis	Path coefficient	*SE*	*t*-value	*p*-value	Decision
TQM → *IP*	0.227	0.085	2.663	0.009	Supported
TQM → *SCI*	0.568	0.103	5.519	0.000	Supported
SCI → *IP*	0.521	0.091	5.748	0.000	Supported

Finally, for the structural model, the path coefficients test and the bootstrap technique were used to determine the strength of the relationship between research hypotheses, and as can be seen in Fig. 3 and Tab. 5, TQM has an impact on *IP* and *SCI*. In addition, the results show that *SCI* affects *IP*.

4.1 Mediation Test

Preacher and Hayes (2008) approach and bootstrap method were used to test the mediation effect of the research [35]. This approach is completely suitable for the PLS-SEM method and has been implemented in Smart-PLS software. First, the research model should be implemented without the presence of the mediating variable, i.e., *SCI*. If the effect of the independent variable (TQM) on the dependent variable (*IP*) is significant, the effect of the mediator variable should be analyzed [24].

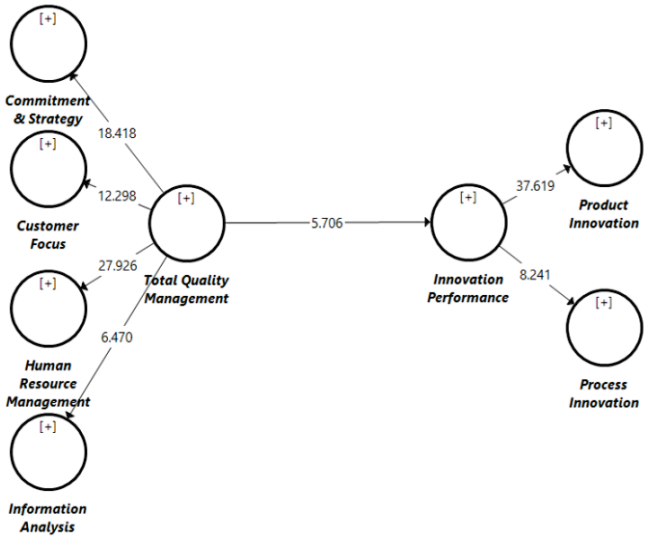


Figure 4 The result of the hypothesis test without the mediator variable

Table 6 The results of the mediator variable test

Hypothesis	Indirect effect	Total effect	*VAF*	Decision
TQM →*SCI*→*IP*	0.296	0.523	0.566	Partial Mediation

As can be seen in Fig. 4, the research hypothesis test was significant without the presence of the mediating variable. Now we have to go to the analysis of the model with the presence of the mediator variable. To measure the effect of the mediating variable in this study, according to Hair et al. (2014), the variance accounted for (*VAF*) was used. Considering that the value of *VAF* was between 0.2 and 0.8, *SCI* plays a partial mediating role in the model [24]. This means that *SCI* accounts for the effect of TQM on innovation performance and mediates it by 56.6%. The results of the mediator variable analysis can be seen in Tab. 6.

5 DISCUSSION

The present research was conducted on SMEs in Golpayegan Industrial Town in Iran. For this purpose, a questionnaire was used to collect data, and the quality managers of these companies were responsive. According to Tab. 5, in the first hypothesis of the research regarding the effect of TQM on *IP*, considering that its *t*-value is 2.663, which is outside the range (–1.96, 1.96), the *p*-value is less than 0.05, and the intensity of the effect 0.227 shows the positive and significant effect of TQM on *IP*. In the second hypothesis of the research about the effect of TQM on *SCI*, considering that its *t*-value is 5.519, which is outside the range (–1.96, 1.96), the *p*-value is less than 0.05, and the intensity of the effect is 0.568. It shows the positive and significant effect of TQM on *SCI*. The third hypothesis of the research is the effect of *SCI* on *IP*, considering that its t-value is 5.784, which is outside the range (–1.96, 1.96) the p-value is less than 0.05, and the intensity of the effect is 0.521. It has a positive and significant effect of *SCI* on *IP*. In addition, according to Tab. 6, it was found that the *SCI* variable plays a mediating role in the relationship between TQM and innovation performance by 0.566.

5.1 Theoretical Implications

The results obtained in this research indicate that for the first time, the mediating role of SCI on the relationship between TQM and innovation performance is studied and considering the importance of SMEs, which has important theoretical implications. The first hypothesis of the research, the effect of TQM on IP, was accepted with an effect rate of 0.227. The results of the research, with the research results of Ahinful et al. (2023), Mushtaq et al. (2022), and Kulenović et al. (2022), are similar. The confirmation of this hypothesis shows that manufacturing companies should support the successful application of TQM practices. TQM contributes to product and process innovation. Innovation and quality are closely related to each other. Hence, TQM helps organizations achieve better innovation performance through commitment to continuous improvement, better decision-making, customer focus, strategic planning, and other quality management practices. This leads to better financial results.

The second hypothesis of the research, the effect of TQM on SCI, was accepted with an effect rate of 0.568. The results of the research, with the results of Yani (2022), Tarigan and Kristianto (2019), and Thai and Jie (2018) are similar. The confirmation of this hypothesis is mainly due to the satisfaction of employees as the key people of organizations. The implementation of TQM in manufacturing companies can facilitate better communication between departments. Management can create good cooperation between the employees of manufacturing companies, so it can create supply chain integration.

The third research hypothesis, the effect of SCI on IP, was accepted with an effect rate of 0.521. The results of the research, with the results of Bwaliez (2021), Kumar et al. (2020), and Somjai and Girdwichai (2019) are similar. Confirmation of this hypothesis shows that integration is needed to improve information processing and ultimately achieve product and process innovation. In addition to using internal integration, companies should use external and customer integration to improve their products and processes. Since knowledge is an important factor in promoting innovation, integration with external partners, especially those in the firm's supply chain, facilitates the flow of knowledge and, therefore, improves innovation performance.

Finally, the fourth hypothesis of the research, namely the mediating role of SCI in the relationship between TQM and *IP*, was accepted with an effect size of 0.566. Companies that

have chosen TQM in their company are looking for new ways to produce and implement internal processes and introduce new and innovative products and services. For this reason, they must always be learning and increasing their knowledge, especially about their supply chain. In this regard, *SCI* can facilitate the connection between TQM and *IP* by focusing on knowledge learning through the integration of internal, external, and customers in the delivery of products and processes.

5.2 Managerial Implications

The researchers' findings provide implications for managers of SMEs. In connection with the first hypothesis of the research and confirming the impact of TQM on *IP*: 1. Paying more attention to TQM, especially the aspects of Information analysis and Customer focus; 2. Holding educational seminars to increase employees' awareness of the benefits of creating innovation in companies, and 3. Helping employees by providing suggestions and accepting criticisms from company managers to improve innovation performance in products and processes.

About the second hypothesis of the research and confirming the effect of TQM on *SCI*: 1. By establishing more connections between industry and academia, from the elite This area should be used to take advantage of their innovative plans and opinions to take an important step towards improving the quality of products, services and supply chain and 2. Increasing the learning of employees at the supply chain level is effective in improving flexible performance and causes an increase in the quality of products, services, and processes in the organization.

In connection with the third hypothesis and confirming the impact of *SCI* on *IP*: 1. Pay more attention to internal integration than supplier integration dimensions; 2. Employees should be encouraged to learn new skills to improve products and processes, and be aware of any changes 3. Form dedicated learning teams and organize group discussions with suppliers and customers to regularly increase their knowledge to improve the quality of services and products.

In connection with the fourth hypothesis of the research and the confirmation of the hypothesis of the mediating role of *SCI* in the relationship between TQM and *IP*: 1. Choosing managers with risk-taking characteristics to improve quality and create innovation, and 2. More communication with internal and external suppliers and customers to create innovative products and processes.

6 CONCLUSION

The purpose of this article was to analyze the mediating role of supply chain integration in the relationship between total quality management and innovation performance. For this purpose, first, by collecting information in the field of theoretical foundations and literature on the subject, library sources, articles, and scientific databases were used, and finally, by testing the proposed model in SME firms in Golpayegan Industrial Town in Iran, the results were discussed. The findings indicated the effect of TQM on SCI and innovation performance. In addition, *SCI* had an impact on innovation performance. In addition, the results showed that *SCI* has a Partial mediating role in the relationship between TQM and innovation performance. The results showed that *SCI,* as a mediating variable, has a greater effect on the dependent variable than TQM as an independent variable. The dimensions of human resource management and commitment, and strategy from TQM and supplier integration from *SCI* have the greatest impact on innovation performance. In addition, information analysis from TQM and internal integration from *SCI* have the least impact on *IP*.

The research topic is theoretically and quantitatively unique compared to similar studies. On the other hand, according to the field findings, the study of the mediating role of *SCI* in the relationship between TQM and innovation performance was conducted for the first time in the manufacturing industries of the world. The comprehensiveness of the investigated variables and the novelty of the research topic were special advantages of the proposed model. In general, it can be concluded that TQM in SMEs is necessary for the development of *SCI* and innovation performance. Just like a system, TQM is an important input, and *SCI* is a key process, so *IP* is a critical output.

One of the limitations of this research can be mentioned that very few studies have examined the research hypotheses. Also, the research location was limited to SMEs in Iran, and therefore, caution should be exercised in generalizing the findings to other companies. The time of data collection was in the winter of 2024, and therefore, caution should be exercised in generalizing the findings to other times. Researchers are advised to conduct more studies on variables affecting innovation performance in future research. Future researchers are suggested to study the effect of TQM and *SCI* on quality performance in the current model. It is also suggested to investigate the mediating role of supply chain learning and supply chain agility in the relationship between total quality management and innovation performance.

7 REFERENCES

[1] Yusr, M. M. (2016). Innovation capability and its role in enhancing the relationship between TQM practices and innovation performance. *Journal of Open Innovation: Technology, Market, and Complexity*, *2*(1), 1-15. https://doi.org/10.1186/s40852-016-0031-2

[2] Robertson, J., Caruana, A. & Ferreira, C. (2023). Innovation performance: The effect of knowledge-based dynamic capabilities in cross-country innovation ecosystems. *International Business Review*, *32*(2), 101866. https://doi.org/10.1016/j.ibusrev.2021.101866

[3] Masoudi, E. (2021). The impact of total quality management on innovation in small and medium-sized enterprises with the mediating role of organizational learning. *Quarterly Journal of Industrial Technology Development*, *19*(43), 77-92. https://doi.org/10.22034/jtd.2021.244744

[4] Masoudi, E. & Shahin, A. (2021). Structural Relation between Green Entrepreneurial Orientation and Green Innovation by Role of Supply Chain Learning as a Mediator in SMEs. *Journal of Entrepreneurship Development, 14*(3), 521-539. https://doi.org/10.22059/jed.2021.316511.653554

[5] Chen, H., Amoako, T., Quansah, C. E., Danso, S. A. & Jidda, D. J. (2023). Assessment of the impact of management commitment and supply chain integration on SMEs' innovation performance: Moderation role of government support. *Heliyon, 9*(5). https://doi.org/10.1016/j.heliyon.2023.e15914

[6] Kulenović, M., Veselinović, L., Šunje, A. & Cero, E. (2022). Understanding the Mechanism of Influence of TQM Practices on Financial Performance: the Mediating Effect of Innovation Performance. *Zagreb International Review of Economics & Business, 25*(1), 171-198. https://doi.org/ 10.2478/zireb-2022-0010

[7] Zeng, J., Zhang, W., Matsui, Y. & Zhao, X. (2017). The impact of organizational context on hard and soft quality management and innovation performance. *International Journal of Production Economics,* 185, 240-251. https://doi.org/10.1016/j.ijpe.2016.12.031

[8] Kanapathy, K., Bin, C. S., Zailani, S. & Aghapour, A. H. (2017). The impact of soft TQM and hard TQM on innovation performance: the moderating effect of organisational culture. *International Journal of Productivity and Quality Management, 20*(4), 429-461. https://doi.org/10.1504/IJPQM.2017.082831

[9] Bwaliez, O. M. (2021). Supply chain integration and manufacturing firm performance: the mediating role of innovation performance. In Full Paper Proceeding of the International Conference on Business, Economics. *Social Science & Humanities*, *6*(120), 1-15. https://doi.org/10.4236/tel.2019.97151

[10] Thai, V. & Jie, F. (2018). The impact of total quality management and supply chain integration on firm performance of container shipping companies in Singapore. *Asia Pacific Journal of Marketing and Logistics*, *30*(3), 605-626. https://doi.org/10.1108/APJML-09-2017-0202

[11] Ahinful, A. A., Opoku Mensah, A., Koomson, S., Nyarko, F. K. & Nkrumah, E. (2023). A conceptual framework of total quality management on innovation performance in the banking sector. *The TQM Journal, 36*(4), 1193-1211. https://doi.org/10.1108/TQM-11-2022-0334

[12] Mushtaq, N., Akhter, Y. & Nadeem, H. (2022). An Exploratory Empirical Investigation on the Intervening Role of TQM & Big Data Analytics between Industry 4.0 and Firms Innovation Performance. *Journal of Development and Social Sciences, 3*(2), 685-699. https://doi.org/10.47205/jdss.2022(3-II)62

[13] Yani, A. (2022). Total Quality Management and Supply Chain Integration on Firm Performance: A Study of Cosmetic Industry in Jakarta. *European Journal of Science, Innovation and Technology*, *2*(4), 90-99.

[14] Tarigan, Z. J. H. & Kristianto, I. (2019). The impact TQM system on supply chain performance through supply chain integration and employee satisfaction. *International Journal of Business Studies*, *2*(1), 8-17.

[15] Kumar, V., Jabarzadeh, Y., Jeihouni, P. & Garza-Reyes, J. A. (2020). Learning orientation and innovation performance: the mediating role of operations strategy and supply chain integration. *Supply Chain Management: An International Journal*, *25*(4), 457-474. https://doi.org/10.1108/SCM-05-2019-0209

[16] Somjai, S. & Girdwichai, L. (2019). Exploring the nexus among the supply chain integration, supply chain learning and the innovation performance of agribased firms in Indonesia. *Polish Journal of Management Studies*, *20*(2), 491-501. https://doi.org/ 10.17512/pjms.2019.20.2.41

[17] Masoudi, E. & Shahin, A. (2022). The influence of the quality criteria on the quality cost of suppliers in SMEs. *Benchmarking: An International Journal*, *29*(7), 2313-2333. https://doi.org/10.1108/BIJ-05-2021-0238

[18] Khan, M. S. (2024). Investigating the impact of supply chain integration on operational performance with a mediating role of supply chain capabilities of the SME sector in Pakistan. *South Asian Journal of Operations and Logistics*, *3*(2), 198-223. https://doi.org/10.57044/SAJOL.2024.3.2.2438

[19] Panitsettakorn, W., Ongkunaruk, P. & Leingpibul, T. (2023). The present state of the cosmetics supply chain in Thailand and the prospective role of Independent Quality Assurance Verifiers (IQAVs) within the supply chain. *Heliyon*, *9*(10). https://doi.org/10.1016/j.heliyon.2023.e20892

[20] Arshad Ali, A. & Mahmood, A. (2023). How Do Supply Chain Integration and Product Innovation Capability Drive Sustainable Operational Performance? *Sustainability*, *16*(1), 277. https://doi.org/10.3390/su16010277

[21] Mehregan, E., Sanaei, S., Manna, M., Bozorgkhou, H. & Heidari, S. (2023). The role of SCM practices in competitive advantage and firm performance: a mediating role of supply chain innovation and TQM. *Tehnički glasnik*, *17*(4), 516-523. https://doi.org/10.31803/tg-20221223200658

[22] Vanichchinchai, A. & Igel, B. (2011). The impact of total quality management on supply chain management and firm's supply performance. *International Journal of Production Research*, *49*(11), 3405-3424. https://doi.org/10.1080/00207543.2010.492805

[23] Escrig-Tena, A. B., Segarra-Ciprés, M., García-Juan, B. & Beltrán-Martín, I. (2018). The impact of hard and soft quality management and proactive behaviour in determining innovation performance. *International Journal of Production Economics*, *200*, 1-14. https://doi.org/10.1016/j.ijpe.2018.03.011

[24] Hair, J. F., Hult, G. T. M., Ringle, C. M. & Sarstedt, M. (2014). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM),* Sage, Thousand Oaks, CA.

[25] Hair, J. F., Sarstedt, M., Pieper, T. M. & Ringle, C. M. (2012). The use of partial least squares structural equation modelling in strategic management research: a review of past practices and recommendations for future applications. *Long range planning*, *45*(5-6), 320-340. https://doi.org/10.1016/j.lrp.2012.09.008

[26] Byrne, B. M. (2013). *Structural equation modelling with Mplus: Basic concepts, applications, and programming* (1st ed.). Routledge. https://doi.org/10.4324/9780203807644

[27] Chin, W. W. (2010). How to write up and report PLS analyses. *In Handbook of Partial Least Squares*. Springer, Berlin, Heidelberg, 655-690.

[28] Hair, J. F., Anderson, R. E., Tatham, R. L. & Black, W. C. (2010). *Multivariate data analysis (7th ed.)*. New Jersey: Prentice Hall.

[29] Bagozzi, R. P. & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the academy of marketing science*, *16*, 74-94. https://doi.org/10.1007/BF02723327

[30] Fornell, C. & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research*, *18*(1), 39-50. https://doi.org/10.1177/002224378101800104

[31] Henseler, J., Ringle, C. M. & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modelling. *Journal of the academy of marketing science*, *43*, 115-135. https://doi.org/10.1007/s11747-014-0403-8

[32] Aldás, J. (2016). *Modelización Estructural con PLS-SEM: Constructos de Segundo Orden,* ADD Editorial, Madrid.

[33] Hair, J. F., Ringle, C. M. & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice*, *19*(2), 139-152. https://doi.org/10.2753/MTP1069-6679190202
[34] Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M. & Lauro, C. (2005). PLS path modeling. *Computational statistics & data analysis*, *48*(1), 159-205. https://doi.org/10.1016/j.csda.2004.03.005
[35] Preacher, K. J. & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior research methods*, *40*(3), 879-891. https://doi.org/10.3758/BRM.40.3.879

Authors' contacts:

Ehsan Masoudi, MA
(Corresponding author)
Quality Management Research Group, University of Isfahan,
Azadi Square, Isfahan, Iran
E-mail: Ehsanma74@yahoo.com

Neda Rajabani, PhD
PhD in Industrial Management, Operations Research, University of Tehran,
16th Azar St, Enghelab Square, Tehran, Iran
E-mail: neda_rajabani@ut.ac.ir

Arash Shahin, Prof.
Department of Management, University of Isfahan,
Azadi Square, Isfahan, Iran
E-mail: shahin@ase.ui.ac.ir

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20250415140751

Preliminary communication
Received: 2025-04-15, Accepted: 2025-05-16

Smart Mini Greenhouse for Eco-Friendly Agriculture

Tomislav Šarić, Elizabeta Tedeško, Goran Šimunović, Sara Havrlišan*

Abstract: This paper presents a proposal for the design, development, and validation of a smart mini greenhouse intended for small-scale agricultural production. The main objective is to explore the feasibility of a cost-efficient smart mini greenhouse prototype. This paper briefly analyses the parameters that influence the process of plant cultivation. Parameters such as air temperature, air and soil humidity, solar radiation and amount of carbon dioxide are analysed. Based on the analysis of the successful process of plant cultivation, sensors are selected for the condition monitoring in the mini greenhouse. By selecting various sensors, actuators and the control unit (Arduino), a system for condition monitoring and dosing of the necessary elements in the process of plant growth (agricultural production) is formed. With the help of computer design tools, a proposal for the construction of a mini-greenhouse is made. The proposed components for growth support and condition monitoring are integrated into the defined construction of the mini-greenhouse. Testing is carried out by designing the control programme and implementing it. By building a mini greenhouse and integrating the selected components (sensors, actuators and other elements) with the control programme, the complete smart mini greenhouse was validated. The validation positively confirmed the proposed and built prototype model of a smart mini greenhouse for production.

Keywords: control monitoring; design; smart mini greenhouse

1 INTRODUCTION

Gardening, the cultivation of plants, is one of the most popular hobbies among individuals living in urban environments. Gardening has a calming and stress-relieving effect and is a healthy and effective way to spend free time. Some citizens want to have fresh produce (vegetables, flowers) all year round and grow it themselves, and there is usually not enough space for gardening in urban areas. The solution may lie in small greenhouses. Citizens in urban areas usually have a lot of work and other activities, so they do not have enough time to monitor and control the condition of the growing plants. The automation of small greenhouses that provide optimal conditions for growing plants can be seen as a solution. The cultivation of plants and flowers requires the setting of cultivation parameters for optimal growth. The integration of various sensors facilitates the monitoring of key parameters essential for plant cultivation. As we live in the era of Industry 4.0, functions such as the Internet of Things, Big Data, the application of artificial intelligence and the like are being used in this area. The Internet of Things, or IoT for short, is a technology that connects electronic devices, sensors and the internet to manage data and applications. The Internet of Things can be used in agriculture for crop management as a medium for monitoring and control, especially in greenhouses, and is referred to as precision farming [1]. One of the key elements of the Automated Urban Greenhouse (AUG) is the development of a network of smart sensors [2]. This network of smart sensors communicates with the AUG control interface and enables automatic control and monitoring of lighting, heating, irrigation and ventilation. The data processing of the sensors as well as the control and monitoring of the processes is done with different microprocessors, one of the most commonly used being the Arduino [3-6]. Various monitoring and control systems are available on the market, which are generally expensive. In the paper [7], the authors compare the systems based on various criteria: Price, light control, soil moisture control, humidity control, temperature control, availability of application and ventilation. One of the reasons for researching and planning a small greenhouse is the cost and an educational approach to developing solutions to problems. Cars, homes, factories are getting smarter, so the aim of this work is to realise a Smart Mini Greenhouse for agricultural production. In order to achieve the set goal, various knowledge in the field of computer-aided design (CAD), monitoring and automatic control, sensors and of course a good knowledge of the process of plant cultivation must be applied. This paper is a presentation of the final thesis in the undergraduate course of mechanical engineering, which was successfully realised in the phase of analysis, design and implementation of the Smart Mini Greenhouse [8].

2 DESIGN OF THE GREENHOUSE AND PARAMETERS OF THE PLANT CULTIVATION PROCESS

A greenhouse is an object that enables the protected and controlled cultivation of plants. The construction consists of walls and a roof, which is usually made of transparent material. The inside of the greenhouse, which is exposed to sunlight, is warmer than the outside temperature. Plants grown in a greenhouse require regulated climatic conditions. A small greenhouse is customised to the type and size of the plants. A small greenhouse is also known as a cold frame. The shape of the greenhouse is shown in Fig. 1, and the flat roof is the most commonly used form of greenhouse [1].

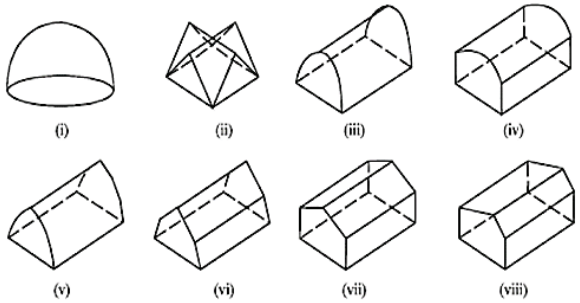


Figure 1 Classification of greenhouses based on the shape of the roof [1]

Various parameters are usually measured in the greenhouse, including heat (air temperature), humidity, soil moisture, solar radiation and the concentration of carbon dioxide (CO_2). Different plants require different conditions for optimum growth. Below you will find a brief description of selected parameters that are important for efficient greenhouse operation.

Heat is expressed by measuring the temperature (°C). The required temperature varies from plant to plant and is one of the key parameters for plant growth. Seed germination takes place at precisely defined temperatures. It is therefore important to maintain the temperature required for each plant species. Temperature can be measured using various thermometers (an example of a thermometer in a greenhouse is shown in Fig. 2) or sensors. Plants consume carbon dioxide (CO_2) and release oxygen (O_2), generating heat. The release of heat can be beneficial for plants that respond to higher temperatures, but for plants that do not require heat, it can lead to stunted growth and disease.

Figure 2 Thermometer [9]

Humidity is important to keep the climate in the greenhouse healthy. Humidity is the amount of water vapour contained in the air and can be measured with a hygrometer, as shown in Fig. 3. When carbon dioxide (CO_2) is processed, the oxygen that is emitted is warm and moist.

Figure 3 Hygrometer [10]

High moisture levels can lead to the growth of mould and various diseases that are harmful to plants. It is important to monitor the humidity in the greenhouse and regulate it regularly.

Soil moisture is also an important parameter for plant cultivation. Water supplies the plants with nutrients. It is important to supply the plants with a sufficient amount of water.

The amount of sunlight is a necessary parameter for plant growth, as plants need it for the photosynthesis process. Some plants require sunlight, others prefer to grow in the shade. In the greenhouse, the amount of light can be controlled with various curtains, blinds and UV protection covers. If the natural light is not sufficient, special lighting (UV light sources) can be installed in the greenhouse.

The measurement of carbon dioxide (CO_2) concentration is often neglected, although carbon dioxide is a critical parameter for plant photosynthesis. A carbon dioxide level that is too low hinders plant growth, but a carbon dioxide level that is too high is also not favourable for the plants.

2.1 Selection of Components for Smart Greenhouse Control

A smart greenhouse significantly simplifies plant care by automating environmental control. For a greenhouse to be smart, it must be able to replace humans in almost all aspects of plant cultivation. A smart greenhouse must be able to control the optimum temperature. It should also be able to recognise whether the plants need to be supplied with water. If the amount of water is insufficient, a message is sent for corrective action.

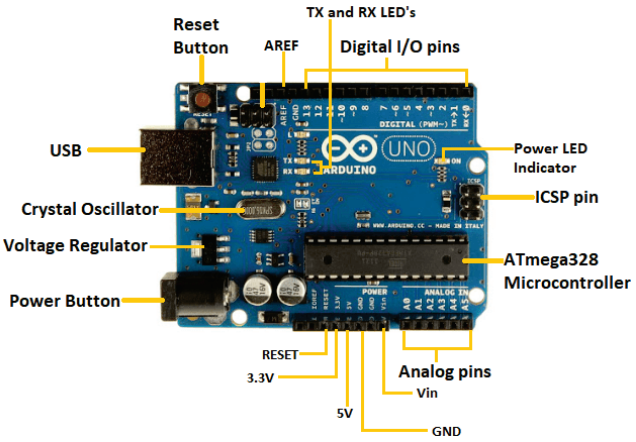


Figure 4 Arduino UNO [11]

Above you will find a brief overview of the devices selected and installed when planning the smart greenhouse. Mainly low-cost components were selected to monitor and control the processes of the smart greenhouse. Arduino UNO is a standard Arduino control board. Arduino UNO is one of the most popular control boards. It is easy to use compared to other control boards such as the Arduino Mega control board and is strongly supported by the Arduino user community. The Arduino UNO [11] control board (Fig. 4) is a practical and cost-effective solution for users entering the world of mechatronics. The Arduino UNO control board is based on the Atmel ATmega328P microcontroller.

The Arduino UNO control board has 6 analogue inputs, 14 digital input/output pins (6 of which can be used as PWM

outputs), a voltage regulator, a USB port, a power port, a power LED indicator, a 16 MHz crystal, a reset button and an ICSP (In-Circuit Serial Programming) port. The In-Circuit Serial Programming pin or ICSP pin allows the user to programme using the Arduino board's firmware. Digital input/output pins can have the value High or Low. They are labelled with numbers from D0 to D13. The analogue pins are labelled A0 to A5. Their function is to read the analogue sensor signal. They can also be used as GPIO pins (General Purpose Input Output). The AREF or analogue reference pin is used to supply the control board with a reference voltage from an external power supply.

The power LED indicator shows when the power is switched on (the LED lights up) and when the power is switched off, the LEDs do not light up. The TX and RX LEDs indicate successful data flow between the computer and the control boards. The reset button is used to reset the connection or the control panel itself. The USB port allows the Arduino UNO control board to be connected to the computer and is required for its programming, it can also serve as a power source. The crystal oscillator has a frequency of 16 MHz which makes the Arduino UNO Control Board powerful. The voltage regulator converts the input voltage into a voltage of 5 V. The GND pins are ground pins and serve as 0 volt pins. Vin is the input voltage. Two Arduino UNO control boards are used in conjunction with the upgrade to control all the actuators to be installed in the greenhouse.

2.2 Selecting the Humidity and Air Temperature Sensor

The DHT11 [12] is a digital temperature and humidity sensor (Fig. 5). It is a reliable and stable sensor characterised by good quality, fast response, protection against interference and economy. Calibration has been carried out beforehand under controlled laboratory conditions so that the sensor is ready for use immediately after purchase.

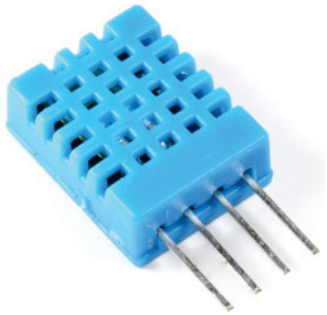


Figure 5 DHT11 sensor [12]

The temperature is measured with a surface-mounted NTC temperature sensor - a thermistor. The thermistor is a variable resistor made of semiconductor material. Changing the temperature changes its resistance. The PTC thermistor (Positive Temperature Coefficient) has a positive temperature coefficient with a measuring range of –50 to +220 °C.

The soil moisture sensor [13] measures the moisture content of the surrounding soil. The sensor is manufactured using the immersion gold process, which protects the nickel surface during the oxidation process. The soil moisture sensor (Fig. 6) is a fork-shaped probe with two exposed conductive plates. The resistance of the probe is inversely proportional to the soil moisture. If there is a lot of water in the soil, the current flows more easily, the conductivity is better and the resistance is lower. Dry soil conducts the current less well and therefore offers a higher resistance. The output voltage of this sensor is calculated by measuring the resistance and the value of the moisture content in the soil is obtained. There is a module on the sensor for connection to the Arduino control board. Such sensors have at least three pins: Vcc, GND and AO. Vcc is the pin that supplies the sensor with power. It is recommended to select a value between 3.3 V and 5 V. GND is the ground pin. AO (Analogue Output), an analogue output pin, generates an output voltage that is proportional to the moisture content. Two analogue soil moisture sensors were used in this work.

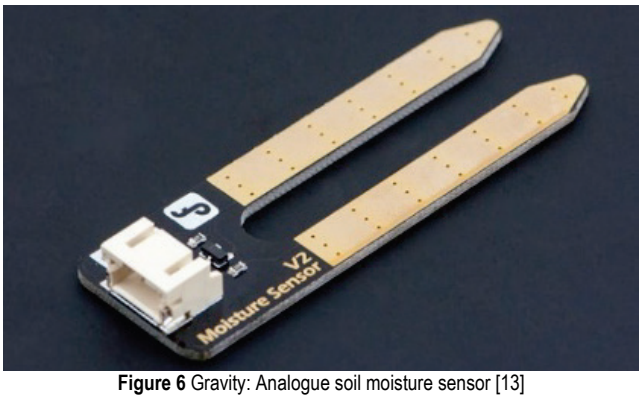


Figure 6 Gravity: Analogue soil moisture sensor [13]

2.3 Selection of Actuators and Motor Drivers

The actuators required to control the greenhouse are a stepper motor and a servomotor. The stepper motor from Moons, type MS17HD2P417A-01, was used in this work. This drive is used to open and close the greenhouse cover. A spindle is connected to its shaft, which converts the rotation of the stepper motor into a linear movement of the lever, which is connected to the greenhouse cover. The servo motor used in this project is the TowerPro MicroServo SG90. Since it is sometimes not necessary for the roof of the greenhouse to be open because the temperature in the greenhouse is not critically high, it is sufficient to open the window/flap to allow airflow. This flap is located at the front of the greenhouse, opposite the fan. This servo motor is used to open and close the flap.

The Arduino UNO [14] controller board cannot fulfil the parameters for operating the selected actuators. It is necessary to upgrade the Arduino L293D Motor Driver Shield (Fig. 7). The L293D shield is one of the most popular motor driver shields capable of controlling motors without additional modules. It can control up to four bidirectional DC motors, two stepper motors and two servo motors.

To communicate with the user, the system is upgraded with an Arduino GSM/GPRS SIM800F shield that enables the use of the GSM mobile phone network. General Packet Radio Service (GPRS) is a protocol that enables wireless data

transmission. In order for the GSM/GPRS SIM800F Shield to work with the formatted programme, a special library must be loaded that enables sending and receiving text messages, making and receiving calls and connecting to the Internet.

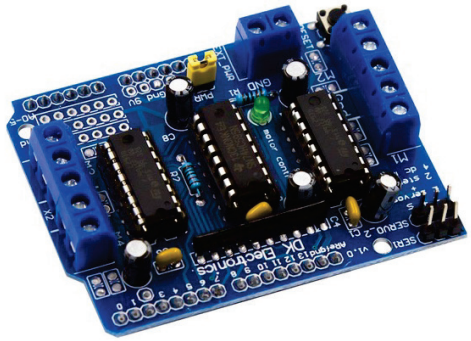


Figure 7 Arduino L293D motor driver shield [14]

3 DESIGN OF THE GREENHOUSE WITH IMPLEMENTATION OF THE COMPONENTS

The moulded greenhouse is designed for small-scale production. The chosen final dimensions of the greenhouse are 280 × 380 × 220 mm. The construction of the greenhouse is shown in Fig. 8 with a view of the opening of the greenhouse cover. A fan is installed at the rear. The fan is used in combination with a damper as a source of airflow in the greenhouse. The fan and the stepper motor are connected to the Arduino Motor Driver shield L293D. On an Arduino UNO control board there are not enough pins for all the motors, sensors and other components. Sensors, buttons and LEDs are connected to the main board of the Arduino UNO. If the DHT11 temperature and humidity sensor detects a temperature that is too high, it sends a signal to the Arduino UNO main control board. A decision is made as to whether the temperature is too high or whether it has reached a critical value. These intervals are determined empirically and entered into the Arduino code depending on the plants planted in the greenhouse.

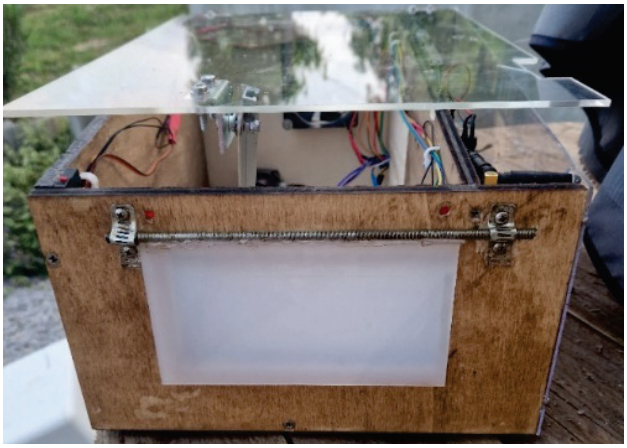


Figure 8 Structure of a greenhouse with a movable cover and flap [8]

A flap (window) is attached to the front of the greenhouse, which is used to ventilate the greenhouse when the lid is closed and the temperature is not critically high. A gear wheel is mounted on the shaft of the servomotor, one tooth of which is extended to a lever that opens the flap by pressing it. When the servomotor turns in the other direction, the lever moves away and the flap closes under its own weight.

Figure 9 Stepper motor with a spindle for lifting the greenhouse cover [8]

A stepper motor is installed inside the greenhouse, to which the shaft is connected (Fig. 9). Activating the stepper motor raises the greenhouse cover until a limit switch is activated, which sends a signal to stop the spindle rotation. The cover is closed by activating the stepper motor in the opposite direction until the limit switch on the greenhouse housing is activated and the cover is closed.

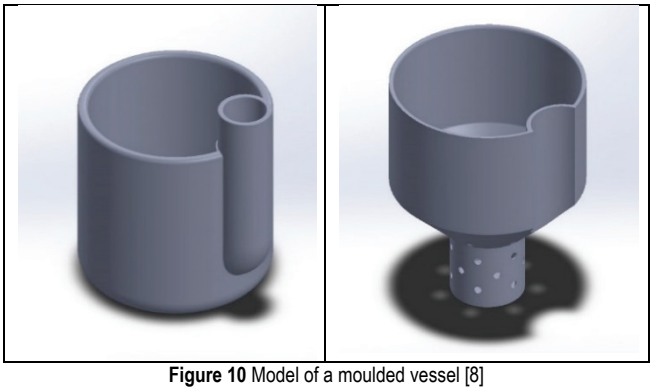


Figure 10 Model of a moulded vessel [8]

The problem of dosing water for plants was solved by designing containers in the SolidWorks programme. The containers were modelled with the aim of extending the time between watering the plants. The containers consist of two parts (Fig. 10). The outer shell is a cylinder that has an additional pipe at one point through which water is poured. The inner container, in which the plant is planted, consists of a larger and a smaller cylinder connected by a cone. The cone and the smaller cylinder are provided with holes. The soil comes into contact with water through the holes and the plant draws exactly as much water through the roots as it needs, significantly extending the time between waterings.

The control system is powered by four AA batteries. The housings are located on the outside of the greenhouse near the Arduino boards, which can be seen in Fig. 11. Attention must be paid to how the batteries are discharged and they must be replaced in good time. Instead of batteries, it is possible to realise a permanent power supply via a rectifier. The GSM/GPRS SIM800F uses additional energy from the charger.

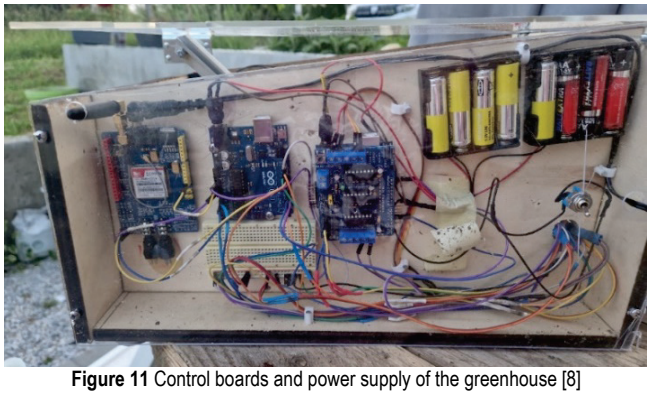


Figure 11 Control boards and power supply of the greenhouse [8]

3.1 Implementation of the Programme Support for the Smart Greenhouse

The system operates by transmitting all sensor-recorded parameters to the Arduino UNO control board, which contains a driver programme (Arduino code) that checks the status of the sensors. So that the driver programme can process the received values, the so-called Arduino libraries must be loaded.

Two management programmes, "Master" and "Slave", were developed to operate the greenhouse. The master programme is used to read sensors, start servomotors and send SMS messages. The slave programme receives the commands from the master programme to control the fan and the stepper motor.

The master Arduino programme uses "if - then - else" loops. The conditions that are constantly checked are the humidity in the greenhouse, the air temperature in the greenhouse and the soil moisture of the plants in the greenhouse.

If the air temperature in the greenhouse is above the set value of +25 °C, the greenhouse is ventilated by opening the flap and switching on the fan. If the temperature continues to rise and exceeds the value of +35 °C, the greenhouse cover is also opened. The subordinate Arduino programme uses a do-while loop.

The DHT11 sensor detects whether the temperature in the greenhouse has risen above +25 °C and sends the information to the main Arduino board with the master programme. The programme then sends a signal to the servomotor and opens the valve. It also sends a signal to the additional Arduino board with the slave programme, which switches on the fan. If the temperature is higher than +35 °C, the master sends an additional signal to the slave programme that further action is required. The slave programme then starts the stepper motor and opens the greenhouse lid. The lid and flap are closed and the fan is stopped according to the same principle, whereby a drop-in temperature is recognised.

Another condition that is constantly monitored is the level of soil moisture in the pots with plants. If the exposed circuit boards of the gravity sensor for soil moisture detect a drop-in moisture below the permissible limit, a signal is sent to the Arduino UNO main control board. A red LED is activated as a visual warning of a change in status. When the soil moisture status changes, the SIM800F shield establishes a connection to the GSM. Once connected, the SIM800F Shield sends an SMS message to the mobile phone with information on which container needs to be refilled. After the plant has been watered, the Gravity soil moisture sensor recognises a change in soil moisture and the red LED assigned to the pot goes out. The realised, tested and implemented programme code can be found in [8].

After selecting and installing all components and implementing the software control system, functional testing of the complete smart greenhouse was conducted. Various scenarios were simulated, and the tests confirmed that the system meets the predefined project requirements.

4 CONCLUSION

This study presents the design and implementation of a smart greenhouse system tailored for small-scale agricultural production. The main parameters for the process are listed, and the parameters for the design of the task in question are particularly emphasised and explained. Based on the analysed parameters, the components required to monitor and smart management of the process in the greenhouse were selected. The Arduino interface and control boards were chosen for their cost efficiency and relative ease of use. They were implemented in a small greenhouse and the hardware and software solutions were tested. After the tests were completed, the realised project of a small smart greenhouse for agricultural production was confirmed. The installed management system can be relatively easily customised in appearance and size to the new greenhouse, depending on the available space and the size of the plants. Depending on the requirements, the number of sensors or motors can be increased or decreased. If desired, additional sensors can be added, e.g. a light sensor. If the greenhouse is located near a water connection, it is also possible to implement an automatic irrigation system which minimises the need for human intervention in the greenhouse. In order to reduce the overall project cost, cost-effective components and materials were selected for the construction of the greenhouse.

Acknowledgements

This research was funded by the University of Slavonski Brod, Republic of Croatia (project SmartPRO).

5 REFERENCES

[1] Kumar, A., Tiwari, G. N., Kumar, S., & Pandey, M. (2006). Role of Greenhouse Technology in Agricultural Engineering. *International Journal of Agricultural Research*, *1*(4), 364-372. https://doi.org/10.3923/ijar.2006.364.372

[2] Meah. K., Forsyth, J., & Moscola, J. (2019). A Smart Sensor Network for an Automated Urban Greenhouse. *The IEEE International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST2019)*, 21 February 2019, Dhaka, Bangladesh. https://doi.org/10.1109/ICREST.2019.8644079

[3] Shimpi, M., Thorat, V., Gavade, S., Pawar, S., & Rajule, N. (2022). Smart Greenhouse Automation and Monitoring System. *The 6th International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pimpri*

Chinchwad College of Engineering (PCCOE), Aug 26-27 2022, Pune, India.
https://doi.org/10.1109/ICCUBEA54992.2022.10010745
[4] Ardiansah, I., Bafdal, N., Suryadi, E., & Bono, A. (2020). Greenhouse Monitoring and Automation Using Arduino: A Review on Precision Farming and Internet of Things (IoT). *International Journal on Advanced Science Engineering Information Technology*, *10*(2), 703-709. https://doi.org/10.18517/ijaseit.10.2.10249
[5] Mahfuz, N., Jahan, R., Islam, Md. M. I., Nigar, M., & Karmokar, S. (2020). Microcontroller Based Intelligent Greenhouse Environment Monitoring and Controlling System. *IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, 26-27 December 2020, Bhubaneswar, India. https://doi.org/10.1109/WIECON-ECE52138.2020.9397991
[6] Kirci, Y. C. P., Erdinc, E., & Celik, Y. (2021). Smart greenhouse and smart agriculture. *Conference of Open Innovations Association*, FRUCT Oy, 455-459.
[7] Sahaidak, T. & Huzynets, N. (2021). Investigation of Greenhouse Monitoring and Control system. *Journal Advances in Cyber-Physical Systems*, *6*(1), 54-62. https://doi.org/10.23939/acps2021.01.054
[8] Tedeško. E. (2023). Smart Mini Greenhouse for Agricultural Production. *Undergraduate Thesis*, University of Slavonski Brod, Slavonski Brod. (in Croatian)
[9] See https://www.greenhousemag.com/article/2020-structures-guide-5-tips-for-temperature-control-in-your-growing-environment/
[10] See https://www.msschippers.com/thermometer-with-hygrometer-analogue-4305813.html
[11] See https://www.javatpoint.com/arduino-uno
[12] See https://ardubotics.eu/hr/senzori/1104-senzor-temperature-i-vlage-dht11.html
[13] See https://www.dfrobot.com/product-599.html
[14] See https://www.engineersgarage.com/arduino-l293d-motor-driver-shield-tutorial/

Authors' contacts:

Tomislav Šarić, full professor
University of Slavonski Brod,
Mechanical Engineering Faculty in Slavonski Brod,
Trg I. B. Mazuranic 2, 35000 Slavonski Brod, Croatia
tsaric@unisb.hr

Elizabeta Tedeško, bachelor student
University of Slavonski Brod,
Mechanical Engineering Faculty in Slavonski Brod,
Trg I. B. Mazuranic 2, 35000 Slavonski Brod, Croatia
etedesko@unisb.hr

Goran Šimunović, full professor
University of Slavonski Brod,
Mechanical Engineering Faculty in Slavonski Brod,
Trg I. B. Mazuranic 2, 35000 Slavonski Brod, Croatia
gsimunovic@unisb.hr

Sara Havrlišan, assistant professor
(Corresponding author)
University of Slavonski Brod,
Mechanical Engineering Faculty in Slavonski Brod,
Trg I. B. Mazuranic 2, 35000 Slavonski Brod, Croatia
shavrlisan@unisb.hr

ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20230925081813

Subject review
Received: 2023-09-25, Accepted: 2024-04-10

The Impact of Social Media on the Sustainability of Fashion Industry Marketing

Ivana Bolanča Mirković*, Katarina Itrić Ivanda, Zdenka Bolanča, Marina Vukoje

Abstract: In the second decade of the 21st century, there were significant changes in the strategies for implementing marketing campaigns. The fashion industry quickly accepted and improved the ways of conducting marketing campaigns. The market requires quick-changing trends, which is evident both in the rapid production of clothing items and in new marketing campaigns. Customers eager for interactivity and viewing trends at a time that suits them, often view fashion novelties on social media. The desire for the representation of fashion brands on as many social network applications as possible greatly contributes to increasing the impact of reducing the quality of the environment. Users need large amounts of energy to browse new trends. Posts of trendy fashion items are often shared in groups or circles of friends, increasing the energy needed. Systems for storing media content require a large amount of energy needed to store data from social media and to cool the system. All the above contributes to the emission of greenhouse gases and the reduction of non-renewable energy sources because unfortunately, renewable energy sources are not yet sufficient for all the listed needs. In research, the habits of respondents related to the use of social media related to fashion items are studied. In addition to the aforementioned, knowledge related to the awareness and knowledge of respondents regarding the influence of social media on the sustainability of marketing drips of the fashion industry was collected. The mentioned topic is not close to the respondents, although most of the respondents were undergraduate graphic design students who may participate in marketing campaigns in the future.

Keywords: fashion industry; marketing; social media; sustainability

1 INTRODUCTION

There are many ways of digital marketing campaigns in the fashion industry, some of the most common and effective ones are content marketing, social media, PPC (Pay-Per-Click) advertising, SEO (Search Engine Optimization), email marketing, influencer marketing, affiliate marketing, video marketing and others [1, 2]. The most successful campaigns often combine several different tactics to achieve maximum effect. [3] Campaign planning aims to ensure alignment with goals, target audience and budget. Social media community is created (such as Facebook, Twitter, Instagram, LinkedIn, etc.) to promote fashion brands, collections and products and communicate with consumers [4-6].

In marketing campaigns with fashion content, the visual component of the advertised products is important because the fashion industry is based on visuality, therefore high-quality and attractive visual content is an indispensable part of campaigns. The elements of the campaign therefore often contain photos and videos showing clothes, and fashion accessories presented by famous models or influencers. In addition to the visual elements of the campaign, a story or text message that expresses values, brand style and lifestyle is important, which can stimulate an emotional connection with customers [7, 8]. Everything mentioned should be expressed originally and creatively, which will further highlight the campaign on social media. Often unconventional approaches can attract attention and set a brand apart from the competition. Followers can play a role in the campaign and be encouraged to share photos of themselves wearing clothes from the collection, further strengthening the brand's social media presence [9]. It should be emphasized that cooperation with influencers can expand the brand's visibility and increase engagement with the target audience [10-12]. All of the above certainly contribute to the quality of the marketing campaigns of a certain fashion brand, but every visual, and textual content, as well as additional content, contributes to an increase in energy consumption.

Social media use data centres to store data, which are mostly supplied with electricity from the public grid [13, 14]. To increase their sustainability and possibly reduce their costs, more and more data centres are maximizing the use of renewable energy sources. To optimize energy consumption, data centres can use sophisticated energy management technologies [15, 16]. Optimization of total energy consumption. System cooling, server load management and the use of energy-efficient equipment can be implemented at multiple business levels. System cooling represents a significant contribution to energy consumption because servers generate heat that disrupts the optimal conditions for server operation [17, 18]. Efficient cooling systems and locating data centres in geographically colder areas can contribute to reducing energy consumption and reducing greenhouse gas emissions [19].

Virtualization can contribute to server load optimization by running multiple virtual machines on a single physical server. In this way, the servers are used at their full capacity, reducing the need for additional physical servers. Consolidation can connect more applications or services to fewer servers, which can contribute to reducing energy consumption and overall costs [20]. Another way of optimization is the load balancing service, which distributes the load balance between several servers to ensure optimal use of resources and prevent overloading of individual servers [21]. This method of optimization can also contribute to reducing the risk of service failure. Automatic server scaling enables dynamic increase or decrease of server resources according to real needs, thus reducing unnecessary consumption of resources [22, 23].

The optimum organization of data on the server can contribute to reducing energy consumption. The caching technique stores frequently used data in fast memory to speed up access to that data, which reduces server load due to

infrequent access to long-term data sources. Prioritizing between different applications reduces the load on less important tasks [24].

In addition to optimizing server performance, applications can be optimized, which can significantly contribute to reducing server load and energy consumption. Optimization usually involves optimizing databases, reducing unnecessary queries, or optimizing algorithms. By combining the mentioned methods depending on your specific needs and data centre environment [25]. It is important to regularly monitor performance and optimize resource management to achieve the best results.

Knowing how to increase energy consumption when creating content on social media and further improving the energy consumption of data centres can significantly influence the reduction of greenhouse gas emissions. The conscience and habits of social media users can additionally contribute to the goal of optimising energy consumption with good marketing results and interesting promotional content [26]. Habits and subject knowledge were investigated in this research to promote the sustainability of textile industry marketing on social media.

2 METHODOLOGY

In this paper, scientific literature and research related to fashion marketing campaigns on social media were followed. The literature search was unbiased and included books, journals and internet data. The key search terms were marketing campaigns, energy consumption, social media, opportunities to reduce energy consumption when publishing marketing content and data centre operations, behavioural styles of individuals on social media and carbon footprint.

In addition to reviewing the literature, survey research was conducted in the period from June 1 to June 30, 2023. The survey research was conducted using Google Forms, and the questions in the survey were composed in such a way that the respondents were offered answers. The survey consisted of two parts. The first part of the survey was aimed at finding out the socio-anthropological data of the respondents. By looking at the mentioned data, patterns of behaviour of different ages, educational or other groups of respondents can be observed. The second part of the survey research consisted of questions that provided information about the respondents' knowledge related to the knowledge of terms related to the sustainability of fashion marketing on social media.

121 respondents participated in the survey. Most of the respondents belonged to the age group of 18 to 23 years (82.6 %) and had a high school education (74.4 %). In the survey is a large representation of the mentioned age group is included in the survey, because the special emphasis of the research was to be placed on the knowledge of sustainable design topics of undergraduate graphic design students. Other respondents were classified into age groups where the age of the respondents increased in intervals of 5 years and the oldest age group consisted of respondents who were over 64 years old. The representation of respondents in the mentioned age groups is equal (about 3.3 %), except in the age group from 58 to 63 years (0.8 %), while the level of education differs. 12.4 %, 5 % and 8.3 % of respondents have undergraduate, and graduate education and a doctorate in science. It is evident from the results that the rest of the respondents have mostly undergraduate education.

3 RESULTS AND DISCUSSION

Fashion campaigns are spreading to more and more social media to reach as many interested customers or collaborators as possible. Every social media application consumes energy for data storage in databases, and database operation, and the user consumes energy when viewing and possibly sharing data related to the fashion campaign. By understanding the habits of the respondents, it is possible to gain insight into the justification for using different applications to achieve better marketing results. Most respondents, 94.2 %, stated that they visit social media several times a day, so according to this statement, the success of campaigns published in the mentioned manner can be assumed. Most respondents find time in the day and want to visit social media. Most respondents (64.5 %) spend two or more hours a day on social media. Some respondents, 5 % of them, spend as much as 6 hours a day. The obtained results additionally emphasize the importance of conducting campaigns on social media, but on the other hand, they emphasize the importance of a sustainable way of conducting campaigns to reduce greenhouse gases. To compare with trends in the world, the data obtained from the survey was used to calculate the average time that an individual spends on social network applications. The research data obtained are 7.46 % less than those obtained on the Global Web Index from July 2021 and amount to 2 hours and 13 minutes. [27] As the data obtained in the research related to respondents in the Republic of Croatia, it can be concluded that habits differ from country to country, but that the differences are not significant. Countries with more inhabitants have residents who probably have more acquaintances, friends and followers with whom they can share promotional materials or socialize through social media, which can result in spending more time on applications.

When the opinion on the most viewed content on social networks is examined, it can be seen that fashion is not the most viewed content, but music and humour (Fig. 1). Fashion content is most represented in the 3rd, 4th and 8th grades. These results show that the interest in fashion is average. The reasons for such answers can be found in the fact that a certain part of the respondents likes fashion and content related to it, but the average customer needs to be more interested in viewing fashion content on social networks. Such results are not correlated with the increase in sales of fast fashion in e-commerce and its increase in sales in the world fashion market, which could indicate that the influence of social media in Croatia is not yet significant. The global market of fast fashion in 2022 was about 60.50 billion USD in 2022, and it is predicted to grow by 2030 to about 179.50 billion USD [27]. Fast fashion is usually associated with cheap and popular clothing items that must reach the commercial market as quickly as possible. Such fashion is often associated with excessive consumerism, which was further emphasized by the introduction of the e-commerce sector, which enabled easier access to fast fashion. The

aforementioned contributes to an additional increase in the carbon footprint. E-stores have brought clothing closer to customers with the availability of benefits that meet customer requirements, such as ease of return or exchange policies, door-to-door delivery, and more. The well-known European fashion house from Spain increased its sales by 68.3 % between 2019 and 2020 [27]. One of the reasons for this success is certainly the quarantine of the population caused by the COVID-19 disease. Recently, fast fashion companies have based their sales growth on the strategy of using the consumer bases of large online sales platforms. In the aforementioned partnership, the benefit has sellers of fast fashion and online sales platforms too. Such cooperation was concluded in August 2023 between the Forever 2021 fast fashion chain and the Sheinom online platform, which has around 150 million customers [28].

Although the respondents were offered options for longer periods, the respondents posted a maximum of two hours a day on social media (Fig. 2). When studying which social media are represented in the selection of respondents in 2 hours in Fig. 2, it can be seen that TikTok and Pinterest have the largest number of respondents (6 and 3 respondents), YouTube and Instagram have the same number of respondents, i.e. 2 respondents, and Twitter (1 respondent). The presented results show that the respondents do not visit social media in large numbers or for long periods. This question brings out the fact that most respondents do not visit social media related to fashion content for more than 15 minutes a day, and the majority of respondents stated that they do not visit them daily. When creating fashion drips, this fact must certainly be considered to modernize the content of the campaigns and attract a larger number of followers, that is, to look at which social applications have the most attendance. Removing fashion content that does not have a high attendance rate, contributed to the reduction of energy consumption.

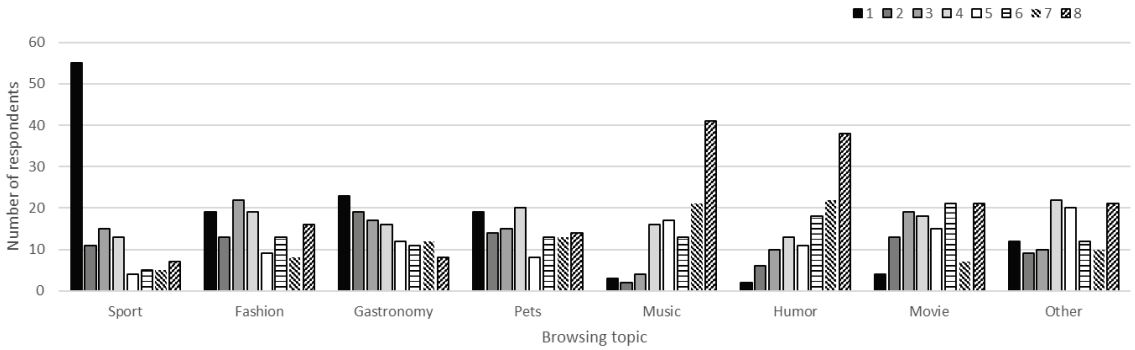


Figure 1 Respondent's statement about reviewed topics on social media (1 is the least frequent, and 8 is the most frequent)

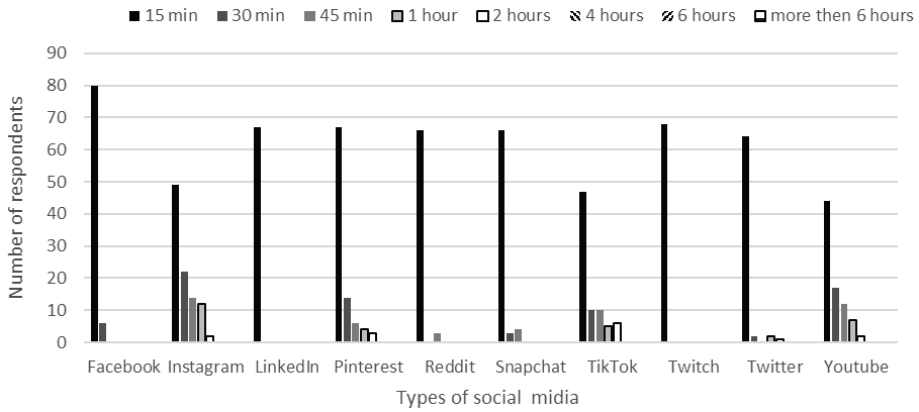


Figure 2 Respondents' statements about daily time spent on social media applications for browsing clothes (1 is the least frequent, and 10 is the most frequent)

In the study, respondents' habits were studied to examine the justification for creating a fashion campaign on social media. Surprisingly, only 5 % of respondents stated that they browse social media with fashion content, and only 1.7 % of respondents stated that they additionally use filters to view fashion-related content. The obtained results indicate that the respondents' habits do not justify the selection of fashion campaigns on social media. This conclusion is supported by the fact that 38.7 % of respondents choose clothes in a store and 29.4 % on websites. The results show that 68.1 % of respondents should apply other methods of fashion drips. Of course, there is always a group of potential customers who are not significantly connected to campaigns, in this survey 20.2 % of respondents cherish their style and are not interested in trends, and 5 % of respondents believe that they are up to date with trends. A good fashion campaign should attract this group of respondents who are not interested or think they are not interested, and campaigns reach them.

Regardless of the respondents' low interest in social media applications, the respondents were pleased with their advantage when presenting fashion clothing products. As many as 44.3 % of respondents stated that they have a better insight into the actual appearance of clothes on a person through social media. The obtained results indicate that the

respondents accessed social media to view other content and that in the meantime they did not change their habits related to fashion browsing content, but that they are aware of the benefits of social media. This statement is supported by the fact that an additional 25.2 % of respondents stated that e-commerce websites are not sufficiently interactive, and 10.4 % of them would like a musical background in the video material. This is a total of 79.9 % of the respondents, which could indicate that campaigns on social media still have room for promotion, they just need to use new ways of gaining an audience.

When creating promotional fashion content on social media applications, it is important to balance the sustainability of the promotion with the maximization of the effects of the promotion, a good selection of the most used applications, removing content from applications where such content is not followed, choosing a desirable and interactive look without using unnecessary effects. Such a task is difficult to perform, especially with little connection to the impact of promotion on the respondents' social media, especially those who could participate in the creation of promotional materials in the future (graphic design students). This is confirmed by the results of the research because 23.1 % of the respondents do not think that social media posts affect their carbon footprint. The obtained result could be influenced by educating students in higher years of study on ways to reduce the consumption of energy when designing promotional materials. The population should be involved in education to acquire sustainable habits.

When respondents were asked to choose the possible options that contribute the most to increasing energy consumption, it can be noted that the answers given were quite different (Fig. 3). Time and effects received the largest number of highest marks. It must be emphasized that time as an element of the presentation received the highest number of lowest ratings from the respondents. The mentioned elements of posts on social media significantly affect energy consumption. The recording time should certainly be reduced to the smallest possible extent while respecting the transparency of the presented product. The effects certainly attract the attention of the observer, but the simplicity of the presentation usually goes with elegance. Background colours and dimensions were selected by the respondents as the elements that have the least impact on energy consumption. Usually, the background colour is white to best emphasize the cuts and colours of the textile products. It is known that the white colour contributes the most to energy consumption. The size of the record should be formed so that the product is easily visible on all screens.

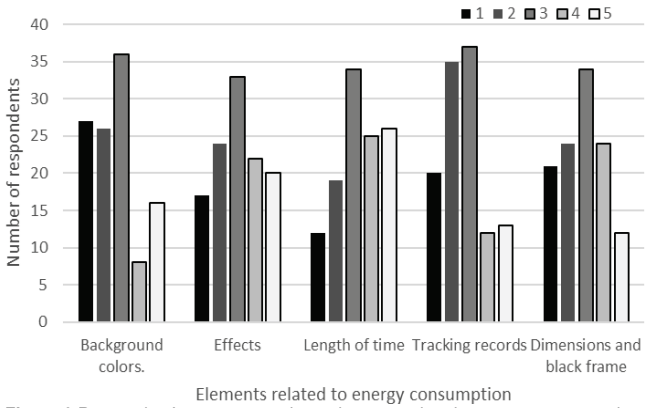


Figure 3 Respondent's statements about elements related to energy consumption on social media applications for browsing clothes (1 is the least frequent, and 5 is the most frequent)

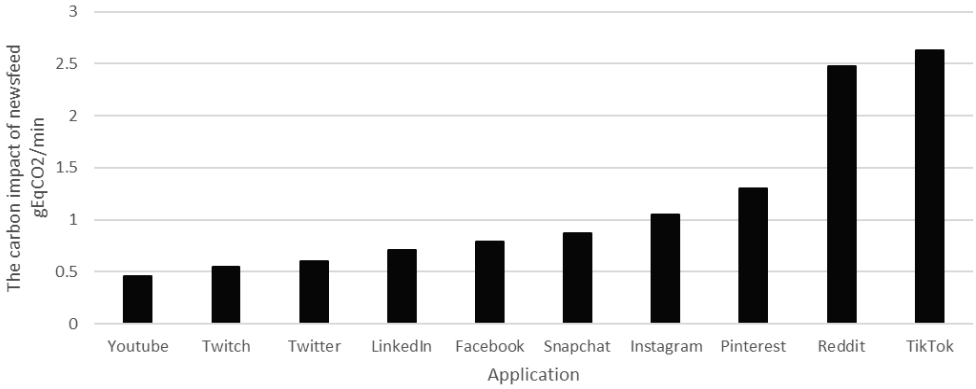


Figure 4 The carbon impact of the newsfeed on applications [27]

Research has confirmed that the impact of applications on the carbon footprint is significantly different. TikTok has more than five times the carbon footprint contribution of YouTube (Fig. 4) [29]. According to the Global Web Index, the average carbon impact in 10 measured applications in a time of 1 minute is 1.15 $gEqCO_2$. In our research, it was determined that an individual spends 2 hours and 13 minutes a day on social media, which would mean that the respondents emit an average of 152.95 $gEqCO_2$ per day. Knowing the effect of applications on the carbon footprint, it is possible to create a campaign that is more sustainable with the same marketing effect.

According to the study by N. Lövehagen and al., carbon footprint emissions differ when using devices that can be used to view social networks smartphones (6 in display) 78 MT CO_2e, tablets (11in display, no keyboard) 21 MT CO_2e, Laptop PCs (14in display) 58 MT CO_2e, desktop PCs (no all-in-one desktops) 22 MT CO_2e and PC displays (25in display) 50 MT of CO_2e [30]. The study shows that smartphone emissions are like PCs, while laptop emissions are

significantly lower. If the data were more available to the wider population, there is a possibility that it would influence their lifestyle and contribute to the reduction of the carbon footprint.

The increasing influence of social media is starting to affect some individuals in a way that they don't want to be part of the consumer who owns mass fashion products. Such thinking is probably fueled by the growing trend of consumers refraining from repeating items of clothing or fearing they will become irrelevant. This thinking could become a new trend. It can contribute to the development of the trend by lowering the prices of slow fashion, which is often of better quality and less common among the population, which could contribute to the individualization of clothing, or by adopting the habit of reducing the number of clothing items purchased, which contributes to reducing the carbon footprint. According to recent reports, the fast fashion industry contributes nearly 10.1 % of all annual carbon emissions, making it one of the biggest polluters of the environment [27].

4 CONCLUSION

Given the diversity of social media, fashion brands often use different platforms to be successful in different parts of the target audience. For example, Instagram is popular for visually presenting collections and collaborating with influencers, while LinkedIn can be useful for connecting with a professional audience in the fashion industry. Monitoring social media analytics can give a good insight into the success of the campaign and enable fashion brands to identify good and bad parts of the campaign to eliminate the types of posts on less successful parts of the campaign. Such a sustainable responsible policy can contribute to the reduction of the carbon footprint with the same marketing results.

Knowing the contribution to the carbon footprint of individual elements in social media posts can greatly reduce the emission of greenhouse gases caused by the use of non-renewable energy sources. In addition, the habits of social network followers can further reduce the carbon footprint. From the conducted survey, it is evident that few respondents are familiar with the negative impact of social networks on the environment. Acquaintance of followers with applications that have a larger carbon footprint can encourage certain applications to increase sustainability or reduce their carbon footprint. It is evident from the results of carbon equivalent emissions that applications offering the same types of content have significantly different levels of emissaries. Similarly, followers can by boycotting applications that are not sustainable encourage applications to invest in increasing sustainability. By monitoring the statistics of campaign success, it is possible to focus campaign announcements on more visited social network applications and on increasing the interactivity of e-commerce websites to reduce announcements on less visited social network applications. In the mentioned way, the promotional material located on the servers will be reduced. Influencers can additionally contribute to increasing sustainability if they make their posts clear, short and interesting without unnecessary effects. The use of a black background on posts could be part of new sustainable campaigns that will make followers aware of the importance of sustainability. An active approach to monitoring the success of promotional campaigns with new and innovative announcements can attract customers.

5 REFERENCES

[1] Chaffey, D. & Ellis-Chadwick, F. (2019). *Digital marketing: strategy, implementation and practice*. Pearson UK.

[2] Ryan, D. (2016). *Understanding digital marketing: marketing strategies for engaging the digital generation*. Kogan Page Publishers.

[3] Baldus, B., Voorhees, C. & Calantone, R. (2015). Online brand community engagement: Scale development and validation. *Journal of Business Research, 68*(5), 2. https://doi.org/10.1016/j.jbusres.2014.09.035

[4] Habibi, M., Laroche, M. & Richard, M. (2014). Brand communities based in social media: How uniqueare they? Evidence from two exemplary brand communities. International Journal of Information Management, 34(2), 123-132. https://doi.org/10.1016/j.ijinfomgt.2013.11.010

[5] Dessart, L., Veloutsou, C. & Thomas, A. (2015). Consumer engagement in online brand communities:A social media perspective. *Journal of Product and Brand Management, 24*(1), 28-42. https://doi.org/10.1108/JPBM-06-2014-0635

[6] Helal, G. & Ozuem, W. (2017). Social Identity Matters: Social Media and Brand Perceptions in the Fashion Apparel and Accessories Industries. In Ozuem, W. & Azemi, Y. (Eds.), *Digital Marketing Strategies for Fashion and Luxury Brands*. Hershey, PA: IGI Global, 326-361. https://doi.org/10.4018/978-1-5225-2697-1.ch016

[7] Manic, M. (2015). Marketing engagement through visual content. *Bulletin of the Transilvania University of Braşov, Series V: Economic Sciences, 8*(57), 2.

[8] Gamble, S. (2016). *Visual Content Marketing*. Wiley.

[9] Arli, D. (2017). Does social media metter? Investigating the effect of social media features on consumer attitudes. *Journal of Promotion Management, 23*(4), 521-539. https://doi.org/10.1080/10496491.2017.1297974

[10] Alvarez-Monzoncillo, J. M. (2022). *The Dynamics of Influencers Marketing*. Routledge. https://doi.org/10.4324/9781003134176

[11] Carter, S. & Yeo, A. (2018). Internet-enabled collective intelligence as a precursor and predictor of consumer behaviour. *Economics, management, and financial markets, 13*(4), 11-38. https://doi.org/10.22381/EMFM13420181

[12] Brown, D. & Fiorella, S. (2013). *Influence marketing: how to create, manage and measure brand influencers in social media marketing*. Indianapolis, Que Publishing.

[13] Ibrahim, I. M., Ameen, S. Y., Yasin, H. M., Omar, N., Kak, S. F., Rashid, Z. N., Salih, A. A., Salim N. O. M. & Ahmed D. M. (2021). Web Server Performance Improvement UsingDynamic Load Balancing Techniques: A Review. *Asian Journal of Research in Computer Science 10*(1), 47-62, Article no. AJRCOS.70184. https://doi.org/10.9734/ajrcos/2021/v10i130234

[14] Shang, W., Liu, D., Zhu, L. & Feng, D. (2017). An improved dynamic load-balancing model. *International Journal of Software Innovation, 5*(3), 33-48. https://doi.org/10.4018/IJSI.2017070103

[15] Rapado, M. A. & Hernandez, J. A. (2019). Automation Platform as an Advanced Energy Management System. *The 7th Eur. Conf. Ren. Energy Sys.*, 10-12 June 2019, Madrid, Spain.

[16] Gough, C., Steiner, I. & Saunders, W. (2015). *Energy Efficient Servers Blueprints for Data Center Optimization*. Apress. https://doi.org/10.1007/978-1-4302-6638-9
[17] Capozzoli, A. & Primiceri, G. (2015). Cooling Systems in Data Centers: State of Art and Emerging Technologies. *Energy Procedia, 83*, 484-493. https://doi.org/10.1016/j.egypro.2015.12.168
[18] Cho, J., Yang, J. & Park, W. (2014). Evaluation of air distribution system's airflow performance for cooling energy savings in high-density data centers. *Energ Buildings, 68*, 270-279. https://doi.org/10.1016/j.enbuild.2013.09.013
[19] Lin, M., Shao, S., Zhang, X., Vangilder, J. W., Avelar, V. & Hu, X. (2014). Strategies for data center temperature control during a cooling system outage. *Energ Buildings, 73*, 146-152. https://doi.org/10.1016/j.enbuild.2013.12.015
[20] Monjur, A. (2013). Physical Server and Virtual Server: The Performance Trade-offs. *European Scientific Journal, 9*(12).
[21] Bourke, T. (2001). *Server Load Balancing, First Edition*. O'Reilly & Associates, Inc
[22] Ahn, Y. W., Cheng, A. M. K., Baek, J., Jo, M. & Chen, H. (2013). An Auto-Scaling Mechanism for Virtual Resources to Support Mobile, Pervasive, Real-Time Healthcare Applications in Cloud Computing. *IEEE Network*. https://doi.org/10.1109/MNET.2013.6616117
[23] Roy, N., Dubey, A. & Gokhale, A. (2011). Efficient Autoscaling in the Cloud Using Predictive Models for Workload Forecasting. *Proc. 2011 IEEE Int'l. Conf. Cloud Computing*, 500-507. https://doi.org/10.1109/CLOUD.2011.42
[24] Voras, I. (2011). Cache server for distributed applications adapted to multicore systems. *Doctoral dissertation*, University of Zagreb Faculty of Electrical Engineering and Computing.
[25] Liao, S., Hung, T., Nguyen, D., Chou, C., Tu, C. & Zhou, H. (2009). Machine learning-based prefect optimization for data center applications. SC '09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis. https://doi.org/10.1145/1654059
[26] Sardianos, C., Varlamis, I., Chronis, C., Dimitrakopoulos, G., Alsalemi, A., Himeur, Y., Bensaali F. & Amira, A. (2021). Reshaping Consumption Habits by Exploiting Energy-Related Micro-moment Recommendations: A Case Study. *Communications in Computer and Information Science book series, 1217*. https://doi.org/10.1007/978-3-030-68028-2_4
[27] Market research report, Fast Fashion Industry Prospective, Zion, (2024) https://www.zionmarketresearch.com/report/fast-fashion-market, available 11.3.2024.
[28] Holman, J. (2023). Shein Forever 21 Team Up in Fast-Fashion Deal, The New York Times, https://www.nytimes.com/2023/08/24/business/shein-forever-21.html, available 11.3.2024.
[29] Derudder, K. (2021). What is the environmental footprint for social media applications? https://greenspector.com/en/social-media-2021/, available 11.6.2023.
[30] Lövehagen, N., Malmodin, J., Bergmark, P. & Matinfar, S. (2023) Assessing embodied carbon emissions of communication user devices by combining approaches. *Renewable and Sustainable Energy Reviews, 183*, 113422. https://doi.org/10.1016/j.rser.2023.113422

Authors' contacts:

Ivana Bolanča Mirković, prof. dr. sc.
(Corresponding author)
University of Zagreb, Faculty of Graphic Arts,
Getaldićeva 2, 10000 Zagreb, Croatia
ibolanca@grf.hr

Katarina Itrić Ivanda, doc. dr. sc.
University of Zagreb, Faculty of Graphic Arts,
Getaldićeva 2, 10000 Zagreb, Croatia

Zdenka Bolanča, prof. dr. sc.
Croatian Academy of Engineering,
Ulica Andrije Kačića Miošića 28,
10000 Zagreb, Croatia
zbolanca@hatz.hr

Marina Vukoje, doc. dr. sc.
University of Zagreb, Faculty of Graphic Arts,
Getaldićeva 2, 10000 Zagreb, Croatia
marina.vukoje@grf.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.

Pictures 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

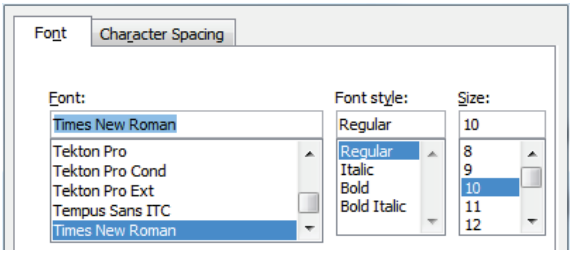


Figure 1 Text under the figure [1]
(Style: Arial Narraw, 8pt, Align Centre)

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

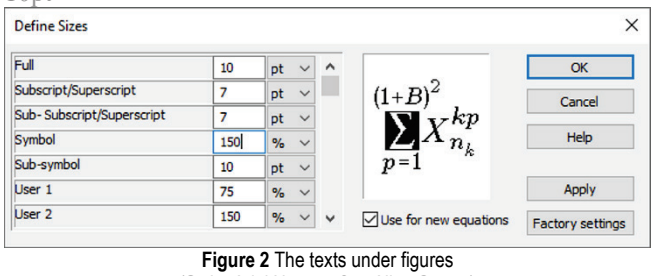


Figure 2 The texts under figures
(Style: Arial Narraw, 8pt, Align Centre)

10pt

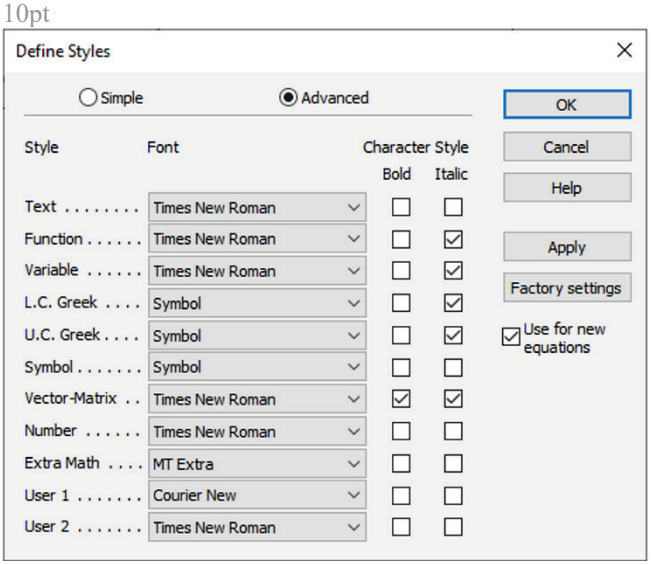


Figure 3 The texts under figures
(Style: Arial Narraw, 8pt, Align Centre)

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.

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.

CCEAI 2026

Buenos Aires, Argentina

2026 10th International Conference on Control Engineering and Artificial Intelligence

Jan. 4 to 7, 2026

www.cceai.org

Committees

General Chairs

Prof. Cees de Bont, Loughborough University, UK

Dan Zhang, Hong Kong Polytechnic University, China

Conference Committee Chair

Zhengtao Ding, University of Manchester, UK

Program Committee Chairs

Jixin Ma, University of Greenwich, UK

Marek Ogiela, AGH University of Krakow, Poland

Jiann-Shiou Yang, University of Minnesota Duluth, USA

Yong Yue, Xi'an Jiaotong-Liverpool University, China

Advisory Committee

Sos Agaian, The City University of New York, USA

Huosheng Hu, University of Essex, UK

Organizing Committee Chairs

Gabriel Gomes, University of campinas, Brazil

Lau Bee Theng, Swinburne University of Technology, Malaysia

Speakers

Prof. Sos Agaian

(Keynote Speaker)

IEEE Fellow

The City University of New York, USA

more is TBA...

Submission Info.

Submission System

https://cmt3.research.microsoft.com/CCEAI2026

Important Dates

Full paper Submission Deadline: July 01, 2025

Notification of Acceptance: Aug. 01, 2025

Registration deadline: Sept. 01, 2025

scan to submit

CFP Topics

ARTIFICIAL INTELLIGENCE

- Algorithms;
- Artificial Intelligence Tools & Applications;
- Bioinformatics;
- Natural Language Processing;
- CAD Design & Testing;
- Computer Vision and Speech Understanding;
- Data Mining and Machine Learning Tools;
- Computational Theories of Learning;

CONTROL & AUTOMATION

- Adaptive Control;
- Automated Guided Vehicles;
- Control Theory and Application;
- Control Theory and Applications Optoelectronics;
- Estimation and Identification;
- Factory Modeling and Automation;
- Fault Detection;

more: https://www.cceai.org/CFP.html

Publication & Indexing

1) All the registered and presented papers will be included in the volume of conference proceeding (online publishing), the publisher will submit articles to Engineering Village, Scopus, Web of Science and other databases for review and indexing after publication.

2)Selected papers with great extension will be recommended to publish in international journal.

CCEAI 2024/2023/2022/2021/2020/2019/2018/2017 conference proceeding have been indexed by EI compendex and Scopus!

See history: https://www.cceai.org/Pub.html

Ms. Yang

Phone: +86-17723329879

E-mail: cceai@apise.org

WeChat: APISE17358663189

Website: www.cceai.org

2026 International Symposium on Computer Vision and Artificial Intelligence (CVAI 2026)

Mar. 27 - 29, 2026 | Shanghai, China **CVAI 2026**

CFP Topics

Track 1: Machine and Deep Learning
- Big data visualization
- Cellular computing
- Cloud computing
- Cognitive intelligence
- Complex Systems
- Deep learning
- Expert systems
- Feature elicitation

Track 2: Computational Intelligence
- Combinatorial and numerical optimization
- Differential evolution
- Evolutionary computing
- Fuzzy quadratic programming
- Fuzzy systems
- Gene expression
- Genetic algorithm
- Genetic programming

Track 3: Robotics and Automation
- Agricultural Robotics
- Autonomous Robotic Systems
- Computer Vision and Image Processing
- Control Architectures and Programming
- Cooperative Perception
- Cooperative Planning and Task Allocation
- Dexterous Manipulation and Grasping

Track 4: Mining and Data Analysis
- Classification and Clustering
- Consistent Data Model
- Data Access
- Data and Knowledge Representation
- Data Mining Applications
- Data Streams Mining, Graph Mining
- Databases

Track 5: Image Processing and Computer Vision
- Activity Detection/ Recognition
- Biometrics, Forensics, Content Protection
- Computational Imaging
- Compressed Image/ Video Analytics
- Document Image Analysis
- Document and Synthetic Visual Processing
- Human Computer Interaction

more: https://cvai2026.org/CFP.html

Important Dates

Submission Deadline:
September 25, 2025
Notification of Acceptance:
October 25, 2025
Registration deadline:
November 25, 2025

APISE
Asia Pacific Institute of Science and Engineering

Submission Method

Scan this QR Code

https://cmt3.research.microsoft.com/CVAI2026

About Shanghai

Shanghai is a luxurious playground for the well-heeled, with Michelin-star dining, high-end fashion houses, and over-the-top hotels. The Huangpu River splits the city into two districts: Pudong and Puxi. The Pudong skyline looks like it was ripped from the Jetsons; on the Puxi side, you can walk the Bund riverside district to get a taste of old Shanghai. The food scene is phenomenal; take advantage of tours that focus on local eats, like dumpling houses. Also, don't miss the chance to fill your suitcase with custom-made clothing from bespoke shops.

APISE LinkedIn APISE WeChat

E-mail:cvai@apise.org
Tel: +86-17723329879
WeChat: 17358663189
Website: https://cvai2025.org/

TEHNIČKI GLASNIK / TECHNICAL JOURNAL – GODIŠTE / VOLUME 19 – BROJ / ISSUE 3

RUJAN 2025 / SEPTEMBER 2025 – STRANICA / PAGES 341-508

Sveučilište Sjever

SVEUČILIŠTE SJEVER / UNIVERSITY NORTH – CROATIA – EUROPE

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

TEHNICKIGLASNIK@UNIN.HR – HTTP://TEHNICKIGLASNIK.UNIN.HR