Doktoranduszok Fóruma

WHAT DOES "GOOD DESIGN" MEAN FOR A GENERATED SHELF HOLDER?

Alireza Aghakhani[1], ***Ágnes Takács***[2]
[1]PhD Student, [2]Associate Professor
[1,2]*Faculty of Mechanical Engineering and Informatics,*
Institute of Machine and Product Design
[1]alirezaaghakhani90@gmail.com, [2]agnes.takacs@uni-miskolc.hu

Abstract

This paper explores the concept of good design in an era dominated by computational tool like Topology Optimization (TO) and Generative Design (GD), reframing engineers as curator of generated design solutions. Tracing five era of design evolution from pre-industrial to AI driven era, it highlights that current evaluation practices prioritize quantitative metrics like weight reduction and structural safety, while neglecting qualitative aspect such as usability, simplicity and manufacturability. A steel shelf holder case study (78 g initial, 60 % target) contrast conservative (39.7 % to 47 g) and exploratory (39.5 % to 46 g) TO strategies, exposing trade-off. Integrating Rams "less but better", Sullivan "form follows function" and DFX, it proposes a balanced framework for human-centred, practical outcomes.

1. INTRODUCTION

The question of what makes a good design has evolved dramatically over time. With computational design tools and artificial intelligence now generating multiple design solutions automatically, this question becomes more complex and important. This study examines how we should evaluate designs when machines, rather than humans, serve as the primary generators. It traces the development of design philosophy, explores the role of computational technologies in modern engineering, and presents practical case study on optimized shelf holders. In the era of artificial intelligence this will offer a critical analysis of design methodology that challenges traditional view of design quality while proposing a framework that combines the computational and efficiency with human centered design principles.

2. THE EVOLUTION OF DESIGN METHODOLOGY AND THE DESIGNER'S CHANGING ROLE

Design practice has undergone profound transformation across five eras as shown in Figure 1. In the pre-industrial era, design was rooted in tactic knowledge and handcrafted expertise with designers functioned as a craftsperson who created unique objects using tool centered approaches [1]. The industrial era introduced standardization and mass production replacing individual craftsmanship; designer role shifted to the process management, and manufacturing optimization for scale, while remaining tool centered [1, 2]. The early 20th century Bauhaus movement Early 20th century introduced the principle "form follows function" articulated by American architect Louis Sullivan [3, 4], positioning designers as problem solvers focused on human centered design moving away from arbitrary ornamentation to functional honesty [5]. The digital era brought computer aided design (CAD), Simulation tools and digital prototyping [6], with design from a drawing based to a simulation-driven work and designers became software-centered digital problem solvers [7]. On the last era, which we are still in it, artificial intelligence and generative algorithms produce multiple design alternatives simultaneously, shifting designer's primary role from creator to curator, who select the and refine options generated

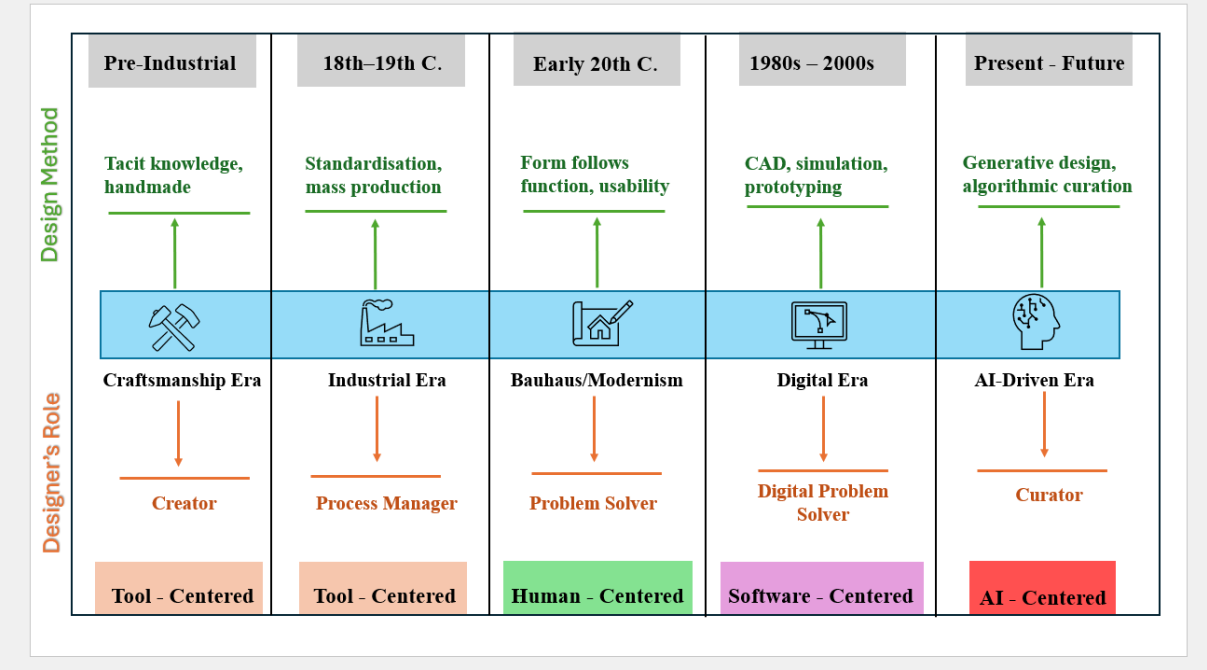


Figure 1: Top Section (green) shows the historical progression of Design methodology while the bottom section (orange) illustrates the designer's role over time

by algorithm rather than starting designs from scratch [8]. This perspective shift carries significant implications. When AI generates design options human designers must evaluate, select and modify these outputs using criteria beyond computational optimization, such as usability and manufacturing. This requires new framework for evaluating design quality that extends beyond traditional engineering metrics.

3. COMPUTATIONAL DESIGN TECHNOLOGIES: TOPOLOGY OPTIMIZATION AND GENERATIVE DESIGN

Engineering design workflows are currently dominated by two main computational technologies: Generative Design (GD) and Topology Optimization (TO).

Topology Optimization (TO): By carefully removing material from a design space while meeting load conditions and constraints, a mathematical technique known as Topology Optimization (TO) finds a single optimal structure. Iterative refinement and Finite Element Analysis (FEA) simulation are used in the process, which results in a single optimized geometry with performance metrics [9, 10].

Generative Design (GD): A broader strategy is represented by Generative Design (GD), which produces a number of design candidates that meet predetermined objectives and limitations. GD uses cloud computing and algorithms to investigate a variety of options, providing designers with trade-off metrics to guide selection, in contrast to TO, which converges to a single solution [11, 12].

TO and GD both are my primary research work. Topology Optimization takes your design space, applies a load, and outputs one optimized geometry. Generative Design does the same thing but generates multiple design candidates to choose from. Both shift the designer's role from creator to curator. You're no longer designing from scratch, you're evaluating what the algorithm produces and selecting the best solution [13, 14].

Despite the sophistication of these computational tools, a critical gap exists in current design evaluation practices. Design selection typically prioritizes quantitative performance metrics,

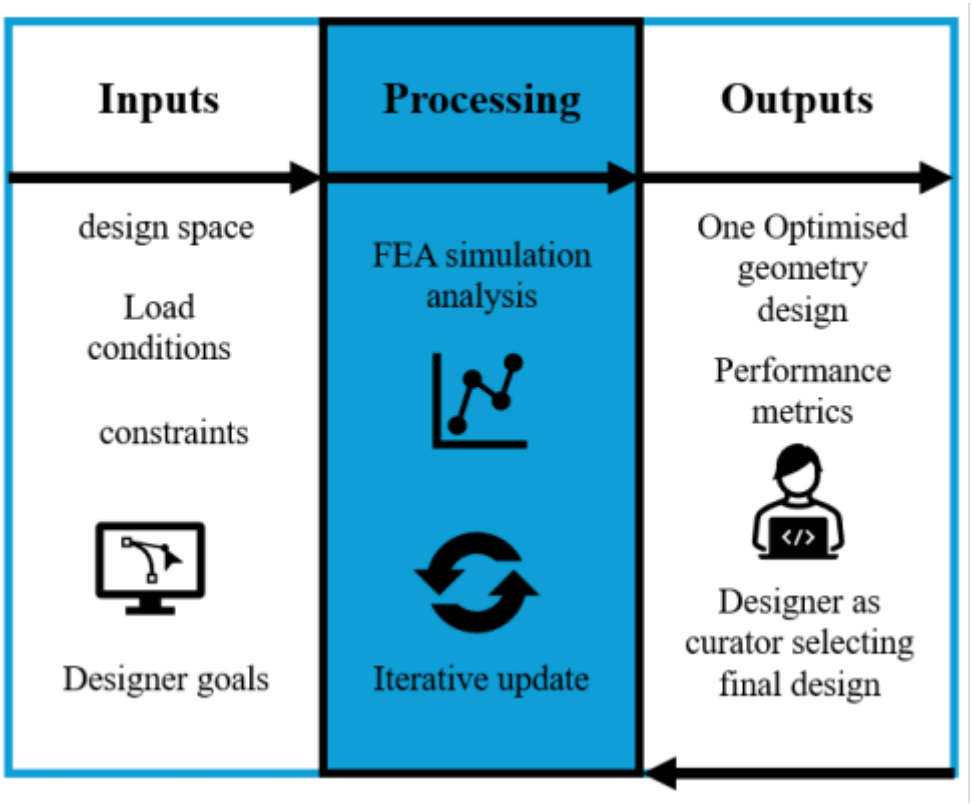


Figure 2: Workflow Schematic for Topology Optimization in Fusion 360

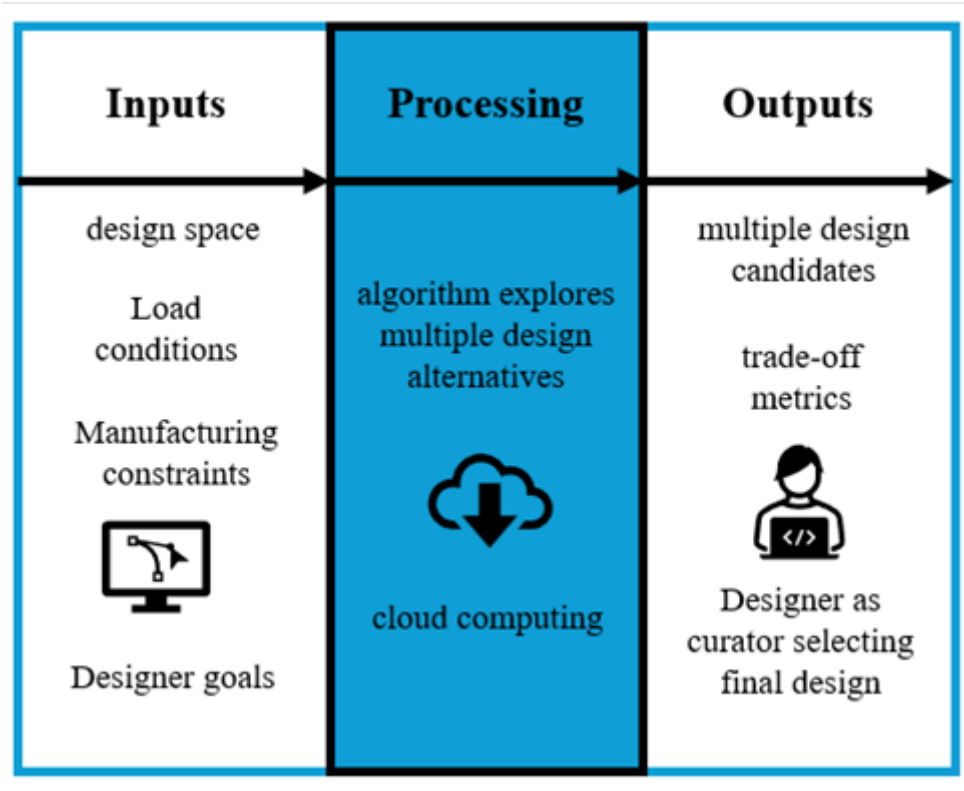


Figure 3: Interface for GD showing input and output

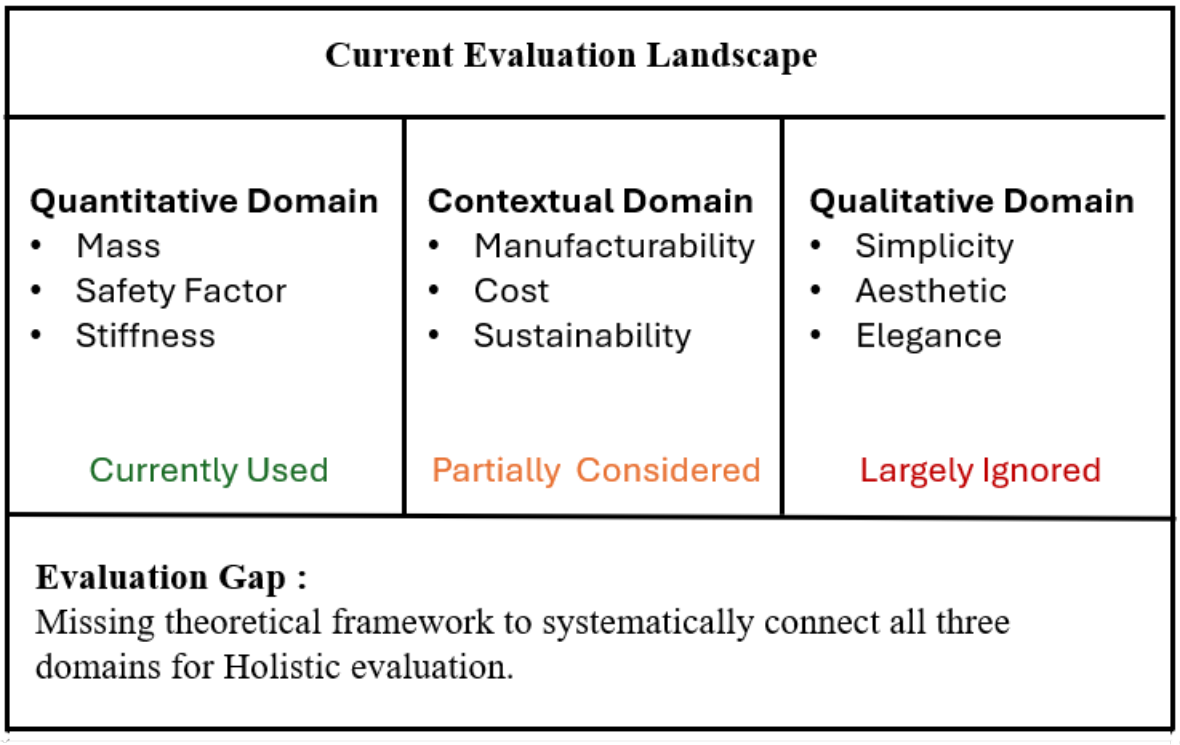


Figure 4: Summary of evaluation domains and gaps in Assessing TO and GD Design Outcomes.

weight reduction and structural safety, while neglecting qualitative aspects such as usability, simplicity, and manufacturability.

This gap creates designs that are "efficient but unbuildable" ordisconnected from real-world human needs. Therefore, the challenge is developing comprehensive evaluation frameworks that balance quantitative and qualitative criteria, ensuring generated designs remain practical and human-centered.

4. EVALUATING TO/GD AGAINST CLASSICAL DESIGN PRINCIPLES

There is more to evaluating generated design then just computational optimization. It requires careful integration of well-established design principles, using realistic engineering methods. This process is shaped by three main bases. Dieter Rams principle of "less but better" which emphasize, clarity, honesty and long-lasting functionality by removing unnecessary components [15]. Louis Sullivan's statement that "form follows function" which guarantees that designs follow naturally from their intended use rather than excessive ornamentation [16]. Designed for X (DFX) approaches, particularly Designed for Manufacturability (DFM) and Design for Assembly (DFA), bridge the gap between theoretical refinement and real-world application alongside these philosophical foundations [17]. DFM addresses production efficiency and cost while DFA improves assembly and labor demands [18]. Together these perspectives, Rams, Sullivan and DFX provide a holistic framework that balances aesthetic integrity, functional purpose and manufacturability, ensuring that design innovation remains both human centered and possible.

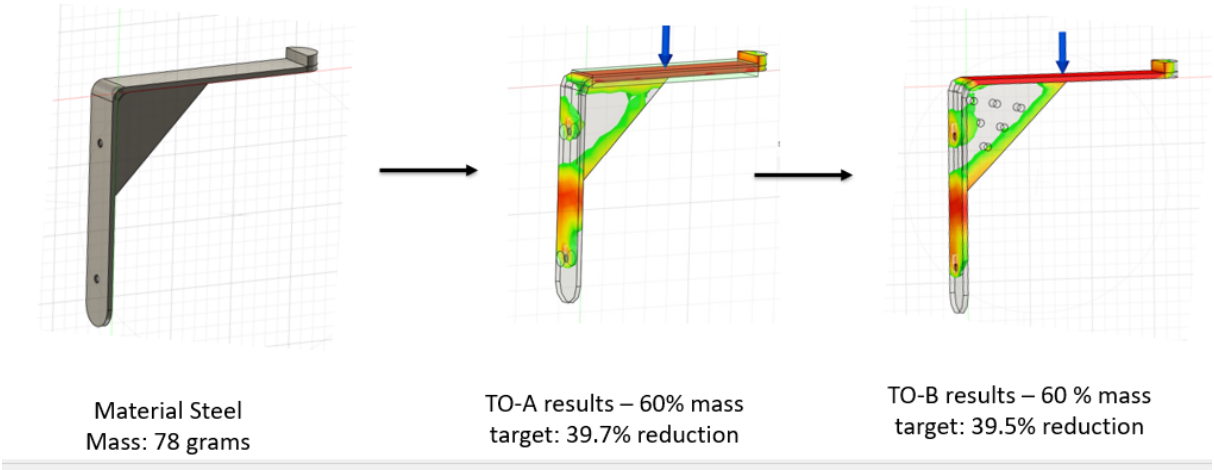


Figure 5: MISSING CAPTION

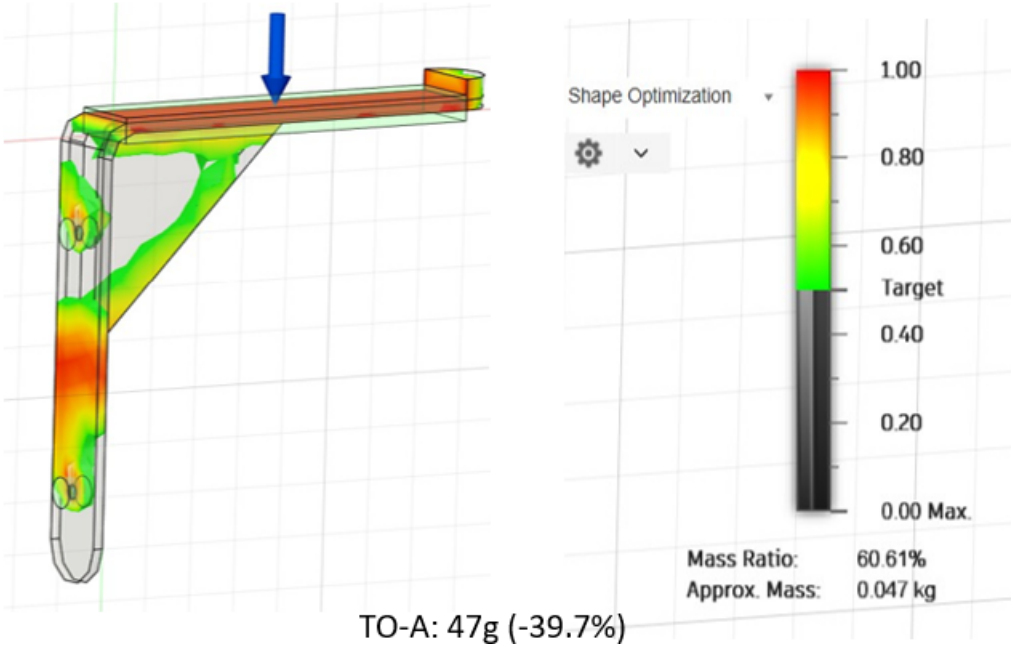


Figure 6: MISSING CAPTION

5. CASE STUDY: GENERATED SHELF HOLDER AND OPTIMIZATION ANALYSIS

This research applied topology optimization to a standard steel shelf holder initially weighing 78 grams with a target mass with preserving 60 % comparing two distinct optimization strategies.

Optimization Strategy A (Conservative Approach) achieved a mass ratio of 60.61 % (47 grams), representing a 39.7 % mass reduction while maintaining a solid triangular support structure with moderate manufacturability and moderate complexity.

In contrast, Optimization Strategy B (Exploratory Approach) achieved marginally better results with a mass ratio of 59.87 % (46 grams) a 39.5 % mass reduction by incorporating distributed holes throughout the structure that improved manufacturability while increasing geometric complexity.

These findings show the fundamental trade-off that comes with computational optimization: although both approaches consistently resulted in structurally safe, mass-efficient designs

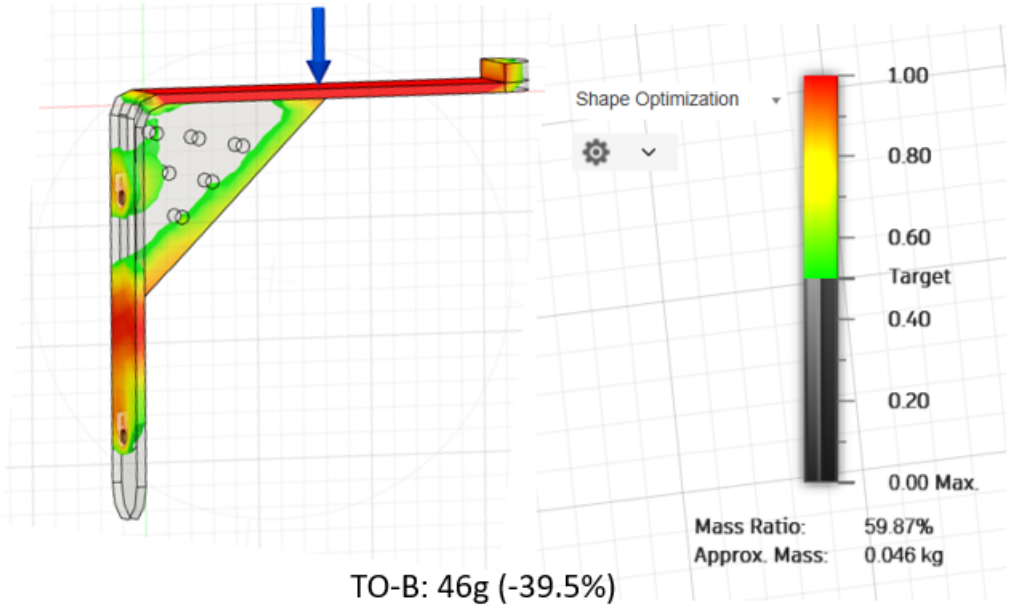


Figure 7: MISSING CAPTION

with performance gains that were almost equal (an average mass reduction of about 39.6 %), their implications for manufacturing were very different. While Strategy B followed aggressive material removal through hole distributions, achieving slightly higher efficiency but necessitating more complex manufacturing processes, Strategy A prioritized structural simplicity at the expense of slightly less mass reduction. Neither optimization approach was found to be inherently better; rather, each is a strategic decision that takes into account various constraints and priorities in the design-to-manufacture chain.

6. CONCLUSION

This study shows that topology optimization able to produce shelf holders that are lighter and structurally safe, but metrics alone can never determine what a "Good design" is. Instead, good design is the result of evaluating safety and mass efficiency alongside cost, sustainability, simplicity, and manufacturability, using the DFX methods, Louis Sullivan, and Dieter Rams as a common framework. In this AI-driven world the designers' job has changed from creating forms to curator. They need to carefully select algorithmic results to find those that actually meet human needs and real-world constraints.

REFERENCES

[1] S. Duman: *History of design methodology studies and the first modern design schools: A critical approach relating to the contemporary design landscape*. Istanbul: Research & Reviews in Architecture, Planning and Design, 2021.

[2] B. Duraković, M. Halilovic: *Industry 4.0: The new quality management paradigm in era of the industrial Internet of Things*. International Journal on Informatics Visualization, vol. 7, no. 2, pp. 580–587, 2023. `https://www.joiv.org/index.php/joiv`

[3] Louis H. Sullivan: *The tall office building artistically considered*. Lippincott's Magazine, March 1896.

[4] C. Pektaş: *The "Form follows function": An oath to the material in Bauhaus*. 2020.

[5] Anita Cross: *The educational background to the Bauhaus*. Design Studies, vol. 4, issue 1, pp. 43–52, 1983. ISSN: 0142-694X, DOI: 10.1016/0142-694X(83)90007-8

[6] L. Stephen Wolfe: *How CAD became universal*. Annals of the History of Computing, vol. 47, pp. 14–25, 2025. https://api.semanticscholar.org/CorpusID:280154484

[7] J. G. Pereira, A. Ellman: *From CAD to physics-based digital twin: Framework for real-time simulation of virtual prototypes*. Proceedings of the Design Society: DESIGN Conference, vol. 1, pp. 335–344, 2020. DOI: https://doi.org/10.1017/dsd.2020.47

[8] X. Li, H. Liu, X. Zhang, R. Cai, Y. Yin, S. Wu, C. Chai: *The evolving roles of modern designers: Through the lens of design behavioral patterns within work environments enhanced by generative AI*. In C. Gray, E. Ciliotta Chehade, P. Hekkert, L. Forlano, P. Ciuccarelli, P. Lloyd (eds.), DRS2024: Boston, 23–28 June, Boston, USA, 2024. DOI: https://doi.org/10.21606/drs.2024.600

[9] Martin P. Bendsøe, Ole Sigmund: *Topology optimization: Theory, methods and applications*. Springer Verlag, Berlin Heidelberg, 2003. ISBN: 3-540-42992-1, DOI: https://doi.org/10.1007/978-3-662-05086-6

[10] M. A. Kütük, İ. Göv: *A Finite element removal method for 3D topology optimization*. Advances in Mechanical Engineering, vol. 5, no. 12, January 1, 2013. DOI: https://doi.org/10.1155/2013/413463.

[11] A. L. Hatchuel: *What is generative in generative design tools? Uncovering topological generativity with a c-k model of evolutionary algorithms*. Proceedings of the Design Society, vol. 1, pp. 3419–3430, July 27, 2021. DOI: https://doi.org/10.1017/pds.2021.603.

[12] Francesco Buonamici, Monica Carfagni, Rocco Furferi, Yary Volpe, Lapo Governi: *Generative Design: An Explorative Study*. CAD-journal, vol. 18, no. 1, pp. 144–155, 2021. DOI: https://doi.org/10.14733/cadaps.2021.144-155.

[13] D. Vlah, R. Žavbi, N. Vukašinović: *Evaluation of topology optimization and generative design tools as support for conceptual design*. Proceedings of the Design Society: DESIGN Conference. vol. 1, pp. 451–460, Cambridge University Press, 2020. DOI: 10.1017/dsd.2020.165

[14] Q. Fossier: *Efficacy of generative design and topology optimization compared to traditional design processes using the case study of a cantilevered beam*. ASEE IL-IN Section Conference. Rose-Hulman Institute of Technology, 2025. Retrieved from https://www.rose-hulman.edu/asee-conference

[15] I. Prochner: *Call to move beyond Dieter Rams' Ten principles of good design*. Design and Culture, vol. 17, no. 2, pp. 225–235, 2025. DOI: https://doi.org/10.1080/17547075.2025.2488565.

[16] Y. Wang, Louis H. Sullivan: *His national architecture*. Journal of Art and Design, vol. 4, 2024. www.scipublications.org/journal/index.php/jad, DOI: 10.31586/jad.2024.966

[17] D. A. Gatenby, G. Foo: *Design for X (DFX): Key to competitive, profitable products*. AT&T Technical Journal, vol. 69: pp. 2–13, 1990. DOI: 10.1002/j.1538-7305.1990.tb00332.x

[18] A. R. Ismail, S. C. Abdullah, A. H. H. A. Manap, K. Sopian, M. M. Tahir, I. M. S. Usman, D. A. Wahab: *DFM and DFA approach on designing pressure vessel*. In Proceedings of the 7th WSEAS International Conference On System Science and Simulation in Engineering (ICOSSSE), pp. 147–151, November, 2008.

ENHANCING DIMENSIONALITY REDUCTION WITH FORMAL CONCEPT ANALYSIS

Fadi Atallah[1], ***László Kovács***[2]
[1]PhD Student, [2]Full Professor
[1,2]*Faculty of Mechanical Engineering and Informatics,*
Institute of Information Technology
[1]fadi.atallah@student.uni-miskolc.hu,
[2]laszlo.kovacs@uni-miskolc.hu

Abstract

This paper investigates the application of Formal Concept Analysis (FCA) as a method for feature reduction in machine learning tasks. High-dimensional datasets often contain redundant and correlated attributes that negatively affect model performance and computational efficiency. We propose an FCA-based reduction approach that leverages concept lattices and implication bases to identify dispensable and weak features while preserving the underlying conceptual structure of the data. The method was evaluated on a mushroom classification dataset using a multilayer perceptron model, where FCA successfully removed approximately 30 % of the features without reducing classification accuracy. However, experimental results show that FCA does not significantly improve training time or memory usage, and it struggles to detect negatively correlated attributes. Although FCA effectively identifies attribute implications, its computational cost and limited impact on overall model efficiency restrict its practicality for large-scale feature reduction tasks.

1. INTRODUCTION

In machine learning, the effectiveness of classification and regression depends strongly on the quality and dimensionality of input features. Real-world datasets often contain redundant, noisy, or irrelevant features due to correlation or overlap, which can degrade accuracy, increase computational cost, and complicate data visualization [1]. Including such features may lead to the curse of dimensionality, resulting in lower learning performance and unnecessary storage and processing overhead [1].

To address these issues, feature reduction techniques aim to remove or transform irrelevant inputs, improving accuracy, efficiency, model simplicity, and interpretability [2, 3]. This is particularly important for neural networks, which frequently operate on high-dimensional data and are prone to overfitting. Feature reduction generally includes feature selection, which identifies informative subsets of variables, and feature extraction, which constructs new composite features [4]. Feature selection methods are commonly classified as filter, wrapper, or embedded approaches, each balancing computational cost and selection quality while seeking minimal information loss [1].

Extensive research shows that feature reduction improves performance, speed, and generalization in high-dimensional settings. In neuroimaging, where datasets often contain far more features than samples, feature reduction is essential to prevent overfitting and enhance predictive accuracy [3]. Even advanced models such as deep neural networks benefit from prior feature selection. Studies have shown that deep models often achieve higher accuracy when trained on a reduced feature set, confirming that eliminating irrelevant inputs can simplify learning and produce more robust models [3, 5].

In this work, we focus on analyzing the efficiency of Formal Concept Analysis (FCA) for feature reduction. Building on existing theory and related studies, we propose an FCA-based approach that identifies and removes dispensable attributes while preserving the essential conceptual structure of the data. By exploiting the concept lattice and implication bases, our method

aims to improve the efficiency and interpretability of an MLP model using a smaller set of features. The remainder of this paper presents the necessary FCA background, describes the proposed reduction method, and evaluates its performance against established feature selection techniques.

2. FORMAL CONCEPT ANALYSIS (FCA)

Formal Concept Analysis (FCA), introduced by Rudolf Wille and Bernhard Ganter in the 1980s, is a field of applied mathematics based on applied lattice theory and order theory [6]. FCA's primary objective is to mathematically formalize the classical philosophical definition of a concept, thereby structuring and visualizing the inherent relationships between objects and attributes in a dataset [7]. FCA is built upon abstract algebraic structures derived from Galois connections between object sets and attribute sets. The fundamental input structure for FCA is the Formal Context K, defined as a triplet $K = \langle G, M, I \rangle$, where G is the finite set of Objects (or data elements), M is the finite set of Attributes (or features), and I is a binary relation $I \subseteq G \times M$, which explicitly depicts which objects possess which attributes.

The formal concept is defined by two complementary parts: extent and intent. The intent is the set of all attributes that are common to those objects:

$$A' = \{m \in M \mid \forall g \in A, (g,m) \in I\}, \text{ where } A \subseteq G. \tag{1}$$

The extent is the set of all objects that share a given group of attributes:

$$B' = \{g \in G \mid \forall m \in B, (g,m) \in I\}, \text{ where } B \subseteq M. \tag{2}$$

A Formal Concept is defined as a pair (A,B) where $A \subseteq G$ and $B \subseteq M$, satisfying the crucial closure conditions: $A' = B$ and $B' = A$ [8]. The set of all formal concepts generated from a context K, denoted $\mathfrak{B}(K)$, forms a Concept Lattice when equipped with the partial order relation defined by concept inclusion. The order relation dictates that (A_1, B_1) is less than or equal to (A_2, B_2) if $A_1 \subseteq A_2$ (meaning $B_2 \subseteq B_1$) [9]. This lattice structure systematically visualizes conceptual hierarchies, providing a robust structure for pattern discovery and knowledge representation, fundamentally linking abstract mathematical concepts with practical data analysis.

3. IMPLEMENTATION AND RESULTS

3.1. Methodology

We considered that FCA is a powerful method for selecting features to remove from the dataset because the concept lattice explicitly reveals relationships and redundancies among attributes. In the FCA concept lattice, if an attribute does not contribute to distinguishing any new concept, it is essentially redundant. This attribute is considered dispensable, and removing it from the dataset (context) does not change the lattice's concept structure. FCA provides mechanisms to identify and eliminate these redundant features. For example, algorithms exist to find a minimal attribute subset that defines the same set of concepts as the full attribute set (the reduced context has a concept lattice with the same shape as the concept lattice of the original context) [10]. By finding this minimal generating set of attributes, one achieves maximal dimensionality reduction while preserving all the original knowledge in the data. Furthermore, the implication base produced by FCA is invaluable for feature reduction. It encodes all attribute dependencies in the data in an irredundant way [11].

As the FCA required the input data to be in $I \subseteq G \times M$ form, as I is a binary relation, our method starts by checking the input data and transforming non-binary features into binary ones.

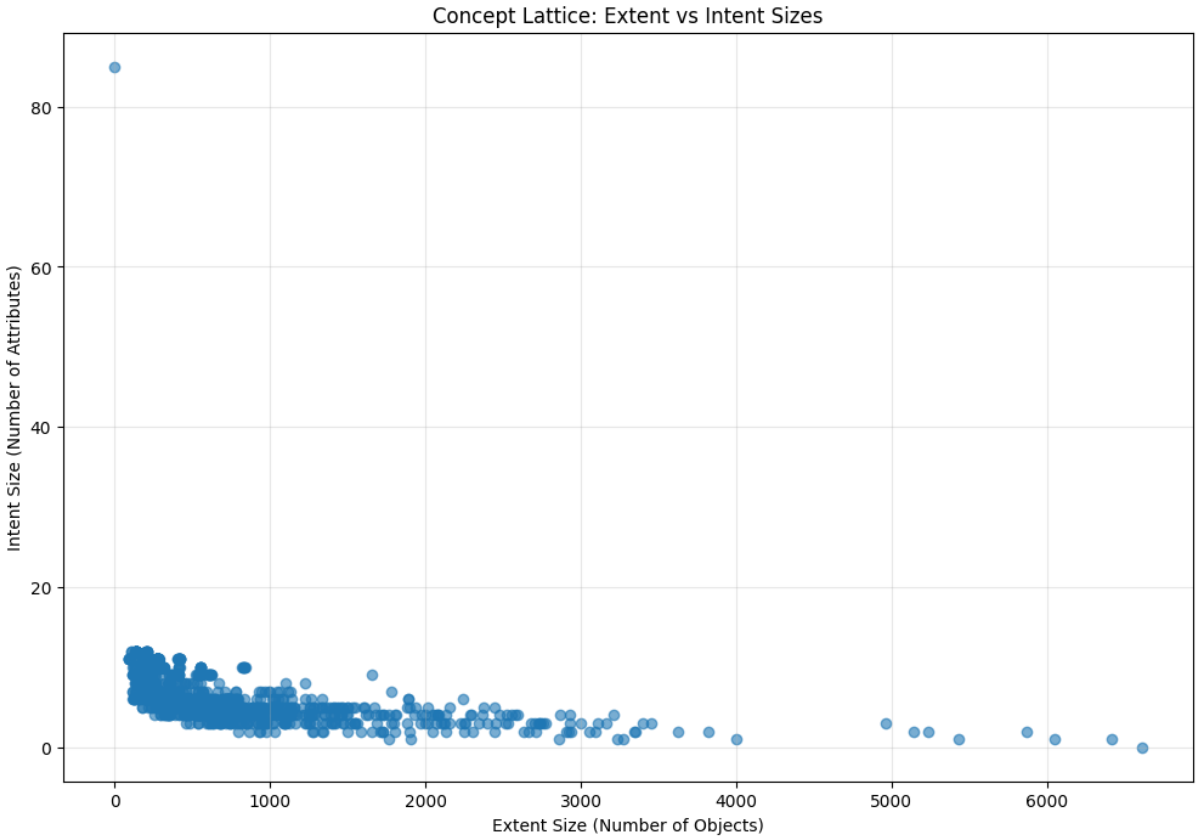


Figure 1: Concept Lattice Extent and Intent Size

Then we build the formal context to construct the concept lattice. After that, we go through the concept lattice, node by node, and compare each concept with its sub-concepts (children) in two approaches:

1. Check if two or more features were added to one of the sub-concepts at the same time. That means these features always occur together. In a mathematical sense, these features imply one another $((A \rightarrow B) \wedge (B \rightarrow A))$. Because of that, we can delete one of these features.

2. Check if the new feature in the sub-concept is a weak feature by checking the number of reduced objects. If the difference is less than a specific amount, we consider this feature weak because it provides little information and remove it from the dataset.

In the end, we evaluate the effect of feature reduction on the neural network training process, focusing on time and memory usage.

3.2. First Test (Test the effect of the feature reduction)

For our first test, we used a mushroom classification dataset [12] containing around 6000 mushroom types. After checking and modifying the dataset to be ready to use with FCA, we built the formal context and the concept lattice. Figure 1 shows the relation between the extents and the intents in the concept lattice (the more intent you have, the less extent you get).

Using the previously explained algorithm, the FCA removed 26 of 85 features (about 30 % of the total). Then we created a classification MLP model and trained it twice: once on the original dataset and once on the reduced dataset to see how the model's accuracy, training time,

Table 1: Compare the Original and Reduced Datasets

Dataset	Number of Features	Accuracy	Time	Memory
Original	85	1.0	11.878 s	17.09 MB
Reduced	59	1.0	11.864 s	16.92 MB

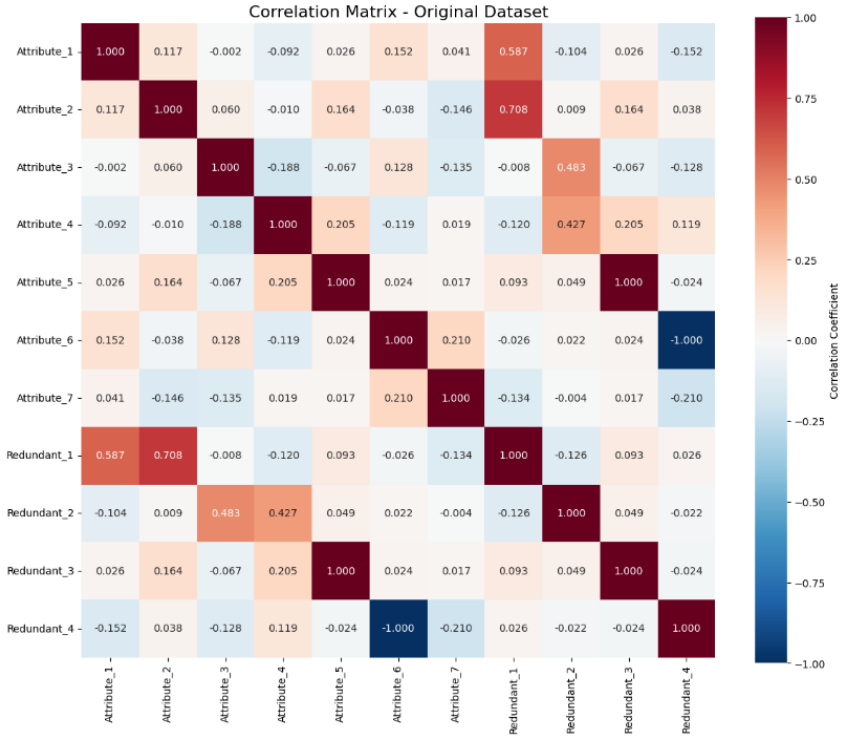


Figure 2: Correlation Matrix

and memory usage changed. Table 1 shows the mean values of the model's accuracy, time, and memory usage after multiple tests. We can see that the feature reduction did not affect the model's accuracy, as it was high in both tests. Also, there was no difference in time or memory usage between the original and reduced datasets.

3.3. Second Test (compare FCA with Correlation)

In the second test, we wanted to check whether the FCA could find correlated and redundant features. For that, we created a random binary matrix where the probability of 0 is 0.4 and for 1 is 0.6. After that we added some redundant features. Figure 2 shows the correlation matrix of the features, where we can see that we have two features that are entirely positively correlated, and another two are fully negatively correlated.

By applying FCA, the algorithm detected the highly positively correlated features. It removed them from the dataset, but it failed to detect the negatively correlated features, which makes FCA ineffective at detecting the implication between a feature and the complement of another feature ($A \rightarrow B'$).

4. CONCLUSION

In this article, we tested the effect of applying FCA in feature reduction. We found that using FCA is suitable for detecting feature implications in the dataset. Still, it is not effective for feature reduction, as it takes a very long time to build the formal context and the concept lattice

(this time increases very rapidly as the dataset grows), with no effect on the model's training time or memory usage. Also, we found that FCA is weak at detecting negatively correlated features.

REFERENCES

[1] H. K. Palo, S. Sahoo, A. K. Subudhi: *Dimensionality reduction techniques: Principles, benefits, and limitations*. Data Analytics in Bioinformatics, John Wiley & Sons, Ltd, pp. 77–107, 2021. DOI: 10.1002/9781119785620.ch4

[2] S. Khalid, T. Khalil, S. Nasreen: *A survey of feature selection and feature extraction techniques in machine learning*. 2014 Science and Information Conference, pp. 372–378, Aug. 2014. DOI: 10.1109/SAI.2014.6918213

[3] B. Mwangi, T. S. Tian, J. C. Soares: *A review of feature reduction techniques in neuroimaging*. Neuroinformatics, vol. 12, no. 2, pp. 229–244, Apr. 2014. DOI: 10.1007/s12021-013-9204-3

[4] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, J. Saeed: *A Comprehensive review of dimensionality reduction techniques for feature selection and feature extraction*. Journal of Applied Science and Technology Trends, vol. 1, no. 1, pp. 56–70, May 2020. DOI: 10.38094/jastt1224

[5] Z. Chen, et al.: *Feature selection may improve Deep Neural Networks for the bioinformatics problems*. Bioinformatics, vol. 36, no. 5, pp. 1542–1552, Mar. 2020. DOI: 10.1093/bioinformatics/btz763

[6] B. Ganter, R. Wille: *Formal Concept Analysis: Mathematical foundations*. Cham: Springer Nature Switzerland, 2024. DOI: 10.1007/978-3-031-63422-2

[7] J. Galasso: *Formal Concept Analysis: a structural framework for variability extraction and analysis*. arXiv: arXiv:2508.06668, Aug. 8, 2025. DOI: 10.48550/arXiv.2508.06668

[8] R. Wille: *Formal Concept Analysis as mathematical theory of concepts and concept hierarchies*. Formal Concept Analysis, vol. 3626, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–33, 2005. DOI: 10.1007/11528784_1

[9] A. K. Sarmah, S. M. Hazarika, S. K. Sinha: *Formal Concept Analysis*: Current trends and directions. Artif Intell Rev, vol. 44, no. 1, pp. 47-86, Jun. 2015. DOI: 10.1007/s10462-013-9404-0

[10] E. Zhang: *A fast algorithm for attribute reduction of Formal Concept Analysis*.

[11] V. Snasel, M. Polovincak, H. M. Dahwa, Z. Horak: *On concept lattices and implication bases from reduced contexts*.

[12] *Mushroom classification*. Available: `https://www.kaggle.com/datasets/uciml/mushroom-classification`

ADVANCEMENTS IN KOLMOGOROV-ARNOLD NETWORKS: A SYSTEMATIC REVIEW OF ARCHITECTURES AND APPLICATIONS

Moenes Benaissa[1], ***László Kovács***[2]
[1]PhD Student, [2]Full Professor
[1,2]*Faculty of Mechanical Engineering and Informatics,*
Institute of Information Technology
[1]moenesb7@gmail.com, [2]laszlo.kovacs@uni-miskolc.hu

Abstract

Kolmogorov Arnold Networks (KANs) are the new replacement for the conventional multiplayer perceptrons (MLPs), creating a huge drift in the neural network design [12, 14]. While the supremacy of the MLP is based on the universal approximation theorem, utilizing fixed activation functions on nodes and linear weights on edges. KAN uses the Kolmogorov-Arnold theorem [1], by swapping the fixed weights with learnable univariate functions normally configured as B-splines on the edge of the network. Resulted in KAN having a better parameter efficiency and "white box" explainability [14, 21]. This article gives a structured review of the KAN environment including its architectural, applications, and the shortcomings.

1. INTRODUCTION

The "black-box" nature of deep learning is still a big hurdle in sensitive fields like genomics and physics-informed modeling. Traditional MLPs offer high accuracy yet no clear mathematical explanation for their internal logic [14]. The emergence of KANs in the early years addressed this by permitting the network to learn the best activation shape for each connection. Since the original framework's release, research has quickly expanded the KAN family to overcome its initial limitations, especially when it comes to computational speed and scalability [1, 13]. This review compiles 21 key papers to present how these networks have evolved from theoretical concepts to practical tools for scientific discovery.

2. THEORATICAL FOUNDATION OF KANS

The unique structure of KANs is based on the elimination of the linear weight matrix. Instead, a KAN layer is formed of a matrix of 1D functions. For a layer with inputs and outputs, the operation is presented as a sum of univariate functions applied to each input component [1, 14].

2.1. The Kolmogorov-Arnold Representation Theorem

The mathematical foundation for KANs says that any continuous multivariate function f on a bounded domain can be represented as [1, 14]:

$$f(x_1,\ldots,x_n) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right). \tag{1}$$

In this formulation, $\phi_{q,p}$ represents the inner univariate functions (edges) and Φ_q represents the outer functions [14, 20]. Unlike MLPs, where the activation function is static (e.g., ReLU), KANs treat these functions as learnable splines that can adapt to the local geometry of the data distribution [14, 17].

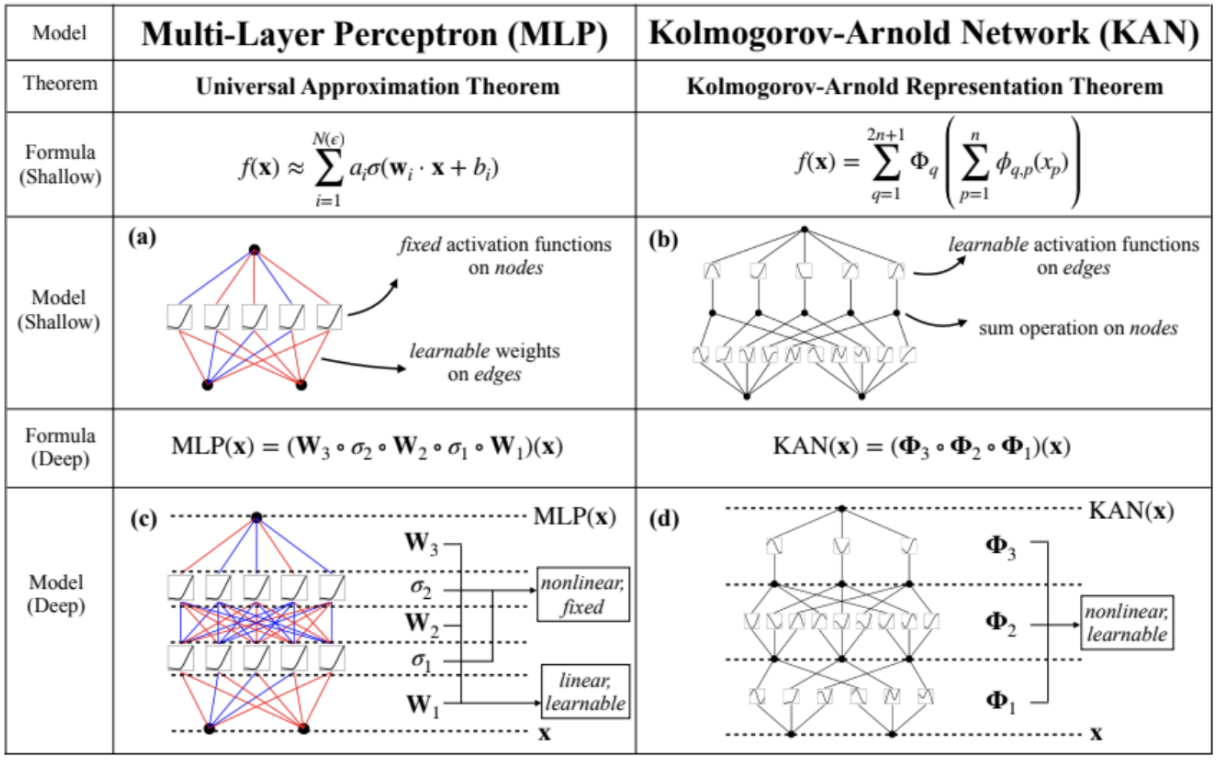


Figure 1: Multi-Layer Perceptrons (MLPs) vs. Kolmogorov-Arnold Networks (KANs) (adapted from [12]).

In the beginning, KAN implemented each univariate function ϕ as a B-spline [12, 14]. This allowed local control over function approximation, where the function's shape is determined by an ensemble of control points on a grid [12]. Unlike global polynomial approximations, B-splines are piecewise, which prevents small local changes from influencing the function's behavior, as a result improving stability during the learning process [1, 14].

2.2. Grid Dependency and Neural Scaling Laws

A critical feature of KANs is their relationship with the underlying grid [14]. The approximation accuracy of a KAN can be refined by increasing the number of grid intervals, a process referred to as grid extension [1, 14]. This allows KANs to follow faster neural scaling laws than MLPs [14, 21]. Specifically, for a smooth -dimensional function, KANs reaches an approximation error of $O(N^{-4/d})$ with a spline order of 3. Meanwhile, MLPs are usually limited by the curse of dimensionality [14]. This makes KANs are particularly effective for scientific tasks where high precision is required with a minimal number of parameters [11, 14].

2.3. ARCHITECTURAL SOLUTIONS AND EVOLUTIONS

Because of the non-parallelizable nature of B-spline recursions, even despite the theoretical advantage, the first generation of KANs struggled with practical issues, such as slow training and inference speeds [13, 19]. To minimise this gap, a lot of architectural variants were developed between 2024 and 2025 to optimize the core KAN operation for modern hardware [1, 13]. Efficiency and Speed Optimization.

2.4. Efficiency and Speed Optimization

The first breakthrough was the development of FastKAN [8]. This variant demonstrates that 3rd-order B-splines can be approximated very well by Gaussian Radial Basis Functions (RBFs) [8, 21]. By treating KANs as RBF networks with fixed centers, FastKAN implementation is significantly faster and more compatible with existing deep learning frameworks [8]. Further improvements in parallelization led to MatrixKAN and ReLU-KAN [13, 19]. MatrixKAN uses Toeplitz matrices to show polynomial functions, so it can treat the calculations of B-splines as standard matrix operations [13]. This method reported speedups of approximately 40x relative to the original implementation [13]. Similarly, ReLU-KAN simplified the architecture by replacing complex splines with a combination of matrix additions, dot products, and ReLU activations, enabling efficient GPU parallelization [19].

2.5. Advanced Basis Functions

Function approximation's stability in high-degree space still a very important to investigate. ChebyKAN tackle the inherent oscillations of high-degree B-splines by using Chebyshev polynomials (T_n) as basis functions [4]. These polynomials are defined by the recurrence relation:

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{2}$$

Thanks to the minimax approximation property, ChebyKAN minimizes the maximum deviation from the target function in the interval $[-1,1]$, improving the numerical stability during high-order fitting [4]. Furthermore, to ease the extraction of multi-scale features, Wav-KAN integrates wavelet transforms into the KAN framework [17]. The univariate functions are constructed through a mother wavelet ψ that undergoes dilation and translation:

$$\phi(x) = \sum_{j,k} w_{j,k}\psi_{j,k}(x), \quad \psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k). \tag{3}$$

This multiresolution analysis allows the architecture to capture both high-frequency transients and low-frequency trends more effectively than traditional spline-based KANs [17].

3. SPECIALIZED APPLICATIONS OF KANS

The capability that KANs possess in being able to break down complex functions involving multiple variables to interpretable univariate components, which may have contributed to their quick uptake in scientific contexts [11, 14]. As opposed to MLP networks, which are often described as a "black box", a KAN aid in visualizing the activation functions learned, enabling symbolic regression and the discovery of hidden physical laws [9, 11].

3.1. Genomics and Computational Biology

In the realm of genomic studies, KANs have been employed for sequence classification and generation tasks [7]. Linear KANs (LKANs) and Convolutional KANs have the capacity to outperform the current best CNN models on genomic benchmarks, such as the Flipon and Genome Understanding Evaluation datasets [7, 12]. This success is attributed to KANs for their superior capacity in capturing non-linear biological motifs with fewer parameters [7].

3.2. Physics-Informed Modeling and Dynamical Systems

KANs have brought a great deal of promise in solving partial differential equations (PDEs) and modeling dynamical systems, and much more [9, 20]. Physical constraints are being directly integrated into the neural network's loss function by KINN (informed Kolmogorov-Arnold). Which it did provide high-precision solutions for forward-inverse problems in fluid mechanics and elasticity [20]. Add to that, KAN-ODEs generalize and then extend the architecture to time-dependent systems and perform the reconstruction of trajectories for complex models, such as the Lotka-Volterra predator-prey system with higher accuracy and better application performance, scaling compared to traditional Neural ODEs [9].

3.3. Time Series Forecasting

For Temporal data, variants like T-KAN and AR-KAN were developed to counter the obstacles of concept drift and non-commensurate frequencies [3, 6]. AR-KAN, more specifically, incorporates an auto-regressive component for handling temporal memory into a KAN for non-linear representation, which performs equally well as the classic ARIMA model on almost-periodic functions while simultaneously preserving the flexibility of deep learning [3].

3.4. Graph Neural Networks and Transformers

The Kolmogorov-Arnold Network for Graphs (KANG) did benefit the non-Euclidean data by expanding its flexibility [16]. KANG consistently outperforms architectures like GCN (Graph Convolutional Network) and GAT (Graph Attention Network) on node classification tasks by incorporating learnable splines into the message-passing framework, which solves the over-smoothing issue common in conventional Graph Neural Networks (GNNs) [16]. Additionally, the Kolmogorov-Arnold Transformer (KAT) substitutes Gated-Rational KAN layers for the conventional MLP layers in the Transformer architecture [14]. This integration has shown a higher performance on larger-scale benchmarks in compression to Vision Transformer (ViT) notably in ImageNet. Where KAT achieves higher accuracy with fewer parameters compared to the Vision Transformer (ViT) [14, 15].

4. SHORTCOMINGS AND CHALLENGES

Despite the rapid progresses in KAN, certain challenges hinder their adoption in replacing MLPs in deep learning. These drawbacks can be broadly classified into computational overhead and optimization complexity, and hardware compatibility [1, 13]. One of the main drawbacks is *the computational cost* involved in the evaluation of the B-spline. Although variants of it, like MatrixKAN and FastKAN, provide substantial speed enhancements, base KAN architectures tend to be slower than MLPs in terms of training and inference speed [1, 13, 21]. This issue is further complicated by the fact that there are no direct spline operations, as the current hardware is mainly optimized for the dense matrix multiplications in MLPs [1, 19]. From an *optimization* point of view, KANs are sensitive to hyperparameters such as the size of the grid or the order of the splines [2, 11]. Contrary to the case with MLPs, KAN typically requires careful data-aligned initialization to prevent vanishing or exploding gradients [16]. Moreover, KANs prove to be very effective on a small-scientific tasks, their scalability to "foundation model" scales is unproven, as the number of univariate functions tends to grow dramatically with the size of the network in both width and depth [1, 7, 12].

5. CONCLUSION

Kolmogorov-Arnold Networks represent the most important step to date toward "white-box" deep learning. By transferring the learning capacity from nodes to edges, KAN provides a unique synergy between symbolic science and connectionist AI. Though not as advanced in training speed or hardware acceleration currently, their superior performances in genomics, physics, and time-series analysis make them a requisite tool in the near future for interpretable artificial intelligence.

REFERENCES

[1] S. Somvanshi, et al.: *A survey on Kolmogorov-Arnold Network*. arXiv preprint, 2024.

[2] H. Ta, A. Tran: *AF-KAN: Activation Function-based Kolmogorov-Arnold Networks*. arXiv preprint, 2025.

[3] C. Zeng, et al.: *AR-KAN: Autoregressive-Weight-Enhanced KAN for time series*. arXiv preprint, 2025.

[4] S. S. Sidharth, et al.: *Chebyshev Polynomial-based KANs*. arXiv preprint, 2024.

[5] A. D. Bodner, et al.: *Convolutional Kolmogorov-Arnold Networks*. arXiv preprint, 2024.

[6] K. Xu, et al.: *KAN for Time Series: Predictive power and interpretability*. arXiv preprint, 2024.

[7] O. Cherednichenko, M. Poptosva: *KANs for genomic tasks*. bioRxiv, 2024.

[8] Z. Li: *Kolmogorov-Arnold Networks are Radial Basis Function Networks*. arXiv preprint, 2024.

[9] B. C. Koenig, et al.: *KAN-ODEs: Learning Dynamical Systems*. arXiv preprint, 2024.

[10] F. Alesiani, et al.: *Variational Kolmogorov-Arnold Network*. arXiv preprint, 2025.

[11] Z. Liu, et al.: *KAN 2.0: Kolmogorov-Arnold Networks meet science*. arXiv preprint, 2024.

[12] Z. Liu, et al.: *KAN: Kolmogorov-Arnold Networks*. arXiv preprint, 2024.

[13] C. Coffman, L. Chen: *MatrixKAN: Parallelized Kolmogorov-Arnold Network*. arXiv preprint, 2025.

[14] X. Yang, X. Wang: *Kolmogorov-Arnold Transformer*. arXiv preprint, 2024.

[15] L. Hu, et al.: *Incorporating matrix group equivariance into KANs*. arXiv preprint, 2025.

[16] G. De Carlo, et al.: *Kolmogorov-Arnold Graph Neural Networks*. arXiv preprint, 2025.

[17] Z. Bozorgasl, H. Chen: *Wav-KAN: Wavelet Kolmogorov-Arnold Networks*. arXiv preprint, 2024.

[18] G. Lin, et al.: *Energy-Dissipative Evolutionary KANs (EvoKAN)*. arXiv preprint, 2025.

[19] Q. Qiu, et al.: *ReLU-KAN: Parallel Computation KANs*. arXiv preprint, 2024.

[20] Y. Wang, et al.: *Kolmogorov-Arnold-Informed Neural Network (KINN)*. arXiv preprint, 2024.

[21] A. Polar, M. Poluektov: *Probabilistic Kolmogorov-Arnold Network*. arXiv preprint, 2024.

TÖBBOSZTÁLYÚ ANOMÁLIAÉRZÉKELÉS IPARI ROBOTKARBAN

Bodnár Dávid[1], ***Jármai Károly***[2]
[1]PhD hallgató, fejlesztőmérnök, [2]Egyetemi tanár
[1]*Emerson Automation FCP Kft., Eger*
[1,2]*Gépészmérnöki és Informatikai Kar, Energetikai és Vegyipari Gépészeti Intézet*
[1]bodnar.david2@gmail.com, [2]karoly.jarmai@uni-miskolc.hu

Absztrakt

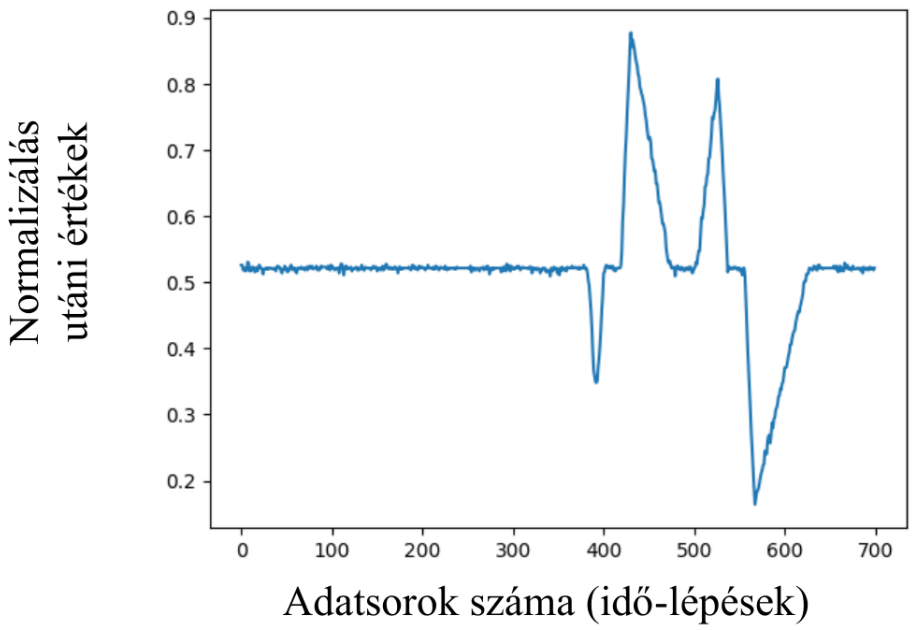
A rendellenességek észlelésére szolgáló hagyományos mélytanulási modellek gyakran jelentős számítási erőforrásokat igényelnek. Ezzel szemben az itt bemutatott PINC-modell csatornánkénti párhuzamos jellemzőkivonást alkalmaz a többszintű fizikai rendellenességek nagy pontosságú azonosítására. Az eredmények azt mutatják, hogy a PINC modell hatékonyan rögzíti a rezgési jellemzőket anélkül, hogy nagy késleltetésű felhőalapú feldolgozásra lenne szüksége. Az eredmények egy skálázható és hatékony hardveres megoldást kínálnak, amely lehetővé teszi a lokalizált, valós idejű hibajelzést az ember és robot által megosztott munkaterületeken.

1. BEVEZETÉS

A kollaboratív robotok integrálása a modern gyártásba rugalmasabb, emberközpontú gyártósorokat tett lehetővé. [1] A Universal Robots UR3e egy piacvezető platform ezen a területen, amelyet gyakran használnak olyan feladatokhoz, mint például a precíz összeszerelés és a laboratóriumi automatizálás. Azonban ezeknek a gépeknek a kollaboratív jellege, amely gyakori feladatváltásokat és az emberi operátorok közelségét jelenti, egyedülálló kihívásokat jelent a hagyományos állapotfigyelési és karbantartási módszerek számára. Ezen rendszerek mechanikai állapotának biztosítása elengedhetetlen a váratlan leállások minimalizálása és a működés folytonosságának fenntartása szempontjából.

Az ipari manipulátorok esetében a hagyományos karbantartási protokollok gyakran időszakos kézi ellenőrzéseken vagy a belső motoráram-érzékelők egyszerű küszöbérték-alapú riasztásain alapulnak. Bár ezek a módszerek beváltak, előfordulhat, hogy nem elég érzékenyek ahhoz, hogy a hardver meghibásodásához vezető apró fizikai rendellenességeket észleljék. A mesterséges intelligencia területén elért legújabb eredményeknek köszönhetően az 1D konvolúciós neurális hálózatok (1D-CNN) automatizált vibráció elemzésre alkalmas eszközként kerültek bemutatásra. [2] Az inerciális mérőegység (IMU) nyers adataiból közvetlenül tanulva, ezek a modellek képesek azonosítani a mechanikus igénybevételt vagy működési hibákat jelző mintákat. [3, 4, 5, 6]

A robotcellákban alkalmazott mélytanulási modellek ígéretesek, de gyakran korlátozzák őket a rendelkezésre álló számítási kapacitások. Ez a tanulmány az UR3e platformhoz kifejlesztett Parallel Inception-stílusú 1D-CNN (PINC) architektúra bemutatásával, teljesítményének vizsgálatával kíván hozzájárulni a területet érintő kutatáshoz. A CASPER [7] adatkészlet felhasználásával a kutatás azt vizsgálja, hogy a csatornánkénti párhuzamos megközelítés hogyan képes többszintű rezgési jellemzőket rögzíteni. A javasolt modell egy egyszerű diagnosztikai segédeszközként kerül bemutatásra, amely illeszkedni igyekszik az ipari peremeszközök memóriakapacitásának korlátjaihoz. A következő szakaszok részletesen ismertetik az IMU-n keresztül rögzített jelek szegmentálásának és normalizálásának módszertanát, valamint a modell fizikai rendellenességek azonosításában nyújtott teljesítményének értékelését.



1. ábra. Egy tipikus ablak az adathalmazban

2. ADATFELDOLGOZÁS ÉS VALIDÁCIÓ

Ez a tanulmány a CASPER 1 és 2 IMU [7] adatkészleteket használja. A bemeneti jellemzők így hat csatornából állnak, amelyek közül három (Ax, Ay, Az) a gyorsulásmérőből származik. Továbbá egy giroszkóp szolgáltatja a szögsebesség adatokat (Gx, Gy, Gz). Az érzékelőadat a robot IMU-ján keresztül nyerhetőek ki.

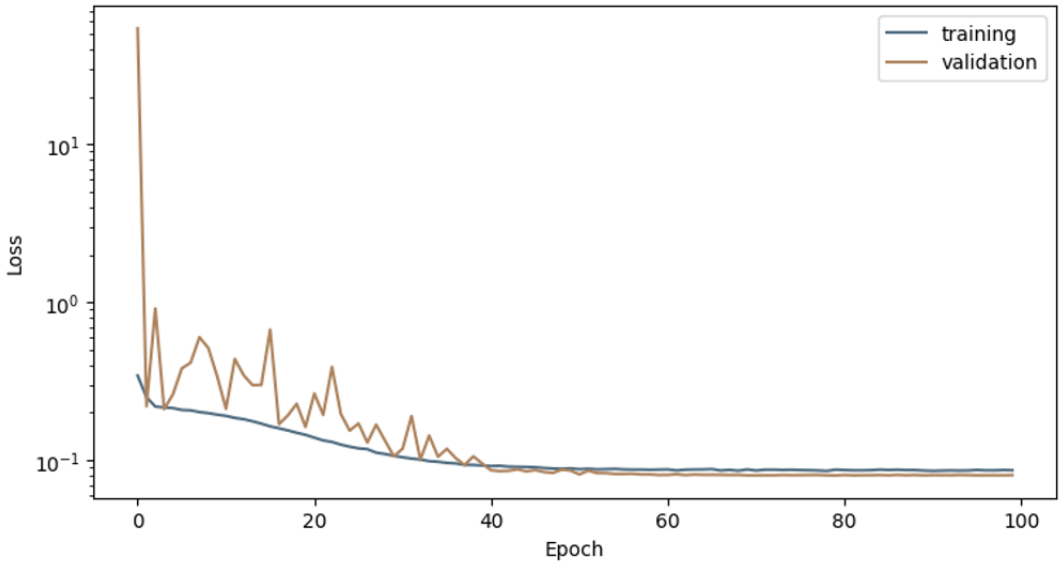
A nyers érzékelőadatok feldolgozásához hat csatorna mindegyikét függetlenül normalizálni szükséges. Ez biztosítja, hogy a különböző gyorsulási és szögsebesség-skálák ne befolyásolják a modellt. A második szakasz az ablakosítás. A folyamatos idősoros adatokat 700 időlépésből álló, egymást átfedő ablakokra osztom. Egy ilyen ablakot normalizálás után az első ábra mutat be. Ezt az ablakméretet úgy választottam meg, hogy mind az átmeneti, mind a hosszabb ideig tartó rezgések lehetőleg teljes mértékben egyetlen bemeneti tenzorba legyenek beágyazva. Minden ablakot „Normális” vagy „Rendellenes működés” címkével látok el az ablak utolsó időlépésében rögzített állapot alapján. Az adatokból ezekkel a lépésekkel a gépi tanulás számára olvasható és értelmezhető adathalmazt képeztem.

A párhuzamos Inception-stílusú konvolúciós (PINC) modell úgy lett kialakítva, hogy minimális paraméterszámmal is működőképes legyen. A szekvenciális architektúráktól eltérően, amelyek egyetlen szűrőhalmazon keresztül dolgozzák fel az összes csatornát, a PINC modell csatornánkénti párhuzamos megközelítést alkalmaz. Minden IMU-csatorna feldolgozása kezdetben a saját párhuzamos Inception-stílusú blokkján keresztül történik. Minden blokk 1×1-es szűk keresztmetszetű konvolúciót használ, amelyet tömeges normalizáció és ReLU aktiválás követ. Ez a struktúra hatékonyan építi fel az temporális mélységet, miközben alacsonyan tartja a paraméterek számát. A csatornánkénti jellemzők felismerését követően a modell globális átlagos összegzést (GAP) alkalmaz. A nagyobb architektúrákban található globális Inception blokkok kihagyásával a modell komplexitása 38 946 paraméterre csökken, ami 152 KB memóriát igényel ebben az esetben.

A modellt a Keras API [8] segítségével, TensorFlow [9] háttérrel építettem fel. A lokális minimumokba ragadás elkerülése céljából exponenciálisan csökkenő tanulási sebességet alkalmaztam, amely 0,01-től indult és 20 epoch alatt 0,1-gyel csökkent. A modellt 100 epoch alatt, 32-es kötegmérettel tanítottam, a tanulási és validációs halmazokat $70/30$ arányban elosztva. Az adathalmaz elosztása azért szükséges, hogy a modell lehetőleg valódi összefüggéseket talál-

1. táblázat. A bemutatott modell adatai

Kategória	**Paraméter**	**Részletek**
	Típus	1D-CNN - párhuzamos
Modell felépítés	Paraméterek száma	38946
	Szükséges memória	152 KB
Adatok	Bemenet	700×6 [adatsor lépések $\times$ csatornák]
Teljesítmény	Pontosság	99,1 %



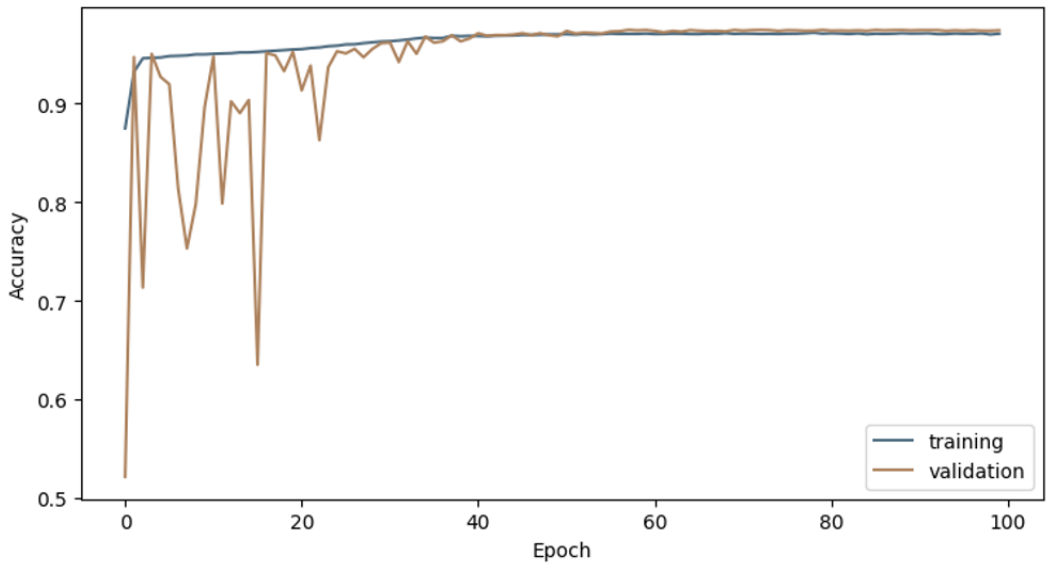
2. ábra. Veszteségfüggvény

jon, és ne csak a validációs halmazban található mintákat tanulja meg. A bemutatott módszerrel felépített modell adatait az 1. táblázat mutatja be.

A PINC modell 99,1 %-os validációs pontosságot ért el a CASPER adatkészleten. A pontosság és veszteségfüggvény értékeinek változását a tanítási folyamat során a harmadik ábra mutatja be. Ez a magas pontosság azt jelzi, hogy a csatornánkénti párhuzamos architektúra sikeresen képes kimutatni a rendellenességekhez kapcsolódó jeleket. Amint az a képzési előzményekből is látható, az exponenciálisan csökkenő tanulási sebesség elősegítette a stabil konvergenciát, megakadályozva, hogy a modell az utolsó epochok során a lokális minimumok körül ingadozzon.

A standard szekvenciális 1D-CNN-ekhez vagy nagyobb keretrendszerekhez képest a PINC modell jelentősen kisebb helyigényű lehet a gyakorlatban. Az állapotfelügyelet területén ez a hatékonyság kritikus fontosságú. A legtöbb ipari IoT-átjáró és robotvezérlő egység korlátozott RAM-mal rendelkezik a nem kritikus háttérfeladatok számára. A 152 KB-os modell közvetlenül a gyártócellában futtatható, így csökken a nagy sávszélességű adatátvitel igénye a központi szerverre, és szinte azonnali anomáliajelzés válik lehetővé. Ehhez természetesen minden esetben szükséges a bemutatott modell újratanulása az adott cellára jellemző adatok figyelembevételével.

Karbantartási szempontból a modell másodlagos megfigyelőként működhet. Míg az UR3e ugyan rendelkezik belső biztonsági rendszerekkel, de a bemutatott modell az integrált rendszert kiegészítve képes lehet azonosítani a hasznos teher nem várt eltéréseit vagy a kisebb ütközéseteket, amelyek még nem is biztos, hogy meghaladják a robot saját, beépített védelmi küszöbértékeit. Ezenkívül, mivel a PINC modell párhuzamosan dolgozza fel az IMU csatornákat, megőrzi



3. ábra. Pontosság és veszteségfüggvény

az érzékelőadatok térbeli integritását. Ez lehetőséget adhat arra, hogy a jövőben nemcsak az anomáliák észlelésére használhassuk, hanem annak meghatározására is, hogy melyik tengely vagy csukló terhelése a legnagyobb.

3. ÖSSZEFOGLALÁS

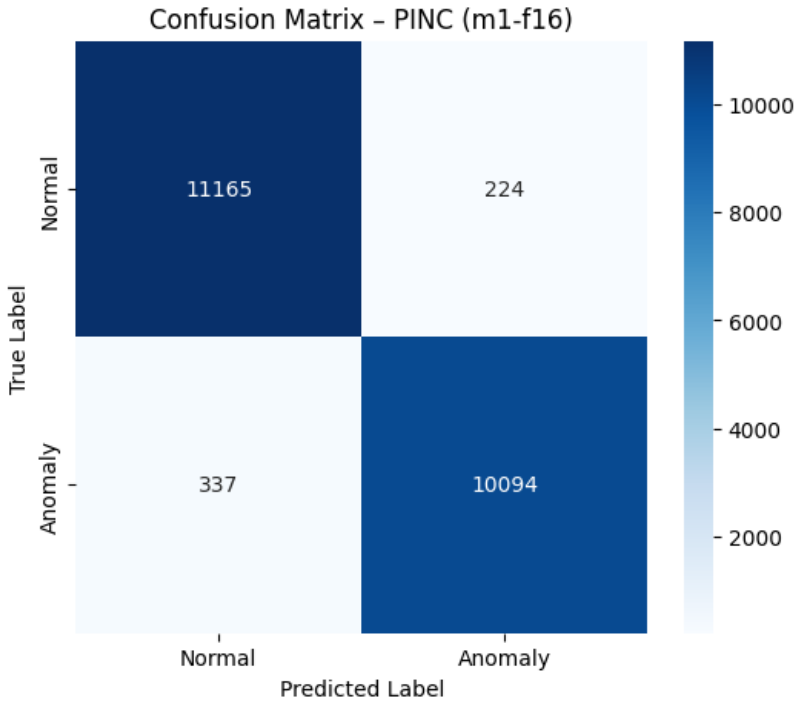
A bemutatott párhuzamos Inception stílusú konvolúciós (PINC) architektúra alkalmas az Universal Robots UR3e kollaboratív platform állapotának figyelemmel kísérésére. A csatornánkénti, párhuzamos stratégia alkalmazásával a modell sikeresen észleli a fizikai rendellenességek jellemzőit, miközben a memória szükséglete minimális.

A CASPER IMU adatkészleten validált kísérleti eredmények azt mutatják, hogy a PINC modell 99,1 %-os diagnosztikai pontosságot ér el, 38 946 paraméter felhasználásával. A 700 időlépéses ablakos stratégia és az exponenciálisan csökkenő tanulási ráta alkalmazása hatékonynak bizonyult. Ipari szempontból a PINC modell alkalmas lehet peremhálózati telepítésre. Alacsony követelményei lehetővé teszik, hogy akár ipari mikrokontrollerekre is telepíthető legyen, megkönnyítve a valós idejű monitorozást a külső felhőalapú feldolgozással járó késleltetés és biztonsági aggályok nélkül. Bár nem helyettesíti az elsődleges biztonsági rendszereket, hasznos diagnosztikai eszközként szolgálhat a megfelelő alkalmazásban.

További kutatással a modell kiterjeszthető zaj, a hasznos teher eltolódás és kopás megkülönböztetésére. Ezenkívül a cél azt is megvizsgálni, hogy az UR3e-n előre betanított PINC modell hogyan alkalmazható nagyobb platformokra, mint például amilyen az UR10e.

4. KÖSZÖNETNYILVÁNÍTÁS

A C2307874 számú projekt a kulturális és innovációs minisztérium nemzeti kutatási fejlesztési és innovációs alapból nyújtott támogatásával, a KDP-2023 pályázati program finanszírozásában valósult meg.



4. ábra. Konfúziós mátrix

IRODALOMJEGYZÉK

[1] M. E. Moran: *Evolution of robotic arms*. In Journal of Robotic Surgery, Springer London, pp. 103–111, 2007.

[2] C. Szegedy et al.: *Going deeper with convolutions*, September 2014. [Online]. Elérhető: `http://arxiv.org/abs/1409.4842`

[3] D. Łuczak: *Data-Driven machine fault diagnosis of multisensor vibration data using synchrosqueezed transform and time-frequency image recognition with Convolutional Neural Network*. Electronics (Switzerland), vol. 13, no. 12, June 2024.

[4] Y. W. Kim, K. L. Joa, H. Y. Jeong, S. Lee: *Wearable imu-based human activity recognition algorithm for clinical balance assessment using 1D-CNN and GRU ensemble model*. Sensors, vol. 21, no. 22, November 2021.

[5] K. Vijayalakshmi, A. Rajakannu, M. Kamarudden, K. P. Ramachandran, A. V. Sri Rajkavin: *Intelligent fault diagnosis of rotating machinery using Deep Learning algorithms: A comparative analysis of MLP, CNN, RNN, and LSTM*. SSRG International Journal of Electrical and Electronics Engineering, vol. 11, no. 9, pp. 294–315, September 2024.

[6] A. Athar, M. A. I. Mozumder, S. Ali Abdullah, H. C. Kim: *Deep learning-based anomaly detection using one-dimensional Convolutional Neural Networks (1D CNN) in Machine CenTers (MCT) and Computer Numerical Control (CNC) machines*. PeerJ Comput. Sci., vol. 10, 2024.

[7] Hakan Kayan: *Industrial robotic arm - IMU data* (CASPER 1 & 2) – `https://www.kaggle.com/datasets/hkayan/industrial-robotic-arm-imu-data-casper-1-and-2`

[8] InceptionV3 model. Elérhető: https://keras.io/api/applications/inceptionv3/inception_v3_model/

[9] API Documentation, TensorFlow v2.16.1. Elérhető: https://www.tensorflow.org/api_docs

EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR TIME SERIES FORECASTING AND ANOMALY DETECTION IN HEALTHCARE: REVIEW

Kawtar Dhaidah[1], ***Samad Dadvandipour***[2]
[1]PhD Student, [2]Associate Professor
[1,2]*Faculty of Mechanical Engineering and Informatics,*
Institute of Information Technology
[1]kawtardhaidah48@gmail.com, [2]samad.dadvandipour@uni-miskolc.hu

Abstract

Time series forecasting and anomaly detection are major challenges in healthcare, as they enable the early detection of clinical deterioration, more efficient resource allocation, and the implementation of proactive interventions. While all these deep learning models, especially the LSTM, GRU, and Transformer architectures, have demonstrated remarkable predictive performance in these two tasks, their lack of interpretability limits their use in clinical practice. Explainable Artificial Intelligence has been recently proposed as a way to overcome such limitations by providing understandable frameworks that make the decisions of complex models understandable and justifiable. This work offers a comprehensive literature review of XAI methods applied to time series forecasting and anomaly detection, with a particular focus on medical applications. This analysis highlights several important shortcomings, including the limited applicability of most techniques to sequential healthcare data, the lack of accurate evaluation of the quality and reliability of interpretation, and the insufficient integration of forecasting and anomaly detection frameworks. In this context, we discuss future research directions that aim to design interpretable models that combine high predictive accuracy with actionable and clinically meaningful interpretations.

1. INTRODUCTION

Healthcare systems constantly generate vast amounts of time-series data. Extracting useful information from this data is highly important for monitoring patient conditions, predicting medical outcomes, and identifying anomalies among patients. Therefore, effective analysis of healthcare data series is essential for ensuring patient safety, supporting informed medical decisions, and improving the overall quality of healthcare.

Traditional statistical models, such as ARIMA, SARIMA, and Holt-Winters models, have long been the standard for analyzing linear and stationary time series [1]. However, the nonlinear, complex, and high-dimensional dynamics that characterize real-world clinical data are still a challenge for these models to capture. These limitations are partly overcome by approaches of classical machine learning, which rely explicitly on modeling long-term dependencies or using feature engineering. More recently, deep learning models, such as Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Transformer, have emerged as powerful solutions able to learn complex temporal patterns directly from data [2–4]. Despite their strong predictive performance, all these models often operate as "black boxes," making it difficult to interpret their predictions or understand from them what the cause could be of an anomaly alert, at least not in clinical domains where transparency and accountability are indispensable.

As a result, Interpretable Artificial Intelligence (XAI) has received increasing attention with the aim of improving the interpretability of complex prediction models. Some popular XAI frameworks, such as SHAP and LIME [5], explain why a model makes a particular prediction, and additional information about the behavior of deep learning models can be obtained using saliency maps and correlation propagation techniques [6,7]. However, most current methods are designed to be useful for tabular data or still images, but they are unable to measure the temporal

information contained in time-series data [8]. Furthermore, prediction tasks are usually considered separate from anomaly detection tasks, but in clinical settings, both are critically important. In response to these challenges, this work provides a comprehensive review of interpretable artificial intelligence approaches applied to time series prediction and anomaly detection, with a focus on healthcare applications.

2. TIME SERIES FORECASTING AND ANOMALY DETECTION IN HEALTHCARE

A crucial aspect of time series data is its use in various applications within the healthcare industry. These include patient vital signs, such as heart rate, blood pressure, continuous glucose monitoring, ICU telemetry, and wearable sensors, among others, where data is continuously produced. Therefore, predicting patient trajectories with reasonable accuracy, as well as detecting anomalies such as sudden changes in their condition, is crucial for early intervention, cost reduction, and improved clinical outcomes.

Traditional statistical methods, such as ARIMA and SARIMA, are effective in modeling linear time trends and have a long history of use in prediction tasks. However, these methods face significant limitations when the underlying data are highly nonlinear, multidimensional, and influenced by complex time dependencies and external covariates [9, 10]. Machine learning algorithms, like Random Forest and Support Vector Machines (SVM), can be more accurate at modelling a nonlinear relationship in a dataset, but often ignore the temporal property of the sequence [9]. More recently, deep learning models, such as LSTM, GRU, and transformer architecture, have shown success in learning sequences with long-term dependencies in a dataset [9, 10].

However, these approaches in healthcare are still limited by the fact that they are essentially "black boxes"; although predictive performance may be high, the logic behind their predictions is often inaccessible, limiting their trust and clinical adoption [5, 11]. In parallel, anomaly detection in healthcare time series data becomes ever more important as identifying abnormalities in behavioral patterns, like arrhythmias or sensor malfunctions, or even acute deterioration in patient condition, enables preventative medical intervention [12–14]. While many approaches exist, including statistical approaches, clustering, autoencoders, and recurrent networks, the explanation of what triggers an anomaly and the ability to explain it remain underdeveloped because the need for an explanation of the triggering factors of the decision cannot be considered a "luxury" but rather a fundamental requirement [11, 15].

3. EXPLAINABLE ARTIFICIAL INTELLIGENCE IN HEALTHCARE TIME SERIES

Explainable artificial intelligence seeks to open the "black box" of complex models by providing human-understandable explanations of their predictions. In healthcare, where decisions directly impact human lives and where the greatest regulatory, ethical, and trust concerns exist, there is a particular need for explainability [15]. Surveys in the field indicate that, despite the considerable attention XAI has received, most current work still focuses on analyzing images or static (tabular) data rather than sequential time-series data [15, 16]. Common XAI techniques include feature-mapping methods (such as SHAP and LIME) [17], attention mechanisms embedded within models [18], prototype-based interpretations and counterfactuals [8], and domain-specific visualization techniques.

For time-series applications, specialized methods have been proposed, such as transforming signals, for example, through virtual inspection layers, with Fourier transforms, to present

the data in more interpretable ways. Relevance propagation can be done using local explanation techniques [17,19]. In healthcare time-series research, some studies have applied gradient-based methods to highlight clinically relevant time windows, such as monitoring rehabilitation exercises [11, 18]. However, a set of persistent challenges has been echoed across existing reviews. More specifically, the clinical validation process, which involves comparing explanations to expert knowledge or performing consistency tests to ensure that explanations align with the data, is still lacking. Furthermore, objective and standardized evaluation indicators, as well as human-centered evaluation strategies, are still underdeveloped [15, 19].

In other words, despite the existence of various techniques, interpretability is still relatively in its early stages compared to predictive performance, particularly in high-risk time-series applications such as those in the healthcare sector.

4. FORECASTING, ANOMALY DETECTION AND EXPLAINABILITY

Recent research in healthcare has started integrating time-series forecasting and anomaly detection with XAI. For example, studies on electronic health records and physiological monitoring systems use explainable artificial intelligence (XAI) methods to analyze patient outcome predictions, although these methods often analyze data at single points in time or simplified sequenced frames [16, 18]. There has been a growing number of reviews of machine learning and deep learning literature in the field of electronic health records for interpretable AI, but only 76 studies from 2017 to 2023 met the criteria for explainable AI in electronic health record environments, and many of them lacked a rigorous evaluation of the interpretations obtained [15, 16].

In the field of non-health time series, progress has been made in evaluating interpretation methods. For example, one study used perturbation analysis of attribution maps across time-series classification tasks to assess the quality of interpretation [17, 19]. Nevertheless, the direct application of these methods to predicting or detecting anomalies in healthcare remains limited. Although some studies monitor anomalies or forecast patient states using deep models, the interpretability component is frequently an afterthought or inadequately evaluated [11, 14, 18].

Another important consideration is the dual nature of forecasting and anomaly detection. Many studies address these tasks separately; however, in clinical time-series data, there is natural overlap: forecasting future patient vitals or states, detecting deviations from expected trajectories, and explaining both simultaneously. Despite this, studies that simultaneously address prediction, anomaly detection, and interpretability in healthcare time series are still rare [14–16].

5. DISCUSSION AND CONCLUSION

This work is based on a narrative review of recent literature addressing explainable artificial intelligence in the context of time series forecasting and anomaly detection for healthcare applications. The discussion draws on peer-reviewed journal articles, conference proceedings, and survey studies published in recent years, with particular emphasis on deep learning models applied to sequential clinical data. Rather than aiming for a systematic or quantitative evaluation, the purpose of this review is to synthesize existing knowledge, identify key limitations, and highlight open research challenges relevant to interpretable and clinically meaningful time-series analysis.

The analysis highlights important trends and several unresolved challenges at the intersection of time series forecasting, anomaly detection, and XAI in healthcare. While deep learning models have demonstrated strong predictive performance in many clinical applications, their lack of transparency remains a major barrier to adoption. Healthcare professionals consistently emphasize the need for interpretability, particularly in sensitive environments such as intensive

care units, remote monitoring, and early warning systems, where algorithmic decisions must be justifiable and clinically relevant.

A critical observation from the literature is that most existing explainability methods were not designed specifically for time series data. Approaches such as SHAP and LIME provide useful local explanations for tabular data but struggle to capture the dynamic dependencies inherent in sequential datasets. Similarly, methods based on salience maps or relevance propagation offer only partial insights into learned characteristics and rarely represent temporal significance clearly, limiting their usefulness in clinical settings where a patient's temporal trajectory is diagnostically important.

Furthermore, prediction and anomaly detection are typically treated as separate tasks, although they are intrinsically complementary in clinical monitoring. Predictive models must anticipate future trajectories while anomaly detection approaches must provide interpretable rationales for deviations. The absence of integrated frameworks limits the development of comprehensive decision support systems.

Evaluation remains a significant challenge. Many studies focus on predictive performance at the expense of the quality and clinical relevance of explanations. In the absence of standardized metrics and consistent evaluation protocols, comparing methods or assessing their utility in real-world settings is difficult, representing a key gap and a promising direction for future research.

Finally, the limited availability of accessible, high-quality sequential clinical datasets hinders validation of XAI methods. Many approaches are tested on synthetic or simplified data, reducing their clinical applicability. Strengthening collaborations between researchers and medical institutions would facilitate the creation of representative benchmarks, promoting the development of models that provide actionable and interpretable explanations in real-world healthcare settings.

Overall, advancing XAI for time-series health data requires not only methodological innovation but also a deeper integration of clinical constraints, human-centered interpretation, and interdisciplinary collaboration. Future research should focus on designing interpretable models capable of maintaining high predictive performance, evaluating explanations rigorously, and ensuring practical clinical utility.

REFERENCES

[1] R. J. Hyndman, G. Athanasopoulos: *Forecasting: Principles and practice*, 3rd ed., 2021.

[2] S. Hochreiter, J. Schmidhuber: *Long Short-Term Memory*. Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] J. Chung, C. Gulcehre, K. Cho, Y. Bengio: *Empirical evaluation of gated recurrent neural networks on sequence modeling*. arXiv preprint arXiv: 1412.3555, 2014.

[4] A. Vaswani et al.: *Attention is all you need*. In Advances in Neural Information Processing Systems, 2017.

[5] S. M. Lundberg, S.-I. Lee: *A unified approach to interpreting model predictions*. Advances in Neural Information Processing Systems, vol. 30, 2017.

[6] G. Montavon, W. Samek, K.-R. Müller: *Methods for interpreting and understanding Deep Neural Networks*. Digital Signal Processing, vol. 73, pp. 1–15, 2018. DOI: 10.1016/j.dsp.2017.10.011

[7] L. Arras, G. Montavon, K.-R. Müller, W. Samek: *Explaining Recurrent Neural Network predictions in sentiment analysis*. In Proceedings of the 8th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp. 159–168, 2017.

[8] W. Samek, T. Wiegand, K.-R. Müller: *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*. arXiv preprint arXiv: 1708.08296, 2017.

[9] Y. Zhang, Q. Sun, D. Qi, J. Liu, R. Ma, O. Petrosian: *ShapTime: A general XAI approach for explainable time series forecasting*. In Lecture notes in networks and systems: Intelligent systems and applications, Springer, pp. 659–673, 2024.

[10] T. Hall, K. Rasheed: *A survey of machine learning methods for time series prediction*. Applied Sciences, vol. 15, no. 11, p. 5957, 2025.

[11] M. H. Lee, Y. J. Choy: *Exploring a gradient-based explainable AI technique for time-series data: A case study of assessing stroke rehabilitation exercises*. In ICLR 2023 workshop, 2023.

[12] J. Backhus, A. R. Rao, C. Venkatraman, C. Gupta: *Time series anomaly detection using signal processing and deep learning*. Applied Sciences, vol. 15, no. 11, p. 6254, 2025.

[13] Z. Zamanzadeh Darban, G. I. Webb, S. Pan, C. C. Aggarwal, M. Salehi: *Deep learning for time series anomaly detection: A survey*. arXiv preprint arXiv: 2211.05244, 2022.

[14] I. Ćirić, N. Ivačko, S. Cvetković: *Application of explainable AI algorithms in machine learning models for time series forecasting: A survey*. In ICIST 2024, pp. 330–337, 2024.

[15] F. Di Martino, F. Delmastro: *Explainable AI for clinical and remote health applications: A survey on tabular and time series data*. Artificial Intelligence Review, vol. 56, pp. 5261–5315, 2023.

[16] G. M. Nagamani, K. C. Kiran: *Design of an improved graph-based model for real-time anomaly detection in healthcare using hybrid CNN-LSTM and federated learning*. Heliyon, vol. 10, no. 24, p. e41071, 2024.

[17] R. Alkhanbouli, H. M. Almadhaani, F. Alhosani: *The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions*. BMC Medical Informatics and Decision Making, vol. 25, p. 110, 2025.

[18] H. Ma, W. Liu, K. McAreavey: *Time series XAI methods: Benchmarking and clinical evaluation*. arXiv preprint arXiv: 2408.10314, 2024.

[19] H. Ma, K. McAreavey, W. Liu: *TSFeatLIME: An online user study in enhancing explainability in univariate time series forecasting*. arXiv preprint arXiv: 2409.15950, 2024.

BIOMECHANICAL ASSESSMENT OF LUMBAR DISC PROSTHESES: THE ROLE OF MATERIAL PROPERTIES IN FUNCTIONAL PERFORMANCE

Katreen Ebrahem[1], ***Szabolcs Szávai***[2]
[1]PhD Student, [2]Associate Professor
[1,2]*Faculty of Mechanical Engineering and Informatics,*
Institute of Matchine and Product Design
[1]katreen.ebrahem@student.uni-miskolc.hu,
[2]szabolcs.szavai@uni-miskolc.hu

Abstract

Material properties of components of lumbar disc prostheses, particularly inlay (core) and articulation surfaces, are critical determinates of biomechanical performance under physiological loading. This study evaluates the influence of several polymeric and metallic material properties on the functional biomechanics of lumbar total disc replacement (LTDR). FE models of $L_3 - L_5$ were developed to simulate articulation responses of three representative TDR designs. Models were subjected to physiological compression during level-walking. The results demonstrated that intradiscal pressure (IDP) of adjacent disc remains in the normal range for all three models. As well as, the viscoelastic and compliant characteristics of polymeric materials of PCU help in avoiding stress shielding phenomena on the bone endplates in contact to the artificial disc, and using PEEK-on-PEEK bearing results in the highest contact pressure in the implant with existing stress concentration on upper keel of endplate and this should be taken in consideration to predict the life span of implant. These findings support the integration of elastomeric materials in lumbar disc replacements to improve clinical outcomes and further research is recommended.

1. INTRODUCTION

Lumbar Disc Prostheses (also called lumbar total disc replacement (LTDR)) are engineered implants designed to replace the degenerated intervertebral disc while preserving physiological motion between vertebras. The implant of effective LTDR must fulfil the following main requirements: A solid, non-destructive interface with the adjacent vertebral bodies, providing mobility to mimic the range of motion of the natural disc, wear and tear resisting in the body to reduce wear debris in the body, and having the ability to absorb shock and distribute loads evenly and effectively [1]. Successful long-term performance depends on the artificial disc design (geometry, articulation mechanics) and biomaterial selection– a key determinate of biomechanical behaviour, biocompatibility, and wear resistance [2]. In the context of the lumbar spine and disc prostheses, the used material properties matter in biomechanics of implants because they determine how tissues and implant behave under real physiological loads, which is an especially critical for functional performance [3].

2. THE PHYSICAL MODEL & SOLUTION METHOD

The analysis of biomechanical behaviour of lumbar disc arthroplasty (TDA) in lumbar spine was conducted using the finite element method (FEM). For this purpose, the real object has to be simplified but enough to create a numerical model. In this study, a patient-specific model for a 33-year-old and 75 kg-body weight male with disc degeneration at the $L_3 - L_4$ spinal level was generated, consisting of three vertebras in the $L_3 - L_5$ segment as shown in Figure1. Utilizing a series of CT (Computed Tomography) scans converted to a DICOM (Digital Imaging

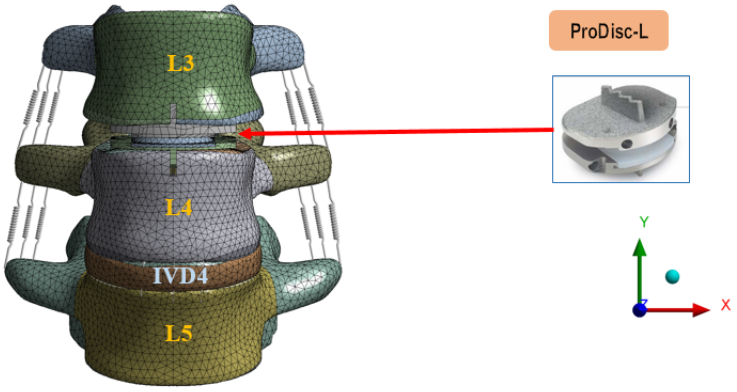


Figure 1: A schematization of the biomodel $L_3 - L_5$ analysed in this study.

and Communications in Medicine) file format, the $L_3 - L_5$ model was generated in Materialize MIMICS software. Subsequently, the model was imported into SolidWorks CAD (Version 2025; Dassault Systèmes SolidWorks Corporation, Vélizy-Villacoublay, France) in STEP file format, forming the basis for the finalized element model used in this study. The intervertebral disc replacement was ProDisc-L from Synthes Spine which is ball-on-socket design approved by food and drug administration (FDA), and the 3D model with size-L of this artificial disc was crafted using SolidWorks (Version 2025; Dassault Systèmes SolidWorks Corporation, France).

Various regions were assigned to vertebrae, including the cortical bone, cancellous bone, posterior bone, and cartilaginous endplates. Additionally, the normal intervertebral discs were modified to include the central part which represents the nucleus pulposus (NP), and six layers of annulus fibrous part (AF). As well as, seven essential ligaments connecting the vertebral bodies were incorporated into the model. These ligaments included the anterior longitudinal ligament (ALL), posterior longitudinal ligament (PLL), ligament (LF), capsular ligament (CL), supraspinal ligament (SS), interspinous ligament (IS), and transverse ligament (TL). Finally, once the bone structures, intervertebral discs components, and the artificial disc were assembled, the model was saved in STL format. Subsequently, the complete geometry was imported into Ansys software (v.19.2, ANSYS, Inc., Canonsburg, PA, USA) to conduct finite element analysis (FEA). In order to assess the behaviour of several materials for ProDisc-L design, three models were analysed; Model 1: Polyetheretherketone bearing (PEEK-on-PEEK), Model 2: Cobalt-Chromium Molybdenum Alloy on Polycarbonate Urethane bearing (CoCrMo-on-PCU), Model 3: the original bearing of ProDisc-L, which is Cobalt-Chromium Molybdenum-on- Ultra-high-molecular-weight polyethylene (CoCrMo-on-UHMWPE). The material behaviour was considered to be homogenous and isotropic linear elastic (viscoelastic for PCU) properties. The data of Table 1 were sourced from the literature [4, 5]. A steady-state friction coefficient (COF) for surface bearings (CoCr-UHMWPE) of 0.08 was used for UHMWPE core sliding against metallic endplates [6], which represents the mean friction coefficient in bovine calf serum conditions (a physiological lubricant) to reach more realistic simulation. For (PEEK-PEEK) surface bearings, COF was of 0.12 [7], and for (CoCrMo-PCU) was 0.03 [8]. Regarding the mesh of the studied components, a mesh with tetrahedral elements was used for the bony structures and the nucleus pulposus, while the main spinal ligaments were modelled as tension-only nonlinear springs, hexahedral elements were utilized for PEEK and CoCrMo components of ProDisc-L, while quadrilateral elements were employed for the polymeric core components (UHMWPE, PCU), The mixed element technique was used to facilitate the construction of the annulus fibrosus of the normal disc, where the six layers were modelled as quadric elements (primary

Table 1: Material properties used in finite element analysis of $L_3 - L_5$ models with artificial disc

Geometry	**Material**	**Density** [g/cm^3]	**Young Modulus** [MPa]	**Poisson Ratio**
Bone Structure	Cortical bone	1.7	12000	0.3
	Cancellous bone	1.1	100	0.3
	Posterior bone	1.4	3500	0.3
	Upper & Lower endplate	1.7	12000	0.3
The normal IVD	NP	1.02	1	0.499
	AF_1	1	550	0.45
	AF_2	1	495	0.45
	AF_3	1	440	0.45
	AF_4	1	420	0.45
	AF_5	1	385	0.45
	AF_6	1	260	0.45
Spinal ligaments	LF	1	50	0.3
	TL	1	50	0.3
	IS	1	28	0.3
	SS	1	28	0.3
	CL	1	20	0.3
	ALL	1	20	0.3
	PLL	1	70	0.3
Artificial Disc	PEEK	1.3	3600	0.35
	CoCrMo	8.9	210000	0.3
	UHMWPE	0.94	1200	0.46
	PCU	1.2	25	0.49
		Instantaneous Shear Modulus (High Rate) **G0** [MPa]		15
		Viscoelastic Decay Constant [s^{-1}]		1.66

Table 2: The number of nodes and elements of finite element models of $L_3 - L_5$ models

The Model	No. of Elements	No. of Nodes
Model 1	145 349	245 362
Model 2	143 073	238 599
Model 3	122 460	215 520

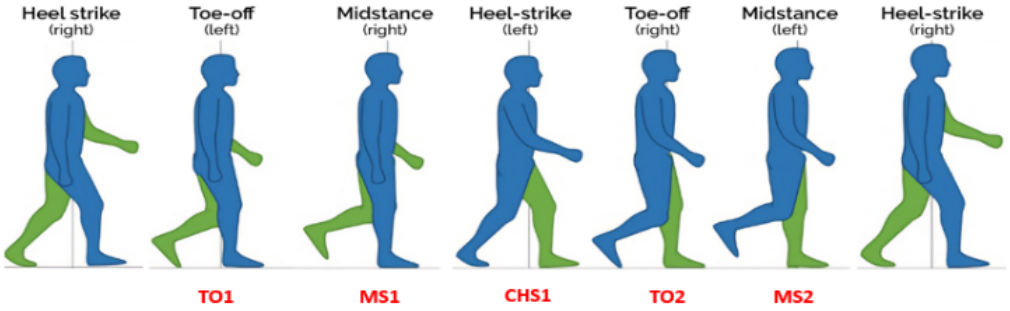


Figure 2: Gait cycle instances

element) and then as tetrahedron elements (secondary element). The number of elements and nodes was determined after a prior convergence study was performed on the model to achieve a compromise between the number of elements and the numerical cost (The convergence criteria were set such that the variation in the results reached less than 1 % with respect of the change in mesh size). Table 2 provides the number of the nodes and elements for the three models.

In this study, level-walking loading was applied to gait instances. During a gait cycle as shown in Figure 2, which is from initial to final left heel strike, major temporal instances occur around 10 % (TO1), 32 % (MS1), 50 % (CHS), 60 % (TO2) and 80 % (MS2) of the gait cycle. [9] A compressive loading was applied on the top of L_3 vertebrae for each instance as shown in Table 3, while the bottom of L_5 vertebrae was been fixed support.

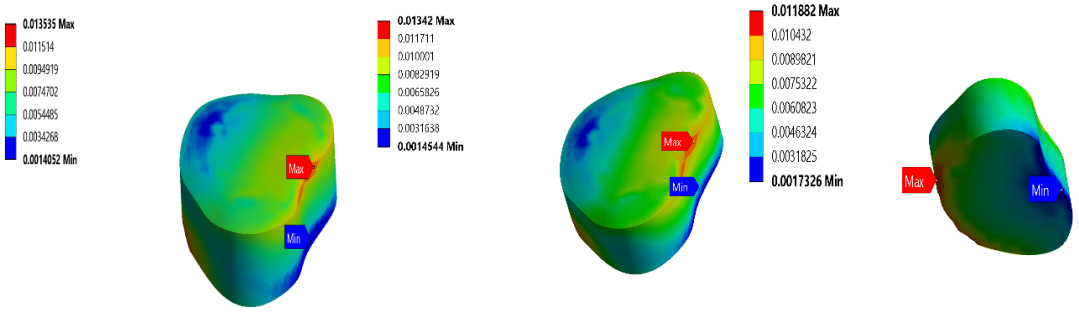
3. RESULTS

3.1. Intradiscal pressure (IDP) of adjacent levels (Disc4)

Changes in implant stiffness change load transfer to adjacent discs-an important predictor of adjacent-segment degeneration (ASD), and (IDP) refers to the internal pressure within the inter-vertebral disc, most commonly measured in the nucleus pulposus, Figure 3 shows the Intradiscal pressure of IVD 4 in studied models.

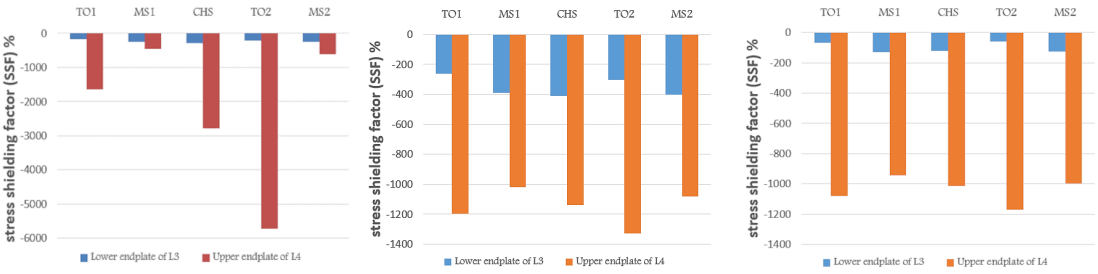
Table 3: Lumbar loads during gait cycle of level-walking

Phase	Gait %	Compression [N]
TO_1	10	368
MS_1	32	1 250
CHS	50	800
TO_2	60	300
MS_2	80	1 100



(a) Model 1 (b) Model 2 (c) Model 3

Figure 3: Intradiscal pressure (IDP) of IVD 4.



(a) Model 1 (b) Model 2 (c) Model 3

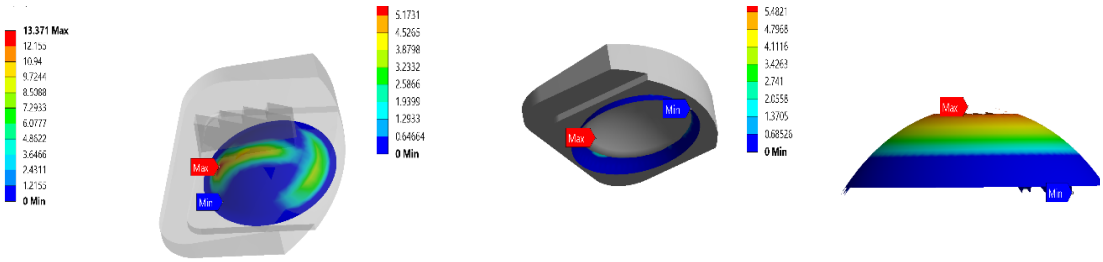
Figure 4: The values of stress shielding factor (SSF) during gait instances.

3.2. Stress in the adjacent bony structures

An artificial intervertebral disc that stresses the bony structures differently than the intact model may cause stress shielding and bone density loss, because the bone near the metallic endplates, where the highest stresses are exchanged, is not stimulated, and hence may resorb according to Wolff's law [10]. To define the Stress Shielding Factor (SSF) in %, the average stress ($\sigma^{M}_{\text{Intact}}$) was calculated and was compared with its counterparts counterparts ($\sigma^{M}_{\text{Implant}}$) for the models analysed in this study. A positive SSF value indicates that the stress developed at the selected locations is lower in the prosthesis model than in the intact (baseline) model. A negative SSF value indicates that stress is lower in the intact model. Figure 4 demonstrate the values of stress shielding factor during gait instances in implantation locations at endplates of L_3 and L_4 vertebras.

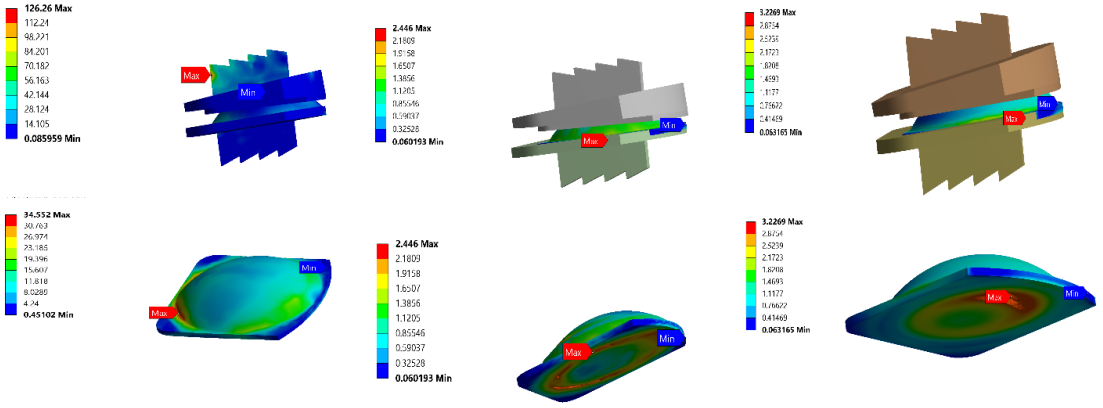
3.3. Contact pressure in the implant

Pressure distribution at inlay-endplates interface effects on friction/lubrication behaviour, consequently on wear, fatigue, and short and long-term performance. It also controls the subsidence risk. Figure 5 shows the resulted values of contact pressure in articulation parts of artificial disc in the analysed models.



(a) Model 1 (b) Model 2 (c) Model 3

Figure 5: The values of contact pressure at ball-on-socket bearing surfaces.



(a) Model 1 (b) Model 2 (c) Model 3

Figure 6: Equivalent von-Mises stress of ProDisc-L

3.4. Stress within the implant

In order to ensure the life span of the artificial discs, equivalent von-Mises stresses of all three discs were been assessed as shown in Figure 6.

4. DISCUSSION

Regarding intradiscal pressure (IDP) of adjacent disc, the equivalent von-Mises stress reached to the maximum value of 0.0135 MPa in Model 1, following 0.0134 MPa in Model 2, and 0.0118 MPa in Model 3 which represents the ProDisc-L model. All obtained results are within the normal range of (IDP) which are 0.53∼0.65 MPa according to the in vivo measurements of pressure in the literature [11]. These findings indicate that the materials employed have exert a negligible influence on adjacent intervertebral discs. All values of stress shielding factor (SSF) in all the gait cycle instances were negative, meaning that there is not a stress shielding on the bone/implant interfaces, despite of that, there is a difference in magnitudes values. For Model 1, with properties of elastic modulus of PEEK material closer to that of cortical bone than metallic materials, the implant can reduce stress on the adjacent bones and promote stress distribution [12]. Additionally, Model 1 demonstrated the highest contact pressure of 13.37 MPa, and may this effect on the implant stability and long-term clinical performance, as well as, maximum von-Mises stress occurs on the upper keel of disc and corresponds to 126.26 MPa, and 34.55 MPa on the inlay, from published material properties sheets of medical grade PEEK, this material can stand until 160 MPa according to ISO 604 under compression. One of the notable observations from Model 2 is the highest magnitude of SSF, so the use of PCU material as an in-

lay between metallic endplates helps in shock absorption, as well as contact pressure within the bearing reached to 5.81 MPa due to the low COF. Due to visco-elastic properties of this material it can mimic the natural disc's function, and because of its mechanical properties ($E, \upsilon, \ldots$ etc.) it can become a promising alternative to UHMWPE [13] with taking in account the life span of artificial disc as shown in figure 7 for Model 2, the maximum von-Mises stress was on the bottom of PCU inlay and equals to 2.446 MPa and the compressive strength of medical grade PCU is 5$\sim$60 MPa. Model 3 which represents the real ProDisc-L, can helps in avoiding stress shielding phenomena similar to Model 1, with contact pressure of 6.16 MPa on the UHMWPE inlay pole and maximum von-Mises stress of 3.22 MPa on the anterior area of core and it's below the yield stress value of UHMWPE corresponds to 21 MPa [14]. So implant in Model 3 with high safety factor followed by implant in Model 2, then the implant in Model 1.

5. CONCLUSION

Material properties play a critical role in load distribution and restoring the functional performance of IVD. This study highlights the numerical investigation of different biomaterials combinations, each studied model needs further research to simulate more realistic state in current articulation designs with the significant role of choosing of exact lubricant to minimize wear overlong-term performance. FEA can provides several classes of materials models ranging from simple but in accurate to very accurate and complex ones to reach the correct simulation. The results obtained from this study can be used before manufacturing and the models can be extended to include the whole lumbar spine, and to investigate long-term performance during daily activities.

REFERENCES

[1] J.-M. Vital, L. Boissière: *Total disc replacement*. Orthopaedics & Traumatology: Surgery & Research, vol. 100, no. 1, pp. S1–S14, 2014.

[2] G. Song, Z. Qian, K. Wang, J. Liu, Y. Wei: *Total disc replacement devices: Structure, material, fabrication, and properties*. Progress in Materials Science, vol. 140, 2023.

[3] S. Taksali, J. N., A. R.: *Material considerations for intervertebral disc replacement implants*. The Spine Journal, vol. 4, no. 6, p. 231S–238S, 2004.

[4] R. Natarajan, G. Andersson: *Modeling the annular incision in a herniated lumbar intervertebral disk to study its effect on disk stability*. Computers & Structures, vol. 64, no. 5-6, pp. 1291–1297, 1997.

[5] C. Wen-Ming, P. ChunKun, L. KwonYong, L. SungJae: *In situ contact analysis of the prosthesis components of Prodisc-L in lumbar spine following total disc replacement*. Biomechanics, vol. 34, no. 20, pp. E716–E723, 2009.

[6] C. M., R. Vicars, R. M., T. D.: *Preferential superior surface motion in wear simulations of the Charité total disc replacement*. European Spine Journal, vol. 21, no. 5, pp. 700–708, 2010.

[7] H. Xin, R. Liu, L. Zhanga, J. Jia, J. Jiaa: *A comparative bio-tribological study of self-mated PEEK and its composites under bovine serum lubrication*. Biotribology, vol. 26, 2021.

[8] H. Barber, C. N., B. Abar, N. Allen, S. B., K. Gall: *Rotational wear and friction of Ti-6Al-4V and CoCrMo against polyethylene and polycarbonate urethane*. Biotribology, vol. 26, 2021.

[9] R. Arshad, L. Angelini, T. Zander, F. Di Puccio, M. El-Rich, H. Schmidt: *Spinal loads and trunk muscles forces during level walking – A combined in vivo and in silico study on six subjects*. Journal of Biomechanics, vol. 70, pp. 113–123, 2018.

[10] H. M.: *Wolff's Law and bone's structural adaptations to mechanical usage: An overview for clinicians*. Angle Orthodontist, vol. 64, no. 3, pp. 175–88, 1994.

[11] H. Wilke, P. Neef, M. Caimi, T. Hoogland, L. E.: *New in vivo measurements of pressures in the intervertebral disc in daily life*. Spine, vol. 24, no. 8, pp. 755–762, 1999.

[12] Y.-H. Yu, S.-J. Liu: *Polyetheretherketone for orthopedic applications: A review*. Current Opinion in Chemical Engineering, vol. 32, 2021.

[13] Y. A., R. Verma, S. A.: *Artificial disc replacement in spine surgery*. Annals of Translational Medicine, vol. 7, no. 5, p. S170, 2019.

[14] D. Baker, R. S. Hastings, L. Pruitt: *Study of fatigue resistance of chemical and radiation crosslinked medical grade ultrahigh molecular weight polyethylene*. J Biomed Mater Res, vol. 46, no. 4, pp. 573–81, 1999.

EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR MULTILAYER PERCEPTRONS: A CASE STUDY USING LIME IN TEXT-BASED EMOTION RECOGNITION

Noureddine Hfaiedh[1], ***Erika Baksáné Varga***[2]
[1]PhD Student, [2]Associate Professor
[1,2]*Faculty of Mechanical Engineering and Informatics,*
Institute of Information Technology
[1]samhrjh@gmail.com, [2]erika.b.varga@uni-miskolc.hu

Abstract

Although neural networks perform well in classification and regression tasks [1], sensitive applications are less trustworthy due to their opaque decision-making processes [2]. This work explores explainable artificial intelligence (XAI) for text-based emotion recognition using Multilayer Perceptrons (MLPs). We illustrate the contribution of individual word features to MLP predictions using the Local Interpretable Model-Agnostic Explanations (LIME) framework [3]. Word-level explanations and prediction probabilities are included in a specific example. Although MLPs are the main focus, there is a brief discussion of how explainability might be extended to Recurrent Neural Networks (RNNs) and Knowledge-Augmented Networks (KANs) [4]. The findings demonstrate the usefulness of XAI in enhancing openness and confidence in neural network-based emotion recognition systems.

1. INTRODUCTION

Artificial intelligence (AI) has profoundly impacted healthcare and education, with Neural Networks (NNs) like Multilayer Perceptrons (MLPs) effectively modeling complex non-linear interactions for emotion recognition. However, these "black box" models often lack transparent logic, a restriction addressed by Explainable AI (XAI). Model-agnostic methods like SHAP and LIME allow users to understand feature contributions to specific decisions. This study provides an example-driven explanation of MLP-based emotion identification using LIME, further discussing future extensions to RNNs and Knowledge-Augmented Networks (KANs).

2. BACKGROUND

MLPs are feedforward networks consisting of input, hidden, and output layers, favored for their computational efficiency in text classification. LIME explains these models by approximating local behavior with an interpretable surrogate model, perturbing inputs to estimate feature importance.

3. LITERATURE REVIEW

In recent years, the scope of deep learning has expanded beyond simple classification toward a focus on model interpretability and reliability. While early research demonstrated the high performance of multi-layered architectures [1], it also highlighted the challenge of tracing information flow through complex hidden layers. In the context of affective computing, Picard [5] emphasized that for machines to interact effectively with humans, they must not only identify emotions but do so through processes that align with human cognitive models.

The introduction of the Local Interpretable Model-Agnostic Explanations (LIME) framework by Ribeiro et al. [3] provided a significant methodology for inspecting "black box" models

Table 1: Dataset Overview

Property	Description
Data type	Short text sentences
Number of samples	~ 2000
Emotion classes	Anger, Joy, Love, Fear, Sadness, Surprise
Feature representation	TF-IDF vectors
Input dimension	Vocabulary size

Table 2: MLP Architecture

Layer	Units	Activation
Input	TF-IDF features	-
Hidden Layer 1	128	ReLU
Hidden Layer 2	64	ReLU
Output	6	Softmax

without necessitating changes to their underlying architecture. While contemporary academic discourse frequently focuses on the interpretability of Transformer-based models, recent studies suggest that understanding simpler Multilayer Perceptrons (MLPs) remains critical, particularly in environments with limited computational resources. Furthermore, as noted by Doshi-Velez and Kim [2], explainability is no longer a secondary technical feature; it is an essential requirement for validating that human-centric AI systems operate as intended.

4. PROBLEM STATEMENT

Even though Multilayer Perceptrons (MLPs) are good at text classification, it's hard to tell why they make certain decisions. We often do not know which parts of the text are most important for a prediction, which makes understanding feature contributions tough. It is also tricky to see how the model decides on individual inputs, especially when there are complicated interactions at play. Furthermore, even though MLPs are used extensively, there is not much research showing how to use explainable AI methods with.

5. METHODOLOGY

5.1. Dataset and Preprocessing

We trained our emotion recognition model using a public dataset from Kaggle, called Emotion Dataset for Emotion Recognition Tasks [6]. It is basically a collection of English tweets, each tagged with a specific emotion. As shown in Table 1, every entry has a tweet or sentence and then tells you if it is exhibiting anger, joy, love, fear, sadness, or surprise. To get the data ready for the model, we followed standard preprocessing procedures: changed everything to lowercase, broke sentences into words, removed stop words that do not contribute significant semantic information, and then converted the text into numerical vectors using TF-IDF.

5.2. MLP Architecture

The MLP consists of an input layer, two hidden layers, and an output layer, as detailed in Table 2. The input dimension for the model, which represents the length of the feature vector, is 5000.

Table 3: Overall MLP Performance on Test Set

Metric	Value
Test Accuracy	0.8416
Test Loss	0.8239

5.3. Training and Implementation

The MLP was implemented in Python using the TensorFlow/Keras framework [7]. The model utilizes specific activation functions to handle non-linearities and classification: the two hidden layers employ the ReLU (Rectified Linear Unit) activation function, while the final output layer uses the Softmax activation function to generate probability distributions across the six emotion classes. Training was conducted using the **Adam optimizer** with a learning rate of 0.001 and categorical cross-entropy loss

$$\text{Loss} = -\sum_{i=1}^{n} y_i \cdot \log(\hat{y}_i). \tag{1}$$

5.4. Explainability with LIME

To figure out what our MLP predictions meant, we used LIME. We prepped each sentence and turned it into vectors, following the same procedure used during the training phase. LIME messes with the input words, gets new predictions from the MLP, and then uses a simple weighted model to estimate the numerical contribution of each word to the final classification. We used 500 samples for each explanation and a kernel width of 25. Basically, if a word had a positive weight, it helped the prediction; if it had a negative weight, it went against it.

6. RESULTS

6.1. Dataset Splits and Model Training

We trained the MLP using 16 000 examples, validated it with 2000, and subsequently tested it on a separate set of 2000 examples. The training process consisted of 10 iterations (epochs). We utilized the Adam optimizer with a learning rate of 0.001 and measured performance using categorical cross-entropy loss. The model architecture included an input layer, two hidden layers, and a softmax output layer, as specified in Table 2.

During these 10 iterations, the training loss steadily decreased, indicating that the model was continuously learning and improving its fit to the training data. Concurrently, the validation loss also generally decreased, which suggests that the model was effectively learning patterns that generalized to unseen data beyond the training set.

6.2. Overall Performance Metrics

The MLP achieved competitive performance on the test set. Table 3 summarizes the overall metrics.

6.3. Example Predictions

A few test words and their actual and anticipated emotions are shown in Table 4. It demonstrates that the MLP can accurately predict both positive emotions, such as joy, and nega-

Table 4: Sample Predictions on Test Set

Sentence	True Emotion	Predicted Emotion
im trying to feel out my house style now that im living on my own and have creative carte blanche	Joy	Joy
i feel suffocated and paranoid	Fear	Fear
i like in this world and making a list of them always makes me feel joyful	Joy	Joy
i really feel and i know the devil hates that its always been something he could use against me and im determined not to let him	Joy	Joy

Table 5: Prediction Probabilities for a Single Sentence

Emotion Class	Probability
Happy	0.69
Neutral	0.21
Sad	0.04
Other	0.03

tive emotions, such as feeling ignored or worried.

6.4. Prediction Probabilities for a Single Sentence

The MLP generates the probability displayed in Table 5 for the input text "I am extremely happy with my progress today". In line with human interpretation, the model accurately predicts Happy with the highest likelihood.

6.5. Word-Level Contributions with LIME

We used LIME to correctly sort test phrases, helping us see why certain words led to certain predictions. Figure 1 shows each word's impact on the Happy category. Terms with low emotional valence had a negligible impact on the model's output. But words like extremely and happy, which are clearly emotional, had a strong positive impact. This shows that the model relies on words that actually mean something, not just random connections, giving us a clearer view of individual predictions.

7. CONCLUSION AND FUTURE PERSPECTIVES

This study demonstrates that LIME effectively elucidates MLP decision-making in emotion recognition, fostering user trust. Limitations include the reliance on a single English dataset and the basic nature of the MLP architecture compared to sequence-modeling techniques. Future research will extend these XAI methodologies to complex architectures like RNNs and Knowledge-Augmented Networks. Additionally, incorporating cross-linguistic datasets and quantitative metrics will be essential to validate explanation stability in high-stakes domains like healthcare. Recent studies continue to demonstrate the evolving necessity of these frameworks in complex sentiment analysis tasks [8].

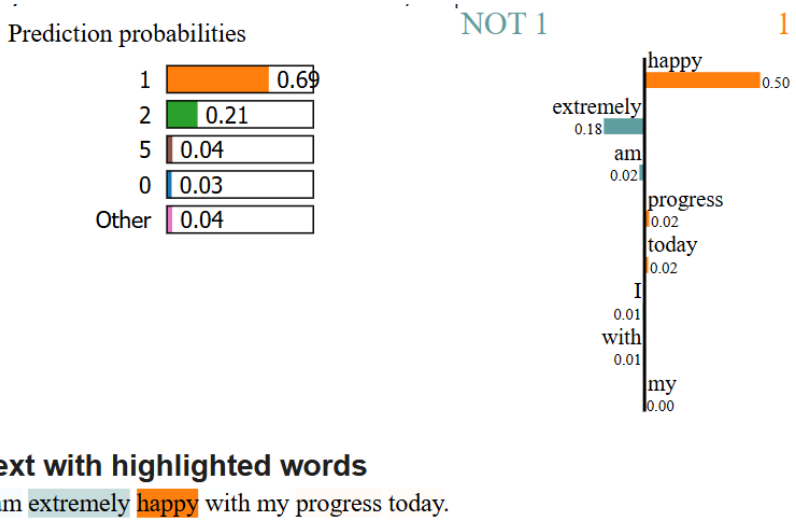


Figure 1: LIME explanation showing word-level contributions for the “Happy” class.

REFERENCES

[1] Y. LeCun, et al.: *Deep Learning*. Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[2] F. Doshi-Velez, et al.: *Towards a rigorous science of interpretable Machine Learning*. arXiv: 1702.08608, 2017.

[3] M. T. Ribeiro, et al.: *Why should i trust you? Explaining the predictions of any classifier*. Proc. 22nd ACM SIGKDD, pp. 1135–1144, 2016.

[4] L. Arras, et al.: *Explaining Recurrent Neural Network predictions in sentiment analysis*. Proc. EMNLP, pp. 159–168, 2017.

[5] R. W. Picard: *Affective computing*. MIT Press, 1997.

[6] P. Pandey: *Emotion dataset for emotion recognition tasks*. Kaggle, 2018. https://www.kaggle.com/datasets/parulpandey/emotion-dataset

[7] M. Abadi, et al.: TensorFlow: *Large-scale Machine Learning on heterogeneous systems*. Proc. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI ’16), pp. 265–283, 2016.

[8] F. Bodria, et al.: *Benchmarking and explaining sentiment analysis*. Scientific Reports, vol. 13, no. 1, pp. 1–15, 2023.

CLOUD WORKLOAD FORECASTING: TRENDS, TECHNIQUES, AND RESEARCH DIRECTIONS

Tasnaim Hosen
PhD Student
Faculty of Mechanical Engineering and Informatics,
Institute of Information Technology
tasnaim.hosen@student.uni-miskolc.hu

Abstract

Cloud workload forecasting has become a critical component of modern cloud resource management, enabling proactive scaling, cost optimization, and improved Quality of Service (QoS). As cloud environments grow increasingly dynamic and heterogeneous, traditional statistical models struggle to capture nonlinear and rapidly fluctuating workload patterns. This paper presents a structured review of contemporary forecasting techniques, including statistical, machine learning, and deep learning approaches, with emphasis on hybrid and clustering-enhanced models. Drawing on recent literature and empirical findings, the study identifies key research gaps, outlines methodological considerations, and proposes future research directions for adaptive, scalable, and uncertainty-aware forecasting frameworks.

1. INTRODUCTION

Cloud computing provides on-demand access to shared computing resources, enabling organizations to deploy applications rapidly and scale flexibly. Cloud service providers offer storage, computation, and hosting services under pay-as-you-go models as shown in Fig. 1, reducing operational overhead and improving service agility [1]. Virtualization plays a central role in cloud data centers, allowing multiple users to share physical resources through virtual machines (VMs) [2].

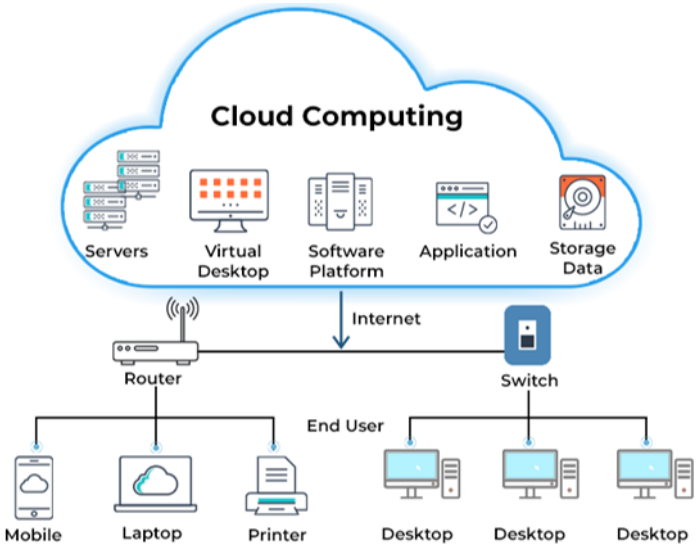


Figure 1: Cloud Computing Architecture (by author)

As cloud adoption accelerates, understanding and predicting workload behavior becomes essential. Cloud workload forecasting refers to predicting future resource demands such as CPU,

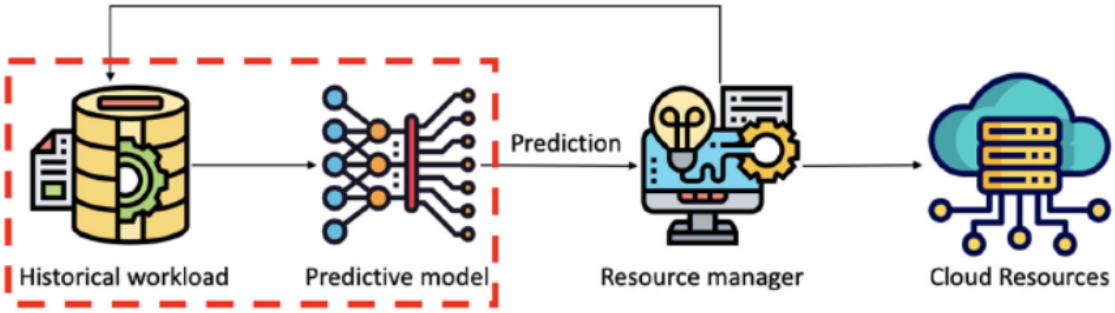


Figure 2: The workload prediction scheme in cloud computing. [4]

memory, and network usage based on historical patterns and current system states [2]. Accurate forecasting enables proactive resource provisioning, reduces service level aggreement SLA violations, and minimizes both over- and under-provisioning. Fig. 2 provides a visual explanation of the workload prediction scheme in cloud computing.

2. MOTIVATION

Cloud workloads are inherently unpredictable due to user behavior, application diversity, and fluctuating traffic patterns. This unpredictability raises several challenges:

- How can cloud providers maintain low cost while ensuring high QoS?
- How can resource allocation remain efficient when workloads exhibit nonlinear, bursty, or seasonal patterns?
- How can forecasting models adapt to workload drift in rapidly evolving cloud environments?

Accurate workload forecasting directly impacts cost efficiency, energy consumption, and system stability. As cloud infrastructures expand globally, the need for intelligent, adaptive forecasting mechanisms becomes increasingly urgent.

3. BACKGROUND AND RELATED WORK

Cloud workload forecasting has evolved significantly over the past decade, driven by the need for accurate, scalable, and adaptive resource management in cloud data centers. Existing approaches can be broadly categorized into statistical models, machine learning techniques, deep learning architectures, hybrid frameworks, and optimization-enhanced forecasting systems.

3.1. Statistical and Time-Series Forecasting Models

Early research relied heavily on classical time-series models such as ARIMA, exponential smoothing, and hybrid statistical frameworks. Devi and Valli [13] proposed a hybrid ARIMA-ANN model that combines linear and nonlinear components to improve CPU and memory utilization forecasting. Their approach reduces multi-step prediction errors and demonstrates the limitations of purely linear models in dynamic cloud environments.

Similarly, Kumar and Singh [2] introduced a workload prediction framework using feed-forward neural networks optimized with adaptive differential evolution, showing improvements over

traditional ARIMA-based methods. These works highlight the strengths of statistical models in capturing short-term trends but also reveal their inability to handle non-stationary, bursty workloads common in cloud systems.

3.2. Machine Learning-Based Forecasting Approaches

Machine learning techniques have been widely adopted to overcome the rigidity of statistical models. Gao et al. [1] proposed a clustering-based prediction framework that groups workloads by behavioral similarity before training specialized models for each cluster. Using Google cluster traces, their method improved CPU prediction accuracy from $\sim$75 % to over 90 %.

Kim et al. [9] introduced CloudInsight, an ensemble-based forecasting framework that dynamically adjusts predictor weights. CloudInsight achieved 13–27 % higher accuracy and reduced provisioning errors by 15–20 %, demonstrating the effectiveness of ensemble learning in real-world cloud environments. Beyond clustering-based and ensemble forecasting methods, several other machine learning techniques have been explored in [16]. An adaptive resource utilization prediction system was proposed in which the forecasting model dynamically switches between ARIMA and AR-NN depending on the statistical characteristics of the workload distribution. This adaptive selection mechanism improves prediction accuracy by choosing the most suitable model for each workload segment, particularly when dealing with heterogeneous resource usage patterns. Swarm intelligence has also been applied to workload forecasting. The Swarm Intelligence-Based Prediction Approach (SIBPA) demonstrated superior performance in predicting CPU, memory, and disk utilization by leveraging collective learning behaviors inspired by natural swarms. Experimental results showed that SIBPA consistently outperformed traditional machine learning models across multiple evaluation metrics, highlighting its robustness in handling noisy and irregular workload traces [19]. Another notable direction in [20] involves pattern-matching-based forecasting, where short-term workload demand is predicted by identifying historical segments that closely resemble the current workload window. By exploiting similarity patterns in past behavior, this method provides fast and interpretable predictions, making it suitable for real-time autoscaling scenarios. However, its effectiveness depends heavily on the availability of representative historical patterns. Collectively, these studies demonstrate that machine learning approaches offer improved flexibility and accuracy compared to classical statistical models. Nevertheless, they often struggle to capture long-term temporal dependencies and may require additional mechanisms such as deep learning or hybrid architectures to handle highly dynamic and non-stationary cloud workloads.

3.3. Deep Learning Models for Cloud Workload Forecasting

Deep learning has become the dominant paradigm due to its ability to model nonlinear, multivariate, and long-range temporal dependencies. Zhang et al. [8] demonstrated that RNN-based approaches achieve high accuracy in cloud cluster workload prediction. LSTM models, in particular, outperform ARIMA due to their ability to learn nonlinear patterns and long-term dependencies. Janardhanan and Barrett, in their study [17], conducted a comparative analysis of LSTM and ARIMA techniques for CPU workload prediction. Their results demonstrated that LSTM models consistently outperform ARIMA, particularly under irregular and highly variable workload patterns, due to their superior ability to capture nonlinear temporal dependencies. Ouhame et al. [14] proposed a CNN-LSTM hybrid model that extracts spatial features using convolutional layers before applying LSTM for temporal forecasting. Their model improved prediction accuracy by up to 10.9 % and reduced error by up to 8.5 %. While Al-Asaly et al. [12] introduced a diffusion convolutional recurrent neural network (DCRNN) capable of

handling nonlinear and inconsistent workloads. Their model achieved a MAPE of 0.18, outperforming existing deep learning approaches. On the other hand recent studies like in [18] incorporate attention mechanisms to focus on the most relevant temporal features. Attention-based CNN-GRU models have shown 2–28 % accuracy improvements over baseline methods. These deep learning models represent a major advancement but often require extensive training data and computational resources.

3.4. Hybrid and Optimization-Enhanced Forecasting Models

Hybrid forecasting models combine multiple analytical techniques to leverage their complementary strengths, particularly when dealing with nonlinear, multivariate, and highly dynamic cloud workloads. A major trend in recent research is the integration of evolutionary optimization algorithms with neural network architectures to enhance convergence speed, improve generalization, and avoid local minima during training. Several studies demonstrate the effectiveness of such hybrid approaches. The esDNN model proposed by Xu et al. [7] employs a supervised deep neural network enhanced with sliding-window preprocessing and a revised GRU architecture. This design enables the model to capture multivariate temporal dependencies more effectively, resulting in improved prediction accuracy across diverse workload traces. Another notable contribution is the BiPhase Adaptive Differential Evolution (BaDE)-optimized neural network introduced by Kumar et al. [3]. This method enhances learning efficiency by dynamically adjusting mutation and crossover strategies across two evolutionary phases. The BaDE-NN framework demonstrates superior convergence behavior and improved forecasting accuracy compared to traditional backpropagation-based neural networks. Optimization-driven hybridization is also evident in the DPSO-GA hybrid model proposed by Simaiya et al. [10], which integrates Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) with a CNN-LSTM forecasting architecture. The combined optimization mechanism fine-tunes hyperparameters and model weights more effectively than standalone optimization methods. Experimental results show significant improvements in dynamic workload provisioning, load balancing, and resource allocation efficiency. A related line of work explores the use of evolutionary optimization to enhance Support Vector Machine-based forecasting. The PSO-enhanced weighted wavelet SVM model [21] applies Particle Swarm Optimization to tune the parameters of a wavelet-based SVM. This hybrid approach outperforms four baseline algorithms including ARIMA, BPNN, standard SVR, and ANFIS both in prediction accuracy and computational efficiency. The incorporation of wavelet decomposition allows the model to capture multi-scale workload patterns, while PSO ensures optimal parameter selection. Collectively, these hybrid and optimization-enhanced models address several persistent challenges in cloud workload forecasting, including slow convergence, sensitivity to initialization, and difficulty escaping local minima. By combining deep learning architectures with evolutionary optimization techniques, these approaches achieve higher accuracy, improved robustness, and better adaptability to complex and rapidly changing cloud environments.

3.5. Autonomic and Self-Adaptive Models

Autonomic and self-adaptive forecasting models represent an emerging direction in cloud workload prediction, aiming to maintain accuracy even when workload patterns shift over time. Unlike static models that require retraining when workload characteristics change, autonomic systems continuously adjust their internal parameters or even switch forecasting strategies in response to workload drift.

One notable contribution in this category is the Self-Directed Workload Forecasting (SDWF) model [22]. SDWF introduces a self-learning mechanism that monitors recent forecasting errors

and dynamically adjusts the model's internal neuron-training process using a blackhole-inspired heuristic. By leveraging recent error trends, SDWF can rapidly correct deviations and refine its predictions without full retraining. Experimental evaluations across six real-world datasets demonstrate that SDWF achieves up to 99.99 % reduction in mean squared error (MSE) compared to state-of-the-art models, highlighting its exceptional adaptability under fluctuating workload conditions. Another important line of research focuses on adaptive prediction models that automatically select the most suitable forecasting algorithm based on the statistical properties of the incoming workload. The study [23] proposes a classifier-driven approach that analyzes historical usage features to determine whether linear, nonlinear, or hybrid predictors are most appropriate for a given workload segment. This dynamic model-selection mechanism enables the forecasting system to respond effectively to changes in workload behavior, achieving 6–27 % accuracy improvements over baseline methods. By avoiding reliance on a single forecasting technique, the model enhances robustness and reduces the risk of performance degradation during workload transitions. Together, these autonomic and self-adaptive forecasting frameworks underscore the importance of adaptability in real-world cloud environments. As cloud workloads become increasingly heterogeneous and unpredictable, models capable of self-adjustment, continuous learning, and dynamic algorithm selection are essential for maintaining high forecasting accuracy and ensuring efficient resource provisioning.

3.6. Workload Profiling, Clustering, and Pre-Forecasting Techniques

Clustering plays a crucial role in improving forecasting accuracy by grouping workloads with similar behavior. Ali and Kecskeméti [15] introduced EFection, an effectiveness-detection clustering technique for cloud workload traces. Their method identifies meaningful workload groups that can significantly enhance downstream forecasting accuracy. One such contribution is the job-pool clustering approach for smart cloud management. In the study [24], Yu et al. propose a method in which incoming jobs are first clustered based on their workload characteristics. A neural network is then trained on each cluster to learn its unique behavioral patterns. This two-stage pipeline significantly improves prediction accuracy by ensuring that the forecasting model is exposed to more homogeneous data distributions, thereby reducing variance and improving generalization. Furthermore, [25] introduces a clustering mechanism that groups virtual machines (VMs) according to their energy consumption patterns. The authors employ Gated Recurrent Units (GRU) combined with transfer learning to enhance forecasting accuracy across clusters with similar energy states. Their method achieves 87.48 % accuracy in classifying VMs by energy consumption patterns and demonstrates strong performance in predicting CPU, memory, and disk usage. This approach highlights the value of clustering not only for workload prediction but also for energy-aware resource management. A third contribution focuses on large-scale workload characterization for Google Cloud [26]. The study analyzes data from over 12,000 servers to identify workload patterns and user–task interactions. By clustering workloads based on behavioral signatures, the authors derive realistic resource utilization models that closely reflect real operational environments, achieving an error margin of less than 5 %. These models are widely used to benchmark forecasting algorithms and simulate realistic cloud conditions, making this study a foundational reference in workload profiling research. These clustering-based approaches demonstrate the importance of workload grouping as a preprocessing step for forecasting. By reducing heterogeneity and capturing latent behavioral structures, clustering enhances the performance of downstream prediction models and supports more efficient, energy-aware, and scalable cloud resource management.

Table 1 summarizes all the covered studies according to their techniques.

Table 1: Summary of Reviewed Studies

Category	Studies	Techniques Used
Statistical Models	[13], [2], [16], [17]	AutoRegressive Integrated Moving Average, ARIMA–Artificial Neural Network, AutoRegressive Neural Network, time series models
Machine Learning	[1], [9], [11], [19], [20], [23]	Clustering with Machine Learning, Ensemble Machine Learning, Swarm Intelligence Based Prediction Algorithm, Pattern Matching, Adaptive Machine Learning
Deep Learning	[7], [8], [12], [14], [17], [18]	Recurrent Neural Network, Long Short-Term Memory, Convolutional Neural Network–Long Short-Term Memory, Diffusion Convolutional Recurrent Neural Network, Convolutional Neural Network–Gated Recurrent Unit with Attention
Hybrid Models	[2], [3], [10], [13], [21]	Artificial Neural Network–Differential Evolution, BiPhase Differential Evolution Neural Network, Convolutional Neural Network–Long Short-Term Memory with Particle Swarm Optimization and Genetic Algorithm, AutoRegressive Integrated Moving Average–Artificial Neural Network, Wavelet Support Vector Machine with Particle Swarm Optimization
Clustering and Profiling	[1], [15], [24], [25], [26]	Behavioral clustering, EFection clustering engine, job pool clustering, energy state clustering, workload profiling
Autonomic and Adaptive	[22], [23], [16]	Self Directed Workload Forecasting, adaptive model selection, dynamic model switching

4. RESEARCH GAP

Despite significant progress, several gaps remain:

- **Statistical models** assume linearity and stationarity, limiting their applicability to dynamic cloud workloads [5].
- **Deep learning models** often treat all historical data uniformly, losing long-term trends and failing to adapt to workload drift [6].
- **Clustering-based** forecasting is underexplored; existing clustering methods are generic and yield limited accuracy improvements [15].
- Most models lack adaptability, requiring full retraining when workload patterns change.
- Few studies address multi-region, heterogeneous cloud fabrics, where workloads differ significantly across geographic zones.

These gaps highlight the need for hybrid, adaptive, and uncertainty-aware forecasting frameworks.

5. RESEARCH OBJECTIVES

5.1. Short-Term Objectives

- Develop a hybrid forecasting pipeline that first clusters workload traces using the EFection clustering engine [15], then feeds cluster-specific summaries into a deep learning model.
- Benchmark the proposed model using publicly available datasets such as Google Cluster and Alibaba traces.

5.2. Long-Term Objectives

- Extend the framework to support multi-region, heterogeneous cloud infrastructures.
- Integrate adaptive reclustering to handle workload drift without full model retraining.
- Enhance the ANN component with online learning to maintain accuracy during rapid fluctuations.

6. Methodology

The proposed methodology consists of the following stages as shown in Fig. 3:

- **Data Collection**: Acquire real-world workload traces from Google Cluster, Alibaba, or PlanetLab.
- **Preprocessing**: Normalize data, remove outliers, and extract temporal features.
- **Clustering**: Apply the EFection clustering engine to group workloads by behavioral similarity.
- **Modeling**: Train a deep learning model (e.g., CNN-LSTM or GRU-Attention) on cluster-specific data.
- **Optimization**: Use evolutionary algorithms (e.g., PSO, GA) to tune hyperparameters.
- **Evaluation**: Assess forecasting accuracy using RMSE, MAE, MAPE, and R^2, and evaluate clustering quality using silhouette scores or Davies–Bouldin index.

7. EVALUATION

Evaluation focuses on four dimensions:

- Forecasting Accuracy: RMSE, MAE, MAPE, R^2,
- Clustering Quality: Internal validity metrics,
- System Performance: Scalability, efficiency, and robustness under workload drift,
- Hybrid models combining clustering and deep learning are expected to outperform single-model approaches, consistent with findings from recent literature.

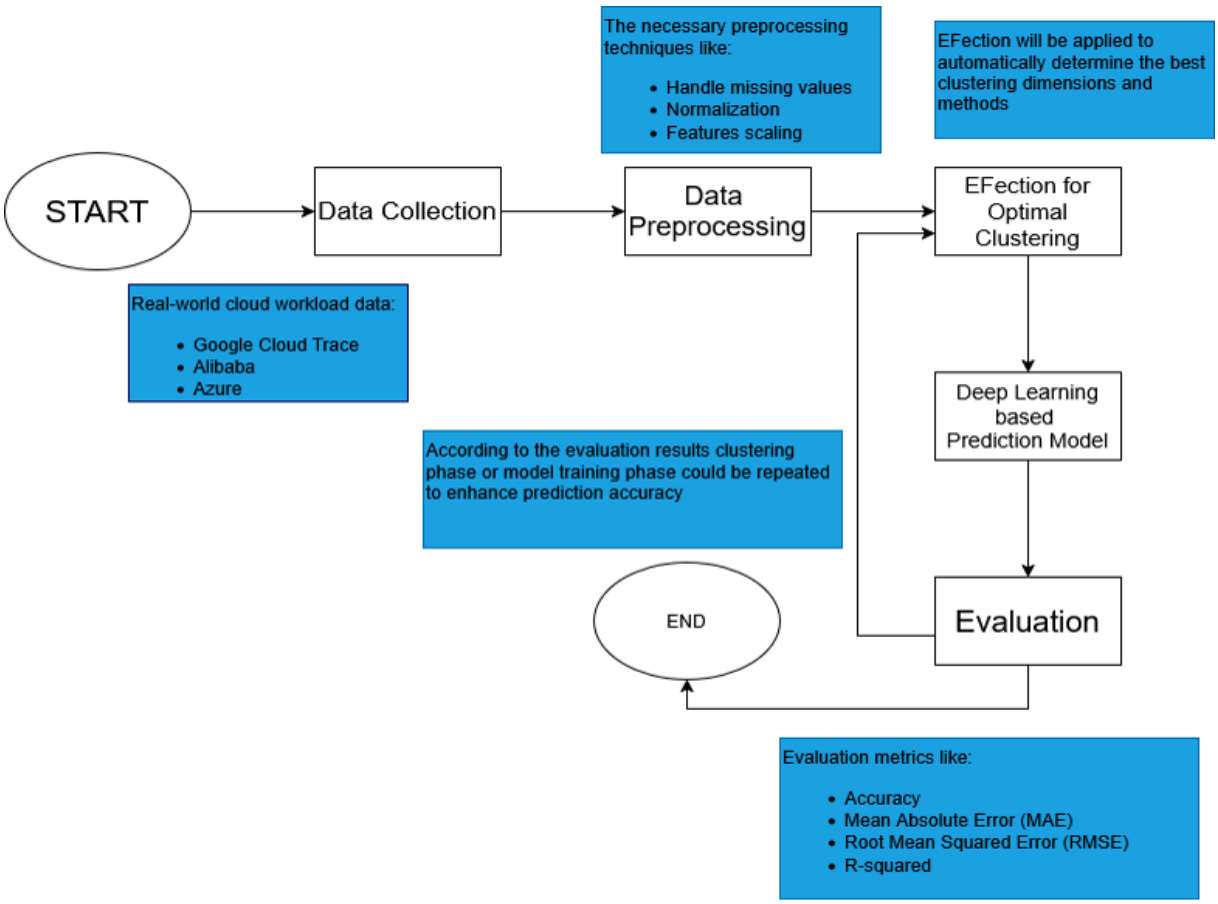


Figure 3: Research methodology work flow (by author)

8. SUMMARY AND RESEARCH DIRECTIONS

This paper reviewed major trends and techniques in cloud workload forecasting, highlighting the evolution from statistical models to advanced deep learning and hybrid approaches. While significant progress has been made, challenges remain in adaptability, multi-region forecasting, and handling workload drift. Future research should focus on:

- Online learning and continuous model adaptation,
- Transfer learning for cross-region workload prediction,
- Uncertainty-aware forecasting,
- Integration of clustering with deep learning at scale.

REFERENCES

[1] J. Gao, et al.: *Machine Learning based workload prediction in cloud computing*. In 2020 29th International Conference on Computer Communications and Networks (ICCCN), pp. 1–9, 2020.

[2] J. Kumar, A. K. Singh: *Workload prediction in cloud using Artificial Neural Network and Adaptive Differential Evolution*. Future Generation Computer Systems, vol. 81, pp. 41–52, 2018.

[3] J. Kumar, et al.: *BiPhase adaptive learning-based neural network model for cloud datacenter workload forecasting*. Soft Computing, vol. 24, pp. 14593–14610, 2020.

[4] A. Rossi, A. Visentin, D. Carraro, et al.: *Forecasting workload in cloud computing: towards uncertainty-aware predictions and transfer learning*. Cluster Computing, vol. 28, no. 258, 2025. DOI: 10.1007/s10586-024-04933-2

[5] G. Zhang, X. He, X. Wang: *PCL-RC: A parallel cloud resource load prediction model based on feature optimization*. Journal of Cloud Computing, vol. 14, no. 45, 2025. DOI: 10.1186/s13677-025-00770-9

[6] S. Deng, H. Zhao, B. Huang, et al.: *Cloud-Native Computing: A survey from the perspective of services*. Proceedings of the IEEE, vol. 112, no. 1, pp. 12–46, 2024. DOI: 10.1109/JPROC.2024.3353855

[7] M. Xu, et al.: *esDNN: Deep Neural Network based multivariate workload prediction in cloud computing environments*. ACM Transactions on Internet Technology (TOIT), vol. 22, no. 3, pp. 1–24, 2022.

[8] W. Zhang, et al.: *Workload prediction for cloud cluster using a Recurrent Neural Network*. In 2016 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI). IEEE, 2016.

[9] I. K. Kim, et al.: *Forecasting cloud application workloads with CloudInsight for predictive resource management*. IEEE Transactions on Cloud Computing, vol. 10, no. 3, pp. 1848–1863, 2020.

[10] S. Simaiya, et al.: *A hybrid cloud load balancing and host utilization prediction method using Deep Learning and optimization techniques*. Scientific Reports, vol. 14, no. 1, 1337, 2024.

[11] S. D. Pasham: *Dynamic Resource Provisioning in cloud environments using predictive analytics*. The Computertech, pp. 1–28, 2018.

[12] M. S. Al-Asaly, et al.: *A Deep Learning-based resource usage prediction model for resource provisioning in an autonomic cloud computing environment*. Neural Computing and Applications, vol. 34, no. 13, pp. 10211–10228, 2022.

[13] K. L. Devi, S. Valli: *Time series-based workload prediction using the statistical hybrid model for the cloud environment*. Computing, vol. 105, no. 2, pp. 353–374, 2023.

[14] S. Ouhame, Y. Hadi, A. Ullah: *An efficient forecasting approach for resource utilization in cloud data center using CNN-LSTM model*. Neural Computing and Applications, vol. 33, no. 16, pp. 10043–10055, 2021.

[15] S. M. Ali, G. Kecskeméti: *EFection: Effectiveness detection technique for clustering cloud workload traces*. International Journal of Computational Intelligence Systems, vol. 17, pp. 202–210, 2024. DOI: 10.1007/s44196-024-00618-1

[16] Q. Z. Ullah, S. Hassan, G. M. Khan: *Adaptive resource utilization prediction system for infrastructure as a service cloud*. Computational Intelligence and Neuroscience, 4873459, 2017.

[17] D. Janardhanan, E. Barrett: *CPU workload forecasting of machines in data centers using LSTM recurrent neural networks and ARIMA models*. In 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST). IEEE, 2017.

[18] J. Dogani, et al.: *Multivariate workload and resource prediction in cloud computing using CNN and GRU by attention mechanism*. Journal of Supercomputing, vol. 79, no. 3, 2023.

[19] H. A. Kholidy: *An intelligent swarm based prediction approach for predicting cloud computing user resource needs*. Computer Communications, vol. 151, pp. 133–144, 2020.

[20] E. Caron, F. Desprez, A. Muresan: *Forecasting for grid and cloud computing on-demand resources based on pattern matching*. In 2010 IEEE Second International Conference on Cloud Computing Technology and Science. IEEE, 2010.

[21] W. Zhong, et al.: *A load prediction model for cloud computing using PSO-based weighted wavelet support vector machine*. Applied Intelligence, vol. 48, no. 11, pp. 4072–4083, 2018.

[22] J. Kumar, A. K. Singh, R. Buyya: *Self directed learning based workload forecasting model for cloud resource management*. Information Sciences, vol. 543, pp. 345–366, 2021.

[23] W. Iqbal, et al.: *Adaptive prediction models for data center resources utilization estimation*. IEEE Transactions on Network and Service Management, vol. 16, no. 4, pp. 1681–1693, 2019.

[24] Y. Yu, et al.: *Improving the smartness of cloud management via machine learning based workload prediction*. In 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), vol. 2. IEEE, 2018.

[25] T. Khan, et al.: *Workload forecasting and energy state estimation in cloud data centres: ML-centric approach*. Future Generation Computer Systems, vol. 128, pp. 320–332, 2022.

[26] I. S. Moreno, et al.: *An approach for characterizing workloads in Google Cloud to derive realistic resource utilization models*. In 2013 IEEE Seventh International Symposium on Service-Oriented System Engineering. IEEE, 2013.

GRÁFALAPÚ ANOMÁLIADETEKTÁLÁS ELOSZTOTT RENDSZEREKBEN

Kiss Áron[1], ***Nehéz Károly***[2], ***Hornyák Olivér***[3],
[1]PhD hallgató, tanszéki mérnök, [2,3]Egyetemi docens
[1,2,3]*Gépészmérnöki és Informatikai Kar, Informatikai Intézet*
[1]kiss.aron@uni-miskolc.hu, [2]nehez.karoly@uni-miskolc.hu
[2]oliver.hornyak@uni-miskolc.hu

Absztrakt

A gráf neurális hálózatokra (GNN) épülő behatolásészlelő rendszerek (IDS) hatékony eszközt kínálnak a hálózati forgalom strukturális mintázatainak modellezésére, ugyanakkor a legtöbb létező módszer nagyméretű, hosszú időszakokból aggregált gráfokra támaszkodik, ami torzíthatja a valós környezetben várható teljesítmény megítélését. A gyakorlatban az IDS-eknek folyamatosan, időben korlátozott adatokon kell működniük alacsony késleltetés mellett. Jelen tanulmány egy GNN-alapú IDS keretrendszert mutat be, amely a hálózati forgalmat rövid, rögzített hosszúságú időablakokból felépített adatforgalmi gráfok segítségével reprezentálja. Az időosztásos gráfkonstruálás és a periodikus osztályozás lehetővé teszi a támadások korai felismerését alacsony észlelési késleltetés mellett. A gráfreprezentáció tanulására egy él-tudatos, figyelmi mechanizmust alkalmazó GNN architektúrát alkalmazunk, amely egyesíti a topológiai és adatfolyam-szintű információkat, akár 9.7%-os F1-érték javulást érve el az E-GraphSAGE alapú referenciamodellhez képest. A módszert a NF-CSE-CIC-IDS2018-v3 adathalmazon értékeltük ki bináris és többosztályos klasszifikációs feladatokban. Az eredmények megerősítik, hogy az időkorlátos gráfalapú modellezés hatékony megoldást jelent online behatolásészlelési feladatokban.

1. BEVEZETÉS

Az informatikai rendszerek gyors fejlődése és egyre komplexebb működése jelentősen megnövelte a számítógépes hálózatok támadási felületét [1]. Ezen támadások nemcsak gyakoribbá, hanem összetettebbé is váltak, és egyre gyakrabban képesek megkerülni a hagyományos védelmi mechanizmusokat [2]. Ebben a környezetben a behatolásérzékelő rendszerek (IDS) a kiberbiztonság alapvető elemeivé váltak, mivel a megelőző eszközöket –például tűzfalakat és hozzáférés-vezérlési megoldásokat– kiegészítve képesek a már folyamatban lévő rosszindulatú tevékenységek felismerésére és jelzésére [3]. Az IDS-ek kulcsszerepet töltenek be a fenyegetések azonosításában, a gyors reagálás támogatásában, valamint az események utólagos elemzésében.

A hatékony behatolásérzékelés alapfeltétele az online működés, amely a hálózati forgalom minél korábbi elemzését teszi lehetővé. A modern támadások gyakran rendkívül gyorsan eszkalálódnak, a késleltetett észlelés jelentősen csökkenti a károk mérséklésének esélyét. Az online IDS-megoldásokkal szembeni legfontosabb elvárás, hogy alacsony késleltetésű észlelést biztosítsanak, lehetővé téve a szakemberek számára a gyors beavatkozást. Ezzel szemben az offline elemzés elsősorban tanítási és utólagos eseményelemzési célokra alkalmas [4].

A hálózati forgalom természetes módon modellezhető gráfként, ahol a csúcsok végpontokat, az élek pedig a megfigyelt kommunikációs kapcsolatokat reprezentálják [5]. Ez a reprezentáció megőrzi a kommunikáció szerkezeti és relációs tulajdonságait, így alkalmas komplex támadási minták azonosítására [6]. Ennek köszönhetően az utóbbi években egyre nagyobb figyelmet kaptak a gráf neurális hálózatokra (Graph Neural Network, GNN) épülő IDS-megoldások, amelyek számos esetben felülmúlják a hagyományos, jellemzőalapú módszereket azáltal, hogy a kommunikáció topológiai mintázatait is képesek hasznosítani [7–13]. Ugyanakkor ezen megközelítések jelentős része nagyméretű, hosszú időtartamot lefedő, aggregált hálózati gráfokon

értelmezi a behatolásérzékelést, gyakran él-szintű predikciós feladatként. Bár ezek a módszerek kedvező eredményeket mutatnak akadémiai környezetben, gyakorlati alkalmazhatóságuk korlátozott, mivel a hosszú megfigyelési intervallumok elfogadhatatlan észlelési késleltetést eredményeznek valós rendszerekben [14].

Jelen tanulmány egy olyan, gráf neurális hálózatra épülő behatolásérzékelő módszert mutat be, amely kifejezetten online működésre képes, és az időben változó hálózati forgalom alapján képes az anomáliák detektálására. A javasolt megoldás rögzített hosszúságú időablakokban végzi a rendszer állapotának periodikus osztályozását, és eltérés esetén riasztást ad. Ez a megközelítés lehetővé teszi a folyamatos monitorozást, miközben kihasználja a gráf-alapú tanulás előnyeit a komplex kommunikációs mintázatok felismerésében. A bemutatott módszer alapot nyújt alacsony késleltetésű, online behatolásérzékelő rendszerek fejlesztéséhez.

2. SZAKMAI HÁTTÉR

2.1. Behatolásérzékelő és behatolásmegelőző rendszerek

A behatolásérzékelő rendszerek (IDS) a modern kiberbiztonsági architektúrák alapvető elemei. Feladatuk a rendszertevékenységek (pl. hálózati forgalom, naplóállományok, erőforrás-hozzáférések) monitorozása, valamint a rosszindulatú viselkedési minták azonosítása [15]. Az IDS-ek különösen fontos szerepet töltenek be azon támadások felismerésében, amelyek a megelőző védelmi mechanizmusokat megkerülve jutnak be a rendszerbe, így támogatva az incidenskezelést és az események utólagos elemzését.

Az IDS-ek elsődlegesen a támadások észlelését és jelzését szolgálják, míg a behatolásmegelőző rendszer (IPS) megoldások erre építve aktív beavatkozásra is képesek. A két megközelítés közötti különbség kutatási és tervezési szempontból releváns.

2.2. Gráf neurális hálózatok

A gráf neurális hálózatok (Graph Neural Networks, GNN) olyan mélytanulási modellek, amelyek gráfstruktúrájú adatok reprezentációjának tanulására szolgálnak, ahol a csúcsok entitásokat, az élek pedig azok kapcsolatait írják le [16]. A hagyományos neurális hálózatokkal szemben, amelyek euklidészi térben értelmezhető bemeneteket feltételeznek, a GNN-ek közvetlenül képesek nem euklidészi struktúrák (pl. változó topológia, fokszám és szomszédsági relációk) kezelésére, ami előnyös hálózati kommunikációs rendszerek modellezésére.

A GNN-ek működésének alapját az ún. „message passing” eljárás képezi. Legyen $G = (V,E)$ egy irányított gráf, ahol V a csúcsok, $E \subseteq V \times V$ az élek halmaza. Minden $v \in V$ csúcshoz egy kezdeti $h_v^{(0)} \in \mathbb{R}^d$ jellemzővektor tartozik. A csúcsreprezentációk rétegenként egy ún. „AGGREGATE-COMBINE” lépésben frissülnek [17]. Minden $l \in \{1,\dots,L\}$ rétegben először a szomszédos csúcsokból érkező információk aggregálása történik:

$$m_v^{(l)} = \text{AGGREGATE}^{(l)}\left(\left\{h_u^{(l-1)} : u \in \mathcal{N}(v)\right\}\right), \tag{1}$$

ahol $\mathcal{N}(v)$ a v csúcs szomszédainak halmaza. Az AGGREGATE operátor tipikusan permutációinvariáns (pl. összeg, átlag vagy maximum), így független a szomszédok sorrendjétől. Ezt követően a csúcs saját reprezentációja frissül az aggregált üzenet felhasználásával:

$$h_v^{(l)} = \text{COMBINE}^{(l)}\left(h_v^{(l-1)}, m_v^{(l)}\right). \tag{2}$$

Ez leggyakrabban egy lineáris és egy nemlineáris transzformáció kombinációjaként valósul meg:

$$h_v^{(l)} = \sigma\left(W^{(l)} \cdot \left[h_v^{(l-1)} \,||\, m_v^{(l)}\right]\right), \tag{3}$$

ahol $W^{(l)}$ az l-edik réteghez tartozó súlymátrix, $||$ a konkatenációt, σ pedig egy aktivációs függvényt jelöl. Gráfszintű feladatok esetén a végső csúcsjellemzőkből egy globális reprezentáció állítható elő a READOUT operátor segítségével:

$$h_G = \text{READOUT}\left(\left\{h_v^{(L)} : v \in V\right\}\right), \tag{4}$$

amely jellemzően szintén valamilyen permutáció-invariáns művelet.

A Graph Attention Network (GAT) architektúra ezt az általános keretrendszert figyelmi mechanizmussal bővíti [18]. A GAT esetében a szomszédos csúcsok nem azonos súllyal járulnak hozzá az aggregációhoz, hanem a modell tanulja meg az egyes kapcsolatok relatív fontosságát. Először minden csúcs jellemzője egy lineáris transzformáción megy keresztül:

$$h'_v = W \cdot h_v, \tag{5}$$

ahol h_v az eredeti jellemzővektor, W pedig egy tanítható súlymátrix. Ezután minden $(v,u) \in E$ élhez egy figyelmi érték kerül kiszámításra:

$$e_{vu} = \text{LeakyReLU}\left(a^T [h'_v \,||\, h'_u]\right), \tag{6}$$

ahol a egy tanítható paraméter vektor. A figyelmi értékek ezt követően softmax normalizáción esnek át:

$$\alpha_{vu} = \frac{\exp(e_{vu})}{\sum_{k \in (v)} \exp(e_{vk})}, \tag{7}$$

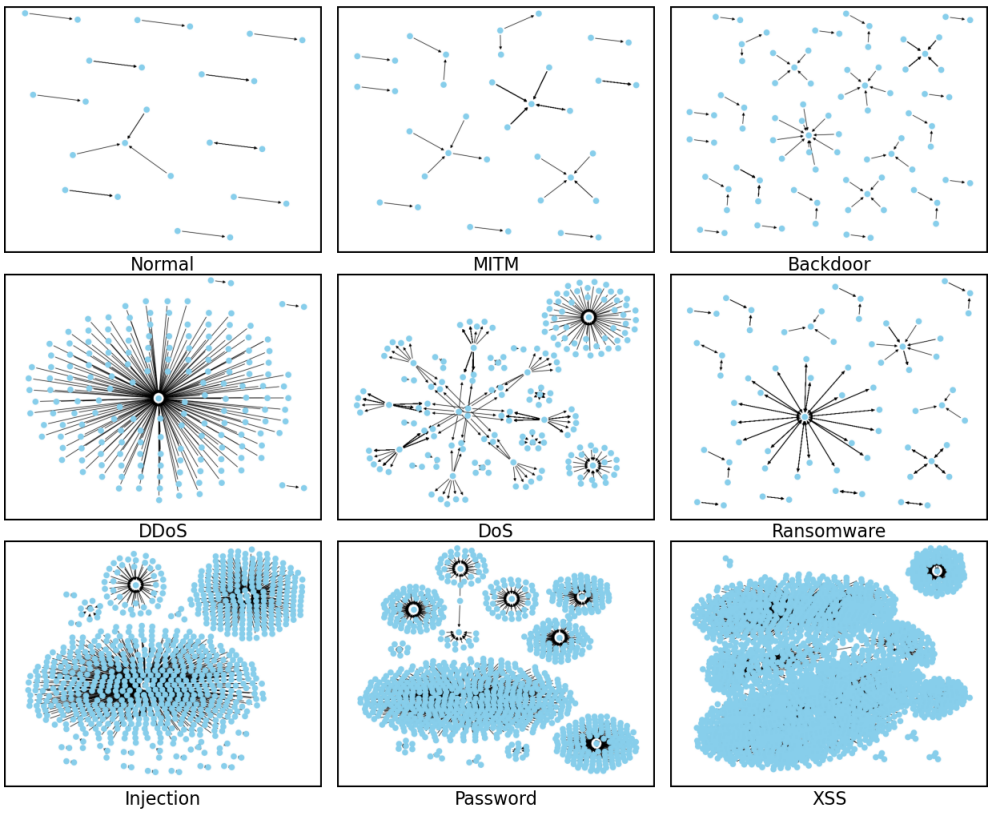
majd a frissített csúcsreprezentáció a szomszédok súlyozott összegeként képződik:

$$h''_v = \sum_{u \in \mathcal{N}(v)} \alpha_{vu} \cdot h'_u. \tag{8}$$

A klasszikus GNN- és GAT-modellek jellemzően a csúcsjellemzőkre és a gráf topológiájára támaszkodnak, miközben az élekhez kapcsolódó attribútumokat csak korlátozottan használják fel [19], ugyanakkor hálózati forgalom elemzésekor a legfontosabb információk az élekhez kötődnek (pl. időzítési vagy adatforgalmi jellemzők formájában). Az él-információk bevonására két fő irány alakult ki: (1) a gráf élgráffá alakítása [20], továbbá (2) az ún. „él-tudatos” (edge-aware) GNN-architektúrák alkalmazása, amelyek közvetlenül integrálják az él-jellemzőket a message-passing folyamatba [21, 22].

3. KAPCSOLÓDÓ MUNKÁK

Az utóbbi években számos kutatás foglalkozott gráf neurális hálózatokra épülő behatolásérzékelő rendszerek fejlesztésével, amelyek több esetben felülmúlták a hagyományos, jellemzőalapú megközelítéseket. A *GraphDDoS* módszer [12] a hálózati forgalmat végpontok közötti kommunikációs gráfokká alakítja, és kifejezetten a DDoS támadásokra jellemző csomag- és folyamszintű mintázatokat ragadja meg. A megközelítés hatékonyan modellezi a kliens–szerver interakciók időbeli és szerkezeti sajátosságait, ugyanakkor erősen specializált egyetlen támadástípusra. A Friji és társai [23] által javasolt keretrendszer általánosabb, folyamalapú gráfreprezentációt alkalmaz, amely a hálózati kommunikáció topológiai tulajdonságait használja ki. A módszer GAT-, GCN- és MLP-alapú komponensek kombinációjával képes több lépésből álló támadások detektálására, és robusztus az IP-spoofing jelenségével szemben. Az egyik leggyakrabban hivatkozott alapmodell az *E-GraphSAGE* [7], amely a *GraphSAGE* architektúra [24] kiterjesztéseként explicit módon bevonja az él-attribútumokat a tanulási folyamatba. A módszer nagyméretű, IoT környezetből származó hálózati forgalmi gráfokon végez él-szintű klasszifikációt, és számos későbbi munka kiindulópontjául szolgált [10, 25].



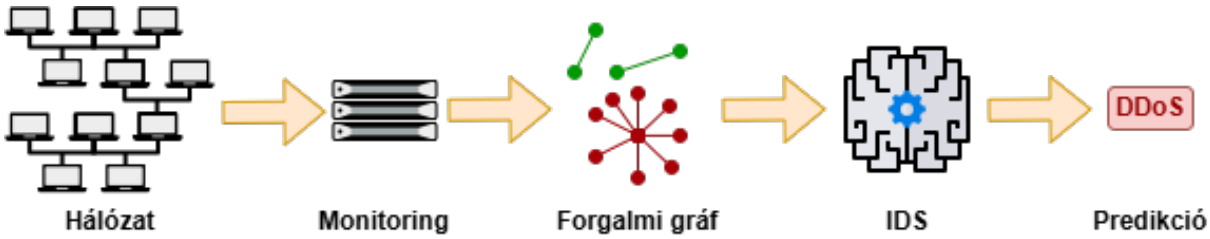
1. ábra. Hálózati forgalmi gráfok normál működés során és eltérő típusú támadások alatt a ToN-IoT adathalmazban.

4. MOTIVÁCIÓ

A mélytanulás alapú behatolásérzékelő rendszerek széles módszertani skálát ölelnek fel. Az elmúlt évtizedben például az LSTM-hálók és Transformer-alapú modellek váltak meghatározóvá, amelyek a hálózati forgalom időbeli sajátosságait használják ki összetett, több lépésből álló támadások detektálására. Az utóbbi években ezzel párhuzamosan egyre nagyobb figyelmet kaptak a gráf neurális hálózatokra épülő IDS-megoldások, amelyek számos esetben jobb teljesítményt mutattak a szekvenciaalapú módszereknél.

Több tanulmány rámutatott ugyanakkor arra, hogy a GNN-alapú IDS-ek értékelési módszertana gyakran nem felel meg a valós üzemeltetési környezet követelményeinek [14]. Sok módszer nagyméretű, hosszú időtartamot (akár több hetet) felölelő hálózati gráfokon végez tanítást és kiértékelést, miközben a tanító és tesztadatok szétválasztása gyakran véletlenszerűen, időbeli korlátok figyelembevétele nélkül történik [7, 10, 13, 26]. Ez információszivárgáshoz vezethet, mivel a modell olyan topológiai vagy időbeli mintázatokat is kihasználhat, amelyek egy online üzemelő IDS számára nem lennének elérhetők. Ennek következtében a kimutatott teljesítmény nem jellemzi jól a valós rendszerekben történő alkalmazhatóságot, valamint a hosszú megfigyelési ablakok elfogadhatatlan észlelési késleltetést eredményeznek. Ezzel szemben a rövidebb időablakok csökkentik a késleltetést, de gyakran teljesítményromlással járnak.

Ezen problémák vizsgálatához egy előzetes, kvalitatív elemzést végeztünk rövid időablakokból konstruált hálózati forgalmi gráfokon. A ToN-IoT Network [27] adathalmaz 20 másodperces szegmenseiből kiválasztott példák (1. ábra) azt mutatják, hogy már ilyen rövid időtartamok esetén is szemmel látható különbségek figyelhetők meg a normál működés és az egyes támadástípusok kommunikációs mintázatai között. Ez a megfigyelés arra utal, hogy a gráfban látható topológiai mintázatok hordozhatnak elegendő információt az anomáliák korai felismeréséhez.



2. ábra. A javasolt IDS folyamat felépítése.

Az előzetes vizsgálat alapján célunk egy olyan IDS-megközelítés létrehozása, amely rövid időablakokra építve, online módon képes a hálózati forgalom osztályozására, és ezáltal alacsony késleltetésű, gyakorlatban is alkalmazható behatolásérzékelést tesz lehetővé.

5. PROBLÉMA DEFINÍCIÓ

Bontsuk a hálózati forgalmat folytonos időben megfigyelve rögzített $\tau > 0$ hosszúságú időablakokra. Egy t időpontban kezdődő időablakot az alábbi intervallum ír le:

$$T_t = [t, t+\tau). \tag{9}$$

A T_t időablakon belül megfigyelt hálózati kommunikációt egy irányított multigráffal modellezzük:

$$G_t = (V_t, E_t, X_t), \tag{10}$$

ahol a $v \in V_t$ csúcsok egyedi hálózati végpontokat, az $e = (u,v) \in E_t$ élek pedig az ezek közötti kommunikációs folyamokat reprezentálják. Minden e élhez egy d-dimenziós jellemzővektor, $x_e \in \mathbb{R}^d$ tartozik, amely a kommunikáció tulajdonságait írja le. Az adott időablak összes éljellemzőjét az X_t halmaz jelöli.

Legyen az osztályok halmaza $\mathscr{Y} = \{0, 1, \ldots, c\}$, ahol a 0 a normál működést, míg az $1, \ldots, c$ értékek különböző támadástípusokat jelentenek. A cél egy olyan

$$M : (V_t, E_t, X_t) \to \mathscr{Y} \tag{11}$$

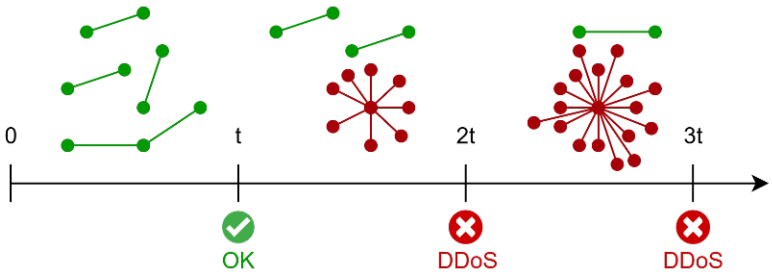
leképezés meghatározása, amely kizárólag az adott T_t időablakon belül rendelkezésre álló információk alapján határozza meg a rendszer állapotát.

6. JAVASOLT MÓDSZER

A javasolt behatolásérzékelő módszer architektúráját a 2. ábra szemlélteti. A rendszer folyamatosan monitorozza a hálózati forgalmat, és előre definiált időközönként elemzést végez annak meghatározására, hogy a vizsgált rendszer normál módon működik-e, vagy valamilyen támadás zajlik épp. Az IDS működése online jellegű: minden döntés kizárólag az aktuális időablakon belül megfigyelt forgalmi adatokon alapul, így alacsony észlelési késleltetést biztosít. A detektált események riasztás formájában továbbíthatók üzemeltetők vagy külső rendszerek felé, és alapot szolgáltathatnak további védelmi mechanizmusok aktiválásához.

6.1. Online monitorozás

A módszer időablakos megközelítést alkalmaz, amelyet a 3. ábra illusztrál. A hálózati forgalom folyamatos rögzítése mellett minden t időpontban a rögzített hosszúságú időablakhoz tartozó



3. ábra. Időablak alapú online behatolásérzékelés.

forgalomból egy hálózati gráf kerül kialakításra. A rendszer célja, hogy a támadásokat –például egy DDoS támadás kezdeti fázisát– a lehető legkorábbi időpillanatban felismerje. Ahogy az idő előrehalad, és a támadás kiterjedtebbé válik, a kommunikációs mintázatok egyre inkább eltérnek a normál működés során megfigyeltektől. A javasolt IDS ezen változások folyamatos nyomon követésére és korai detektálására törekszik.

6.2. Gráf konstrukció

A hálózati forgalmat egy irányított multigráf reprezentálja, ahol a csúcsok egyedi IP:port pároknak felelnek meg, az élek pedig az adott időablakon belül megfigyelt kommunikációs folyamokat jelölik. Minden élhez egy jellemzővektor tartozik, amely statisztikai (pl. küldött és fogadott csomagok száma) és protokollszintű (pl. TCP flag-ek, kapcsolat állapota) információkat foglal magában. Az IP- és portszámok kizárólag a gráf szerkezetének felépítéséhez kerülnek felhasználásra, a modell tanításakor nem jelennek meg bemeneti jellemzőként. Ez a döntés megakadályozza, hogy a modell nem általánosítható korrelációkra építsen, mivel a legtöbb, támadásokat tartalmazó hálózati adathalmazban a rosszindulatú forgalom jellemzően néhány IP-címhez köthető. Lo és társai [7] ajánlásait követve a csúcsjellemzők egységes, konstans vektorokként kerülnek inicializálásra, amelyek dimenziója megegyezik az éljellemzők dimenziójával, a numerikus stabilitás biztosítása érdekében; a csúcsvektorok kezdetben minden komponensükben 1 értéket vesznek fel, míg a forgalom érdemi leírását az élekhez rendelt jellemzők hordozzák.

A portszintű reprezentáció lehetővé teszi az azonos IP-címen futó különböző szolgáltatások elkülönítését, ami finomabb és informatívabb gráfszerkezetet eredményez, mint a pusztán IP-alapú modellezés. Az élek hálózati folyam (traffic flow) alapú reprezentációja egyensúlyt teremt a részletesség és a skálázhatóság között: elegendő információt biztosít az anomáliák felismeréséhez, miközben elkerüli a csomagszintű feldolgozás jelentős számítási és tárolási költségeit.

6.3. Reprezentáció tanulás

A javasolt módszer enkóder-dekóder architektúrát követ, ahol az enkóder feladata a forgalmi gráf tömör és informatív reprezentációjának előállítása gráf neurális háló segítségével. Ennek megvalósítására az EdgeGAT [22] architektúrát alkalmaztuk, amely az üzenettovábbítás során explicit módon integrálja az éljellemzőket. Ez különösen előnyös hálózati forgalmi gráfok esetén, ahol a releváns információk döntően az élekhez kötődnek.

Az EdgeGAT rétegekben egy v_i csúcs frissített reprezentációja a szomszédos csúcsok, valamint az őket összekötő élek információinak együttes figyelembevételével kerül kiszámításra:

$$v_i' = W_s \cdot v_i + \sum_{j \in \mathcal{N}(v_i)} \alpha_{j,i} \left(W_n \cdot v_j + W_e \cdot e_{j,i} \right), \tag{12}$$

ahol $\mathcal{N}(v_i)$ a v_i csúcs szomszédainak halmaza, míg W_s, W_n és W_e rendre a csúcs-, szomszéd- és

éljellemzőkhez tartozó tanítható súlymátrixok. Az $\alpha_{j,i}$ figyelmi súlyok meghatározása a következő módon történik:

$$\alpha_{j,i} = \text{softmax}_i \left(\text{LeakyReLU} \left(a^T \left[W_n \cdot v_i \,||\, W_n \cdot v_j \,||\, W_e \cdot e_{j,i} \right] \right) \right), \tag{13}$$

ahol a egy tanítható paramétervektor, a $||$ jel a konkatenációt, míg a softmax_i a normalizálást jelenti a v_i-be beérkező élek mentén. A figyelmi mechanizmus lehetővé teszi, hogy a modell eltérő súlyt rendeljen az egyes kommunikációs kapcsolatokhoz, és így a döntés szempontjából leginformatívabb forgalmi mintázatokra fókuszáljon.

A GNN-rétegek kimeneteként keletkező csúcsbeágyazásokat „max pooling" művelettel aggregáljuk egy fix dimenziójú gráfszintű reprezentációvá, amely tömören összefoglalja az adott időablak hálózati forgalmának szerkezeti és statisztikai jellemzőit, és alkalmas bemenetet biztosít a későbbi osztályozási lépéshez.

6.4. Osztályozás

A kapott gráfbeágyazás egy többrétegű perceptron (MLP) bemenetét képezi, amely gráfszintű klasszifikációt végez. Az MLP dekóder egyszerű, mégis hatékony megoldást kínál a gráf reprezentációk feldolgozására, és lehetővé teszi annak meghatározását, hogy az adott időablak normál működésnek vagy egy konkrét támadástípusnak felel-e meg.

7. KIÉRTÉKELÉS

Ebben a fejezetben a javasolt módszer teljesítményét elemezzük azt vizsgálva, hogy a modell milyen hatékonysággal képes a hálózati behatolások detektálására.

7.1. Adathalmaz

A kiértékeléshez a NF-CSE-CIC-IDS2018-v3 [28] adathalmazt használtuk, amely a CSE-CIC-IDS2018 benchmark [29] NetFlow v3 alapú, időbeli jellemzőkkel kiegészített változata. Az adathalmaz valósághű hálózati környezetben gyűjtött hálózati adatokat, és különböző temporális statisztikákat tartalmaz. Az adathalmaz kb. 20 millió hálózati flow-ból áll, statisztikai eloszlása közelíti a valós üzemeltetési környezeteket: a forgalom 87 %-a normál működéshez kötődik, míg a fennmaradó 13 % támadási minta 6 különböző támadási osztály között oszlik meg.

7.2. Adatok előfeldolgozása

Az előfeldolgozás során a hálózati folyamokat időrendbe rendeztük, majd 20 másodperces időablakokra bontottuk azokat. Az elemzéshez nem releváns attribútumokat eltávolítottuk, így 46 db éljellemző maradt meg. A forrás- és céloldali socketeket IP:port párokból képeztük, azonban ezek kizárólag a gráf topológiájának felépítésére szolgáltak, a modell bemeneteként nem kerültek felhasználásra.

Az időablakok címkézése a flow-szintű annotációk alapján történt: a kizárólag normál forgalmat tartalmazó ablakok normál címkét kaptak, míg az egyetlen támadástípust tartalmazó ablakokat a megfelelő osztályhoz rendeltük; több támadástípus együttes előfordulása esetén az érintett időablakokat kizártuk. A tanító és teszt halmazt időben elkülönítve 80 %–20 %-os megosztás szerint alakítottuk ki. A kategorikus jellemzőket célkódolással [30], a numerikus jellemzőket z-score normalizálással dolgoztuk fel, kizárólag a tanítóhalmazon illesztett paraméterekkel.

7.3. Hardver és szoftverkörnyezet

A kísérleteket Ubuntu 22.04.1 környezetben végeztük, Windows 11 alapú gazdagépen futó WSL2 alatt. A hardverkörnyezet egy NVIDIA RTX 3050 Ti grafikus kártyát és 32 GB rendszermemóriát tartalmazott. A megvalósítás Python 3.12.11 nyelven készült, az alábbi fő könyvtárak felhasználásával: PyTorch 2.4.0 (CUDA 12.4 támogatással) a mélytanulási modellekhez, DGL 2.4.0 a gráfmodellezéshez, valamint scikit-learn 1.6.1 az adatok előfeldolgozásához és a kiértékeléshez.

7.4. Metrikák

A javasolt módszer teljesítményének értékelésére a széles körben alkalmazott $\text{precision} = \dfrac{\text{TP}}{\text{TP}+\text{FP}}$, $\text{recall} = \dfrac{\text{TP}}{\text{TP}+\text{FN}}$, és $\text{F1} = 2 \cdot \dfrac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$ metrikákat használtuk, ahol a TP a helyesen pozitívnak osztályozott minták számát, az FP a tévesen pozitívként besorolt mintákat, az FN az adott osztályba tartozó, de más osztályként azonosított mintákat, a TN az összes többi, helyesen azonosított, az adott osztályba nem tartozó mintát jelöli. A közölt eredmények osztályonként súlyozott átlagolással kerültek kiszámításra, ahol a súlyok az egyes osztályok adathalmazon belüli gyakoriságának felelnek meg.

7.5. Referenciamodellek

A javasolt módszer teljesítményének értékeléséhez három referenciamodellt implementáltunk, amelyeket bináris és többosztályos klasszifikációs feladatokon egyaránt teszteltünk, majd eredményeiket összevetettük az általunk javasolt megközelítéssel. A referenciamodellek eltérő adatfeldolgozási paradigmákat képviselnek, lehetővé téve a különböző információforrások (statisztikai, szekvenciális, strukturális) összehasonlítását. A betanítást minden modell esetén 5 epoch türelmi idejű korai leállítással végeztük.

Az első referenciamodell egy többrétegű perceptron (MLP) volt, amely nem képes közvetlenül gráfstruktúrák vagy idősorok feldolgozására. Ennek megfelelően az egyes időablakokhoz tartozó hálózati forgalmat statisztikai aggregálással egyetlen jellemzővektorrá alakítottuk. Minden éljellemzőre kiszámítottuk az átlagot, szórást, minimumot és maximumot, így egy 4 dimenziós bemeneti vektort kaptunk. Az MLP két rejtett rétegből állt (rendre 256 és 128 neuron), a tanítást 128-as batch mérettel végeztük.

A második referenciamodell egy *Long Short-Term Memory* (LSTM) háló volt, amely képes szekvenciális adatok feldolgozására. Az adott időablakon belüli forgalmi flow-kat kezdési idő szerint rendeztük, majd sorozatként adtuk át a modellnek. Az LSTM utolsó rejtett állapotát egy teljesen összekapcsolt réteg segítségével transzformáltuk osztályeloszlássá. A modell 64 dimenziós rejtett állapotot használt, és a tanítás 128-as batch mérettel történt.

A harmadik referenciamodell az *E-GraphSAGE* modell volt, amely közvetlenül gráfstruktúrákon működik. A módszert időablakos működésre és gráfszintű klasszifikációra adaptáltuk. A kapcsolódó munkák ajánlásait követve két GNN-réteget és ReLU aktivációt alkalmaztunk, a tanítást pedig 128-as batch mérettel végeztük.

Saját modellünk esetében három GNN-réteget alkalmaztunk, 128 dimenziós rejtett reprezentációval és 8 figyelmi fejjel. A tanítást 128-as batch mérettel, 0.001-es tanulási rátával végeztük.

A kísérleteink 3 célt szolgáltak. Egyrészt azt vizsgáltuk, hogy a flow-szintű reprezentáció mennyiben járul hozzá a pontos behatolásérzékeléshez a durvább, statisztikai aggregáláson alapuló megközelítésekhez képest. Másrészt elemeztük, hogy a hálózati forgalom esetén a kom-

1. táblázat. Bináris osztályozás teljesítménymutatói. A legjobb értékek félkövérrel vannak kiemelve.

Modell	Precision	Recall	F1
Saját módszer	**0,8740**	**0,8555**	**0,8582**
E-GraphSAGE	0,8533	0,7780	0,7820
LSTM	0,8483	0,7748	0,7790
MLP	0,8189	0,7880	0,7856

2. táblázat. Többosztályos osztályozás teljesítménymutatói. A legjobb értékek félkövérrel vannak kiemelve.

Modell	Precision	Recall	F1
Saját módszer	**0,8912**	**0,8243**	**0,8370**
E-GraphSAGE	0,8759	0,7638	0,7838
LSTM	0,5711	0,7169	0,6255
MLP	0,7963	0,6677	0,7028

munikáció strukturális tulajdonságai informatívabbak-e, mint a pusztán időbeli sorrendiség. Továbbá célunk volt annak bemutatása, hogy a javasolt módszer versenyképes, illetve kiemelkedő teljesítményt ér el korszerű referenciamodellekkel összevetve.

7.6. Eredmények

A bináris osztályozás során az összes támadást egyetlen osztályba vontuk össze, míg a normál forgalom külön osztályként szerepelt. Az eredményeket az 1. táblázat foglalja össze. A javasolt módszer 85,82 %-os F1-értéket ért el, amely minden vizsgált referenciamodellt felülmúl. Az E-GraphSAGE és az LSTM modellek teljesítménye közel azonos, míg az MLP alacsonyabb precízió mellett magasabb recall értéket mutatott. A vizsgált modellek viselkedése jól szemlélteti a precision–recall kompromisszumot. Az MLP nagyobb érzékenységgel ismeri fel a támadásokat, azonban ez a pontosság rovására megy, ami több téves riasztást eredményez.

A többosztályos osztályozás eredményeit az 1. táblázat mutatja be. Ebben a beállításban a javasolt módszer 83,7 %-os F1-értéket ért el, amely jelentősen meghaladja mind a szekvenciális (LSTM), mind a statisztikai aggregáción alapuló (MLP) referenciamodell teljesítményét, és szignifikánsan felülmúlja az adaptált E-GraphSAGE modellt is.

Az LSTM modell gyenge precíziója arra utal, hogy a hálózati forgalom esetében az események időbeli sorrendje önmagában nem hordoz elegendő diszkriminatív információt a támadástípusok elkülönítéséhez. Ezzel szemben az MLP pusztán numerikus jellemzők alapján is elfogadható eredményt ért el, ami megerősíti, hogy a flow-szintű statisztikák informatívak. Ugyanakkor mindkét modell jelentősen elmarad a gráf-alapú megközelítésektől, ami egyértelműen jelzi, hogy a kommunikáció szerkezeti tulajdonságai kulcsszerepet játszanak a pontos osztályozásban.

Az adathalmaz osztályeloszlása kiegyensúlyozatlan, ezért az aggregált metrikák mellett a normalizált konfúziós mátrix (3. táblázat) elemzése is indokolt. Az eredmények erős diagonális dominanciát mutatnak, ami azt jelzi, hogy a modell a legtöbb osztály esetében robusztusan teljesít.

A „Bruteforce”, „DDoS”, és „Infiltration” támadások felismerése kiemelkedően sikeres, a modell ezeknél közel 100 %-os pontosságot ér el. Ez arra utal, hogy a nagyméretű, elosztott támadások jellegzetes kommunikációs mintázatai jól elkülöníthetők gráf-alapú reprezentációval.

Jelentős eltérés a „DoS” osztály esetében tapasztalható, ahol a minták egy része normál forgalomként került besorolásra. Ez részben a viszonylag alacsony elemszámnak, részben pedig

3. táblázat. A javasolt módszer normalizált konfúziós mátrixa a vizsgált adathalmazon

Valós címke	Predikált címke							Mintaszám
	Bot	**Brutef.**	**DDoS**	**DoS**	**Infiltr.**	**Web**	**Normal**	
Bot	0,99	0,00	0,00	0,00	0,00	0,00	0,00	203
Brutef.	0,00	1,00	0,00	0,00	0,00	0,00	0,00	114
DDoS	0,00	0,00	1,00	0,00	0,00	0,00	0,00	94
DoS	0,00	0,00	0,00	0,56	0,00	0,00	0,44	75
Infiltr.	0,00	0,00	0,00	0,00	0,98	0,00	0,02	638
Web	0,00	0,00	0,00	0,00	0,00	0,83	0,16	225
Normal	0,00	0,00	0,00	0,00	0,24	0,00	0,76	2470

annak tudható be, hogy a DoS támadások kezdeti fázisa gyakran kevésbé különbözik a normál működéstől. A „Web” alapú támadások esetében a pontosság mérsékeltebb, ami arra utal, hogy ezek forgalmi mintázatai részben átfednek a legitim webes kommunikációval. A normál forgalom esetében a téves besorolások döntően az „Infiltration” osztály irányába történtek, ami jól tükrözi az ilyen támadások célját: a legitim viselkedéshez hasonló mintázatok előállítását.

Az eredmények alapján megállapítható, hogy a javasolt módszer hatékonyan és robusztusan detektálja a hálózati behatolásokat a vizsgált adathalmazon. Mind bináris, mind többosztályos beállításban felülmúlja a korszerű baseline modelleket. A konfúziós mátrix elemzése azt mutatja, hogy a jó teljesítmény nem csupán a többségi osztályokra korlátozódik, hanem a ritkább támadástípusokra is kiterjed. Ennek alapján a módszer alkalmas online IDS-ként történő alkalmazásra, ahol a pontos és gyors észlelés kiemelt jelentőséggel bír.

8. KÖVETKEZTETÉSEK

A bemutatott eredmények alapján megállapítható, hogy az időkorlátos adatokból épített hálózati gráfok is alkalmasak hatékony behatolásészlelésre. A kísérletek igazolták, hogy a gráfalapú modellezés képes megragadni azokat a strukturális kommunikációs mintázatokat, amelyek a pusztán statisztikai vagy szekvenciális megközelítések számára nehezen felismerhetők. Az időosztásos monitorozásra épülő feldolgozás és az él-tudatos, figyelmi mechanizmust alkalmazó GNN-modell következetesen felülmúlta a referenciamodelleket, ami alátámasztja a gráfalapú reprezentációk előnyeit hálózati forgalom elemzésekor. Az eredmények azt is jelzik, hogy az online működéshez szükséges alacsony késleltetés nem jár szükségszerűen jelentős teljesítményromlással, így a megközelítés gyakorlati IDS-ekben is alkalmazható lehet.

A módszer jelenlegi formájában azonban rendelkezik korlátokkal is. Az osztályozás időablakonként egyetlen támadástípust feltételez, ami korlátozza a párhuzamos vagy egymásra épülő támadások felismerését. Emellett a modell kizárólag a tanítás során megfigyelt támadási minták detektálására képes, a korábban nem látott támadások azonosítása még nem megoldott. Skálázhatósági kihívást jelenthetnek továbbá a nagy méretű gráfok, például extrém mértékű DDoS vagy kiterjedt szkennelési támadások esetén, ahol a számítási igény gyorsan megnőhet.

A jövőbeni kutatások egyik iránya a módszer további adathalmazokon történő kiértékelése. Ígéretes fejlesztési lehetőség a többcímkés klasszifikáció bevezetése, amely lehetővé tenné párhuzamos támadások egyidejű felismerését, valamint a modell kiterjesztése korábban nem látott támadások detektálására adaptív vagy inkrementális tanulási stratégiák segítségével. További irányként a gráf-alapú magyarázatgenerálás bevonása vizsgálható, amely elősegítheti a riasztások gyorsabb értelmezését és a támadások hatékonyabb elhárítását. Végül, az extrém nagy gráfok kezelésére mintavételezési és adaptív részgráf-kiválasztási technikák vizsgálata szükséges a módszer skálázhatóságának biztosításához.

IRODALOMJEGYZÉK

[1] L. Diana, P. Dini, D. Paolini: *Overview on Intrusion Detection Systems for computers networking security*. Computers, vol. 14, no. 3, p. 87, March 2025. DOI: 10.3390/computers14030087

[2] R. Prakash, N. Jyoti, S. Manjunatha: *A survey of security challenges, attacks in IoT*. E3S Web of Conferences, vol. 491, p. 04018, Feb. 2024. DOI: 10.1051/e3sconf/202449104018

[3] A. M. Awaad, K. M. A. Alheeti, A. K. A. H. Najem: *Anomaly-Based IDS (Intrusion Detection System) for Cyber-Physical Systems*. Mesopotamian Journal of Big Data, vol. 2024, pp. 230–240, Dec. 2024. DOI: 10.58496/MJBD/2024/017

[4] B.-V. Vasilică, F.-D. Anton, A. D. Ioniță: *Evaluation of IBM Watson Services for IDS implementation*. U.P.B. Scientific Bulletin, Series C, vol. 86, no. 4, pp. 183–198, 2024.

[5] M. Zhong, M. Lin, C. Zhang, Z. Xu: *A survey on graph neural networks for intrusion detection systems: Methods, trends and challenges*. Computers & Security, vol. 141, p. 103821, June 2024. DOI: 10.1016/j.cose.2024.103821

[6] T. Bilot, N. E. Madhoun, K. A. Agha, A. Zouaoui: *Graph Neural Networks for Intrusion Detection: A Survey*. IEEE Access, vol. 11, pp. 49114–49139, 2023. DOI: 10.1109/ACCESS.2023.3275789

[7] W. W. Lo, S. Layeghy, M. Sarhan, M. Gallagher, M. Portmann: *E-GraphSAGE: A Graph Neural Network based Intrusion Detection System for IoT*. In NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium, IEEE, pp. 1–9. Apr. 2022. DOI: 10.1109/NOMS54207.2022.9789878

[8] T. Altaf, X. Wang, W. Ni, G. Yu, R. P. Liu, R. Braun: *GNN-based network traffic analysis for the detection of sequential attacks in IoT*. Electronics, vol. 13, no. 12, p. 2274, June 2024. DOI: 10.3390/electronics13122274

[9] Z. Sun, A. M. H. Teixeira, S. Toor: *GNN-IDS: Graph Neural Network based Intrusion Detection System*. In Proceedings of the 19th International Conference on Availability, Reliability and Security, ACM, pp. 1–12, July 2024. DOI: 10.1145/3664476.3664515

[10] L. Chang, P. Branco: *Embedding residuals in graph-based solutions: the E-ResSAGE and E-ResGAT algorithms. A case study in intrusion detection*. Applied Intelligence, vol. 54, no. 8, pp. 6025–6040, April 2024. DOI: 10.1007/s10489-024-05404-2

[11] W. W. Lo, G. Kulatilleke, M. Sarhan, S. Layeghy, M. Portmann: *XG-BoT: An explainable Deep Graph Neural Network for botnet detection and forensics*. Internet of Things, vol. 22, p. 100747, July 2023. DOI: 10.1016/j.iot.2023.100747

[12] Y. Li et al.: *GraphDDoS: Effective DDoS attack detection using Graph Neural Networks*. In 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE, pp. 1275–1280, May 2022. DOI: 10.1109/CSCWD54268.2022.9776097

[13] D. Pujol-Perich, J. Suarez-Varela, A. Cabellos-Aparicio, P. Barlet-Ros: *Unveiling the potential of Graph Neural Networks for robust Intrusion Detection*. ACM SIGMETRICS Performance Evaluation Review, vol. 49, no. 4, pp. 111–117, June 2022. DOI: 10.1145/3543146.3543171

[14] A. Venturi, D. Pellegrini, M. Andreolini, L. Ferretti, M. Marchetti, M. Colajanni: *Practical evaluation of Graph Neural Networks in Network Intrusion Detection*. In Proceedings of ITASEC 2023: The Italian Conference on CyberSecurity, CEUR Workshop Proceedings, 2023. [Online]. Available: https://ceur-ws.org/Vol-3488/paper29.pdf

[15] A. Shamim, B. Fayyaz, V. Balakrishnan: *Layered defense in depth model for IT organizations*. In 2nd International Conference on Innovations in Engineering and Technology (ICCET'2014) Sept. 19-20, 2014 Penang (Malaysia), International Institute of Engineers, pp. 21–24, 2014. DOI: 10.15242/IIE.E0914047

[16] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini: *The Graph Neural Network Model*. IEEE Transactions on Neural Networks, vol. 20, no. 1, pp. 61–80, Jan. 2009. DOI: 10.1109/TNN.2008.2005605

[17] C. Ying, et al.: *Do transformers really perform bad for graph representation?*. November, 2021.

[18] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio: *Graph Attention Networks*. February, 2018.

[19] M. T. Tahmid, T. A. Zaman, M. S. Rahman: *GraFusionNet: Integrating node, edge, and semantic features for enhanced graph representations*. November, 2024. DOI: 10.1101/2024.11.22.624875

[20] F. Harary, R. Z. Norman: *Some properties of line digraphs*. Rendiconti del Circolo Matematico di Palermo, vol. 9, no. 2, pp. 161–168, May 1960. DOI: 10.1007/BF02854581

[21] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, J. M. Solomon: *Dynamic Graph CNN for Learning on Point Clouds*. ACM Transactions on Graphics, vol. 38, no. 5, pp. 1–12, October, 2019. DOI: 10.1145/3326362

[22] T. Monninger et al.: *SCENE: Reasoning about traffic scenes using heterogeneous Graph Neural Networks*. IEEE Robotics and Automation Letters, vol. 8, no. 3, pp. 1531–1538, March, 2023. DOI: 10.1109/LRA.2023.3234771

[23] H. Friji, A. Olivereau, M. Sarkiss: *Efficient network representation for GNN-based Intrusion Detection*. pp. 532–554. 2023. DOI: 10.1007/978-3-031-33488-7_20

[24] W. L. Hamilton, R. Ying, J. Leskovec, *Inductive representation learning on large graphs*. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17), Curran Associates Inc., pp. 1025–1035, 2017. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf

[25] E. Caville, W. W. Lo, S. Layeghy, M. Portmann: *Anomal-E: A self-supervised network intrusion detection system based on Graph Neural Networks*. Knowledge-Based Systems, vol. 258, p. 110030, December, 2022. DOI: 10.1016/j.knosys.2022.110030

[26] Q. Xiao, J. Liu, Q. Wang, Z. Jiang, X. Wang, Y. Yao: *Towards network anomaly detection using graph embedding*. pp. 156–169, 2020. DOI: 10.1007/978-3-030-50423-6_12

[27] N. Moustafa: *A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets*. Sustainable Cities and Society, vol. 72, p. 102994, Sept. 2021. DOI: 10.1016/j.scs.2021.102994

[28] M. Luay, S. Layeghy, S. Hosseininoorbin, M. Sarhan, N. Moustafa, M. Portmann: *Temporal analysis of netflow datasets for Network Intrusion Detection Systems*, March 2025.

[29] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani: *Toward generating a new intrusion detection dataset and intrusion traffic characterization*. In Proceedings of the 4th International Conference on Information Systems Security and Privacy, SCITEPRESS - Science and Technology Publications, pp. 108–116, 2018. DOI: 10.5220/0006639801080116

[30] scikit-learn developers, *TargetEncoder*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.TargetEncoder.html

A FELSZÍNREKONSTRUKCIÓ MATEMATIKAI ÉS IMPLEMENTÁCIÓS MEGKÖZELÍTÉSEI

Molnár-Zékány Szabina[1], ***Árvai-Homolya Szilvia***[2]
[1]PhD hallgató, fejlesztő, [2], 3Egyetemi docens
[1,2]*Gépészmérnöki és Informatikai Kar, Informatikai Intézet*
[2]*Bay Zoltán Alkalmazott Kutatási Közhasznú Nonprofit Kft.*
[1]zekany.szabina@student.uni-miskolc.hu,
[2]szilvia.homolya@uni-miskolc.hu

Absztrakt

A háromdimenziós szkennelési technológiák eredményeként digitális objektumok jönnek létre, amelyek a vizsgált valós felületeket geometriai és topológiai szempontból írják le. Jelen tanulmány egy nyomási fekélyek LiDAR-alapú háromdimenziós pontfelhő, illetve háló felszínrekonstrukciójának megvalósítására, illetve matematikai megvalósításának lehetőségeire fókuszál. A szkennelési folyamat eredményeként létrejövő pontfelhő matematikai algoritmusok segítségével háromszögelt hálóvá alakítható, amely közelíti a valós sebgeometriát. A cikk célja a felszínrekonstrukció matematikai alapjainak, valamint a leggyakrabban alkalmazott algoritmusok implementációs sajátosságainak bemutatása, különös tekintettel a pontfelhő-alapú megközelítésekre és azok gyakorlati alkalmazhatóságára klinikai környezetben.

1. BEVEZETÉS

A háromdimenziós felszínek pontos és reprodukálható rekonstruálása kiemelt jelentőséggel bír a számítógépes grafika, a számítógépes látás, az orvosi képalkotás és a mérnöki tervezés számos területén. Az orvosi alkalmazások esetében a geometriai pontosság nem pusztán vizuális kérdés, hanem közvetlen hatással van a diagnosztikai és monitorozási folyamatokra. A nyomási fekélyek világszerte jelentős egészségügyi terhet jelentenek ezért a seb rekurzív felülvizsgálata nagy jelentőséggel bír. A seb felszínének kvantitatív, reprodukálható 3D rekonstrukciója lehetővé teszi mélységi, felületi és térfogat-alapú mérések elvégzését, amelyek elengedhetetlenek az állapot objektív nyomon követéséhez. [1] A mérések megfelelő elvégzéséhez topológiailag zárt, folytonos felszínmodell szükséges. A hagyományos, manuális mérési módszerek szubjektívek és nehezen reprodukálhatók, ezért az utóbbi években egyre nagyobb hangsúlyt kapnak a digitális, háromdimenziós mérési technikák. Számos tanulmány foglalkozik ezzel a témával a test különféle területein megtalálható sebek méréseivel. [2, 3, 4] A közelmúlt kutatásai például fotogrammetrián és strukturált fényű szkennelési technikákon alapulnak, ugyanakkor a lézerszkennerek alkalmazása is megfelelő megoldást nyújthat. [5] A klinikai környezetben is ideális megoldás az adott eltervezett rendszertől és a piacképes elvárásoktól is függhet illetve a probléma pontos megfogalmazásán is alapul.

Jelen tanulmány egy nagyobb kutatási és fejlesztési projekt részeként a háromdimenziós szkennelés és felszínrekonstrukció matematikai és implementációs kérdéseit vizsgálja. A kutatás célja egy hordozható, költséghatékony, klinikai környezetben is alkalmazható szkennelési megoldás elemzése, amely támogatja az ápolási és sebmonitorozási folyamatokat.

2. GEOMETRIAI ADATGYŰJTÉS ÉS FELSZÍNREPREZENTÁCIÓ

Egy háromdimenziós objektum digitális reprezentációja során a valós felszín diszkrét mintavétele történik, amelynek eredménye egy háromdimenziós pontfelhő. A pontfelhő a felszín térbeli helyzetét írja le, topológiai információt azonban nem tartalmaz. Az adatgyűjtés LiDAR (Light

Detection And Ranging) technológián alapult, amely aktív távolságméréssel határozza meg a felszín pontjainak koordinátáit a kibocsátott és visszavert lézerimpulzusok közötti időkülönbség alapján. A mérések egy iPhone 13 Pro készülék beépített LiDAR szenzorával, a Scaniverse alkalmazás segítségével történtek, kiegészítve az RGB kamera által szolgáltatott képi információkkal. A szkennelés manuális, körbejárásos módszerrel zajlott, egyenletes megvilágítás mellett. A szkennelés eredményeként kapott pontfelhő a vizsgált felszín diszkrét közelítését adja. A folytonos felszín előállításához a pontfelhőből háromszögelt felületi háló generálható, amely csúcsokból, élekből és háromszög alakú felületelemekből épül fel. A hálóhoz rendelt felületi normálvektorok a lokális felszínorientációt írják le, és alapvető szerepet játszanak a geometriai elemzés, a vizualizáció és a további mérési feladatok során. [6] A Scaniverse alkalmazás a rögzített mélységi és képi adatok integrálásával automatikus felszínrekonstrukciót hajt végre, amelynek eredménye egy háromszögelt hálóreprezentáció. Így rendelkezésünkre áll a későbbi mnkák során egy szkennelt adatbázis a beszkennelt betegágy mellett végzett szkenneléseknek köszönhetően. A sebfelszín meghatásozásában segít egy erre kifejlesztett neurális háló. A háló minél több adatot kap feldolgozásra, annál pontosabban működhet a megfelelő beállítások és a minőségi tanítóadatok mellet. A saját adatbázis A saját mérésekből származó adatállomány korlátozott mérete miatt a robusztus tanítás érdekében szükséges további, külső forrásból származó adatok bevonása. Ehhez a közelmúltban kifejezetten háromdimenziós sebfelszínek feldolgozására létrehozott, nyilvánosan elérhető –részben szintetikus– adatbázisok is rendelkezésre állnak. [7]

3. PONTFELHŐ ELŐFELDOLGOZÁS

A háromdimenziós felszínrekonstrukció minőségét alapvetően meghatározza a bemeneti pontfelhő geometriai minősége. A nyers szkennelési adatok zajt, kiugró pontokat és egyenetlen pontsűrűséget tartalmazhatnak, amelyek kedvezőtlenül befolyásolják a hálógenerálás eredményét. Ennek következtében a pontfelhő-előfeldolgozás a feldolgozási lánc alapvető lépését képezi. Az előfeldolgozás során zajszűrés és pontsűrűség-normalizálás alkalmazható a nem reprezentatív pontok eltávolítására és a számítási stabilitás javítására. A redundáns adatok csökkentése voxelalapú ritkítással valósítható meg, amelyben a tér kis térfogatelemekre (voxel) kerül felosztásra, és minden voxelhez egy reprezentatív pont tartozik. Ez az eljárás csökkenti a számítási igényt, miközben megőrzi a felszín globális geometriai jellemzőit. A felszínrekonstrukciót megelőzően meghatározásra került a seb határvonala, majd a nem releváns pontok –beleértve a környező bőrfelületeket és egyéb geometriailag irreleváns elemeket– eltávolításra kerültek. A jelen vizsgálatban a tisztítás manuálisan történt, több feldolgozási lépéshez kapcsolódva. A manuális szegmentálás automatizálása a kutatás jövőbeli irányát képezi. [8] Ennek alapját a szkennelés során rögzített felülnézeti képek képezhetik, amelyek mesterséges intelligencián alapuló módszerekkel támogathatják a sebhatár automatikus detektálását és a pontfelhő tisztítását. [9]

4. FELSZÍNREKONSTRUKCIÓS ALGORITMUSOK ÁTTEKINTÉSE

A pontfelhőből történő folytonos felszín előállítása különböző matematikai alapokon nyugvó algoritmusokkal valósítható meg. Ezek a módszerek eltérően kezelik a pontsűrűséget, a hiányos mintavételezést és a mérési zajt, ezért alkalmazhatóságuk a vizsgált objektum geometriai tulajdonságaitól, valamint a szkennelési és feldolgozási környezettől függ. A jelen tanulmány a leggyakrabban alkalmazott felszínrekonstrukciós eljárásokat tekinti át, hasonlítja össze egy mulázson és egy valódi tesztelt szkennelésen és választja ki a legmegfelelőbbet, ami a munka további részének alapját képezi.

4.1. Poisson rekonstrukció

A Poisson-alapú felszínrekonstrukció implicit módszer, amely a pontfelhőhöz rendelt normálvektorok alapján egy skalármezőt számít, majd annak egy kiválasztott izoértékéhez tartozó izofelületet állítja elő. Az eljárás eredménye egy zárt, folytonos és vízzáró háromszögelt háló. A módszer robusztusan kezeli a hiányos vagy egyenetlen mintavételezést, valamint a mérési zajt, ezért különösen alkalmas olyan esetekben, ahol a topológiai zártság elsődleges követelmény. A beépített simító hatás következtében a finom geometriai részletek részben elsimulhatnak, azonban ez a tulajdonság számos alkalmazásban előnyös a stabil felszínmodell előállítása szempontjából. [10]

4.2. BPA

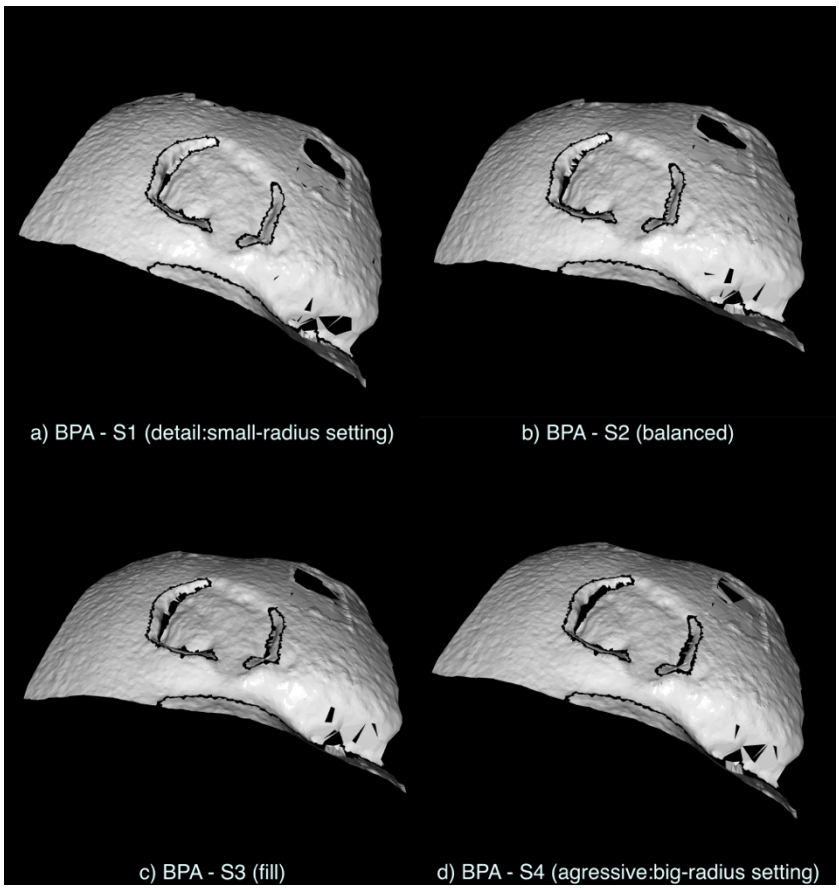
A Ball Pivoting algoritmus lokális geometriai felszínrekonstrukciós módszer, amelyben egy adott sugarú virtuális gömb gördül végig a pontfelhő felszínén, és az érintett pontokból háromszögeket képez. Az eljárás közvetlenül a pontfelhő lokális szerkezetére épül, ami lehetővé teszi az éles geometriai jellemzők pontos megőrzését. Megfelelő pontsűrűség és paraméterválasztás mellett részletgazdag felszínek állíthatók elő, ugyanakkor hiányos vagy egyenetlen mintavételezés esetén a háló lokálisan nyitott maradhat, ami további feldolgozási lépések alkalmazását teheti szükségessé. [11]

4.3. Marching Cubes

A marching cubes egy voxelalapú felszínrekonstrukciós algoritmus, amely egy háromdimenziós skalármező izofelületének háromszögelését végzi el. Az eljárás széles körben alkalmazott az orvosi képalkotásban, különféle adatok feldolgozása során, ahol jól strukturált térbeli adatok állnak rendelkezésre. Pontfelhő-alapú alkalmazás esetén előzetes voxelizáció szükséges, amely felbontás és számítási igény közötti kompromisszumot eredményez, és a finom geometriai részletek mérsékelt torzulásához vezethet, ugyanakkor stabil és jól kontrollálható felszínmodell állítható elő. [12]

5. REKONSTRUKCIÓS ALGORITMUSOK ÖSSZEHASONLÍTÁSA

A felszínrekonstrukciós algoritmusok gyakorlati alkalmazása során a Scaniverse alkalmazás által exportált Polygon File Format (PLY) állományokból származó pontadatok szolgáltak bemenetként. A pontfelhő-alapú rekonstrukció nyílt forráskódú környezetben, az Open3D könyvtár felhasználásával történt. [13] A Poisson Surface Reconstruction (PSR) és a Ball Pivoting algoritmus több paraméterbeállítással került kiértékelésre annak vizsgálata céljából, hogy miként reagálnak az eltérő pontsűrűségre és az egyenetlen mintavételezésre. A vizsgálatok célja nem egy véglegesen optimalizált háló előállítása volt, hanem annak elemzése, hogy az eltérő matematikai megközelítések azonos bemeneti adatok esetén milyen geometriai és topológiai felszíneket eredményeznek. Ennek megfelelően az elkészült hálók kvalitatív összehasonlítása történt meg. Az elemzésbe bevonásra kerültek a Scaniverse alkalmazás által automatikusan generált hálók is. Bár az alkalmazás belső rekonstrukciós algoritmusa nem dokumentált, az eredményül kapott hálók zárt, simított és vízzáró tulajdonságokat mutattak, ami lehetővé tette a közvetlen összehasonlítást az Open3D környezetben előállított rekonstrukciókkal. A különböző felszínek vizuális összehasonlítását szemléltető ábrák mutatják be. A Ball Pivoting algoritmus esetében több gömbsugár- és kapcsolódási paraméterkombináció került vizsgálatra annak elemzésére, hogy miként alakul a részletmegőrzés és a felszín zártsága közötti kompromisszum. A paraméterválasztás hatásait a rekonstruált hálók geometriai és topológiai tulajdonságain az 1. ábra szemlélteti.



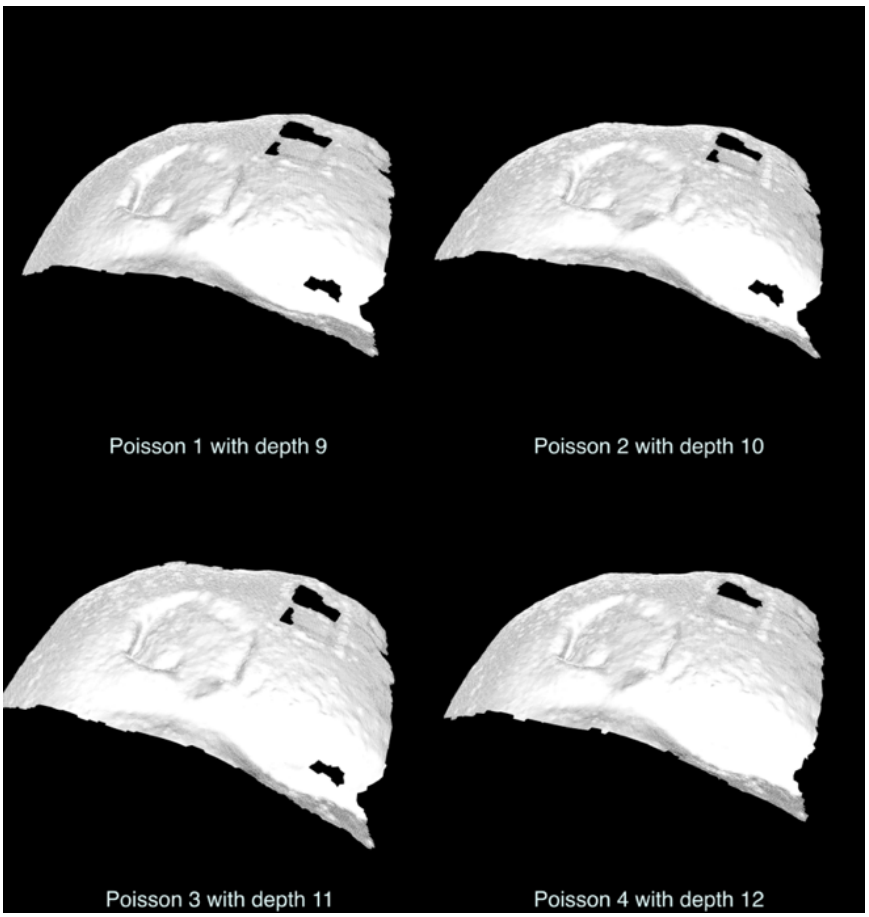
1. ábra. Ball Pivoting algoritmus négy eltérő paraméterbeállítással

A Poisson-féle felszínrekonstrukció több paraméterbeállítással került értékelésre, amelyek esetén a kapott hálók összességében hasonló geometriai minőséget mutattak. A vizsgálatok során egy octree-alapú megfogalmazás került alkalmazásra, ahol a maximális octree-mélység szabályozta a térbeli felbontást, lehetővé téve a geometriai részletesség, a zajérzékenység és a számítási igény közötti kompromisszum elemzését. A lokálisan megjelenő felszíni hiányosságok elsősorban olyan területeken jelentkeztek, ahol a szkennelés során idegen, homogén vagy részben reflektív objektumok voltak jelen, például a betegazonosításhoz használt QR-kódot tartalmazó fehér papírfelületek esetében. Az ilyen alacsony textúrájú, jellegtelen felszínek rekonstruálása kihívást jelent, mivel nem biztosítanak elegendő geometriai vagy vizuális információt a megbízható pontfelhő-generáláshoz, ami lokális geometriai hiányok és rekonstrukciós artefaktumok megjelenéséhez vezethet.

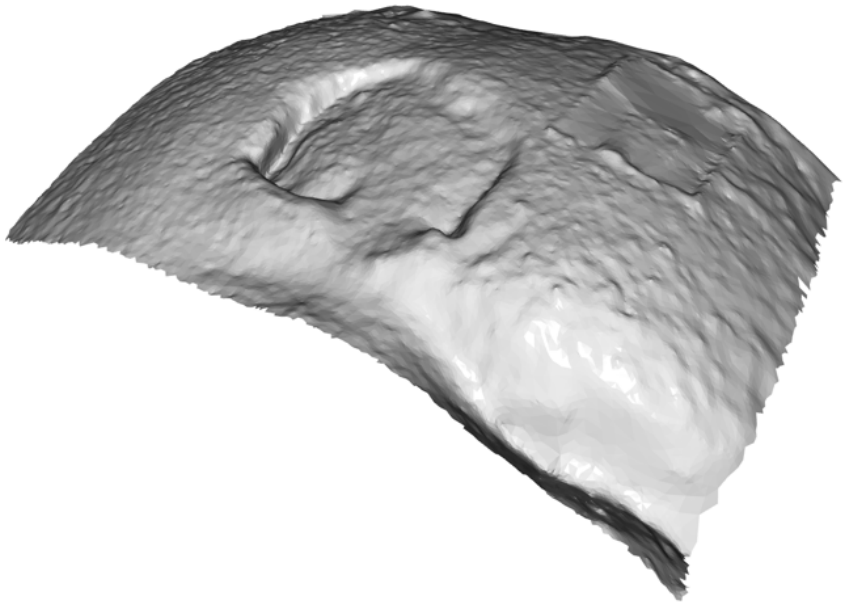
A Scaniverse alkalmazás által automatikusan előállított hálót a 3. ábra mutatja be. A modell geometriai karaktere hasonló a Poisson-alapú rekonstrukciókhoz, ugyanakkor a kapott felszín minden esetben zárt volt, ezért referenciafelületként szolgált a további összehasonlítások során.

6. EREDMÉNYEK

Az azonos bemeneti pontfelhőből különböző felszínrekonstrukciós algoritmusokkal előállított hálók összehasonlítása egyértelműen rámutatott az alkalmazott módszerek geometriai és topológiai eltéréseire. Az értékelés a Poisson-féle felszínrekonstrukció, a Ball Pivoting algoritmus, valamint a Scaniverse alkalmazás által generált hálók alapján történt, eltérő paraméterbeállítások figyelembevételével. A Ball Pivoting algoritmussal végzett rekonstrukciók minden vizsgált paraméterkombináció esetén lokálisan nyitott, hiányos felszínt eredményeztek (1. ábra). A kisebb gömbsugár-beállítások csökkentették ugyan a felszíni hiányok méretét, míg a



2. ábra. A Poisson-algoritmus alkalmazása eltérő paraméterbeállításokkal



3. ábra. A Scaniverse által előállított felszínmodell

nagyobb sugarak kiterjedtebb lyukakat eredményeztek, azonban egyik beállítás sem biztosított topológiailag zárt hálót. Ennek következtében a módszer a vizsgált alkalmazási környezetben korlátozottan bizonyult alkalmasnak további geometriai és térfogat-alapú elemzésekhez. A Poisson-féle felszínrekonstrukció alkalmazása során a kapott hálók túlnyomórészt folytonosak és nagyrészt zártak voltak (2. ábra), és a különböző paraméterbeállítások mellett is hasonló geometriai minőséget mutattak. A lokálisan megjelenő felszíni hiányosságok elsősorban homogén, alacsony információtartalmú területekhez voltak köthetők, és nem érintették a klinikailag releváns felszínt. A Scaniverse alkalmazás által automatikusan generált hálók minden esetben zárt, lyukmentes felszínt biztosítottak (3. ábra). Geometriai karakterük jól illeszkedett a Poisson-alapú rekonstrukciókhoz, ugyanakkor a felszín folytonossága következetesen biztosított volt. Az elért felbontás és részletesség megfelelt a jelen vizsgálatban meghatározott klinikai mérési követelményeknek. Az eredmények alapján a Scaniverse által előállított hálók bizonyultak a legalkalmasabbnak további elemzésekhez, mivel folytonos, topológiailag zárt és klinikailag jól értelmezhető geometriai reprezentációt biztosítottak. A nyílt forráskódú környezetben végzett rekonstrukciók nemcsak az algoritmusok összehasonlítását tették lehetővé, hanem alapot szolgáltatnak a jövőbeli fejlesztésekhez is. A kiválasztott felszínrekonstrukciós megközelítésre építve a valós sebzárási folyamat első változatának kidolgozása már megkezdődött, és további finomítás alatt áll.

7. ÖSSZEFOGLALÁS

A háromdimenziós sebfelszínek előállítására számos felszínrekonstrukciós módszer áll rendelkezésre, azonban alkalmazhatóságuk nagymértékben függ a vizsgált feladattól és a választott paraméterezéstől. Az elvégzett összehasonlító vizsgálatok egyértelműen rámutattak arra, hogy a felszín folytonossága és topológiai zártsága alapvető követelmény a megbízható mélység- és térfogat-alapú mérésekhez, különösen nyomási fekélyek esetében. Az eredmények alapján megállapítható, hogy a mobil LiDAR-alapú szkennelés és az automatizált hálógenerálás kombinációja klinikai környezetben is kellően részletes és konzisztens geometriai reprezentációt biztosít. A Scaniverse alkalmazás által előállított felszínek stabil és gyakorlatban jól használható megoldást nyújtottak az ágy melletti sebmonitorozás követelményeinek teljesítésére. A bemutatott feldolgozási folyamat fontos lépést jelent a strukturált háromdimenziós objektumrekonstrukció irányába, mivel a felszínrekonstrukciós módszer megválasztása közvetlenül befolyásolja a további feldolgozási lépések hatékonyságát. A módszer moduláris felépítése lehetőséget teremt további fejlesztésekre, így az automatizált szegmentáció, a rekonstrukció szabályozott finomítása, valamint kiegészítő adatforrások integrálása révén a jövőben még nagyobb fokú automatizálás érhető el klinikai alkalmazásokban.

IRODALOMJEGYZÉK

[1] C. Bansal, R. Scott, D. Stewart, C. J. Cockerell: *Decubitus ulcers: A review of the literature*. International Journal of Dermatology, vol. 44, no. 10, pp. 805–810, 2005.

[2] C. E. Ananian, Y. S. Dhillon, C. C. V. Gils, D. C. Lindsey, R. J. Otto, C. R. Dove, J. T. P. a. M. C. Saunders: *A multicenter, randomized, single-blind trial comparing the efficacy of viable cryopreserved placental membrane to human fibroblast-derived dermal substitute for the treatment of chronic diabetic foot ulcers*. Wound Repair and Regeneration, vol. 26, no. 3, pp. 259–305, 2018.

[3] R. Chierchia, L. Lebrat, D. Ahmedt-Aristizabal, O. Salvado, C. Fookes, R. S. Cruz: *SALVE: A 3D reconstruction benchmark of wounds from consumer-grade videos*. In IEEE Winter

Conference on Applications of Computer Vision (WACV) 2025, Waikoloa, Hawaii, USA, 2025.

[4] T. Shirley, D. Presnov, A. Kolb: *A lightweight approach to 3D measurement of chronic wounds*. Journal of WSCG, vol. 27, no. 1, pp. 67–74, 2019.

[5] Y. M. Zhao, E. H. Currie, L. Kavoussi, S. Y. Rabbany: *Laser scanner for 3D reconstruction of a wound's edge and topology*. International Journal of Computer Assisted Radiology and Surgery, vol. 16, no. 10, pp. 1761–1773, 2021.

[6] M. Botsch, L. Kobbelt, M. Pauly, P. Alliez, B. Lévy: *Polygon mesh processing*, Natick, MA, USA: AK Peters, 2010.

[7] L. Lebrat, R. Santa Cruz, R. Chierchia, Y. Arzhaeva, M. A. Armin, J. Goldsmith, J. Oorloff, P. Reddy, C. Nguyen, L. Petersson, M. Barakat-Johnson, G. Luscombe, C. Fookes, O. Salvado: *Syn3DWound: A synthetic dataset for 3D wound bed analysis*. In IEEE International Symposium on Biomedical Imaging (ISBI), Nashville, TN, USA, 2024.

[8] R. Honti, J. Erdélyi, A. Kopáčik: *Plane segmentation from point clouds*. Pollack Periodica: An International Journal for Engineering and Information Sciences, vol. 13, no. 2, pp. 159–171, 2018.

[9] C. Wang, D. M. Anisuzzaman, V. Williamson, M. K. Dhar, B. Rostami, J. Niezgoda, S. Gopalakrishnan, Z. Yu: *Fully automatic wound segmentation with deep convolutional neural networks*. Scientific Reports, vol. 10, no. 1, pp. 21897, 2020.

[10] M. Kazhdan, M. Bolitho, H. Hoppe: *Poisson surface reconstruction*. In Eurographics Symposium on Geometry Processing, Cagliari, Italy, 2006.

[11] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, G. Taubin: *The ball-pivoting algorithm for surface reconstruction*. IEEE Transactions on Visualization and Computer Graphics, vol. 5, no. 4, pp. 349–359, 1999.

[12] W. E. Lorensen, H. E. Cline: *Marching Cubes: a high resolution 3D surface construction algorithm*. In ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques, Anaheim, CA, USA, 1987.

[13] Q.-Y. Zhou, J. Park, V. Koltun: *Open3D: A modern library for 3D data processing*. arXiv, 2018.

BEHAVIOR OF PHOTOVOLTAIC TECHNOLOGIES UNDER SUDDEN WEATHER CHANGES: A SIMULATION-BASED ANALYSIS

Dávid Szalánczi[1], ***Péter Bencs***[2]
[1]PhD Student, [2]Associate Professor
[1,2]*Faculty of Mechanical Engineering and Informatics,*
Institute of Energy and Chemical Machinery
[1]szalanczi.david@student.uni-miskolc.hu,
[2]peter.bencs@uni-miskolc.hu

Abstract

This study investigates the temperature and efficiency behavior of three photovoltaic technologies –conventional silicon, PERC, and PERC combined with a UV-protective film– under dynamically changing weather conditions using simulation-based modeling. The analysis covers three typical weather transitions (clouding, rainfall, and overcast-rain-sunshine sequences) across three seasons (summer, autumn, and winter). The results show that the PERC + UV film technology exhibits the smallest performance loss and the most stable thermal response under rapid environmental fluctuations. The findings highlight the decisive role of temperature-dependent behavior in determining long-term efficiency and energy yield stability of PV systems.

1. INTRODUCTION

The performance and efficiency of photovoltaic (PV) systems are strongly influenced not only by the applied technology but also by the surrounding environmental conditions. Temperature fluctuations, variations in solar irradiance, and sudden weather transitions—such as cloud cover, rainfall, or rapid cooling—can dynamically alter the operating behavior of PV modules. Understanding these effects is essential for optimizing long-term energy production and for making informed technological choices when designing PV installations.

Previous studies have shown that modern PV technologies—such as PERC (Passivated Emitter and Rear Contact) and PERC modules enhanced with UV-protective films—exhibit improved thermal characteristics and can reduce efficiency losses caused by overheating. However, the behavior of these advanced technologies under rapid weather changes has been less thoroughly investigated, despite the fact that such transitions may induce significant thermal and electrical disturbances within the panels [1].

The aim of this study is to model, in a simulation environment, the response of three different PV technologies –conventional silicon, PERC, and PERC with UV-reflective film– during sudden weather changes. Special emphasis is placed on analyzing internal temperature variations and the corresponding efficiency fluctuations. The simulations are based on real meteorological data from the city of Miskolc, and three distinct seasons –summer, autumn, and winter– are examined.

Beyond highlighting technological differences, the analysis also reveals how periodic environmental effects influence long-term operational performance. Based on the results for the autumn season, it can be concluded that the findings closely represent spring conditions as well, since the characteristic patterns of temperature and solar irradiance exhibit similar dynamics during these periods.

2. LITERATURE REVIEW

Numerous studies have demonstrated that increases in operating temperature significantly reduce the efficiency of photovoltaic (PV) modules. Due to the thermal sensitivity of silicon-based

semiconductors, even a 1 °C rise in cell temperature can lead to an estimated 0.4–0.5 % decrease in output [2]. Prolonged overheating not only diminishes instantaneous efficiency but also accelerates material fatigue and long-term degradation processes within the module structure [3].

Rapidly changing weather conditions –such as sudden cloud cover, rainfall, or wind– can cause abrupt variations in irradiance and temperature within short time intervals. These fast transitions may induce thermal shock effects, which negatively influence the stability and electrical performance of PV panels. Conventional silicon modules have been shown to be particularly sensitive to such fluctuations, exhibiting higher susceptibility to rapid thermal transients [4].

Other studies have focused on building-integrated photovoltaic (BIPV) systems, examining their thermal and electrical behavior under various environmental conditions. These investigations highlight that the interaction between weather dynamics and PV technology can substantially influence efficiency, especially in systems exposed to frequent or irregular irradiance changes [5].

3. MATERIALS AND METHODS

3.1. Overview of the Examined Technologies

In this study, the thermal behavior and performance of three different photovoltaic (PV) technologies are investigated under rapidly changing weather conditions:

- **Conventional silicon-based PV modules**: The most widely used industrial technology, characterized by relatively high thermal sensitivity.
- **PERC (Passivated Emitter and Rear Contact) technology**: Incorporates a passivated rear surface with enhanced light absorption and improved infrared reflection, enabling better thermal management.
- **PERC with UV-protective film**: An advanced configuration in which an additional UV-blocking layer reduces internal heat load, mitigating material fatigue and improving long-term stability.

These technologies are compared within a simulation environment to determine which system exhibits the most stable performance when subjected to sudden environmental fluctuations.

3.2. Simulation Environment and Weather Data

The simulations were performed using the principles of the Simulink modeling environment. Real meteorological conditions from the city of Miskolc (Hungary) were incorporated to represent different seasonal and diurnal patterns.

The following dates were selected as representative cases:

- **20 July** – a typical summer day with strong irradiance and rapid cloud formation,
- **15 October** – an autumn day characterized by variable weather, moderate sunlight, and frequent cloud or precipitation events,
- **10 January** – a typical winter day with weak solar radiation.

3.3. Simulated Weather Scenarios

Three characteristic weather-change sequences were defined to evaluate the dynamic response of the PV systems:

- Clear sky → 30 minutes of cloud cover → clear sky,
- Clear sky → 30 minutes of rainfall → clear sky,
- Partial cloudiness → rainfall → rapid return to sunshine.

Each scenario was embedded into a 24-hour simulation cycle, modeled with hourly resolution to capture both transient and daily thermal effects.

3.4. Measured Parameters

For each PV technology, the following performance indicators were recorded during the simulations:

- Internal temperature variation (°C),
- Efficiency change (%).

These parameters allow direct comparison of the thermal dynamics and performance stability of the technologies under identical environmental conditions.

3.5. Methodological Notes

The objective of the methodology is not to model an entire year but rather to focus on extreme and rapid weather transitions and their effects on the selected technologies. The simulation relies on linear heat transfer and efficiency-temperature relationships, while the dynamic characteristics of sudden transitions –such as sharp cooling or rapid heating– are treated with special emphasis.

4. RESULTS AND EVALUATION

4.1. Internal Temperature Analysis of PV Panels Across Different Seasons

4.1.1 Case 1: Summer

The results shown in Figure 1 clearly illustrate the temperature increase induced by solar irradiance throughout the day. The conventional silicon module exhibits the most pronounced rise in internal temperature, reflecting its higher sensitivity to thermal loads. In comparison, the **PERC + UV** film technology demonstrates the most stable thermal behavior, maintaining significantly lower and more uniform internal temperatures under the same conditions.

4.1.2 Case 2: Autumn / Spring

Due to the milder solar irradiance and the more frequent cloud cover, the overall temperature rise is noticeably lower compared to summer conditions. Nevertheless, the differences between the examined technologies remain clearly distinguishable, as illustrated in Figure 2.

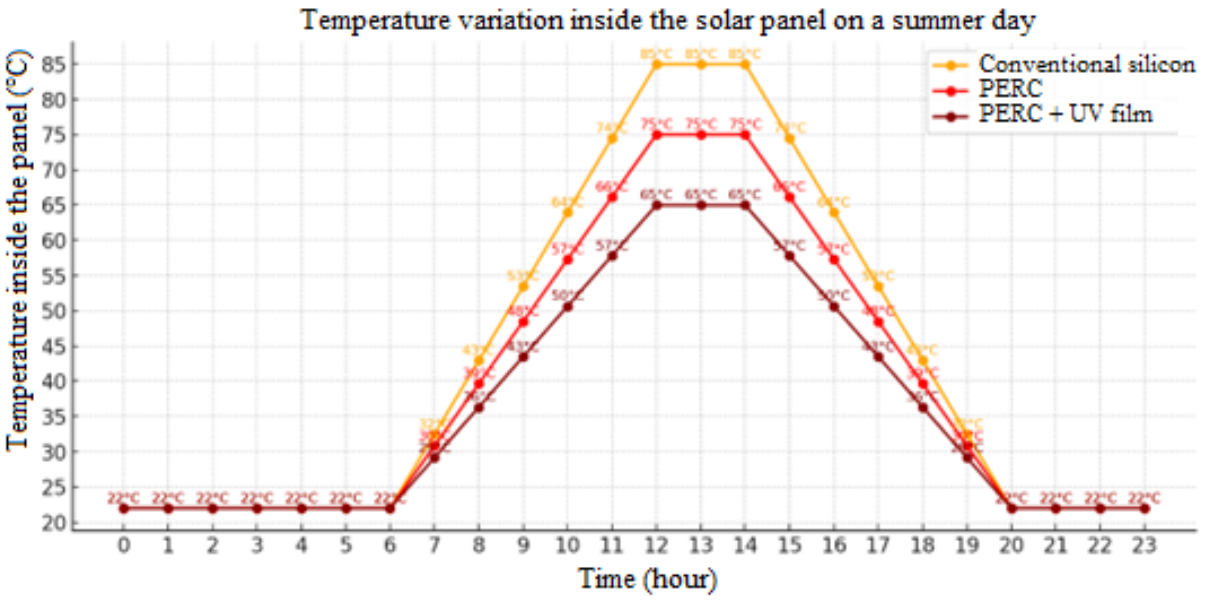


Figure 1: Variation of internal PV panel temperature on a summer day.

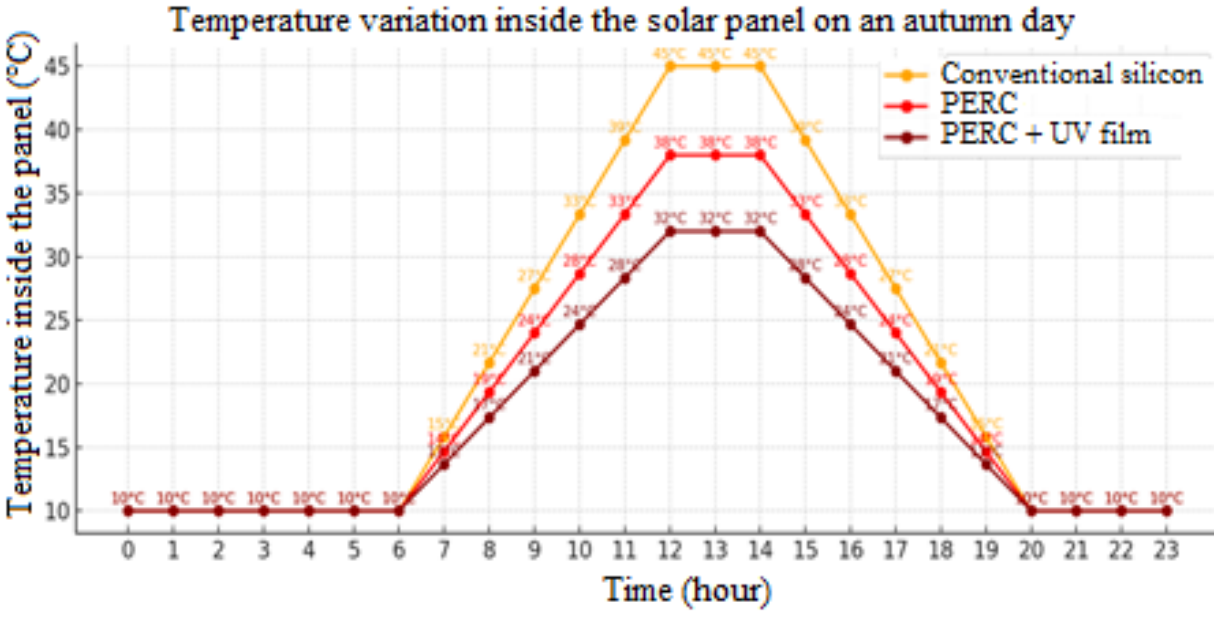


Figure 2: Internal temperature behavior of PV panels under autumn/spring conditions.

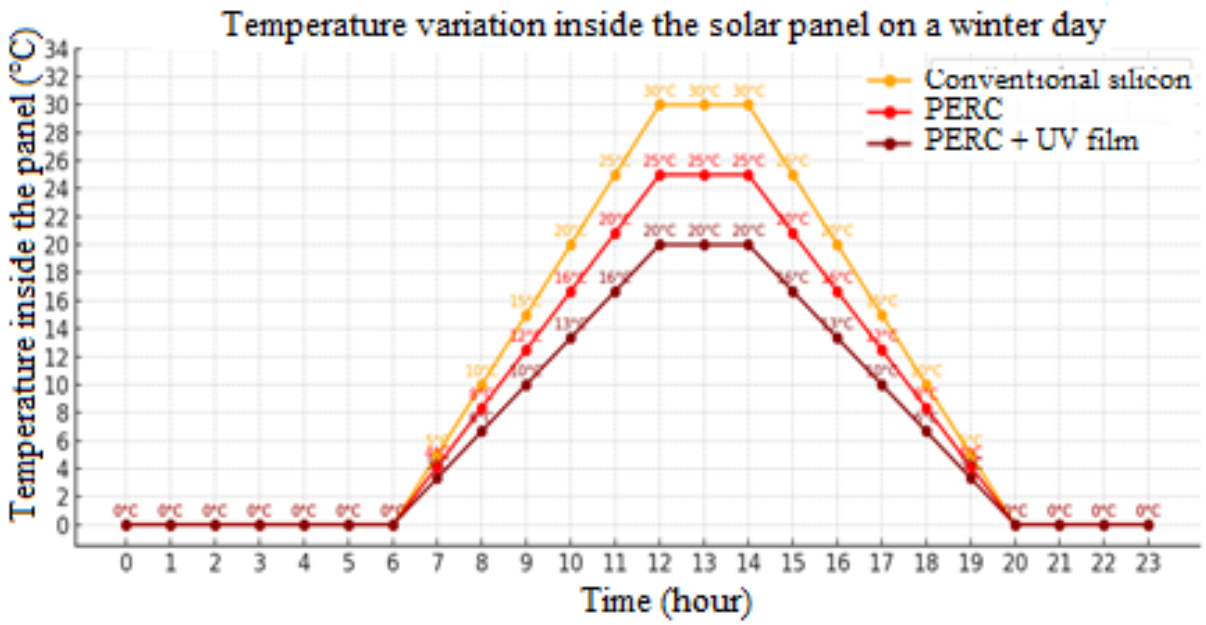


Figure 3: Internal temperature variation of PV panels under winter sunlight conditions.

4.1.3 Case 3: Winter

Because winter sunlight is considerably weaker, the resulting temperature increase within the PV modules remains minimal. Even under these reduced irradiance conditions, the **PERC + UV film** technology exhibits the smallest temperature fluctuations among the examined systems, as shown in Figure 3.

4.2. Seasonal Variation of PV Module Efficiency

4.2.1 Case 1: Summer

The pronounced temperature fluctuations occurring during sunny summer conditions lead to a marked reduction in the efficiency of conventional silicon panels. In contrast, the **PERC + UV film** technology maintains its initial efficiency more effectively, exhibiting considerably smaller performance losses under identical thermal stress, as illustrated in Figure 4.

4.2.2 Case 2: Autumn / Spring

As shown in Figure 5, the efficiency reductions are more moderate during autumn and spring compared to summer conditions. Nevertheless, the relative performance ranking of the technologies follows the same overall pattern, with similar dynamic behavior observed in the magnitude and progression of efficiency losses.

4.2.3 Case 2: Winter

The efficiency variation during winter remains relatively small; however, the conventional silicon panel is still more sensitive to even minor temperature fluctuations compared with the more advanced technologies. This behavior is clearly illustrated in Figure 6.

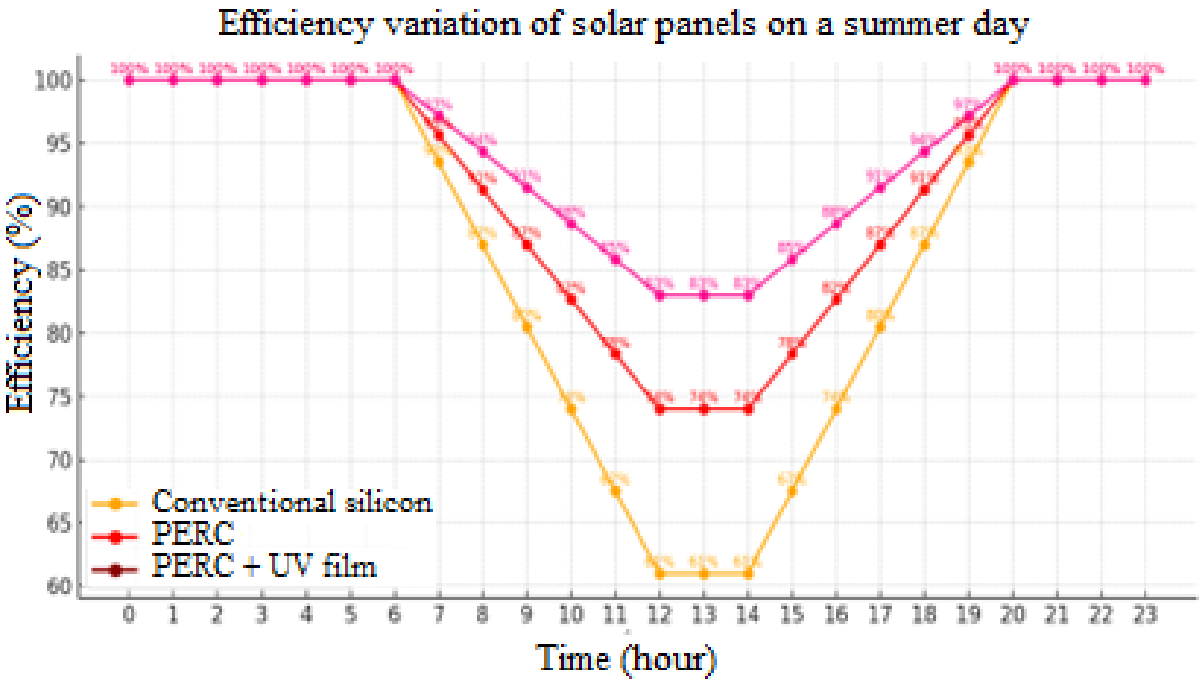


Figure 4: Efficiency variation of PV modules on a summer day.

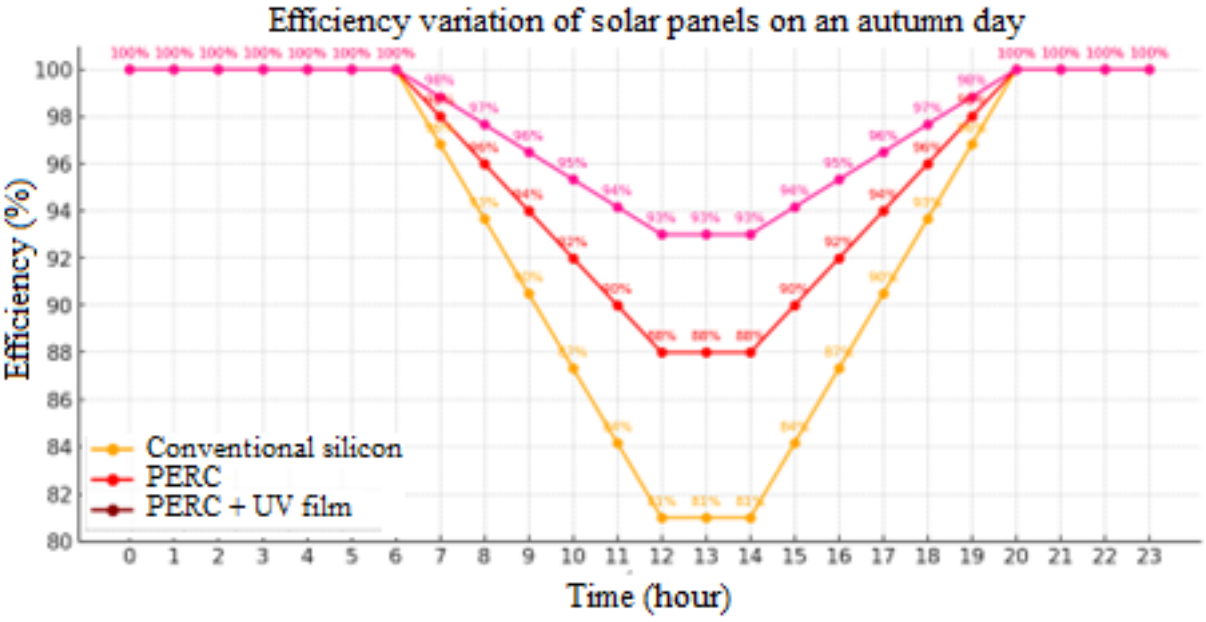


Figure 5: Efficiency behavior of PV modules under autumn/spring conditions.

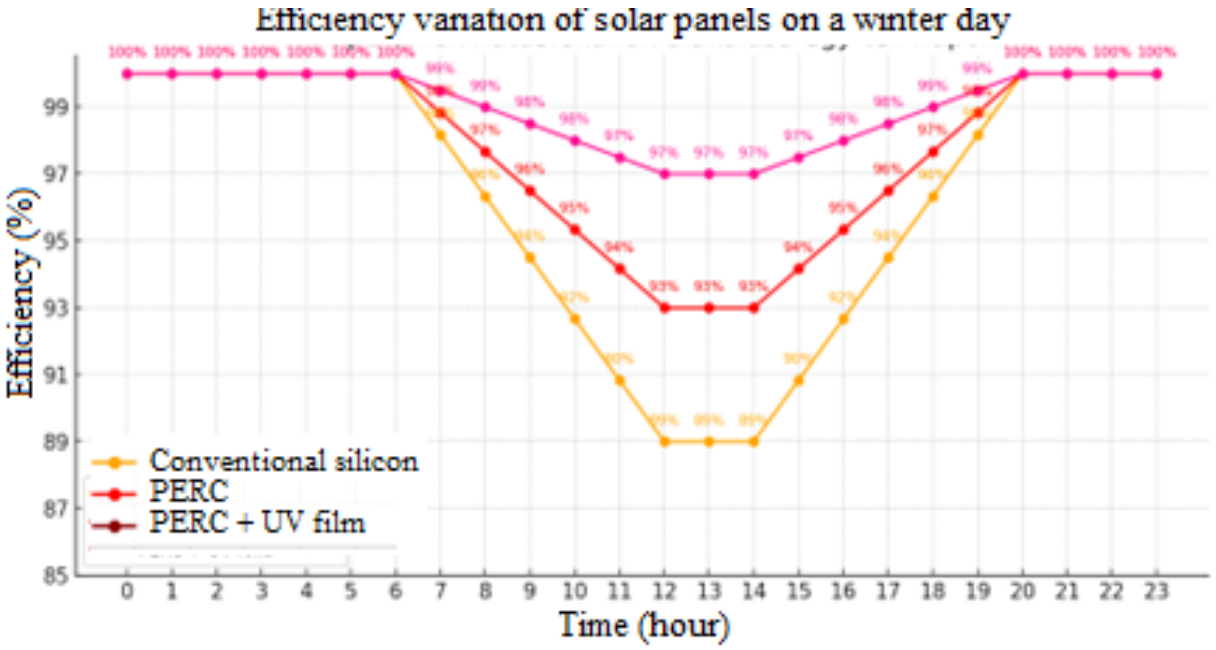


Figure 6: Efficiency variation of PV modules under winter conditions.

4.3. Weather-Change Scenarios Examined in Summer

Since temperature fluctuations and performance losses are most pronounced during summer, the weather-change scenarios were evaluated exclusively under summer conditions.

4.3.1 Case 1: Variation of Internal Panel Temperature on a Summer Day with Sudden Cloud Cover (at 13:00)

The sudden cloud cover occurring at 13:00 has a pronounced effect on the internal temperature of the PV modules, producing a clearly observable drop in the temperature curves, as shown in Figure 7. Prior to this event, the intense solar irradiance causes a continuous rise in panel temperature; however, once the sunlight is abruptly blocked, the temperature decreases rapidly.

The main characteristics of this behavior can be summarized as follows:

- **Morning period (0–6 h)**: All technologies start from approximately the same initial temperature, around 22 °C.
- **After 6 h**: As irradiance increases, the panels begin to heat up, with the rate of temperature rise differing among the technologies.
- **At 13:00 (cloud onset)**: The sudden shading leads to a sharp drop in internal temperature, halting the previously increasing trend.
- **Afternoon period**: The panels gradually cool down, and by evening their temperatures converge again to nearly identical values.

The Figure 8 presents the corresponding efficiency changes observed under this weather-change scenario.

As illustrated in Figure 8, the different photovoltaic technologies respond distinctively to the rapid cooling induced by the sudden cloud cover. The efficiency curves clearly reveal the varying degrees of resilience to temperature fluctuations:

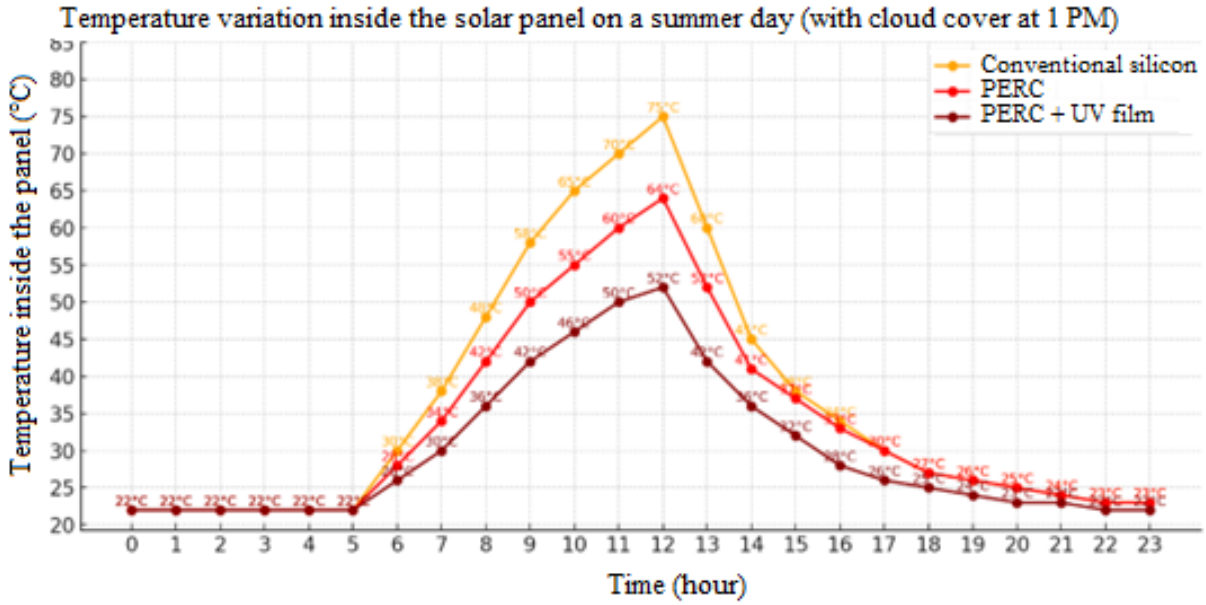


Figure 7: Internal panel temperature variation during sudden cloud cover at 13:00 (1 PM).

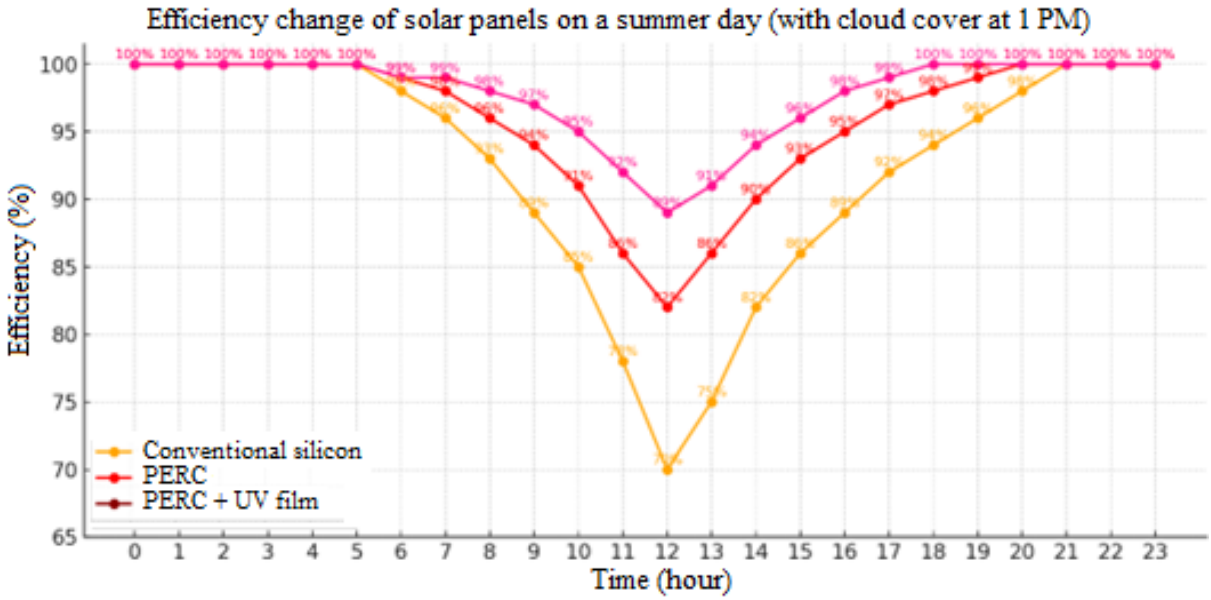


Figure 8: Efficiency response of PV modules during sudden cloud cover.

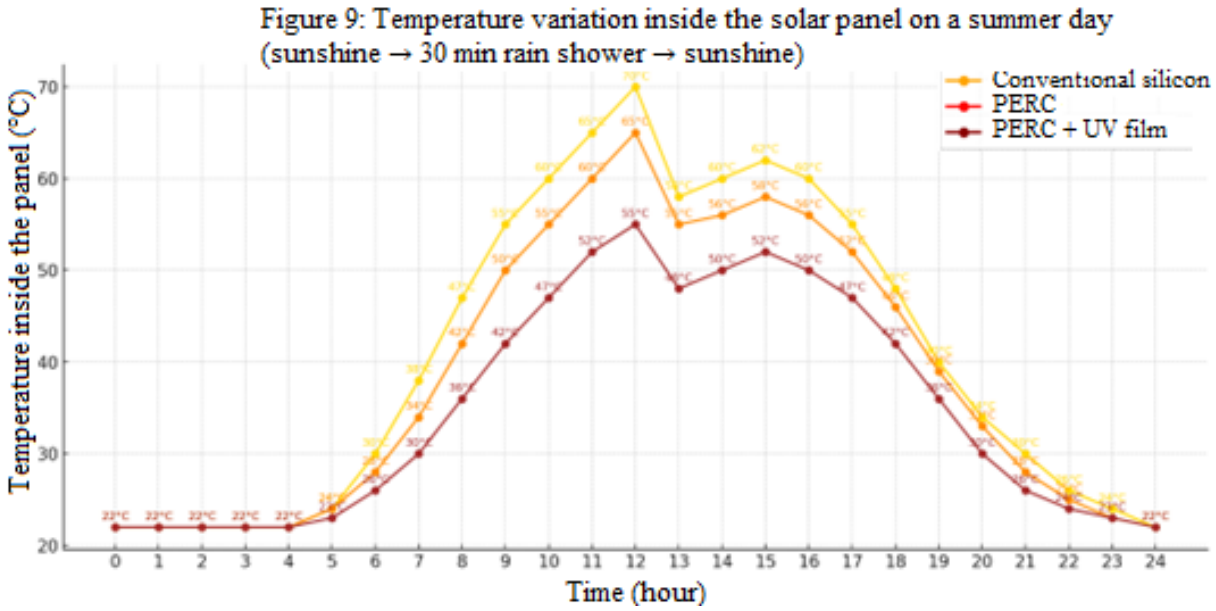


Figure 9: Internal temperature behavior during a rain-induced weather transition.

- **Conventional silicon modules** exhibit a steep decline in efficiency, followed by a slow and gradual recovery.
- **PERC and PERC + UV film technologies** experience noticeably smaller efficiency losses and stabilize more rapidly, demonstrating superior adaptability under dynamic environmental conditions.

4.3.2 Case 2: Second Weather-Change Scenario (Sunshine → 30 Minutes of Rain → Sunshine)

As shown in Figure 9, the onset of rainfall causes a sudden drop in the internal temperature of the PV modules, followed by a relatively rapid reheating phase once the rain subsides. The differences among the technologies become evident not only in their maximum operating temperatures but also in their cooling rates and subsequent ability to stabilize after the disturbance.

As illustrated in Figure 10, the sudden temperature drop caused by the rainfall leads to a temporary improvement in the efficiency of all PV modules. This effect is most pronounced in the conventional silicon technology, where the reduction in internal temperature results in a noticeable efficiency increase. In contrast, the **PERC** and **PERC + UV film** systems exhibit smaller fluctuations and recover more quickly, reflecting their enhanced thermal stability and reduced sensitivity to rapid temperature changes.

4.3.3 Case 3: Third Weather-Change Scenario (Partial Cloudiness → 30 Minutes of Rain → Sunshine)

The temperature variations shown in Figure 11 clearly reflect the impact of dynamic weather conditions. The partially overcast morning results in a slower initial warming of the panels, followed by a sudden cooling during the rainfall period. Once the sunshine returns, the modules exhibit a rapid reheating response, demonstrating the strong influence of fluctuating irradiance on their thermal behavior.

The efficiency trends closely follow the temperature profiles of the panels. The gradual warming observed in the morning leads to a steady decrease in efficiency, while the cooling effect of the rainfall results in a temporary performance improvement. Once sunshine returns

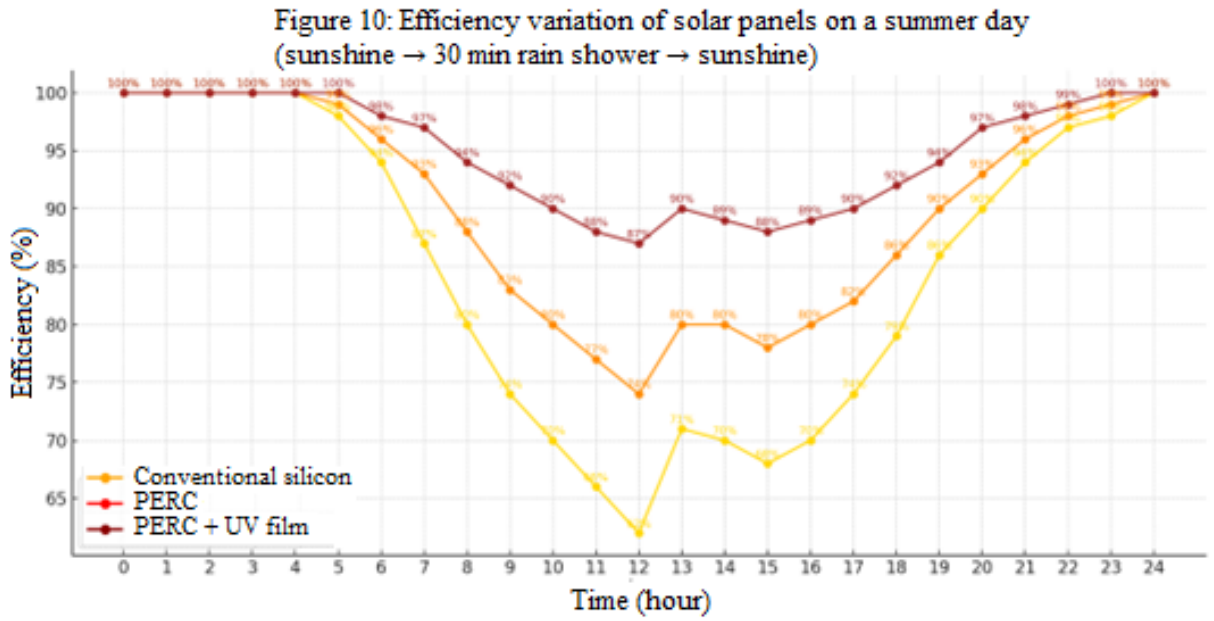


Figure 10: Efficiency variation of PV modules during a short rainfall event.

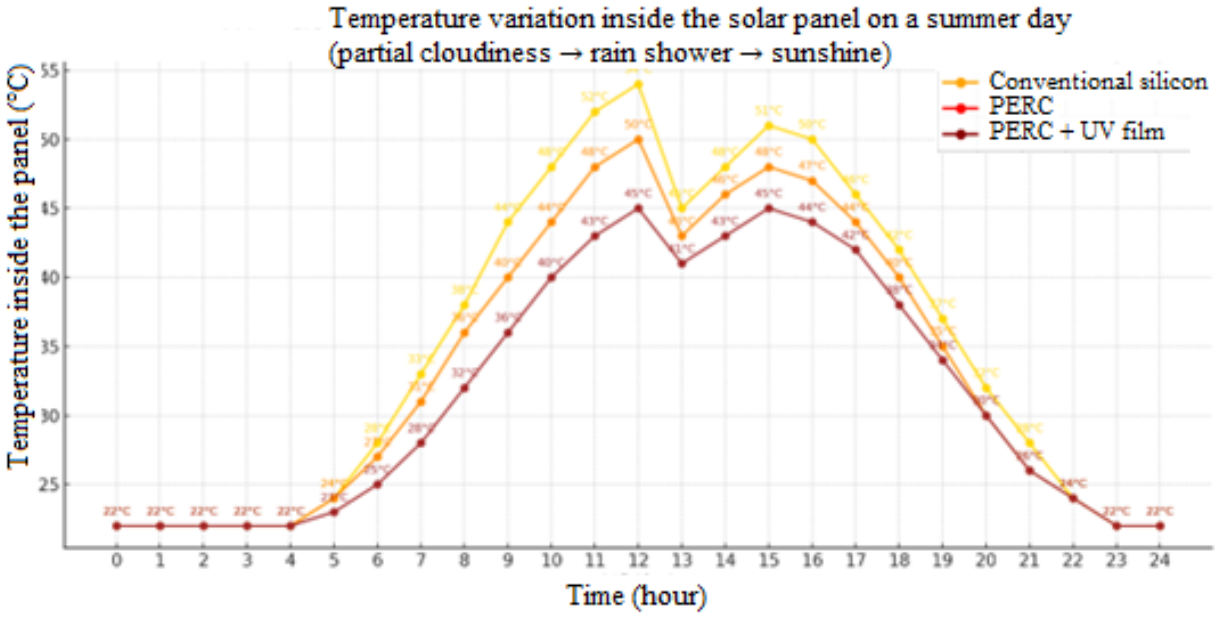


Figure 11: Internal temperature variation during an overcast–rain–sunshine sequence.

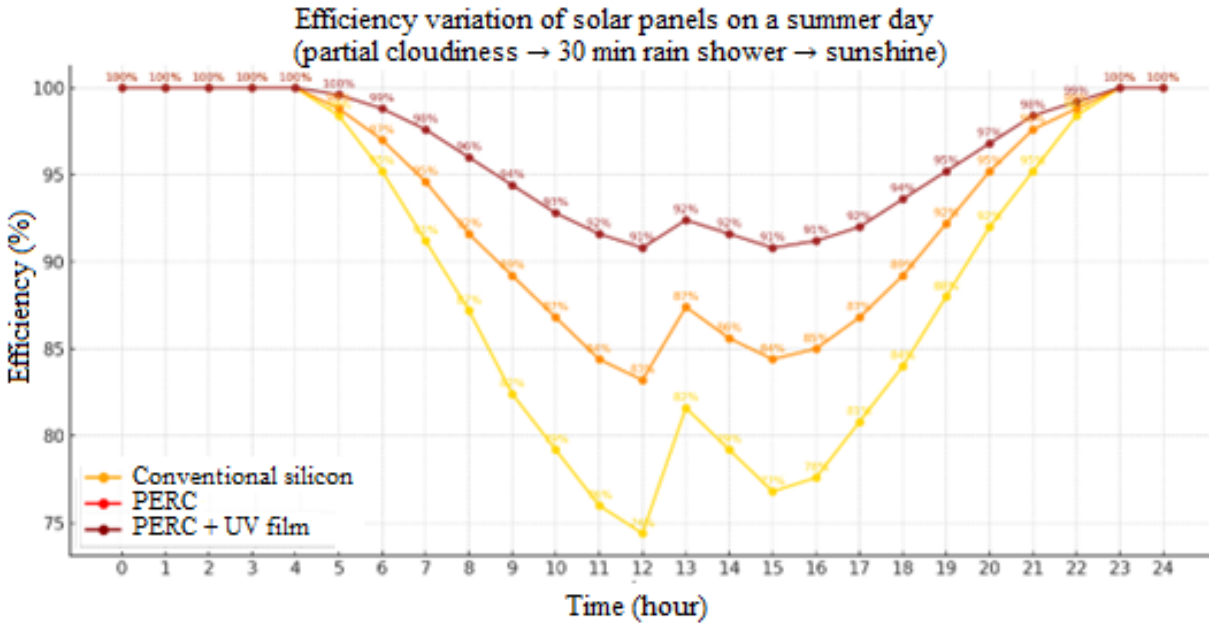


Figure 12: Efficiency variation of PV modules during an overcast–rain–sunshine sequence.

and the panel temperatures rise again, the efficiency decreases accordingly. These dynamics are clearly illustrated in Figure 12.

4.4. Comparative Analysis

4.4.1 General Observations

The simulations evaluated the behavior of three photovoltaic technologies –conventional silicon, PERC, and PERC with UV film– under three distinct weather-change scenarios: sudden cloud cover, short-duration rainfall, and a combined sequence of overcast conditions, rain, and rapid return to sunshine. The results demonstrate that the operational behavior of the PV modules is strongly influenced by the dynamics of sudden temperature changes, particularly when irradiance varies rapidly.

The temperature curves reveal that:

- Conventional silicon modules are the most sensitive to thermal fluctuations, heating up quickly and cooling down more slowly.
- PERC technology displays more balanced thermal responses, as the infrared-reflective rear layer helps reduce excessive heat absorption.
- PERC + UV film modules exhibit the slowest response to environmental changes, resulting in the most stable internal temperature profiles.

The efficiency trends are consistent with these thermal characteristics. Following temperature peaks, conventional modules show the largest reduction in efficiency, while the PERC + UV systems experience smaller losses and recover more rapidly. This highlights the superior resilience of advanced PV technologies under dynamically changing weather conditions.

4.4.2 Magnitude of Performance Losses

The simulations also quantitatively assessed the extent of performance reduction caused by the different weather-change scenarios for each PV technology. Across all cases, conventional

silicon modules exhibited the highest efficiency losses. For example, during a sudden rain event or rapid cloud cover, efficiency could drop by 5–7 % within a short period, recovering only gradually as the panels cooled.

In contrast, PERC modules demonstrated more moderate losses, typically in the range of 3–5 %, with quicker recovery due to their improved thermal management capabilities. The PERC + UV film technology showed the highest stability, with only 2–3 % temporary efficiency reduction and noticeably faster adaptation to the changing environmental conditions.

These findings confirm that advanced thermal protection measures –particularly the UV-reflective film– play a critical role not only in moderating temperature fluctuations but also in stabilizing performance under rapidly varying weather conditions.

4.4.3 Technological Sensitivity to Weather Fluctuations

The simulation results clearly show that the examined photovoltaic technologies exhibit varying levels of sensitivity to rapid weather fluctuations. Conventional silicon modules are the most affected, as both the rapid rise in temperature and the subsequent cooling after rainfall exert a significant influence on their performance. PERC technology demonstrates substantially greater stability due to its infrared-reflective rear layer, which helps moderate thermal loads. The PERC + UV film configuration displays the smallest performance variability, making it the most suitable option for environments characterized by unstable or rapidly changing weather conditions.

4.4.4 The Role of Cooling Efficiency

The findings indicate that the magnitude and rate of cooling following weather events –such as rainfall or sudden cloud cover– play a decisive role in restoring the performance of PV modules. PERC + UV film panels cool more rapidly and effectively, resulting in shorter periods of reduced efficiency. In contrast, conventional silicon modules, due to their higher thermal inertia, respond more slowly, leading to longer-lasting performance losses. Consequently, the cooling dynamics directly influence overall energy yield, particularly in regions where rapid or frequent weather transitions are common.

5. INTERPRETATION OF RESULTS AND CONCLUSIONS

The analyses and simulations clearly demonstrate that the examined photovoltaic technologies respond markedly differently to thermal and weather-related effects. Conventional silicon modules are prone to rapid internal temperature rise, which leads to significant efficiency losses and, over the long term, increases the risk of material fatigue. In contrast, PERC technology provides improved thermal regulation, while PERC + UV film systems exhibit outstanding performance in mitigating temperature fluctuations and reducing emissive losses.

The weather-change scenarios –including sudden cloud cover, rainfall, and partially overcast conditions– revealed that rapid temperature transitions can cause substantial performance degradation in conventional silicon technology. Across all examined conditions, the PERC + UV film modules delivered the most stable performance, enabled by their faster cooling capability and slower rate of reheating, which together help maintain high efficiency over extended periods.

Conclusions:

- The application of advanced thermal-management technologies is essential for ensuring the long-term efficiency of modern PV systems.

- The PERC + UV film configuration not only provides protection against rapid temperature variations but also ensures more stable energy yield under dynamically changing weather conditions.
- The comparative analysis of cooling dynamics and thermal responses indicates that not only the heating rate but also the rate of cooling plays a decisive role in determining daily performance stability.

Future research should incorporate real-time meteorological measurements and more detailed modeling of nonlinear degradation processes to enable more accurate long-term performance predictions. The results presented in this study may contribute to the development of improved PV technologies and the optimization of next-generation solar power systems. Based on the findings, PERC + UV film technology is recommended in regions where rapid and frequent weather changes are common.

REFERENCES

[1] G. Perrakis, A. C. Tasolamprou, G. Kenanakis, E. N. Economou, S. Tzortzakis, M. Kafesaki: *Ultraviolet radiation impact on the efficiency of commercial crystalline silicon-based photovoltaics: A theoretical thermal-electrical study in realistic device architectures*. Optics Express / OSA Continuum, 2020.

[2] A. Idzkowski, K. Karasowska, W. Walendziuk: *Temperature analysis of the stand-alone and building integrated photovoltaic systems based on simulation and measurement data*. MDPI. 2020, August 18. https://www.mdpi.com/1996-1073/13/16/4274

[3] J. A. Del Cueto, S. R. Rummel: *Degradation of photovoltaic modules under high voltage stress in the field: Preprint*, 2010, August. https://www.researchgate.net/publication/255216076_Degradation_of_Photovoltaic_Modules_Under_High_Voltage_Stress_in_the_Field_Preprint

[4] S. Kurtz, K. Whitfield, D. C. Miller, J. Joyce: *Evaluation of high-temperature exposure of rack-mounted photovoltaic modules*, July, 2009. https://www.researchgate.net/publication/224113570_Evaluation_of_high-temperature_exposure_of_rack-mounted_photovoltaic_modules

[5] R. Zhang, P. A. Mirzaei, J. Carmeliet: *Prediction of the surface temperature of building-integrated photovoltaics: Development of a high accuracy correlation using computational fluid dynamics*. Solar Energy, 2017, May. https://www.dora.lib4ri.ch/empa/islandora/object/empa:13845

IMPACT OF PARTIAL AND DYNAMIC SHADING ON THE PERFORMANCE AND THERMAL BEHAVIOR OF VARIOUS PHOTOVOLTAIC TECHNOLOGIES

Dávid Szalánczi[1], ***Péter Bencs***[2]
[1]PhD Student, [2]Associate Professor
[1,2]*Faculty of Mechanical Engineering and Informatics,*
Institute of Energy and Chemical Machinery
[1]szalanczi.david@student.uni-miskolc.hu,
[2]peter.bencs@uni-miskolc.hu

Abstract

This study investigates the effects of partial and dynamic shading on different photovoltaic technologies, including conventional silicon, PERC, and PERC + UV film modules. Using Simulink-based simulations, variations in internal panel temperature and conversion efficiency were analyzed under clear sky, partial shading, severe shading, and dynamic shading conditions.

The results indicate that partial shading induces nonlinear performance effects, with conventional silicon panels experiencing significantly higher temperature rises and efficiency losses. PERC technology demonstrates improved thermal behavior by reducing temperature fluctuations, while PERC + UV film modules provide the highest level of efficiency stability and thermal protection across all simulated shading scenarios.

Dynamic shading effects, such as those caused by moving tree branches, introduce additional challenges but can be effectively modeled. These conditions further highlight the performance differences between the examined technologies, emphasizing the importance of advanced thermal protection in shaded environments.

1. INTRODUCTION

The performance of photovoltaic (PV) systems is influenced by a wide range of environmental factors, among which partial shading is considered one of the most critical. Shading may be caused by trees, chimneys, antennas, power lines, or surrounding building structures, leading to localized power losses and significant temperature non-uniformities within the PV module. Under non-uniform irradiance conditions, certain cells may experience excessive heating, resulting in the so-called hot-spot phenomenon, which can cause long-term damage to the module and reduce its operational lifetime.

In recent years, several advanced PV technologies have been developed with the aim of improving conversion efficiency and enhancing thermal stability. Among these, PERC (Passivated Emitter Rear Cell) technology and its variant combined with a UV-protective film have shown considerable promise. However, the extent to which these technologies can mitigate the localized thermal effects and performance losses caused by partial shading has not yet been fully explored.

The objective of this study is to investigate, through simulation-based analysis, the effects of partial and severe shading on the behavior of three different PV technologies, with particular emphasis on internal temperature distribution and efficiency variation. The examined systems include conventional silicon-based modules, PERC modules, and PERC modules enhanced with UV-protective film.

2. OBJECTIVES

The aim of this study is to analyze, within a simulation environment, the impact of partial shading on the thermal and performance behavior of different photovoltaic technologies. The research focuses on quantifying changes in the internal temperature of the modules and the corresponding reduction in efficiency when portions of the PV panel are subjected to shading during specific periods of the day.

The study compares the following three PV technologies:

- Conventional silicon-based modules,
- PERC (Passivated Emitter Rear Cell) technology,
- Advanced PERC modules combined with a UV-protective film.

The simulations are designed to assess the extent to which modern PV technologies are capable of mitigating localized thermal effects and performance losses induced by partial shading.

3. LITERATURE REVIEW

Partial shading represents one of the most common challenges in the operation of photovoltaic (PV) systems, significantly affecting both performance and system lifetime. Shading not only reduces instantaneous energy output but may also induce the so-called hot-spot phenomenon, leading to localized overheating of PV cells and causing long-term degradation of the module (Pandian et al., 2016) [1].

Several studies have demonstrated that PERC technology is more effective in handling thermal non-uniformities than conventional silicon-based panels (Vogt et al., 2017) [2]. This improved behavior is primarily attributed to enhanced rear-side passivation, which contributes to more uniform heat distribution under non-ideal irradiance conditions.

Research has also highlighted the nonlinear nature of shading effects: even minor or temporary shading can trigger disproportionately large performance losses in photovoltaic systems. This phenomenon is particularly pronounced under partial shading, where only a portion of the panel receives irradiance, yet the series-connected cell structure causes a reduction in the efficiency of the entire module (Yaoba et al., 2022) [3].

4. MATERIALS AND METHODS

The objective of the investigation was to determine how different photovoltaic technologies – conventional silicon-based modules, PERC, and PERC modules combined with a UV-protective film– respond to partial shading conditions, with particular emphasis on temperature variation and performance loss.

4.1. Technological Background

The following PV technologies were examined in the simulation study:

- **Conventional silicon-based PV modules**: Characterized by lower conversion efficiency (approximately 18 %) and higher sensitivity to localized heating under shading conditions.
- **PERC technology**: Exhibits an efficiency of around 20 % and demonstrates improved tolerance to shading effects due to its passivated rear-side layer.

- **PERC + UV film technology**: Features a nominal efficiency of approximately 22 %. The UV-filtering layer reduces heat absorption and promotes more uniform temperature distribution within the module.

4.2. Simulation Environment

The simulations were implemented based on the principles of the Simulink modeling environment and covered a 24-hour cycle using real environmental parameters derived from meteorological data of the city of Miskolc (Hungary). The modeled weather conditions represent a typical summer day, during which partial shading occurs intermittently.

4.3. Shading Scenario

The simulation consists of three distinct phases:

- **Sunny morning period**: Uninterrupted operation between 0:00 and 9:00, with full solar irradiance.
- **Partial shading phase**: Between 9:00 and 14:00, randomly alternating shaded intervals lasting 15-45 minutes, resulting in 25–50 % power reduction.
- **Afternoon recovery period**: From 14:00 to 20:00, the PV module is again exposed to full illumination.

4.4. Evaluated Parameters

The following parameters were analyzed during the simulations:

- Temporal variation of the internal panel temperature,
- Degree of efficiency reduction during shaded periods,
- Percentage of power loss caused by partial shading,
- Comparative evaluation of technological differences under identical environmental conditions.

4.5. Efficiency Values and Initial Conditions

The initial conditions applied in the simulation model are summarized in Table 1. These parameters define the nominal efficiency, initial internal temperature, and thermal conductivity of the examined photovoltaic technologies.

Table 1: Initial conditions of the examined PV module technologies

Technology	**Normal efficiency**	**Initial internal temperature**	**Thermal conductivity** $[W/m \cdot K]$
Conventional silicon	18 %	25 °C	148
PERC	20 %	25 °C	130
PERC + UV film	22 %	25 °C	120

The selected values represent typical characteristics of commercially available PV modules and provide a consistent baseline for comparative analysis under identical shading conditions.

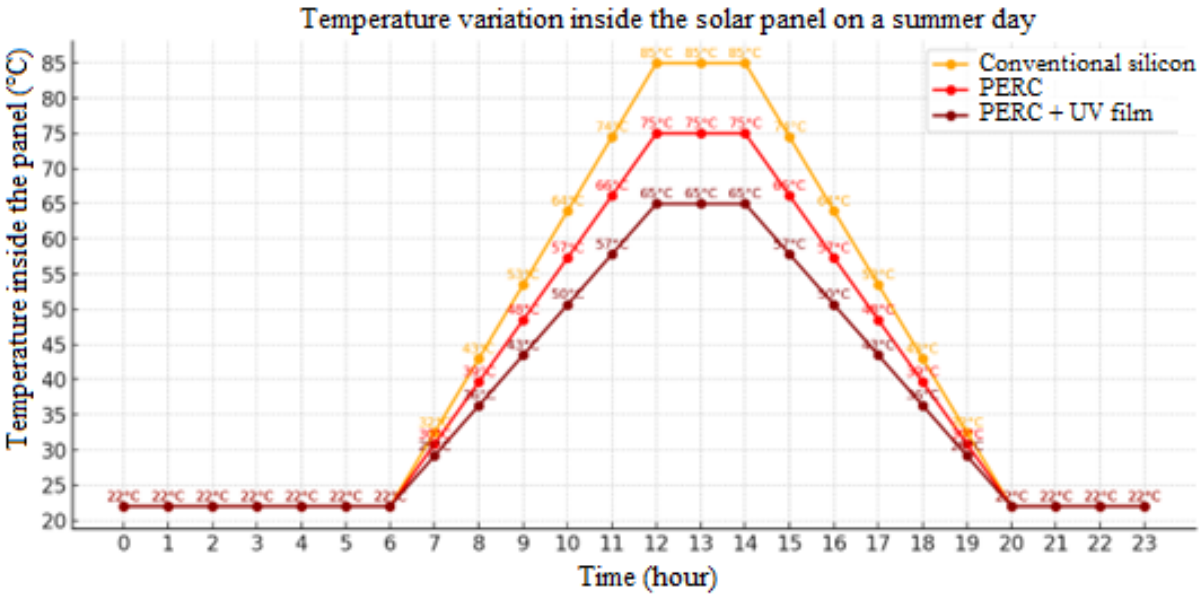


Figure 1: Internal temperature variation of the three PV module technologies under clear weather conditions

4.6. *Definition of Shading Scenarios*

During the simulations, four different shading conditions were evaluated to assess the response of the PV technologies:

- **Clear weather (baseline case)**: Uninterrupted solar irradiance without shading effects. This scenario serves as the reference condition for all technologies.
- **Partial shading**: Approximately 70 % reduction in incident irradiance during a defined time interval (e.g., 9:00–11:00 or 12:00–14:00). This scenario typically represents shading caused by trees, antennas, or other objects that partially cover the module surface.
- **Severe shading**: More intense obstruction, corresponding to a 50 % or greater reduction in incoming solar radiation. This type of shading may affect continuous regions of the module, such as shading from neighboring buildings or roof structures.
- **Dynamic (moving) shading scenario**: A time-varying and non-uniform shading pattern characterized by fluctuating intensity and duration. This scenario more accurately represents real-world conditions caused by natural disturbances, such as moving tree branches or rapidly shifting cloud shadows.

5. RESULTS AND EVALUATION

5.1. *Clear Weather Conditions without Shading*

The Figure 1 presents the internal temperature evolution of three photovoltaic technologies – conventional silicon, PERC, and PERC + UV film– on a clear and uninterrupted summer day. The results are shown with hourly resolution over a 24-hour period (0-24 h).

Observations from the figure:

- **Between 0:00 and 6:00**, all technologies remain at a similar low temperature level of approximately 22 °C.

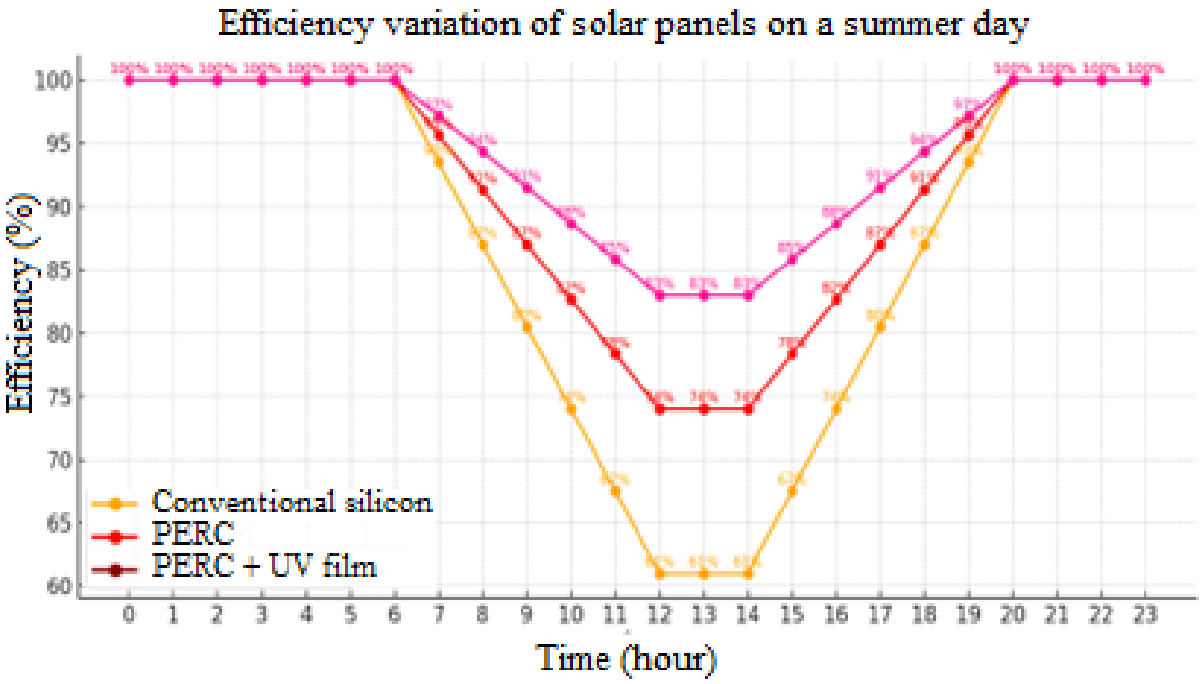


Figure 2: Efficiency variation of the three PV module technologies under clear weather conditions

- **From 7:00 to 11:00**, the heating process intensifies, with distinct rates observed for each technology:
  - The **conventional silicon module** exhibits rapid temperature increase and reaches a peak of approximately 85 °C **between 11:00 and 14:00**.
  - The **PERC module** shows a lower maximum temperature of around 75 °C, representing a reduction of about 10 °C compared to the conventional panel.
  - The **PERC + UV film module** reaches a maximum temperature of only 65 °C, remaining approximately 20 °C below the conventional silicon technology.
- **Between 15:00 and 18:00**, the cooling phase begins:
  - **PERC + UV** system exhibits the fastest heat dissipation.
  - The **conventional silicon module** cools down the slowest due to higher thermal inertia.
- **During the evening period (19:00–24:00)**, all technologies return to a stable baseline temperature of approximately 22-23 °C.

Interpretation: The temperature curves clearly demonstrate that the technological design of PV modules has a significant influence on their thermal behavior. **Conventional silicon** panels are more prone to overheating, which may degrade performance and accelerate aging over long-term operation. In contrast, **PERC + UV film modules** experience lower thermal stress and faster cooling, resulting in improved operational stability and enhanced thermal resilience.

Figure 2 illustrates the daily efficiency profiles of three photovoltaic technologies –conventional silicon, PERC, and PERC + UV film– during a clear summer day. The purpose of the simulation was to investigate how increasing operating temperature affects the efficiency of energy conversion.

Observations from the figure:

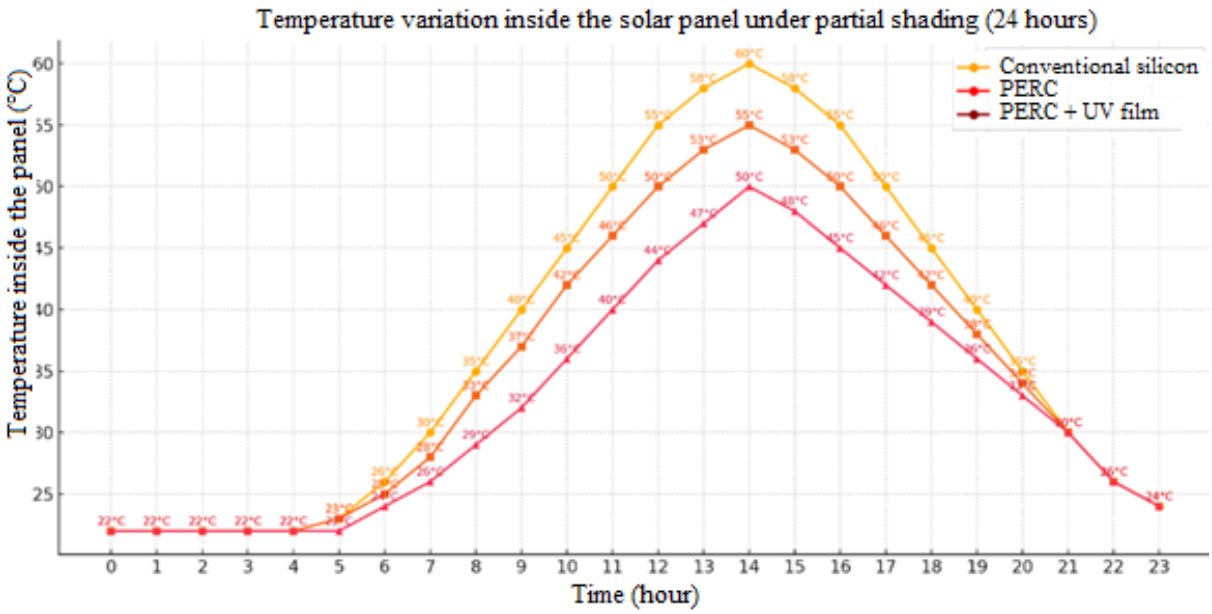


Figure 3: Internal temperature variation of the three PV module technologies under partial shading conditions

- **Between 0:00-6:00 and 19:00-24:00**, all systems operate at efficiency levels close to 100 %, corresponding to low internal temperatures.
- From **7:00 onward**, efficiency begins to decrease, reaching its minimum between **12:00 and 14:00**:
  - The **conventional silicon module** exhibits the most pronounced degradation, with efficiency dropping to approximately 61 % at peak temperature ($\sim$ 85 °C).
  - The **PERC technology** maintains higher efficiency, with a minimum value of around 74 %.
  - The **PERC + UV film system** demonstrates the most stable behavior, with efficiency remaining at approximately 83 %, corresponding to only a 17 % reduction.
- As the modules cool during the afternoon hours, efficiency gradually recovers, with the **PERC + UV film technology** showing the fastest stabilization.

Conclusion: Even under clear summer conditions, significant performance differences can be observed among the examined technologies. While conventional silicon panels suffer substantial efficiency losses due to thermal stress, the **PERC + UV film technology** provides superior performance stability and enhanced resistance to elevated operating temperatures.

5.2. Weak (Partial) Shading Conditions

Figure 3 illustrates the internal temperature evolution of three photovoltaic technologies – conventional silicon, PERC, and PERC + UV film– over a 24-hour period under partial shading conditions. The aim of the analysis was to assess the extent of heating and subsequent cooling exhibited by each technology during the dynamic alternation of sunlight and shading.

Observations from the figure:

- **During the early morning hours (0:00–6:00)**, all technologies maintain a constant and low temperature of approximately 22 °C, as the absence of solar irradiance does not induce heating.
- **From the morning period onward (6:00–10:00)**, a gradual increase in temperature is observed, reaching a peak around midday. The **conventional silicon modules** exhibit the highest temperature rise, reaching a maximum of approximately 60 °C, while **PERC technology** peaks at around 55 °C, and the **PERC + UV film** configuration shows the lowest maximum temperature of approximately 50 °C.
- **During the afternoon period (14:00–20:00)**, the cooling behavior differs among the technologies. Conventional panels cool down more slowly, whereas the **PERC + UV film technology** demonstrates faster heat dissipation, indicating superior thermal management characteristics.
- **In the evening hours (20:00–23:00)**, the temperature curves converge, gradually returning toward their initial baseline values.

Conclusions:

- **Conventional silicon technology** exhibits higher peak temperatures and slower cooling, which may increase the risk of material fatigue and long-term degradation.
- **PERC technology** provides intermediate performance, showing improved thermal behavior compared to conventional panels.
- **PERC + UV film technology** demonstrates the most effective temperature regulation, characterized by lower maximum temperatures and faster cooling. These properties contribute to improved efficiency stability and extended module lifetime.

These results support the applicability of modern PV technologies incorporating thermal-protection layers for ensuring stable energy production under variable environmental conditions.

Figure 4 illustrates how the efficiency of photovoltaic modules changes under partial shading during a typical summer day. The shading effect occurring around midday leads to a noticeable reduction in efficiency, particularly in the case of **conventional silicon technology**. The magnitude of the efficiency loss closely follows the increase in internal panel temperature and the disruption of energy balance caused by shaded cells.

The **PERC technology** exhibits improved tolerance to partial shading, while the **PERC + UV film solution** demonstrates the most stable efficiency behavior, with a less pronounced performance drop. This characteristic is especially advantageous in urban or semi-built environments, where shading is often dynamic or intermittent in nature.

5.3. Severe Shading Conditions

Figure 5 presents the simulated internal temperature evolution of three photovoltaic technologies –**conventional silicon, PERC**, and **PERC + UV film**– under conditions of strong, daylong partial shading. The objective of this analysis was to determine which technology is more capable of mitigating heat accumulation under reduced direct solar irradiance.

Observations from the figure:

- **During the early morning hours (0:00–5:00)**, all modules remain at a similarly low temperature of approximately 22 °C.

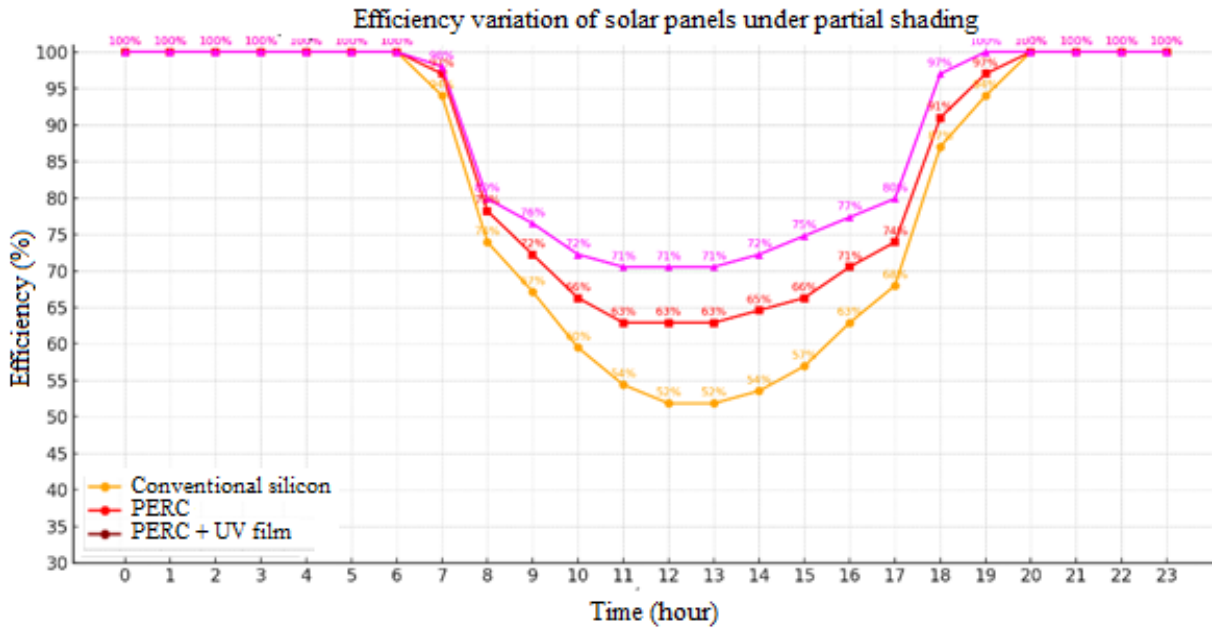


Figure 4: Efficiency variation of the three PV module technologies under partial shading conditions

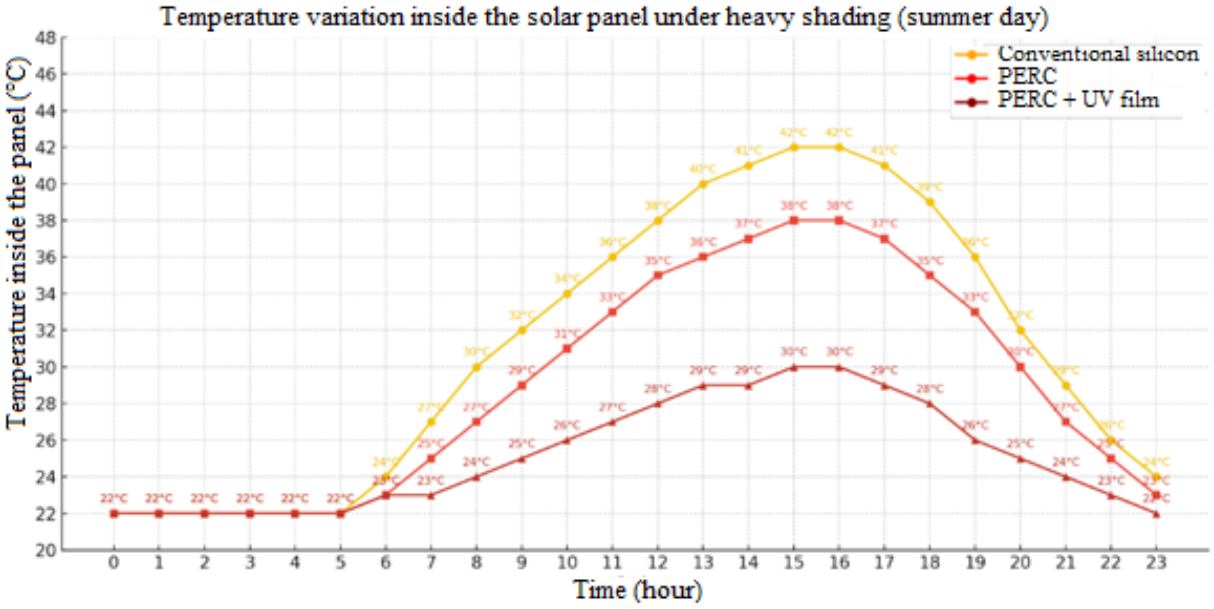


Figure 5: Internal temperature variation of the three PV module technologies under severe shading conditions

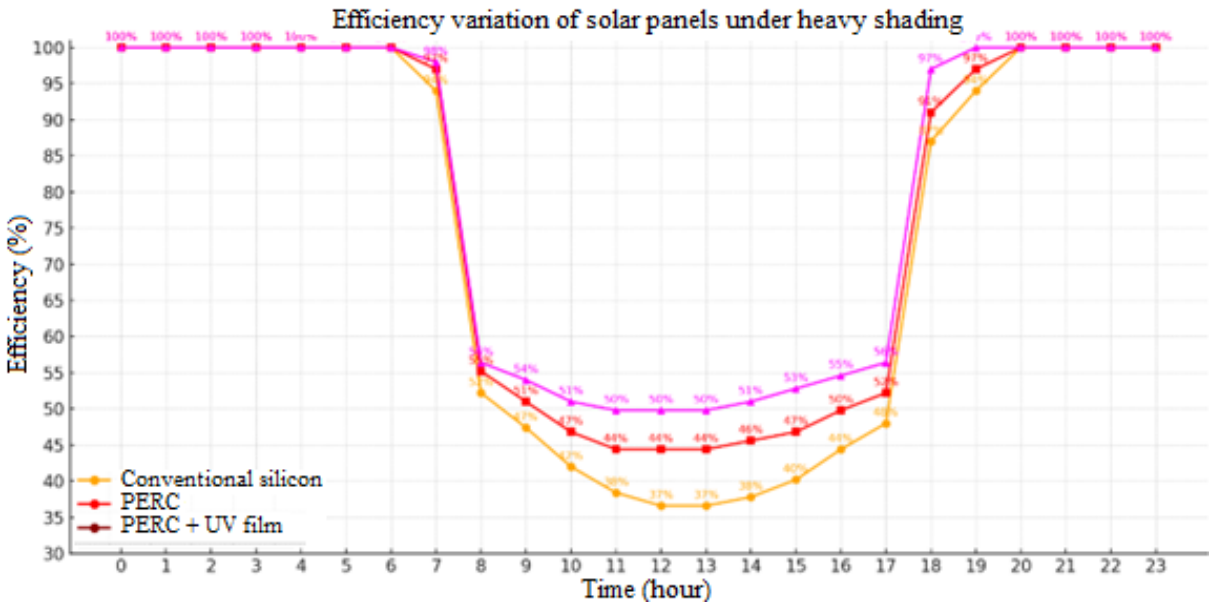


Figure 6: Efficiency variation of the three PV module technologies under severe shading conditions

- **Throughout the daytime period (6:00–16:00)**, differences between the technologies become increasingly pronounced:
  - The **conventional silicon module** reaches a temperature of approximately 42 °C, indicating considerable thermal load even under shaded conditions.
  - The **PERC technology** exhibits a lower maximum temperature of around 38 °C, reflecting improved thermal management.
  - The **PERC + UV film technology** shows the lowest temperature peak at only 30 °C, attributable to UV radiation reflection and reduced heat conduction.
- **During the afternoon and evening hours**, cooling behavior also differs among the technologies:
  - **Conventional modules** cool down more slowly.
  - The **PERC + UV technology** returns to baseline conditions more rapidly.

Conclusions:

- Even under severe shading, noticeable temperature increases occur, particularly in conventional silicon modules.
- **PERC** and especially **PERC + UV film technologies** demonstrate significantly more favorable thermal dynamics.
- This behavior directly affects efficiency stability and long-term module lifetime.

As shown in Figure 6, severe shading leads to a substantial reduction in the efficiency of photovoltaic modules, particularly during the midday hours, when solar irradiance would otherwise be at its highest. Shading not only reduces direct irradiance but also alters the thermal dynamics of the panel, which further contributes to performance degradation.

The **conventional silicon** technology exhibits the most pronounced efficiency loss, with efficiency dropping to as low as 37 %. In comparison, the **PERC technology** maintains a higher

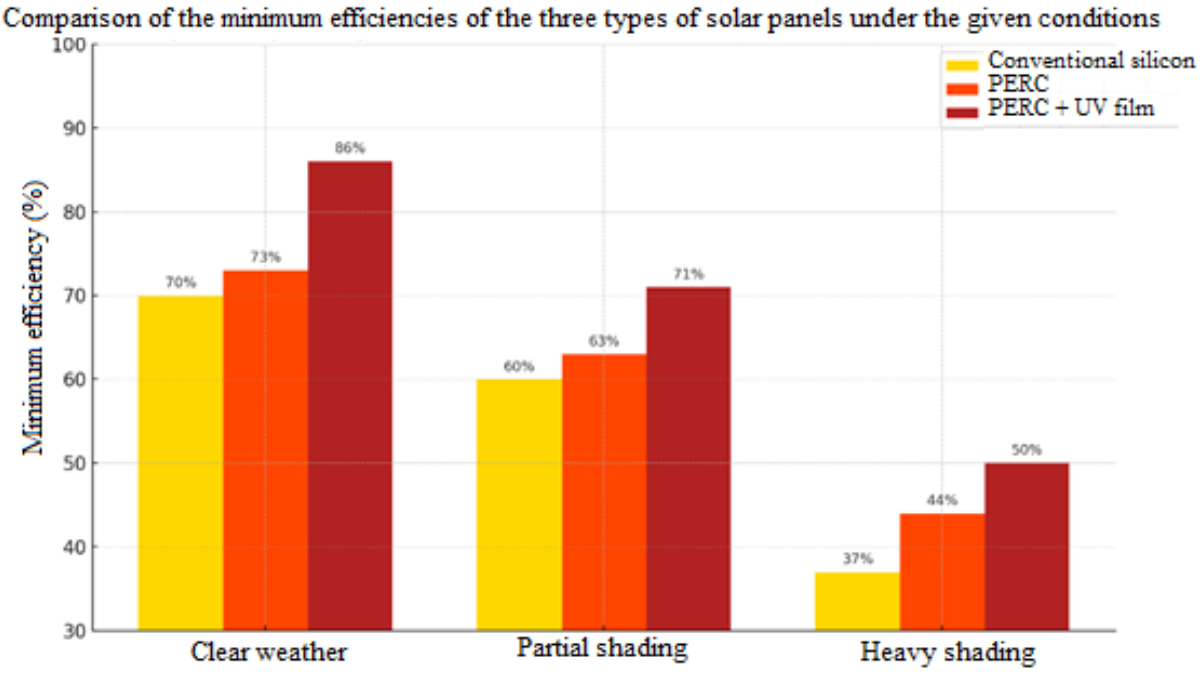


Figure 7: Comparison of minimum efficiency values of the three PV module technologies under different conditions

minimum efficiency of approximately 44 %, while the **PERC + UV film solution** demonstrates the greatest stability, sustaining a minimum efficiency of around 50 % even during the most critical periods.

These differences become especially significant in environments where persistent or intense shading occurs. Under such conditions, technological distinctions manifest not only in efficiency but also in overall reliability. Once again, the **PERC + UV film technology** stands out due to its balanced and resilient performance, making it a more favorable choice under extreme environmental constraints.

5.4. Summary Efficiency Comparison Based on the Three Shading Scenarios

To enable a more detailed comparison of the observed phenomena, separate bar charts were prepared to illustrate the minimum efficiency values and the magnitude of efficiency reduction across the three shading scenarios.

The bar charts shown in Figures 7 and 8 present a comparative analysis of the minimum efficiency values and efficiency reductions of three photovoltaic technologies –conventional silicon, PERC, and PERC + UV film– under three different operating conditions: clear weather, partial shading, and severe shading. The purpose of this comparison is to illustrate the sensitivity of the examined technologies to thermal load and shading effects.

Based on the results, the conventional silicon technology exhibits the largest efficiency degradation across all scenarios, particularly under partial and severe shading conditions. PERC technology demonstrates improved stability but remains noticeably affected by temperature and shading-related stress. In contrast, the PERC + UV film-based system delivers the best performance under all examined conditions, maintaining lower maximum temperatures, smaller efficiency losses, and faster recovery following shading events.

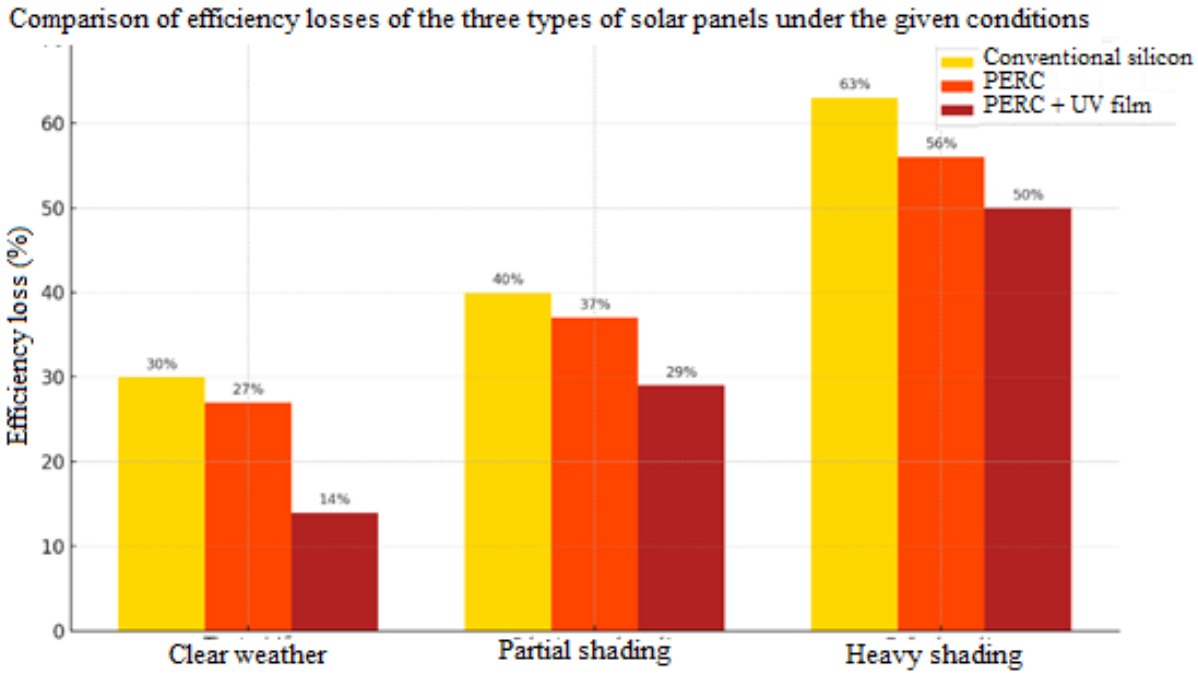


Figure 8: Comparison of efficiency reductions of the three PV module technologies under different condition

5.5. Simulation of Dynamic Shading (Tree Branches)

The objective of this simulation is to investigate the behavior of different photovoltaic technologies in a realistic dynamic shading environment, where solar irradiance varies irregularly due to tree branches moved by wind. The examined time interval spans from 12:00 to 16:00, corresponding to the hottest period of the day, when the PV modules are already exposed to elevated operating temperatures.

The baseline temperature and efficiency profiles were derived from the clear weather scenario, onto which the effects of dynamic shading were superimposed. This approach allows the isolation and detailed evaluation of the impact of time-varying, non-uniform shading on the thermal and performance response of the PV modules.

Figure 9 illustrates that, under dynamic shading conditions, the internal temperature of the PV modules not only decreases but also begins to fluctuate dynamically. This behavior effectively models variations in light transmission caused by wind-driven tree branches, particularly under gusty weather conditions.

Based on the simulated temperature profiles:

- The **conventional silicon module** reaches the highest temperature levels, fluctuating within the range of approximately 76–85 °C, accompanied by pronounced oscillations.
- The **PERC module** exhibits improved stability, with internal temperatures varying between approximately 70–77 °C.
- The **PERC + UV film-based module** demonstrates the most balanced thermal behavior, operating within a narrower range of approximately 62–68 °C, with moderate fluctuations and strong resistance to short-duration shading impulses.

This behavior highlights that, in dynamic shading environments, thermal stability is a critical factor alongside efficiency. Effective temperature regulation plays a key role in mitigating overheating, material fatigue, and long-term degradation of photovoltaic modules.

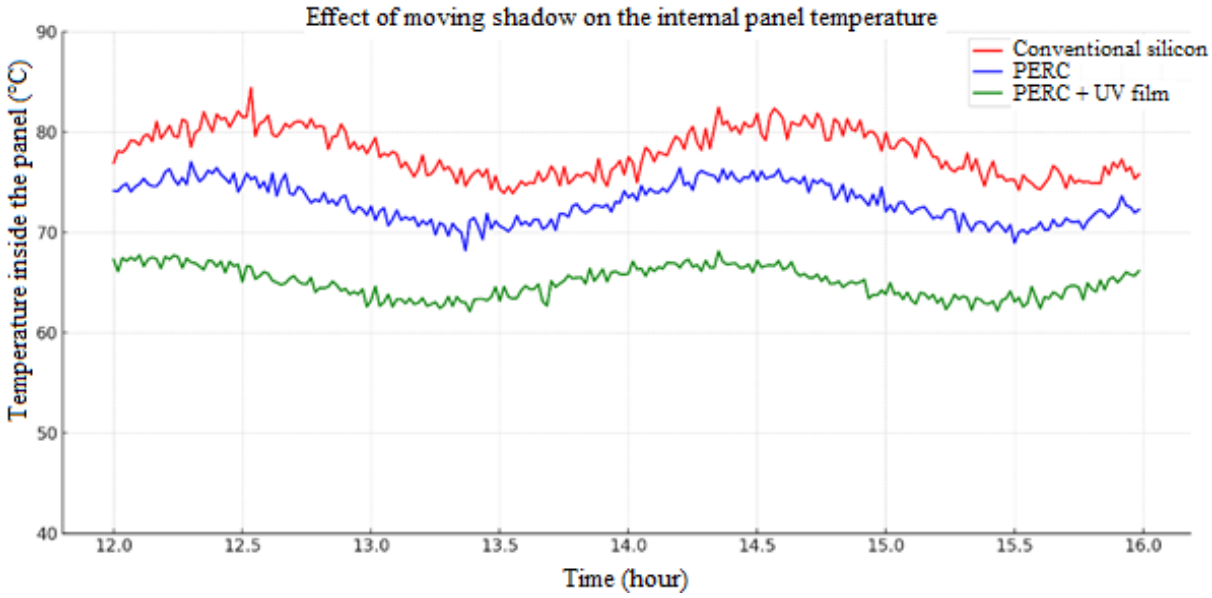


Figure 9: Comparison of minimum internal temperatures of the three PV module technologies under dynamic shading conditions

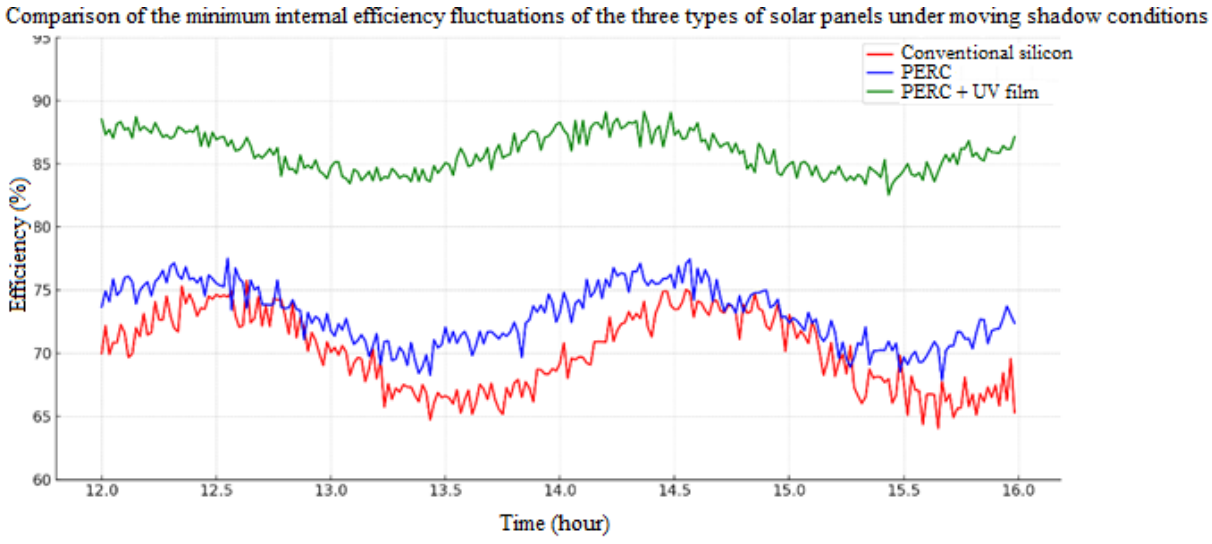


Figure 10: Comparison of minimum efficiency fluctuations of the three PV module technologies under dynamic shading conditions

Based on Figure 10, it is evident that **conventional silicon technology** is the most sensitive to rapid fluctuations in solar irradiance. Its efficiency curve exhibits pronounced oscillations and periodically drops below 65 %, indicating poor adaptability to fast-changing shading conditions.

In contrast, **PERC technology** demonstrates more balanced behavior; however, noticeable efficiency fluctuations are still present.

The application of a **PERC + UV film** significantly mitigates the impact of dynamic shading. In this case, efficiency remains consistently within the range of approximately 84–88 %, showing only minor fluctuations.

This simulation scenario highlights that technological differences become even more pronounced under dynamic shading conditions. Such effects are particularly relevant in urban environments, where shading is often intermittent or mobile, caused by factors such as tree branches, moving vehicles, or surrounding buildings.

6. INTERPRETATION OF RESULTS AND CONCLUSIONS

Based on the simulation results, it can be clearly observed that photovoltaic technologies exhibit significantly different behavior under varying shading conditions. Conventional silicon-based modules experience the highest temperature rise and the most pronounced efficiency degradation, particularly under partial and severe shading. This technology is especially susceptible to hot-spot phenomena and demonstrates slower recovery to its original efficiency level.

PERC technology shows improved tolerance to thermal stress and exhibits more moderate efficiency losses under shaded conditions. However, a noticeable performance gap remains between clear and shaded scenarios, and the recovery time is not optimal in all cases.

The PERC + UV film combination clearly stands out among the examined technologies. Under all investigated conditions, this configuration exhibits the smallest efficiency reduction, the lowest temperature peaks, and the fastest recovery behavior. The thermal-stabilizing effect of the UV-protective layer therefore contributes not only to preventing overheating but also to mitigating performance losses.

In summary, regardless of the type of shading, PERC + UV film-based PV modules prove to be the most robust and balanced solution. From a technological perspective, this configuration appears to be the most suitable choice for applications in partially or intermittently shaded environments.

7. SUMMARY AND RECOMMENDATIONS

The objective of this study was to compare, through simulation-based analysis, the behavior of three different photovoltaic technologies –conventional silicon, PERC, and PERC + UV film– under various shading conditions. The investigation considered three primary scenarios during a summer day: clear weather, partial shading, and severe shading.

The results indicate that technological differences have a substantial impact on PV system performance. Conventional silicon cells exhibit high sensitivity to shading effects, while PERC technology shows more moderate performance degradation. The PERC + UV film systems demonstrate the best overall performance and thermal stability across all examined scenarios.

The simulations also reveal that not only the extent of shading but also the type of shading (uniform versus partial or non-uniform) significantly influences both efficiency and thermal behavior.

Recommended directions for further research include:

- Simulations focusing on winter conditions, where thermal load is lower but the relative impact of shading may be more pronounced.

- Advanced modeling of dynamic shading effects, such as moving clouds or oscillating tree branches.
- Investigation of the role of bypass diodes in different PV technologies under shading conditions.
- Experimental validation of the simulation results through laboratory or field measurements.

The modeling of dynamic shading phenomena introduces additional simulation challenges and may require more advanced predictive tools to achieve higher accuracy in future studies.

REFERENCES

[1] A. Pandian, K. Bansal, D. J. Thiruvadigal, S. Sakthivel: *Fire hazards and overheating caused by shading faults on photo voltaic solar panel*. Fire Technology, vol. 52, pp. 349–364, 2016.

[2] M. R. Vogt, H. Schulte-Huxel, S. Blankemeyer, M. Offer: *Reduced module operating temperature and increased yield of modules with PERC instead of Al-BSF solar cells*. November, 2016. https://www.researchgate.net/publication/309756771_Reduced_Module_Operating_Temperature_and_Increased_Yield_of_Modules_With_PERC_Instead_of_Al-BSF_Solar_Cells

[3] Yaouba, M. Bajaj, C. Welba, K. Bernard, Kitmo, S. Kamel, M. F. El-Naggar: *An experimental and case study on the evaluation of the partial shading impact on PV module performance operating under the sudano-sahelian climate of Cameroon*. Frontiers. June 3, 2022. https://www.frontiersin.org/journals/energy-research/articles/10.3389/fenrg.2022.924285/full

ANALYSIS OF MHD ENTROPY GENERATION OF RADIATIVE FLOW PAST A NONLINEAR STRETCHING POROUS PLATE

Wekesa Simon Waswa[1], ***Winifred Nduku Mutuku***[2], ***Krisztian Hriczó***[3]
[1]PhD Student, [2,3]Associate Professor,
[1,3]*Faculty of Mechanical Engineering and Informatics, Institute of Information Technology*
[2]*Department of Mathematics & Actuarial Science, Kenyatta University, Kenya*
[1]simon.waswa.wekesa@student.uni-miskolc.hu,
[2]mutuku.winifred@ku.ac.ke, [3]krisztian.hriczo@uni-miskolc.hu

Abstract

Improvement of nanofluid can take into account effective heat transfer fluid when developing energy systems. In order to find solutions, engineers either develop new nanofluids or improve the ones that already exist. A two-dimensional radiative nanofluid flow via an unevenly extending sheet is investigated in this work. The strength of the magnetic field is applied perpendicularly. The similarity and space variables are used to nondimensionalize, and linearize the flow equations. The impact of factors such as porosity and magnetic strength on the flow variables is emphasized. Numerically simulated graphs and tabular values were produced using a shooting and Runge-Kutta method in MATLAB. In addition to using entropy to improve image quality, the results allow for the comprehension, control, and regulation of heat energy movement.

1. INTRODUCTION

Nanofluids have a significant role in cooling applications. In solar drying procedures that use porous absorbers, nanofluids flow over the surfaces of solar collectors to improve heat absorption. Combustion chambers and heat exchangers use perforated metal foams that are heated by external radiation sources like furnaces and nuclear reactors to transfer heat. Nanometals can be added to a base fluid to help collect, dissipate and control heat flow in MHD generators, nuclear reactors, molten salts, and metal casting operations. Increased exergy is linked to a decrease in thermal irreversibility, resulting in both internal and exterior thermo-equilibrium. The advancements and future prospects of the mathematical model was proposed by Tiwari and Das (2007), [7]. The context of nanofluid flow across various geometries, accounts for viscosity and thermal conductivity, among other physical properties, which enhances the rate of thermal transfer. The characteristics of entropy optimization in viscous second-grade nanofluid subjected to thermal radiation was investigated utilizing the Tiwari and Das model, [2].

The increase in Reynolds and Brinkman numbers resulted in a rise in the system's entropy. In the exploration of the entropy analysis of MHD hybrid nanoparticles through the application of the OHAM, the effects of viscous dissipation and thermal radiation were considered, [6]. The magnetic parameter and the entropy generation rate contributed to an increase in the velocity profile. OHAM technique was applied in the examination of the influence of modified heat and mass fluxes on the thermally radiated boundary layer flow over a stretched sheet, [3]. The findings revealed the temperature profile diminishes with an increase in the Prandtl number.

Least addresses the irregular stretching of porous plate. This research aims to bridge that gap by incorporating solar radiation within the framework of the Tiwari-Das model.

2. METHODOLOGY

The model takes into account the particles' displacement as non-interactive with streamlined and incompressible flow. Heat is produced by the system both inside and outside. In light of the

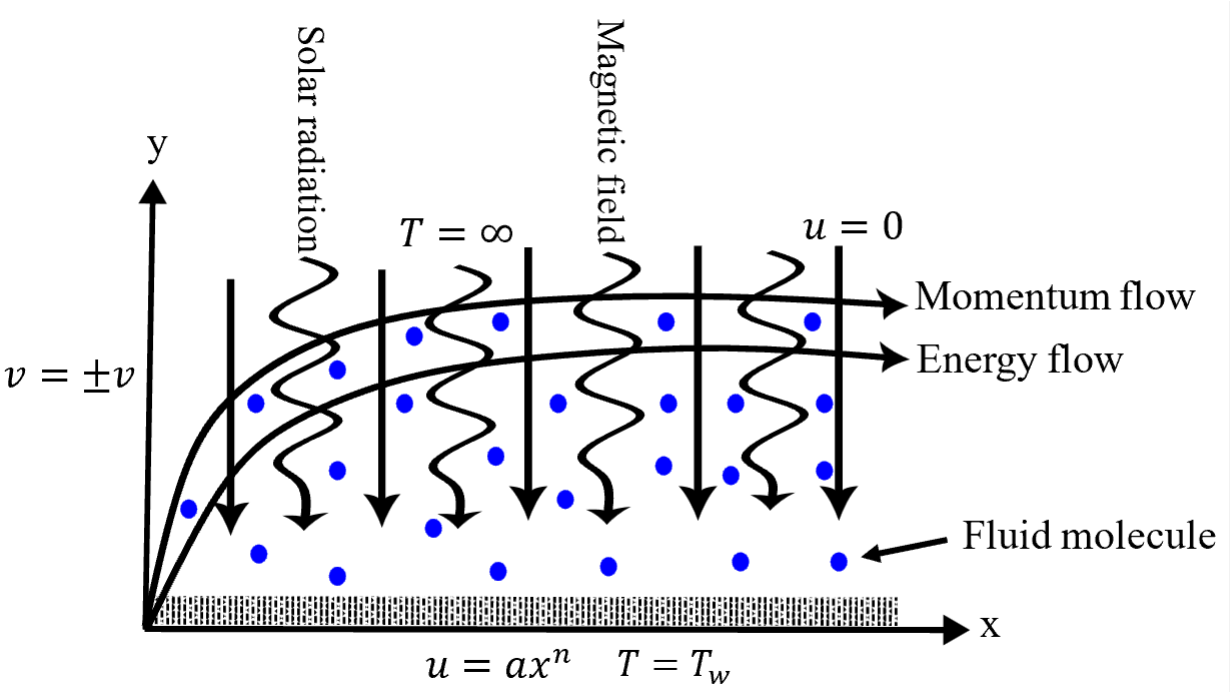


Figure 1: Schematic geometry of the flow

aforementioned assumptions, a schematic geometry (Figure 1) has been constructed to depict the flow behavior and the mass, momentum, and energy equations (1–3) [1] which are expressed along with the appropriate flow constraints (4) as follows.

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \tag{1}$$

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = \frac{\mu}{\rho}\frac{\partial^2 u}{\partial y^2} - \frac{\sigma\beta_0}{\rho}u - \frac{\mu u}{k_p}, \tag{2}$$

$$u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y} = \alpha\frac{\partial^2 T}{\partial y^2} + Q_r'' - \frac{\mu u}{\rho C_p}\left(\frac{\partial u}{\partial y}\right)^2 + \frac{\sigma\beta_0^2}{\rho C_p}u^2, \tag{3}$$

$$\left.\begin{array}{l} u = u_w = cx^n, v = \pm v_0 x^n, T = T_w = T_\infty + Ax^n \text{ at } y = 0 \\ u \to 0, T \to \infty \text{ as } y \to \infty \end{array}\right\}, \tag{4}$$

where the parameters denoted by (v,u) represents velocity along the x and y axes, (x,y) are horizontal and vertical lengths, μ is the kinematic viscosity, ρ is density, σ is magnetic permeability of free space, β_0 is the transverse magnetic strength, k_p is permeability of porous material, T is the temperature, α is thermal diffusivity, Q_r'' is the source radiation modeled by [3], C_p is specific heat capacity, n is nonlinearity parameter, v_0 is velocity scaling constant, (c,A) are scaling constants, T_w is the wall temperature, T_∞ is far field temperature.

The transformative similarity variables applied to contract equations (1), (2), (3) and the boundary conditions (4) to a set of nonlinear dimensionless equation (5) and constraint (6).

$$\left.\begin{array}{l} f''' + \dfrac{n+1}{2}ff'' - nf'^2 - Mf' - Kf' = 0, \\ \theta'' + Pr\left(\dfrac{n+1}{2}f\theta' - nf'\theta\right) = -S_0 + PrEcf'' + PrM_\theta f'^2, \end{array}\right\} \tag{5}$$

$$\left.\begin{array}{l} f = f_0, f' = 1, \theta = 1, \eta = 0, \\ f' = 0, \theta' = 0, \eta \to \infty. \end{array}\right\} \tag{6}$$

with M as Hartmann number, K as porosity, Pr as Prandtl number, S_0 as heat source entropy, Ec as Eckart number, M_θ as joule heating, f_0 as dimensionless suction parameter, η is the dimensionless length f and θ are dimensionless velocity and temperature variables.

The entropy production (S_1) per unit volume and the characteristic scale (S_2) are expressed by [4] as equation (7) and (8).

$$S_1 = \frac{k}{T^2}\left(\frac{\partial T}{\partial y}\right)^2 + \frac{\mu}{T}\left(\frac{u}{\partial y}\right) + \frac{\sigma\beta_0^2 u^2}{T} + Q_0 I_0 e^{-\beta y} + \frac{\mu u^2}{k_p T} = 0, \tag{7}$$

$$S_2 = \frac{k(T_w - T_\infty)^2}{T_\infty^2 L_x^2} = 0, \tag{8}$$

where β is the absorption coefficient, k is thermal conductivity, Q_0 is the heat generation coefficient, I_0 is the initial intensity, L_x is the characteristic length.

The ratio of equation (7) to (8) caliminates into unitless form by the use of suitable similarity variables, stream and equipotential functions. The conduction, radiation and joule heating over total entropy production can be examined as Bejan number by

$$Be = \frac{\theta'^2 + Se^{-\lambda\eta} + PrM_\theta f'^2}{\frac{1}{(1+\varepsilon\theta)}\left(\theta'^2 + PrEcf''^2 + PrM_\theta f'2 + \frac{PrEc}{K}f'^2\right) + Se^{-\lambda\eta}}, \tag{9}$$

where S is the radiation source, ε is dimensionless temperature ratio and λ is the heat absorption coefficient.

The skin friction and the rate of energy transfer can be stated as equation (10) and nondimensionalized to equation (11) and (12) using Newton's law of viscosity and Fourier law of heat transfer.

$$Cf_x = \frac{\tau_w}{\frac{1}{2}\rho u^2}, \quad Nu_x = \frac{xq_w}{k(T_w - T_\infty)}, \tag{10}$$

$$Cf_x Re_x^{\frac{1}{2}} = (2(n+1))^2 f'', \tag{11}$$

$$Nu_x Re_x^{-\frac{1}{2}} = -\left(\frac{n+1}{2}\right)^{\frac{1}{2}} \theta', \tag{12}$$

Cf_x is skin friction, τ_w is shear stress and q_w is the heat flux near the wall, Nu_x and Re_x are Nusselt and Reynolds numbers respectively.

3. RESULTS AND DISCUSSION

The deployment of state variable on to equation (5), a shooting technique together with Runge-Kutta methods applied on unitless forms of boundary condition (6), and equations (11–12) computed numerical results in MATLAB. The graphs were generated and values tabulated.

The selected parameter $K = 0.5$ falls within the typical range of 0.3 to 0.9, which corresponds to commonly used porous materials such as metal foams and fibrous media. The Prandtl number, representing the ratio of thermal diffusion to momentum diffusion, was chosen between 1 and 2 to investigate the dominance of heat transfer mechanisms in liquid metals and

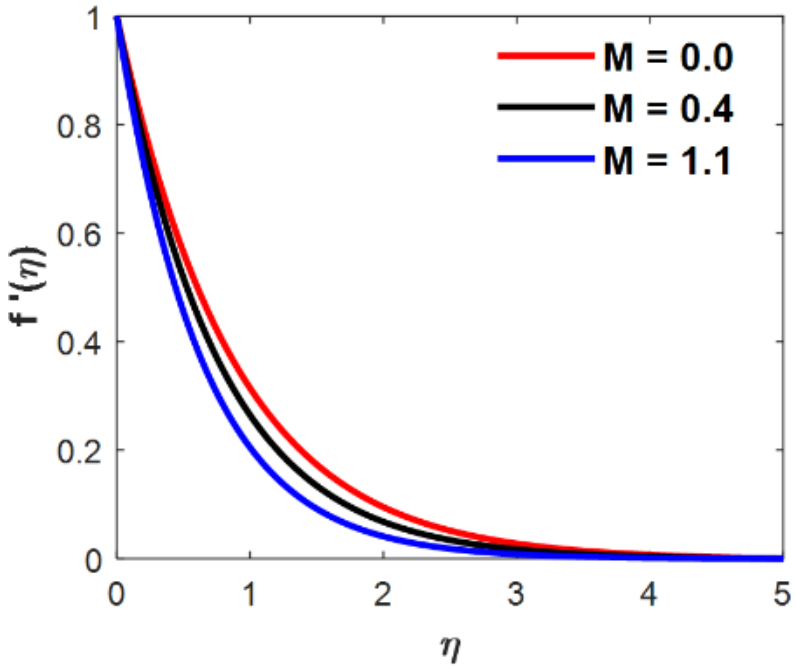


Figure 2: Effect of Magnetic field on velocity ($n = 0.5, K = 0.5, Pr = 0.72, S = 0.2, Ec = 0.01, M_\theta = 0.1, f_0 = 0.2$)

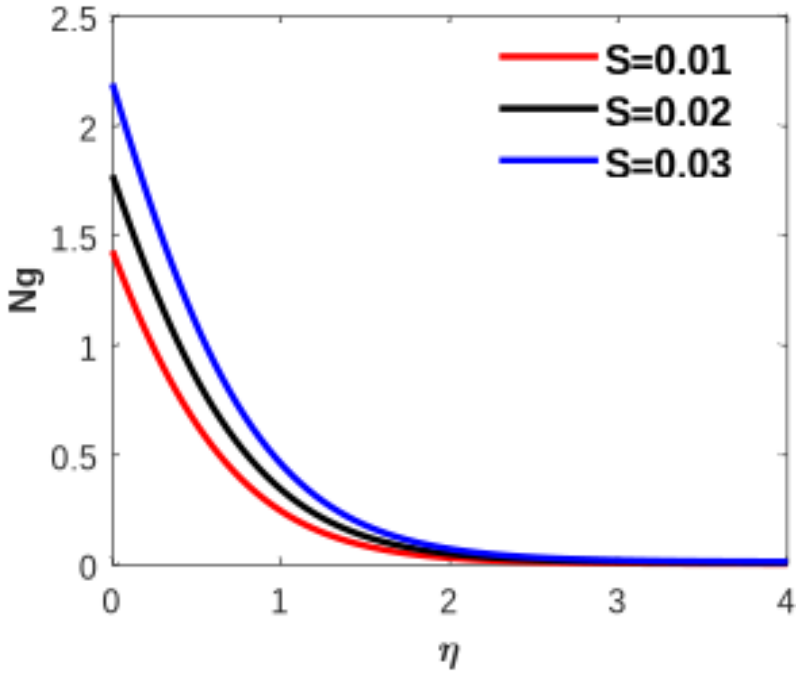


Figure 3: Effect of solar radiation on entropy ($n = 0.2, f_0 = 0.2, Pr = 1.62, Br = 0.1, \varepsilon = 0.2, \lambda = 0.2, Ec = 0.5, M = 0.2, M_\theta = 0.4, K = 0.5$)

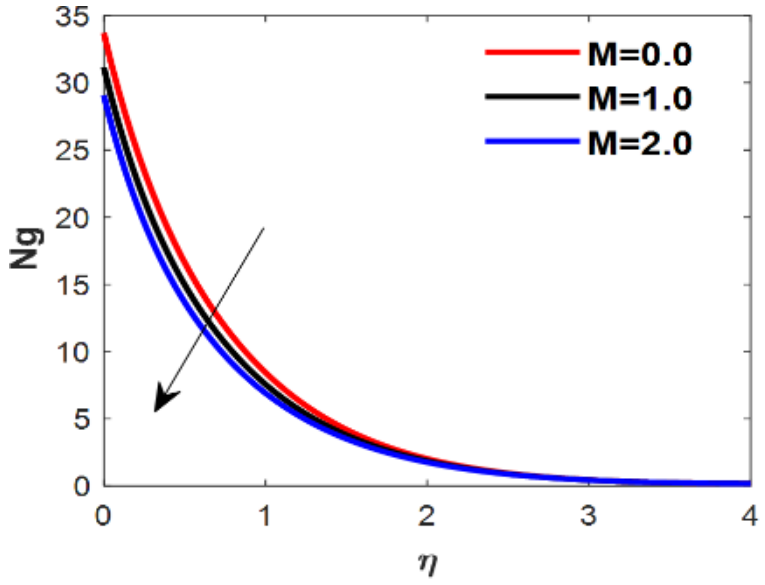


Figure 4: Effect of magnetic field on entropy ($n = 0.4, f_0 = 0.2, K = 0.5, Pr = 0.92, Br = 0.4, M_\theta = 0.1, \varepsilon = 0.2, \lambda = 0.3, S = 0.6, Ec = 0.5$)

Table 1: Effect of porosity on stickiness and rate of energy transfer ($n = 0.2, f_0 = 0.2, Pr = 0.7, Ec = 0.15, M_\theta = 0.15, S = 0.02$)

M	K	$Cf_x Re_x^{\frac{1}{2}}$	$Nu_x Re_x^{-\frac{1}{2}}$
0	0.1	-5.7600	0.30000
1	0.2	-7.5061	0.30014
2	0.3	-9.7092	0.30039
3	0.4	-11.4870	0.30068
4	0.5	-13.0190	0.30094

nanofluids. The Eckert number provides insight into the influence of fluid momentum on thermal behavior. Its value was varied from 0 to 1 to represent conditions ranging from minimal to significant velocity effects on the flow. The Brinkman number, which quantifies the relative significance of viscous dissipation compared to heat conduction, was selected to reflect a moderate level of viscous heating.

Joule heating included captures the strong thermal effects produced by electric currents within the flowing medium. The values are consistent with those reported in the referenced studies such as the mixed convective and viscous heating effect of electromagnetic Tiwari-Das nanofluid model with permeable wall conditions.The heat source parameter, varied between 0 and 6, highlights the relevance of energy generation in nanofluid applications. The nonlinearity parameter characterizes the intensity of irregular variations in both the governing flow equations and boundary constraints. The remaining parameter values were chosen empirically for analytical purposes.

4. CONCLUSION

The entropy generation analysis for radiative flow through porous media with nonlinear stretching plate has been conducted. Constraint equations were formulated, nondimensionalized and analyzed through the described numerical procedures; results were expressed in graphs and tables. The probe demonstrates the reciprocal motion of fluid nanoparticles by magnetic force generating current. Both internal and exterior fields opposed the deceleration of fluid's velocity,

as shown in Figure 2.

Table 1 reveals that an increase in magnetization and porosity enhances the energy transfer rate with reduction of stickiness. Lorentz force lets pass more fluid particles and heat energy through big holes. Both the heat migration and nanoparticle diffusion rates grow with pore sizes. Magnetic strength reduces entropy production, as shown in Figure 2. Viscosity produces less heat when the mechanical speed is decreased. Figure 3 shows how solar radiation increases entropy by adding more heat to the system. These insights can be used in solar energy systems, reduce material heating, analyze signal loss in optic fiber, enhance CT and X-ray medical imaging, solar dryers and regulate thermal coolants in nuclear reactors.

REFERENCES

[1] S. Alao, S. O. Salawu, R. A. Oderinu, A. A. Oyewumi, A. A. Yahaya, A. T. Adeosun, A. D. Ohaegbue: *Mixed convective and viscous heating effect of electromagnetic Tiwari-Das nanofluid model with permeable wall conditions*. Partial Differential Equations in Applied Mathematics, vol. 14, 2025. DOI: 10.1016/j.padiff.2025.101181

[2] W. Jamshed, S. R. Mishra, P. K. Pattnaik, K. S. Nisar, S. Suriya Uma Devi, M. Prakash, F. Shahzad, M. Hussain, V. Vijayakumar: *Features of entropy optimization on viscous second grade nanofluid streamed with thermal radiation: A Tiwari and Das model*. Case Studies in Thermal Engineering, vol. 27, p. 101291, July, 2021. DOI: 10.1016/j.csite.2021.101291

[3] W. Jamshed, S. R. Mishra, P. K. Pattnaik, K. S. Nisar, S. Suriya Uma Devi, M. Prakash, F. Shahzad, M. Hussain, V. Vijayakumar: *Features of entropy optimization on viscous second grade nanofluid streamed with thermal radiation: A Tiwari and Das model*. Case Studies in Thermal Engineering, vol. 27, p. 101291, July, 2021. DOI: 10.1016/J.CSITE.2021.101291

[4] T. G. Mayerhöfer, S. Pahlow, J. Popp: *The Bouguer-Beer-Lambert Law: Shining light on the obscure*. In Chemphyschem: a European journal of chemical physics and physical chemistry, vol. 21, issue 18, pp. 2029–2046, 2020. NLM (Medline). DOI: 10.1002/cphc.202000464

[5] M. Sohail, E. Rafique, A. Singh, A. Tulu: *Engagement of modified heat and mass fluxes on thermally radiated boundary layer flow past over a stretched sheet via OHAM analysis*. Discover Applied Sciences, vol. 6, 240, 2024. DOI: 10.1007/s42452-024-05833-1

[6] F. Waseem, M. Sohail, N. Ilyas, E. M. Awwad, M. Sharaf, M. J. Khan, A. Tulu: *Entropy analysis of MHD hybrid nanoparticles with OHAM considering viscous dissipation and thermal radiation*. Scientific Reports, vol. 14, no. 1, pp. 1–18, 2024. DOI: 10.1038/s41598-023-50865-z

[7] M. Zafar, H. Sakidin, M. Sheremet, I. B. Dzulkarnain, A. Hussain, R. Nazar, J. A. Khan, M. Irfan, Z. Said, F. Afzal, A. Al-Yaari: *Recent development and future prospective of Tiwari and Das mathematical model in nanofluid flow for different geometries: A review*. Processes, vol. 11, no. 3, 2023. DOI: 10.3390/pr11030834

EQUATIONS FOR DESCRIBING OF FATIGUE CRACK PROPAGATION IN SELECTED HYDROGEN EXPOSED MATERIALS

Elisha Zaheer[1], ***János Lukács***[2]
[1]PhD Student, [2]Full Professor
[1,2]*Faculty of Mechanical Engineering and Informatics,*
Institute of Materials Science and Technology
[1]elisha.zaheer@student.uni-miskolc.hu,
[2]janos.lukacs@uni-miskolc.hu

Abstract

The role and importance of hydrogen is constantly increasing today, and accordingly, the study of the effect of hydrogen on structural materials is becoming wider and deeper. During the production, transport and storage of hydrogen, various structures are often subjected to cyclic stresses. As a result of these stresses, cracks appear, which can propagate and lead to failure. In this article, limit curves for fatigue crack propagation were presented in the presence of hydrogen using the example of a pressure vessel and a pipeline material (ASME SA-723 Grade 1 and API 5L X52, respectively). It will be demonstrated that different methodologies may be required for different kind of steel and their welded joints.

1. INTRODUCTION

There are many steel structural elements and structures in operation that are subjected to cyclic loading. When designing and/or inspecting such structures, in addition to static and/or dynamic loads, cyclic loads must also be taken into account, which requires reliable material parameters. These material parameters can be lower or upper limit curves, or design curves based on some logic, or their parameters. In the case of high cycle fatigue, lower limit curves can be used (e.g., [1, 2]), and in the case of fatigue crack propagation, upper limit curves can be applied (e.g., [3]). Both limit curves and design curves are based on experimental results on the one hand and statistical approaches on the other hand. Lower limit curves can be "Mean $-$ 2SD", upper limit curves can be "Mean $+$ 2SD" values, where SD means standard deviation, and they can also be based on some survival probability. In the case of design curves, the parameters of the curves carry about identical or different probability contents for each parameter. Welded joints deserve special attention in all cases due to microstructural changes and residual stresses caused by welding process.

The role of hydrogen has become more important nowadays, and its effects on various structural materials and their properties (e.g., embrittlement) have been discussed in numerous publications (e.g., [4, 5]). Pressure vessels and pipelines play an important role in the production, storage, and transportation of hydrogen to its points of use, the latter in the form of either existing natural gas transmission systems (blending, retrofitting) or newly constructed systems [6]. The base materials of high-pressure systems are often unalloyed or low-alloy steels, and they can operate for decades. During long periods of operation, microcracks may form in the structural elements and their welded joints, which can be “activated” in the presence of hydrogen and under cyclic stresses (e.g., internal pressure changes), adversely affecting the integrity of the structural elements.

In this article, limit curves for fatigue crack propagation were presented in the presence of hydrogen using the example of a pressure vessel and a pipeline material. It will be demonstrated that different methodologies may be required for different types of steel and their welded joints.

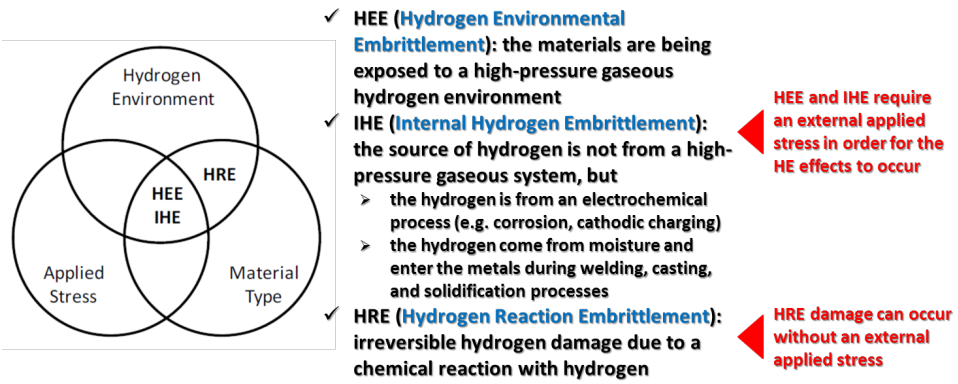


Figure 1: Three form of hydrogen embrittlement (based on [7])

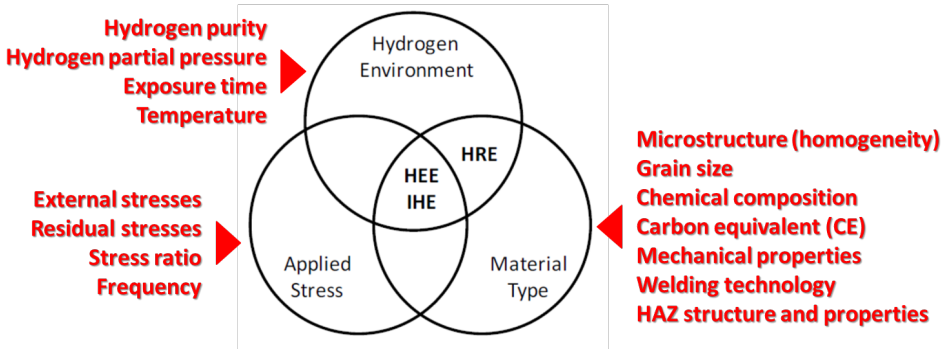


Figure 2: Influencing factors on hydrogen embrittlement (based on [7])

2. ABOUT HYDROGEN EMBRITTLEMENT

There are several forms of hydrogen embrittlement, which are influenced by different factors belonging to three groups (material type, applied stresses, and environment). The characteristic forms of embrittlement and the influencing factors are illustrated in Figs. 1 and 2, respectively [7].

American researchers studied several pressure vessel and pipeline steels in hydrogen and moist air environments [8]. A summary of the different characteristics of environmentally influenced fatigue crack growth at near-threshold and intermediate crack growth rates can be seen schematically in Fig. 3. In the near threshold region, under cyclic loading with low stress ratio (R), the threshold stress intensity factor range (ΔK_0) in hydrogen is found to be smaller than that in air; however at higher stress ratios, the fatigue crack growth curves are identical in hydrogen and air. For intermediate stress intensity factor range values, there is a critical maximum stress intensity factor ($K^T_{\max}$) characterizing the onset of the hydrogen acceleration effect on fatigue crack growth rate. It should be noted that the article was published in 1982 and was based on the material qualities used at that time.

Looking at the characteristic summary shown in Fig. 3, it can be concluded that no single (closed-form) equation can describe the propagation of fatigue cracks in hydrogen-containing environments, especially not in the whole range of the stress intensity factor range (ΔK). In the range close to the threshold value (ΔK_0), a different type of equation can be applied than in the later phases of crack growth. It is also likely that including the stress ratio (R) and/or frequency (ν in Fig. 3) into the equations further complicates their form, thereby reducing their range of application. The same applies to welded joints.

The Paris-Erdogan [9] law (1) is the oldest and most commonly used to describe fatigue crack propagation, or more precisely, the linear range of the "AIR" curve shown in Fig. 3,

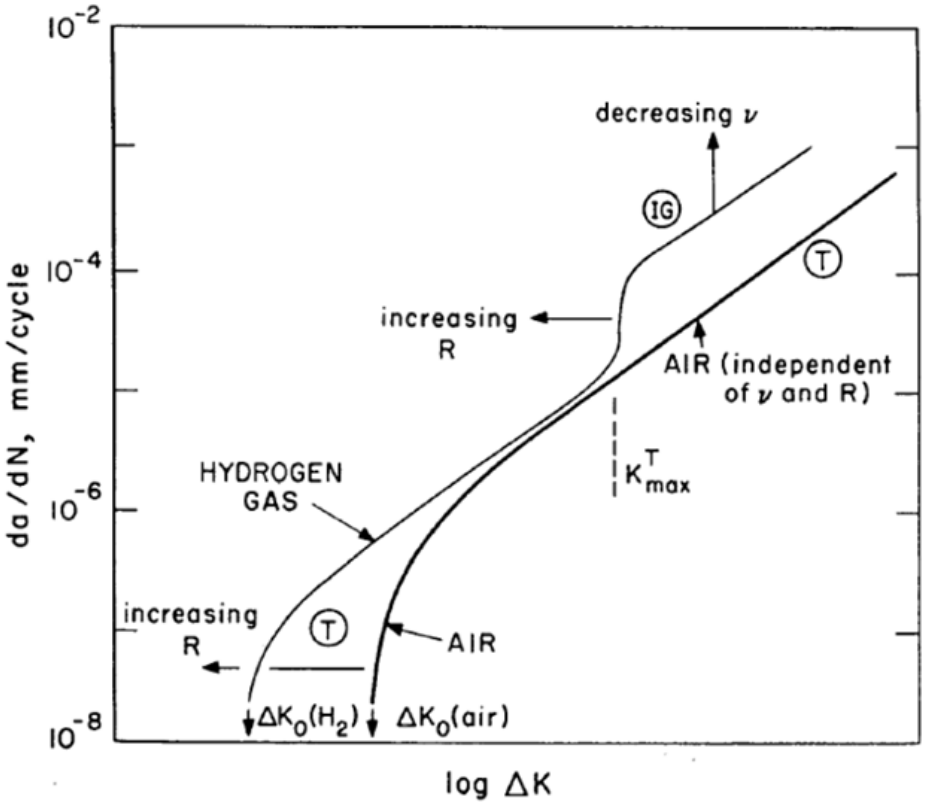


Figure 3: Schematic diagram of the effect of dry gaseous hydrogen on fatigue crack growth in lower strength steels (R = stress ratio, ν = frequency, T = mainly transgranular fracture, IG = mainly intergranular fracture) [8]

where da/dN is the fatigue crack propagation rate, ΔK is the stress intensity factor range, and C and n are material constants.

$$da/dN = C \cdot (\Delta K)^n \tag{1}$$

Based on this law, several additional equations have been developed, and the bilinear model (equations (2a) and (2b)) can also be found in standard [3].

$$da/dN = C_1 \cdot (\Delta K)^{n_1} \tag{2a}$$

$$da/dN = C_2 \cdot (\Delta K)^{n_2} \tag{2b}$$

Standard [10], referring to publications [11] and [12], contains upper limit curves for carbon steels, for design pressures below 20 MPa (200 bar) and stress ratios $R < 0.5$. The minimum specified yield strength (MSYS) of the steel shall not exceed 550 MPa, and the maximum tensile strength of the steel and welded joint shall not exceed 760 MPa. The equation for the upper limit curve is given by (3), where $a_1 = 4.0812 \cdot 10^{-9}, b_1 = 3.2016, a_2 = 4.0862 \cdot 10^{-11}, b_2 = 6.4822, a_3 = 4.8810 \cdot 10^{-8}, b_3 = 3.6147$.

$$da/dN = a_1 \cdot (\Delta K)^{b_1} + [(a_2 \cdot (\Delta K)^{b_2})^{-1} + (a_3 \cdot (\Delta K)^{b_3})^{-1}]^{-1} \tag{3}$$

Equations (4a) and (4b) are very similar to the bilinear equations (2a) and (2b). These two fatigue crack growth relationships represent a pressure-dependent regime at low ΔK referred to as (da/dN) lower, and a pressure-independent regime at high ΔK called (da/dN) higher. The

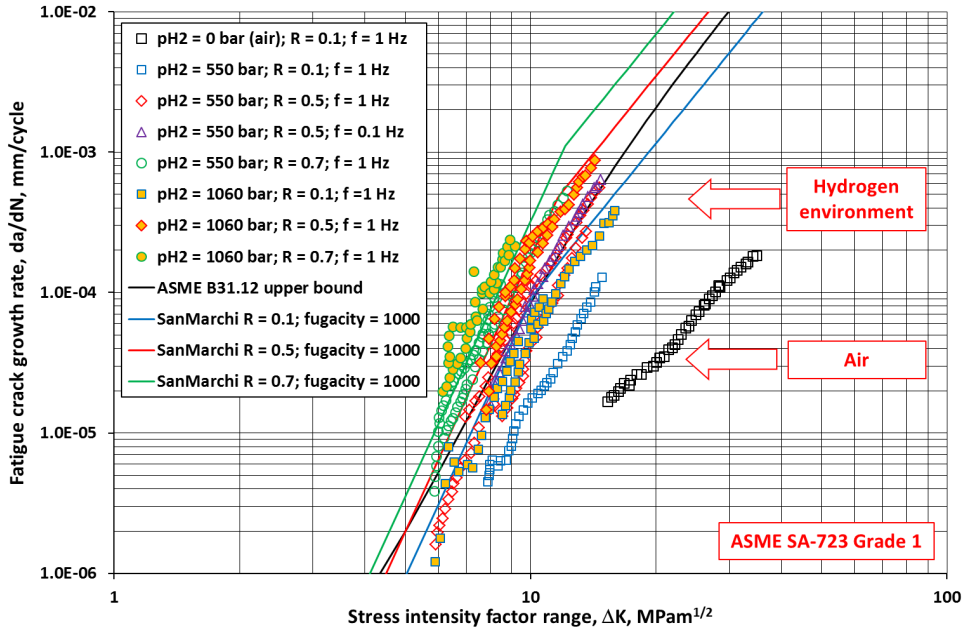


Figure 4: Fatigue crack growth rate test results for ASME SA-723 Grade 1 steel (based on [15])

transition between the two regimes is defined as the value ΔK at which da/dN is equivalent for both relationships [13, 14].

$$(da/dN)_{\text{lower}} = c_1 \cdot ((1 + c_2 \cdot R)/(1 - R)) \cdot (\Delta K)^{d_1} \cdot (\text{fugacity}/c3)^{d_2} \tag{4a}$$

$$(da/dN)_{\text{upper}} = c_4 \cdot ((1 + c_5 \cdot R)/(1 - R)) \cdot (\Delta K)^{d_3} \tag{4b}$$

In the equations (4a) and (4b) the constant values are as follows: $c_1 = 3.5 \cdot 10^{-11}, c_2 = 0.4268, d_1 = 6.5, c_3 = 2110, d_2 = 0.5, c_4 = 1.5 \cdot 10^{-8}, c_5 = 2, d_3 = 3.66$.

3. APPLICATION EXAMPLES

The pressure vessel material is ASME SA-723 Grade 1 steel; the results of fatigue crack growth rate tests performed in air and various hydrogen media are shown in Fig. 4 [15]. The figure also includes the upper limit curves determined by equations (3), (4a) and (4b), where equation (3) is not valid for all stress ratios.

The pipeline material is API 5L X52 steel; the results of fatigue crack growth rate tests performed in air and various hydrogen media on vintage and modern materials are shown in Figs. (5) and (6), respectively. The figures also contain data for welded joints made using different welding processes, and the upper limit curves determined by the relationships (3), (4a) and (4b) are also shown.

4. SUMMARY

The fatigue crack propagation behaviour of the materials examined differs significantly in air and hydrogen media; various equations (models) are suitable for describing this behaviour. The more parameters of the hydrogen environment are taken into account, the more complex the equations (models) required describing fatigue crack propagation become. Upper limit curves determined for one group of materials are not necessarily applicable to another group of materials, even if the validity ranges allow it.

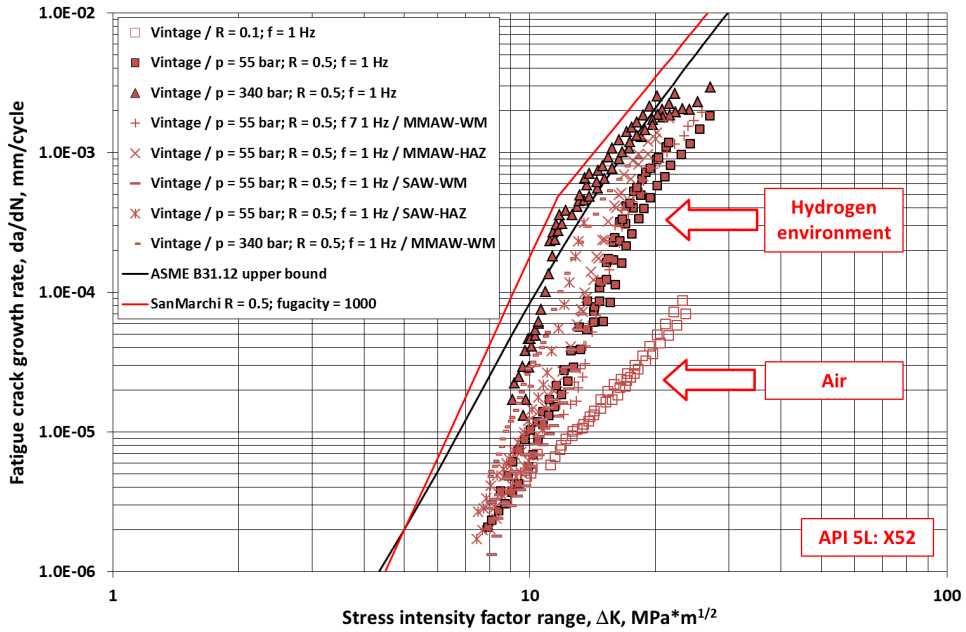


Figure 5: Fatigue crack growth rate test results for API 5L X52 vintage steel (based on [16])

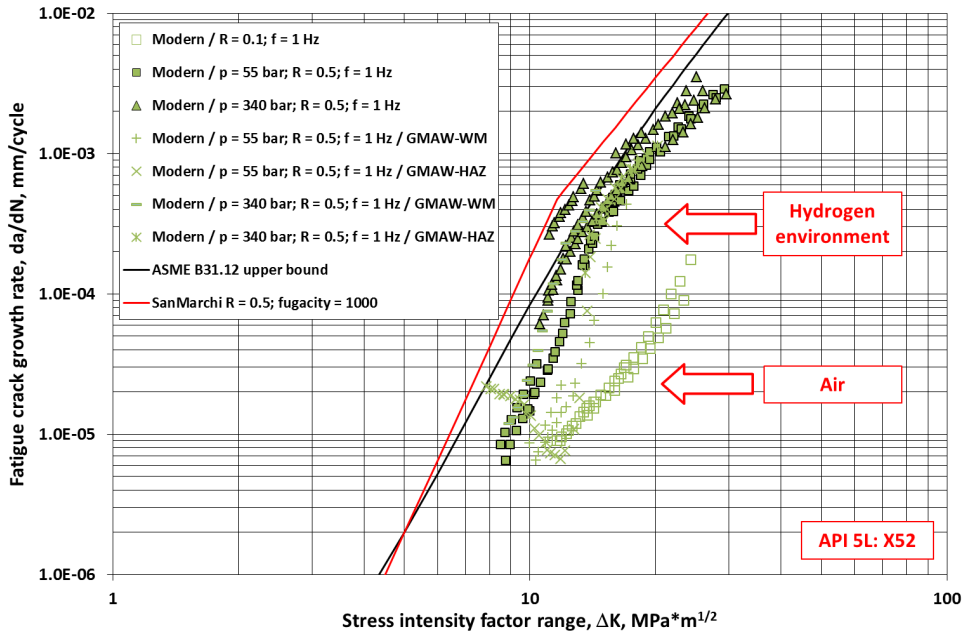


Figure 6: Fatigue crack growth rate test results for API 5L X52 modern steel (based on [16])

REFERENCES

[1] EN 1993-1-9 / AC, 2005 / 2009: Eurocode 3: *Design of steel structures*. Part 1–9. Fatigue.

[2] BS 7608 + A1, 2014 + 2015: *Guide to fatigue design and assessment of steel products*.

[3] BS 7910, 2019: *Guide to methods for assessing the acceptability of flaws in metallic structures*.

[4] C. San Marchi, B. P. Somerday: *Technical reference for hydrogen compatibility of materials*. Sandia National Laboratories, Albuquerque, NM, USA and Livermore, CA, USA Rep. SAND2012-7321, 2012.

[5] M. B. Djukic, et al.: *The synergistic action and interplay of hydrogen embrittlement mechanisms in steels and iron: Localized plasticity and decohesion*. Eng Fract Mech, 216:106528, 2019.

[6] *Global hydrogen review 2024*. Revised version, IEA, Paris, France. October 2024.

[7] J. A. Lee: *Hydrogen embrittlement*. NASA, Marshall Space Flight Center, Huntsville, Alabama, NASA/TM-2016–218602, April 2016.

[8] S. Suresh, R. O. Ritchie: *Mechanistic dissimilarities between environmentally influenced fatigue-crack propagation at near-threshold and higher growth rates in lower strength steels*. Metal Sci, vol. 16, no. 11, pp. 529–538, 1982.

[9] P. Paris, F. Erdogan: *A critical analysis of crack propagation laws*. J Basic Eng, Trans ASME, vol. 85, pp. 528–534, 1963.

[10] ASME B31.12, 2023: *Hydrogen piping and pipelines*.

[11] A. J. Slifka, et al.: *Fatigue measurement of pipeline steels for the application of transporting gaseous hydrogen*. ASME J Pressure Vessel Technology, vol. 140, no. 1, p. 01140, February 2018.

[12] R. L. Amaro, et al.: *Development of a model for hydrogen-assisted fatigue crack growth of pipeline steel*. ASME J Pressure Vessel Technology, vol. 140, no. 2, p. 021403, April 2018.

[13] J. Ronevich, et al.: *Consistency of fatigue crack growth behavior of pipeline and low-alloy pressure vessel steels in gaseous hydrogen*. Int J Hydrogen Energy, vol. 136, pp. 686–694, 2025.

[14] L. Paterlini, et al.: *Mechanical testing methods for assessing hydrogen embrittlement in pipeline steels: A review*. Metals, vol. 15, p. 1123, 2025.

[15] C. Cui, et al.: *Computational predictions of hydrogen-assisted fatigue crack growth. Int J Hydrogen Energy*, vol. 72, pp. 315–325, 2024.

[16] Kovács, J., Lukács, J.: *A fáradásos repedésterjedés leírása hidrogén tartalmú közeget szállító acél csőtávvezetékekben*. In: Lukács, J., Pap, J. (Eds.) Kutatási eredmények a Miskolci Egyetem Gépészmérnöki és Informatikai Karának Anyagszerkezettani és Anyagtechnológiai Intézetében – 2025, Miskolc, Miskolci Egyetem, Gépészmérnöki és Informatikai Kar Anyagszerkezettani és Anyagtechnológiai Intézet, pp. 243–257, 2025.