

Statistical Analysis of Numerical Topological Parameters of Information Systems for Machine Learning Models*

Andriy Makarchuk^{1,*†}, Oleg Barabash^{1,†}, Oleh Kopiika^{2,†}, Volodymyr Tkachov^{3,†} and Bohdan Bilyavskiy^{3,†}

¹ National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Beresteyskyi Prosp. 37, Kyiv, 03056, Ukraine

² Institute of Applied Control Systems of the National Academy of Sciences of Ukraine, Akademika Glushkova Ave. 40, Kyiv, 03187, Ukraine

³ The National Defence University of Ukraine, Air Force Avenue, 28, Kyiv, 03049, Ukraine

Abstract

This paper addresses the problem of selecting numerical input parameters of information system topology for machine learning models used in assessing functional stability. A statistical analysis was conducted on a dataset comprising over 30,000 instances described by eight key topological parameters. Using simulation modeling and correlation matrix analysis, pairwise relationships between the metrics were identified, enabling the detection and elimination of multicollinear variables. The results of the study make it possible to optimize the training process of classifiers by reducing the dimensionality of the feature space without loss of informativeness, thereby improving the accuracy of predicting the functional stability state of distributed systems.

Keywords

Functional stability, network topology, machine learning, multicollinearity, statistical analysis, simulation modeling, information systems, feature selection.

1. Introduction

The use of machine learning for the numerical assessment of functional stability has recently attracted considerable research attention [1-3]. In complex and distributed computing architectures, ensuring uninterrupted operational reliability requires robust quantitative metrics and continuous monitoring paradigms. In particular, a number of studies have conducted comprehensive experiments with various machine learning models, including neural networks [2, 4], decision trees [1], random forests [1], and other ensemble methods [4], to evaluate functional stability indicators and predict potential failures under dynamic load conditions [5]. These predictive frameworks aim to automate decision-making processes, thereby optimizing resource allocation and enhancing the overall resilience of critical infrastructures.

However, despite the evident success of predictive analytics, many of these studies face the fundamental problem of selecting numerical topological parameters of the information system under

2nd Workshop "Intelligent information technologies" UkrProg-IIT'2026 co-located with 16th International Scientific and Practical Programming Conference UkrPROG'2026, May 5-6, 2026, Kyiv, Ukraine

* Corresponding author.

† These authors contributed equally.

✉ andreymakarchuk2@gmail.com (A. Makarchuk); bar64@ukr.net (O. Barabash); okopiika@gmail.com (O. Kopiika); vova.vsk@gmail.com (V. Tkachov); bilyavskiy1977@gmail.com (B. Bilyavskiy)

ORCID 0000-0002-6422-7488 (A. Makarchuk); 0000-0003-1715-0761 (O. Barabash); 0000-0003-0189-3915 (O. Kopiika); 0000-0001-8957-2723 (V. Tkachov); 0009-0006-9036-7229 (B. Bilyavskiy)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

consideration that are free from multicollinearity and high pairwise correlation. The presence of redundant, highly correlated input features often leads to model overfitting, inflates computational complexity during training, and degrades the interpretability and stability of classification results. To mitigate these issues, feature space optimization and dimensional reduction become essential prerequisites for constructing efficient machine learning pipelines.

In this work, the authors employ simulation modeling to identify the main patterns governing the selection of certain normalized numerical characteristics of graphs that describe the topology of an information system. By systematically evaluating a large-scale dataset of generated network graphs, this study provides a formal statistical rationale for identifying non-redundant topological descriptors. The obtained results allow researchers and system architects to eliminate collinear parameters prior to model training, thereby improving classifier precision while minimizing the computational overhead associated with high-dimensional feature processing in distributed information systems.

2. Simulation modeling and analysis of its results

Within the simulation modeling framework, a total of 33,860 graphs representing potential topologies of information systems were generated, and for each of them, eight numerical characteristics were computed, namely:

$$\begin{aligned} x_1 &= \frac{d_{\min}(G)}{n-1}, \quad x_2 = \frac{d_{\text{mean}}(G)}{n-1}, \quad x_3 = \frac{d_{\max}(G)}{n-1}, \\ x_4 &= \min_{i,j=1,2,\dots,n} |\text{fft}(A(G), i, j)|, \quad x_5 = \frac{1}{n^2} |\text{fft}(A(G), i, j)|, \\ x_6 &= \max_{i,j=1,2,\dots,n} |\text{fft}(A(G), i, j)|, \quad x_7 = \frac{\lambda(G)}{n-1}, \quad x_8 = \frac{\kappa(G)}{n-1}, \end{aligned} \quad (1)$$

where G is a graph, what describes a topology of considered information system, n is a number of vertices of this graph, $d_{\min}(G)$, $d_{\text{mean}}(G)$ and $d_{\max}(G)$ are min, average and max degrees of vertices of graph G , $A(G)$ is an adjacency matrix of this graph, $\text{fft}(\bullet, i, j)$ is a coefficient of 2D DFT of matrix, $\lambda(G)$ and $\kappa(G)$ are degrees of vertex and edge connectivity of graph G , respectively.

The results of the described calculations are further analyzed using statistical data analysis methods.

2.1. Detection and drop of anomalies in generated dataset

To begin with, let us assess the presence of anomalies in the obtained dataset. This can be accomplished using Fig. 1.

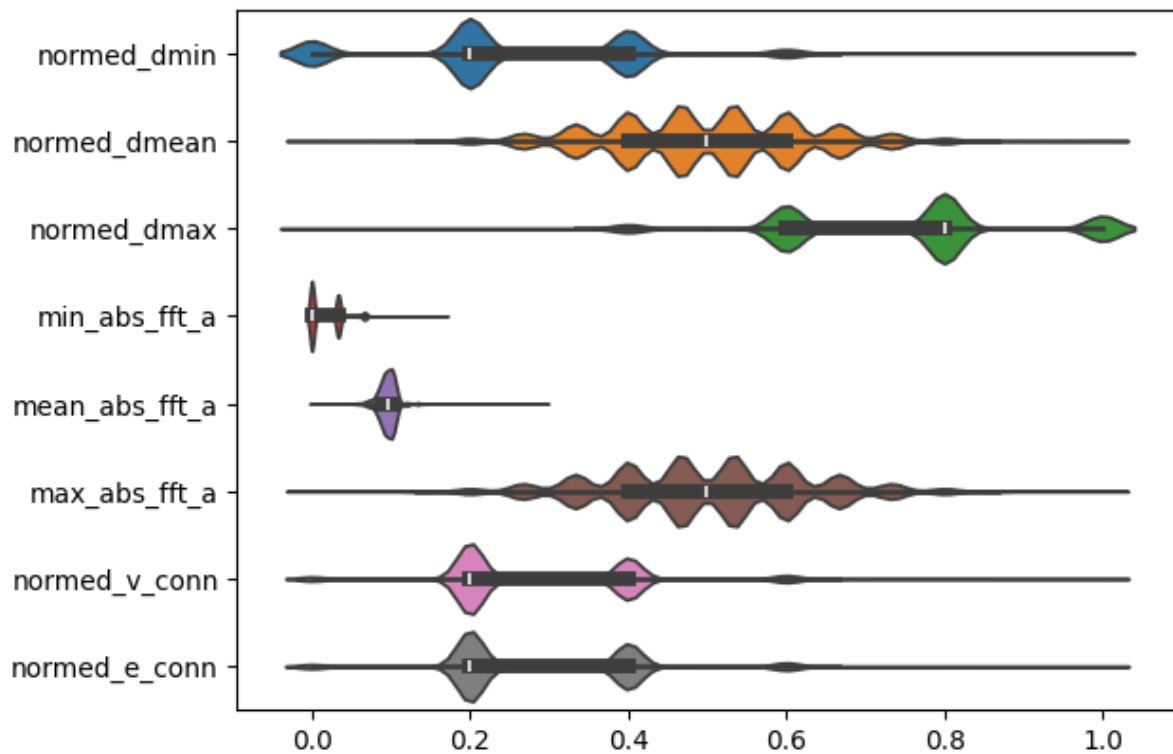


Figure 1: Violinplots for parameters (1) in the dataset

As shown in Fig. 1, the presence of anomalies is clearly observed across all parameters, which can be further confirmed using other types of visualization, such as box plots.

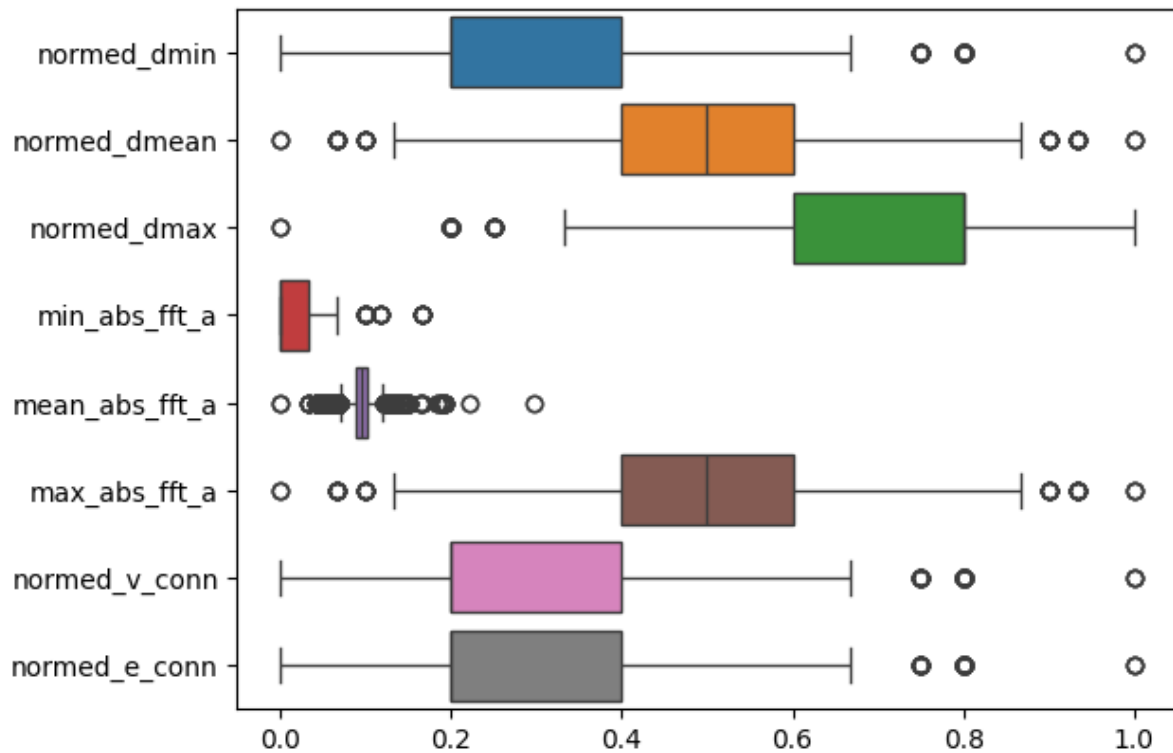


Figure 2. Boxplots for parameters (1) in the dataset

If quantile-based analysis methods are applied, the anomalies observed in Figs. 1–2 can be removed, thereby enabling correlation analysis to be performed without potential distortions.

2.2. Correlation analysis of parameters (2)

For conducting the correlation analysis, it is sufficient to compute three measures:

- the correlation matrix;
- the maximum F-statistic for subsets of parameters of a given size;
- the variance inflation factor (VIF) for each parameter, with all others treated as predictors.

To begin with, a visualization of the correlation matrix for the obtained dataset is constructed.

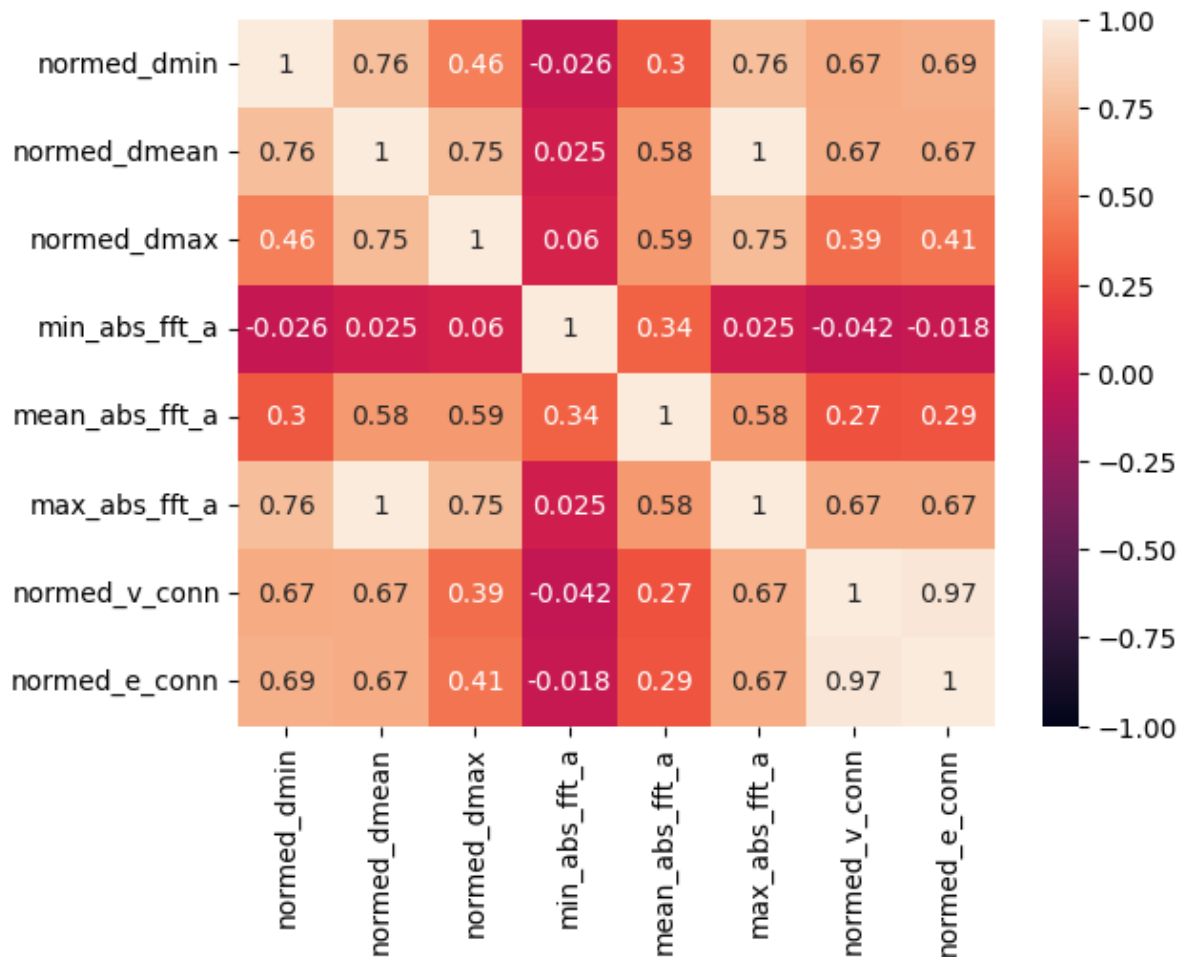


Figure 3: Heatmap of correlation matrix

Fig. 3 clearly shows two pairs of well-correlated parameters, namely:

- x_2 and x_6 ;

- x_7 and x_8 .

In the context of the problem under consideration, a high correlation between these pairs of parameters indicates that only one of them can be used as input to a machine learning model without a significant loss in accuracy. However, to ensure this, additional computations are performed.

Specifically, the maximum F-statistic for subsets of parameters of a given cardinality is evaluated, along with the variance inflation factor (VIF) for each parameter, with all remaining variables treated as predictors.

Table 1

Max f-statistics for given parameters number from (1)

Parameters number	Max f-statistics	Parameters set
2	725830,058	x_3, x_4
3	654906,552	x_3, x_4, x_5
4	394904,2	x_3, x_4, x_5, x_6
5	282348,697	$x_2, x_3, x_4, x_5,$ x_6
6	212332,602	$x_2, x_3, x_4, x_5,$ x_6, x_7
7	168962,907	$x_2, x_3, x_4, x_5,$ x_6, x_7, x_8
8	131740,112	All parameters

Table 2

VIF for parameters with other parameters as predictors

Parameter	VIF
x_2	∞
x_6	∞
<i>const</i>	144,618
x_8	19,363
x_7	19,023
x_1	3,053
x_3	2,679
x_5	2,07
x_4	1,212

On Fig. 3 and tables 1-2 we may see, that there is most probable, that parameters x_3 and x_4 gives most “unique” information in (1) despite high f-statistics for this pair of parameters. So, for enough accuracy of machine learning models parameters x_3 and x_4 must be given as an input. Opposite situation we may see for parameters x_6 and x_8 : these parameters strongly depend on other

parameters in (1) (especially, with x_2 and x_7 , respectively), what specifically good may be seen on fig. 3 and table 2. It means, that if we give as input of model parameters x_2 and x_7 , parameters x_6 and x_8 should not be given as input and vice versa.

As the result, we see, that from parameters (1) we definitely must give as input parameters x_3 and x_4 , but with parameters x_6 and x_8 we need to be as careful as possible and without a need sometimes don't give it as input of machine learning model.

Now, to be sure, that parameters x_6 and x_8 may be potentially ignored, let see on the figure 4 and the figure 5.

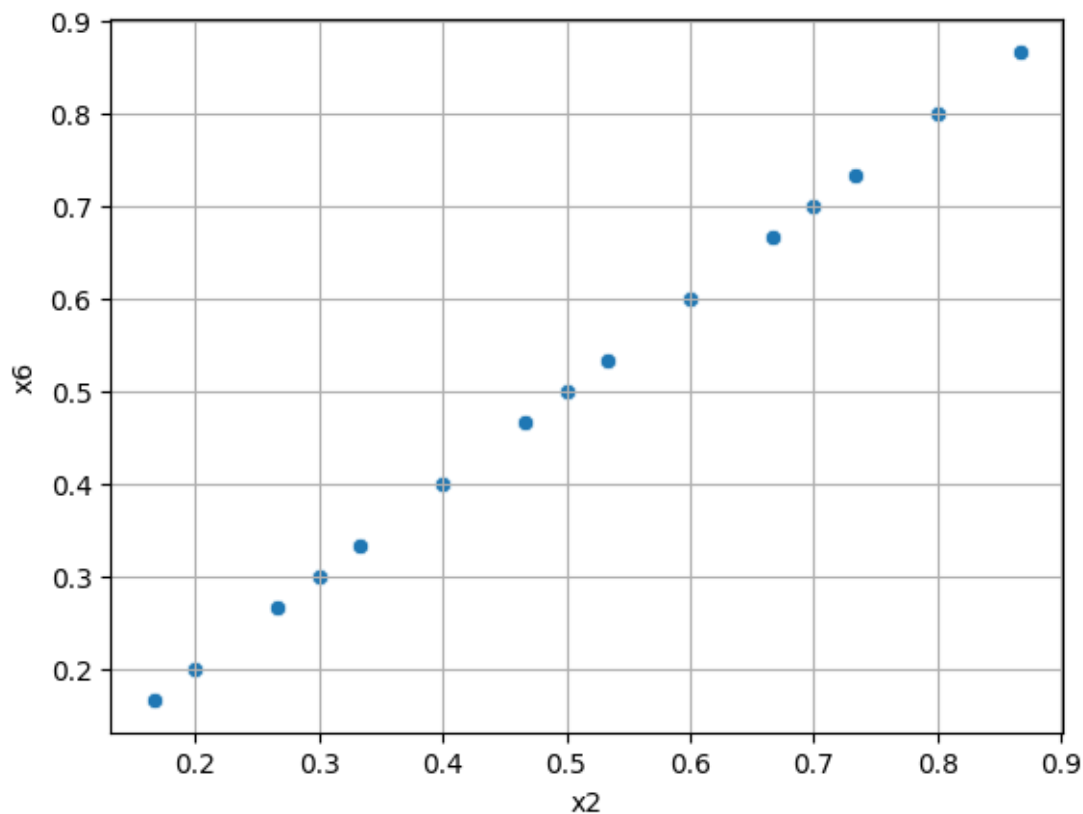


Figure 4: Scatter for parameters x_2 and x_6

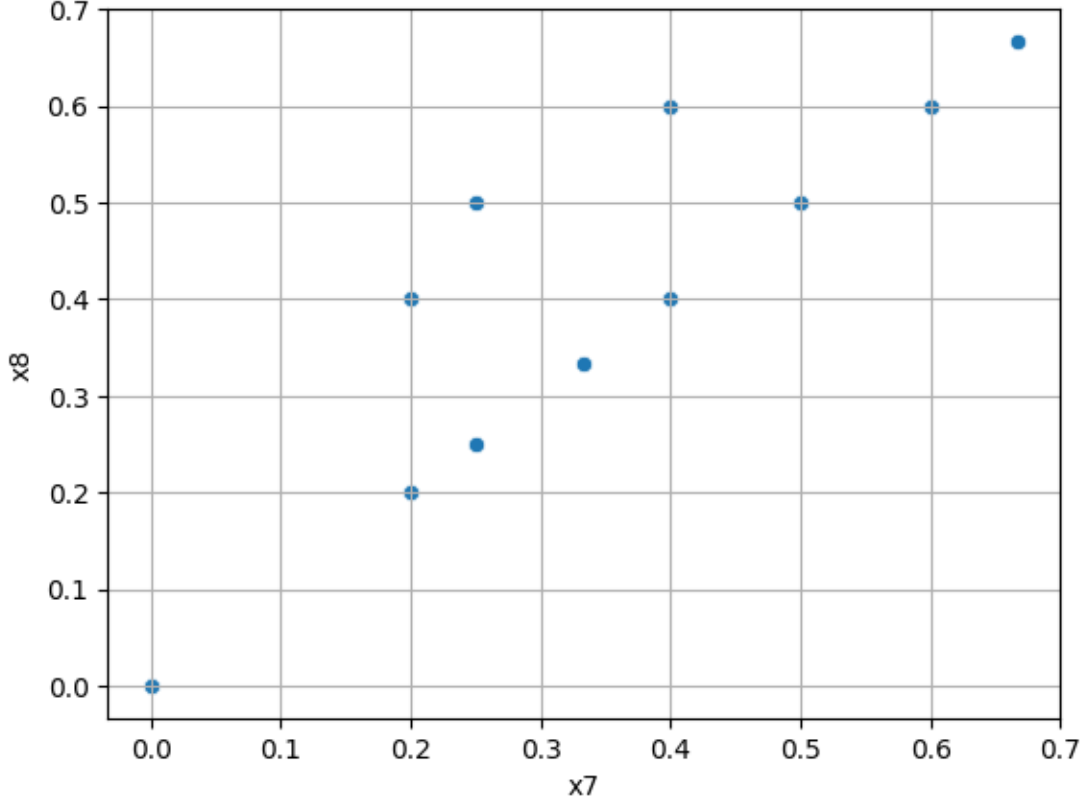


Figure 5: Scatter for parameters x_7 and x_8

As clearly observed in Fig. 4, the joint distribution between the normalized average vertex degree x_2 and the maximum 2D Discrete Fourier Transform coefficient x_6 exhibits a rigorous, near-deterministic linear trajectory. This strict functional relationship demonstrates that the spectral maximum of the adjacency matrix scales directly with the structural density of graph edges. Consequently, retaining both parameters simultaneously introduces duplicate Variance Inflation Factor (VIF) overhead without providing any additional decision-making variance to machine learning models.

Similarly, the scatter plot in Fig. 5 confirms an almost perfect monotonic and discrete alignment between normalized vertex connectivity x_7 and normalized edge connectivity x_8 . This behavior naturally aligns with the fundamental graph theory principle where edge connectivity is strictly bounded below by vertex connectivity. The visual clustering along the primary linear trend underscores that either variable independently captures the structural resilience threshold of the information system graph.

Therefore, these empirical visualizations provide solid justification for pruning x_6 and x_8 from the input feature space, enabling the construction of lightweight, non-redundant classifiers while fully preserving topological information integrity.

3. Conclusions

The study addresses the problem of optimizing the training process of machine learning models for assessing the functional stability of information systems. The main idea of the research is the justified selection of input topological parameters through the identification and elimination of

multicollinearity, which makes it possible to improve the accuracy of system state classification while reducing the computational cost of model training.

During the simulation modeling and statistical analysis of a dataset comprising over 30,000 instances, a correlation matrix was computed and the level of pairwise relationships between parameters was determined. This made it possible to establish that, out of all parameters (1), it is reasonable to provide no more than six as input features to machine learning models, namely:

1. x_1 ,
2. x_2 or x_6 ,
3. x_3 ,
4. x_4 ,
5. x_5 ,
6. x_7 or x_8 .

The obtained results may be applied in various domains, particularly those that rely heavily on information and other distributed systems, including sensor networks [5], data centers [6-8], cybersecurity [9, 10], virtual and cloud systems [11, 12], as well as numerous other application areas [13-18]. In practical deployment scenarios, eliminating multicollinear topological variables allows for the construction of lightweight, computationally efficient classifiers capable of operating in real-time edge and fog environments. Furthermore, reducing the dimensionality of the input feature space enhances the generalization ability of machine learning architectures, preventing model overfitting and ensuring stable performance under non-stationary network loads. Ultimately, the proposed feature selection protocol serves as a foundational step toward building automated diagnostic and control frameworks that continuously monitor infrastructure health, optimize resource utilization, and guarantee high functional stability across dynamic and scalable computing environments.

Acknowledgements

The authors express their gratitude to the departments and institutions represented in this work for providing the necessary computational infrastructure and research environment that enabled the simulation and statistical analysis.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration on Generative AI

The authors declare that this computer-scientific work was prepared without the use of generative Artificial Intelligence tools, adhering strictly to academic integrity standards.

References

- [1] O. Barabash, A. Makarchuk, A. Musienko and O. Maistrov, Prediction of fulfilling the conditions of functional reliability of information systems using machine learning models, In: Proceedings of the 2024 IEEE 7th International Conference on Actual Problems of Unmanned Aerial Vehicles Development (APUAVD), 2024, pp. 179–182. doi:10.1109/APUAVD64488.2024.10765875.
- [2] A. Makarchuk, O. Barabash, O. Kopiika, V. Tkachov and B. Bilyavskiy, The algorithm and software for selection of optimal topology of information systems with limited budget, CEUR Workshop Proceedings 4049 (2025) 101–106. URL: <https://ceur-ws.org/Vol-4049/paper9.pdf>.
- [3] S. Yevseiev, R. Hryshchuk, O. Zakovorotnyi, O. Milov, H. Kuchuk and S. Milevskyi, Intelligent control and security systems models synthesis methodology for critical infrastructure objects, In: Proceedings of the 2024 IEEE 5th International Conference on Advanced Trends in Information Theory (ATIT), 2024, pp. 275–281. doi:10.1109/ATIT64324.2024.11222460.
- [4] O. Barabash, A. Makarchuk, P. Openko and S. Korotin, Application of SVM, FFNNs, k-NN and their ensembles for identifying functionally reliable systems, Axioms 14 (2025) 237. doi:10.3390/axioms14040237.
- [5] O. Barabash, V. Sobchuk, A. Sobchuk, A. Musienko and O. Laptiev, Topological aspects of designing functionally robust wireless sensor networks, Advanced Information Systems 9 (2025) 28–38. doi:10.20998/2522-9052.2025.4.05.
- [6] O. Kopiika, S. Dovgyi and A. Yaremenko, Technological principles for building a network architecture of service data processing centers, Lecture Notes in Networks and Systems 1338 (2025) 111–126. doi:10.1007/978-3-031-89296-7_7.
- [7] L. Raskin, V. Karpenko, Yu. Ivanchykhin and D. Sokolov, Diagnosis of systems under conditions of small initial data sampling, Advanced Information Systems 7 (2023) 39–43. doi:10.20998/2522-9052.2023.3.05.
- [8] V. I. Skurikhin, D. P. Kucherov and L. S. Zhitetskij, Integration of information technologies in decision making training in terminal control problems, Upravlyayushchie Sistemy i Mashiny 5 (2003) 30–42. Kyiv: Institute of Cybernetics.
- [9] S. Yevseiev, S. Milevskyi, V. Sokol, Ye. Melenti, O. Akhiezer and K. Petrenko, Methodological foundations for constructing intelligent information security systems, In: Proceedings of the 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA), 2025, pp. 1–6. doi:10.1109/ICHORA65333.2025.11017207.
- [10] S. Milevskyi, S. Yevseiev, I. Aksonova and S. Dunaiev, Development of a method for evaluating data transmission in cyber-physical systems, Springer Tracts in Additive Manufacturing (2025) 475–487. doi:10.1007/978-3-031-84873-5_37.
- [11] I. Zamrii, V. Vyshnivskyi, V. Sobchuk and V. Vasylenko, The method for ensuring the functional stability of virtual cloud resources through the assessment of their condition and the selection of countermeasures, In: Proceedings of the 2024 IEEE 5th International Conference on Advanced Trends in Information Theory (ATIT), 2024, pp. 173–178. doi:10.1109/ATIT64324.2024.11222495.
- [12] I. Zamrii, V. Vyshnivskyi and V. Sobchuk, Method of ensuring the functional stability of the information system based on detection of intrusions and reconfiguration of virtual networks, CEUR Workshop Proceedings 3654 (2024) 252–264. URL: <https://ceurspt.wikidata.dbis.rwth-aachen.de/Vol-3654/paper21.pdf>
- [13] V. Sobchuk, O. Barabash, A. Musienko, O. Laptiev, V. Kozlovskiy and Yu. Shcheblanin, Evaluation of efficiency of application of functionally sustainable generalized information system of the enterprise, In: Proceedings of the 2022 4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2022, pp. 1–6. doi:10.1109/HORA55278.2022.9799892.
- [14] Yu. Kalinin, I. Koliesnik, H. Kuchuk, N. Kuchuk, S. Polyashenko and M. Medvid, Principles of diagnostics of complex systems using structural automaton algebra and taking into account

- signal time coordination, In: Proceedings of the 2025 IEEE 6th KhPI Week on Advanced Technology (KhPIWeek), 2025, pp. 1–6. doi:10.1109/KhPIWeek61436.2025.11288711.
- [15] Zh. Hu, V. Mukhin, Ya. Kornaga, Ya. Lavrenko, O. Barabash and O. Herasymenko, Analytical Assessment of Security Level of Distributed and Scalable Computer Systems, *International Journal of Intelligent Systems and Applications* 8 (2016) 57–64. doi:10.5815/ijisa.2016.12.07.
- [16] V. Mukhin, Ya. Kornaga, V. Bondarenko, V. Zavgorodnii, O. Herasymenko and O. Sholokhov, Mathematical model for heterogeneous databases parameters estimation in distributed systems with dynamic structure, In: Proceedings of the 2020 IEEE 2nd International Conference on Advanced Trends in Information Theory (ATIT), 2020, pp. 158–161. doi:10.1109/ATIT50783.2020.9349331.
- [17] S. Krepych, I. Spivak, S. Spivak and R. Krepych, The method of assessing the reliability of software systems based on a graphic model of the dependence of methods of the system under test, *Advanced Information Systems* 9 (2025) 58–67. doi:10.20998/2522-9052.2025.2.08.
- [18] D. Kucherov, X. Fang, Zh. Wang, Zh. Xu, X. Fu, Identification of parameters of the mathematical model of a dynamic object of critical infrastructure, *CEUR Workshop Proceedings* 4068 (2024) 135–146. doi: 10.1088/3050-2454/ae9293/pdf.