UDC: 004.7

INTELLIGENT METHODS FOR DATA ANALYSIS IN INFORMATION AND COMMUNICATION SYSTEMS MONITORING PROCESSES

Andrii Senyk[1]*, Volodymyr Kotsun[2], Bohdan Penyukh[3], Bohdan Tsybulyak[4]

[1]Department of Telecommunication, Lviv Polytechnic National University,
12 Stepan Bandera Str., Lviv, 79013, Ukraine
[2]Department of Computer Science and Software Engineering,
Private Higher Education Institution "European University",
16V Acad. Vernadsky Blvd., Kyiv, 03115, Ukraine
[3]Department of Physical and Biomedical Electronics,
[4]Department of System Design,
Ivan Franko National University of Lviv,
50 Drahomanova str., Lviv, 79005, Ukraine

Senyk, A., Kotsun, V., Penyukh, B., Tsybulyak. B. (2025). Intelligent Methods for Data Analysis in Information and Communication Systems Monitoring Processes. *Electronics and Information Technologies*, 31, 53–60. https://doi.org/10.30970/eli.31.5

ABSTRACT

Background. In modern monitoring of information and communication systems (ICS), a key challenge remains the timely detection of anomalies while maintaining a low false positive rate. Classical machine learning or deep learning methods often show a trade-off between high precision and the ability to detect most anomalies, limiting their efficiency in dynamic network environments.

Materials and Methods. This study proposes the Hybrid Adaptive Monitoring Method with Multi-level Anomaly Validation (HAM-MAV), which combines a deep autoencoder for anomaly detection (unsupervised) with a Random Forest classifier (supervised) and an adaptive threshold mechanism. In the first stage, the autoencoder identifies suspicious samples based on reconstruction error. These samples are then refined by the Random Forest, reducing false positives. The threshold is updated dynamically according to the statistics of the latest observation window. The experiments used the NSL-KDD (Network Security Laboratory – Knowledge Discovery in Databases) dataset with preprocessing steps including normalization, one-hot encoding, and feature selection based on correlation criteria.

Results and Discussion. Experimental results show that HAM-MAV achieves Precision of 96.92%, Recall of 62.67%, F1-score of 76.12%, and ROC-AUC (Receiver Operating Characteristic – Area Under Curve) of 0.8003, outperforming Autoencoder, Random Forest, and Isolation Forest in most metrics. The method reduces false positives while improving anomaly detection capability, maintaining a fast processing time. HAM-MAV's key advantage is its balanced performance between precision and recall, which is critical for continuous ICS monitoring.

Conclusion. HAM-MAV provides an optimal combination of precision, recall, and execution speed, outperforming traditional methods in real-time conditions. Its architecture allows effective operation in environments with changing traffic characteristics, making it a promising approach for cybersecurity applications, particularly in automated intrusion detection systems.

Keywords: anomaly detection, deep learning, random forest, adaptive threshold, intrusion detection, NSL-KDD.

INTRODUCTION

Monitoring of information and communication systems (ICS) is a crucial tool for ensuring their reliability, security, and operational efficiency. Given the rapid growth in the volume of data transmitted and processed today, as well as the complexity of ICS architectures, continuous monitoring of transactions and timely detection of any deviations are of paramount importance. Monitoring has become a key component for maintaining service quality, managing risks, and protecting against cyberattacks [1-4].

This topic is also important due to the deep integration of ICS systems into all spheres of life, including business, government, healthcare, and energy. Any disruptions in their operation can lead to significant economic losses and even provoke a crisis. Therefore, effective monitoring methods are crucial for the stable operation of critical infrastructure and maintaining information security. Traditional methods of monitoring data analysis, based on fixed thresholds and simple processing algorithms, have several shortcomings. In modern conditions, the application of intelligent data analysis methods, in particular machine learning and artificial intelligence algorithms, is becoming increasingly important due to their ability to detect complex dependencies and patterns in large volumes of information. Similar hybrid approaches are actively applied in other areas of digital analytics. In particular, [5] presents an example of an effective combination of deep learning methods and classical classification for detecting synthetic content in social media. This confirms the universality of such architectures for anomaly detection tasks and highlights the relevance of the HAM-MAV approach proposed in this study for monitoring information and communication systems.

In the study, HAM-MAV was developed, which combines the capabilities of deep neural networks and classical machine learning algorithms. In particular, at the first stage, an autoencoder was used to detect potentially anomalous samples, which operates in unsupervised mode and estimates the degree of deviation of input data from normal behavior using the reconstruction error. At the second stage, the selected samples are additionally analyzed using the Random Forest model, which allows for significantly reducing the number of false positives. To increase the flexibility of operation, the system is equipped with an adaptive mechanism for changing threshold values, which is updated depending on the statistical characteristics of the last observation window. A comparative analysis was carried out with three well-known methods (Autoencoder, Random Forest, and Isolation Forest) by key model evaluation metrics (Precision, Recall, F1-score, ROC-AUC), as well as by execution time. The experimental results showed the superiority of HAM-MAV in most indicators, which confirms the feasibility of its use in real-time monitoring systems of information and communication systems .

MATERIALS AND METHODS

Traditional methods for analyzing monitoring data typically rely on thresholds and simple statistical measures such as the mean or variance. While these methods work well in stable environments, they have limited ability to detect complex anomalies or emerging threats. Intelligent data analysis methods, including machine learning (ML) and deep learning (DL), are increasingly being used in monitoring systems. Classic machine learning algorithms that have demonstrated high accuracy in anomaly detection include random forests, support vector machines (SVM), and k-nearest neighbors (k-NN). These algorithms have been successfully used to classify cyberattacks and analyze the behavior of system processes. Deep neural networks (DNNs), including long short-term memory networks (LSTMs) and autoencoders, have shown themselves well in time series processing and anomaly detection in network traffic.

The paper proposes a Hybrid Adaptive Monitoring Method with Multi-level Anomaly Validation (HAM-MAV). This method combines deep anomaly detection using autoencoders with online classification using Random Forest and uses a multi-stage validation mechanism to reduce the number of false positives. Its key feature is a multi-stage decision filter. First, an autoencoder (without a teacher) is used to detect anomalies based on the deviation between the reconstructed and true data [6-9].

Potential anomalies are then fed to a Random Forest, pre-trained on labeled data from the same dataset and trained with a teacher, to improve classification performance. In addition, an adaptive threshold is used that changes depending on the statistics of the last observation window (e.g., traffic flow in one minute). This reduces the sensitivity of the system to “noisy” anomalies. Method architecture:

1. Data pre-processing
   - Normalization of features.
   - Unimportant or highly correlated features are removed.
   - Categorical values are converted to numerical values (single coding).
2. Autoencoder for anomaly detection
   - The model learns to reconstruct normal (unsupervised) samples.
   - The mean squared error (MSE) is calculated for each sample.
   - Samples with an error exceeding the adaptive threshold are classified as "suspicious".
3. Random Forest for clarification
   - Receives suspicious samples as input.
   - Operates in a controlled mode, using labeled data from a dataset.
   - Returns the final class: "abnormal" or "normal".
4. Adaptive threshold
   - Dynamically calculates the moving average and standard deviation of the reconstruction error.
   - The threshold is defined as $\mathrm{mean\_error} + \mathrm{k} * \mathrm{std\_error}$, where k is configurable.

The proposed method provides a lower false positive rate due to two-stage validation (autoencoder + Random Forest). Adapts to environmental changes: the threshold is automatically updated based on the latest data. Works with both labeled and unlabeled data: the first stage is unsupervised learning, and the second stage is supervised learning.

Research overview

The section details the development, implementation, and testing of HAM-MAV. This approach combines the autoencoder for anomaly detection with a Random Forest algorithm for classification optimization using an adaptive thresholding mechanism. The study was divided into three main phases: data preparation, model construction and tuning, and experimental testing with results compared to existing methods.

Data usage

To test this approach, the open NSL-KDD dataset (Tavallaee et al., 2009) was used, freely available through the UCI Machine Learning Repository. This dataset contains 125,973 training data samples and 22,544 test data samples, representing typical network traffic and various attack types (DoS, Probe, R2L, and U2R). Each record is described by 41 attributes, including both numeric attributes (e.g., number of bytes in a packet) and categorical attributes (e.g., protocol and service type). NSL-KDD is a classical but not the most recent dataset. Its use in this study is justified by the fact that NSL-KDD is one of the most widely adopted benchmark datasets for comparing the performance of intrusion detection systems. This enables the results of HAM-MAV to be compared with a large number of previous studies, ensuring the objectivity and validity of the conclusions. It should also be noted that, for the initial stage of evaluating a hybrid approach, it is important to work with a dataset that has a balanced class structure and clearly labeled attack types (DoS, Probe, R2L, U2R). Nevertheless, in future research, the HAM-MAV approach is planned to be tested on more recent and large-scale datasets, such as CICIDS2017, UNSW-NB15, and TON_IoT, which will allow the method to be evaluated in environments with a broader spectrum of modern attacks.

Pre-processing includes:

- Normalization of numerical features to the range [0, 1] using the min-max method.
- One-Hot Encoding for categorical variables.
- Removal of insignificant features with a correlation coefficient with the target variable less than 0.05 (Pearson's correlation coefficient for numerical variables and chi-square test for categorical variables).

After data preparation, the autoencoder is trained using only normal samples from the training set [7-11].

Architecture of the HAM-MAV

Level 1 – autoencoder

The autoencoder is designed with three hidden layers, with 64, 32, and 64 neurons, respectively. The hidden layers use the ReLU activation function, and the output layer uses the Sigmoid function. The optimization is performed using the Adam algorithm with a learning rate of 0.001 and a MSE loss function. The training is per-formed for 50 epochs with a batch size of 256. The reconstruction error is calculated as the mean square error between the input value and the reconstructed feature value. The threshold for detecting suspicious samples is calculated adaptively:

$$threshold = \mu_{err} + k \cdot \sigma_{err}, \tag{1}$$

where μ_{err} represents the average reconstruction error within a sliding window of 500 samples, σ_{err} is the standard deviation, and k is an empirically chosen parameter ($k = 1.5$ in the experiments).

Level 2 – Random Forest

The Random Forest algorithm was used to improve the classification of suspicious samples. The number of trees is 200, the maximum depth is 20, and the splitting criterion is Gini impurity. The class_weight='balanced' parameter is used to balance the classes. The training data contains all classes.

Adaptive threshold mechanism

During the test run, the autoencoder threshold was updated every 500 new examples to reduce the impact of time-varying data distribution (conceptual bias).

For each method, Precision, Recall, F1-score, and ROC-AUC were calculated. In addition, the False Alarm Rate (FAR) was analyzed for 1000 test examples.

RESULTS AND DISCUSSION

Statistical analysis

All experiments were performed in Google Colab using TensorFlow 2.12, Scikit-learn 1.3, Pandas 2.0, and NumPy 1.25. Statistical analysis was performed using the SciPy 1.11 Python package. For each performance measure, the mean and standard deviation were calculated based on 5-fold cross-validation (stratified). Paired t-tests were used to test the statistical significance of differences between methods ($p < 0.05$ was considered statistically significant). The experimental results for different models are presented in **Table 1**.

Precision

The Precision metric shows what proportion of predicted anomalies are actually anomalies. Higher precision means fewer false positives.

Conclusion on the effectiveness of HAM-MAV: Precision of HAM-MAV is very high and comparable to that of Autoencoder. That is, the model rarely mistaken, marking normal data as anomalous. In terms of accuracy, HAM-MAV is practically at the level of the best models, and it can be considered effective in avoiding false positives.

Table 1. **Experiment results**

Model	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC	Execution Time (s)
Autoencoder	97.40	8.58	15.77	0.5414	24.94
Random Forest	96.66	61.19	74.94	0.7921	1.88
Isolation Forest	94.55	31.20	46.92	0.6442	0.41
HAM-MAV	96.92	62.67	76.12	0.8003	1.52

Recall

Metric explanation: Recall shows what proportion of real anomalies were detected. A higher value means fewer anomalies were missed.

HAM-MAV has much higher completeness compared to Autoencoder and Isolation Forest, and slightly better than Random Forest. This means that the method is able to find most real anomalies. Overall conclusion: HAM-MAV has a good balance between accuracy and ability to find anomalies, which makes it reliable for detection.

F1-score

The F1-score combines Precision and Recall into a single value, showing the balance between false positives and missed anomalies. Higher F1 means better balance. The F1 of HAM-MAV is the highest among the models, which means that the proposed method provides the best balance between precision and completeness. In terms of comprehensive anomaly detection efficiency, HAM-MAV appears to be the best model.

ROC-AUC

ROC-AUC shows the ability of the model to distinguish between normal and abnormal data at any threshold. A higher score indicates better classification. HAM-MAV has the highest ROC-AUC among all models, confirming its ability to separate anomalies from normal data. Therefore, HAM-MAV demonstrates the best overall classification ability.

Execution Time

The execution time of HAM-MAV is very low and comparable to Random Forest, significantly faster than Autoencoder, although slightly slower than Isolation Forest. Thus, HAM-MAV provides high performance with fast processing, making it practical for real-world applications.

Therefore, according to the conducted experiment, it can be concluded that HAM-MAV demonstrates the best balance between accuracy, completeness, and F1-score, has a high ability to distinguish anomalies (ROC-AUC), and works fast. This makes the proposed method effective and competitive compared to classical approaches.

CONCLUSION

In the field of information and communication system monitoring, anomaly detection is a critical task, as unsuccessful attacks or excessive false positives can have serious consequences. Current methods often suffer from poor accuracy and the ability to detect most real threats. Autoencoders achieve good results in reducing false positives, but they often miss a significant portion of attacks, while traditional methods (such as Random Forest or Isolation Forest) can have low accuracy or be sensitive to noise.

The paper proposes HAM-MAV, a hybrid method that combines an autoencoder for initial anomaly detection and a Random Forest for improved classification. The use of an adaptive threshold that varies depending on the statistics of the last observation window reduces the number of false positives and improves the system's robustness to data changes.

Experiments on the NSL-KDD dataset showed that HAM-MAV exhibits Precision of 96.92% (at the level of the best competitor), Recall of 62.67% (630% higher than Autoencoder and 101% higher than Isolation Forest), F1-score of 76.12% (382% higher than Autoencoder and 62% higher than Isolation Forest), and ROC-AUC of 0.8003 (47% higher than Autoencoder and 24% higher than Isolation Forest). Thus, the proposed method provides a better balance between precision and completeness than other considered approaches. In terms of execution time, HAM-MAV (1.52 s) significantly outperforms Autoencoder (~17 times faster) and runs faster than Random Forest (19%), although slightly slower than Isolation Forest. This makes the method suitable for processing large amounts of data in near real time, which is important for operational monitoring systems of ICS.

The main advantages of HAM-MAV are the ability to reduce the level of false positives, adaptability to traffic changes, and versatility in working with both labeled and unlabeled data. Thanks to multi-level validation of solutions, the method is suitable for critical systems where a balance is needed between threat sensitivity and stability of operation without overloading operators with false alarms. Future extensions are possible, such as the inclusion of deep learning models in the classification, the use of Bayesian methods to refine adaptive thresholds, and their application to large data streams. This will increase the detection efficiency and reduce the cost of computing resources, while maintaining high-quality threat detection in dynamic ICS.

ACKNOWLEDGMENTS AND FUNDING SOURCES

The authors received no financial support for the research, writing, and publication of this article.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any.

AUTHOR CONTRIBUTIONS

Conceptualization, [A.S., V.K.]; methodology, [A.S., B.P.]; validation, [A.S., B.T.]; formal analysis, [A.S., V.K.].; investigation, [A.S., B.P.]; resources, [A.S., B.T.]; data curation, [A.S., V.K.]; writing – original draft preparation, [A.S.]; writing – review and editing, [A.S., B.P.]; visualization, [A.S., B.T.] supervision, [A.S.]; project administration, [A.S.]; funding acquisition, [A.S.].

All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] Senyk, A., Klymash, M., Pyrih, Y., Tsybulyak, B., Penyukh, B., & Shuvar, R. (2024). Improving the cybersecurity monitoring algorithms efficiency using neural networks. In 2024 IEEE 17th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET) (pp. 428–431). IEEE. https://doi.org/10.1109/TCSET64720.2024.10755520.

[2] Klymash, M., Senyk, A., & Pyrih, Y. (2024). Investigation of a context-sensitive cyber security monitoring algorithm based on recurrent neural networks. ICTEE, 4(1), 1–9. https://doi.org/10.23939/ictee2024.01.001.

[3] Feng, S., Yang, Z., Huang, M., & Wu, Y. (2021). Big data analysis of intellectual property service agencies. In 2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI) (pp. 326–330). IEEE. https://doi.org/10.1109/PRAI53619.2021.9551058.

[4] Cai, J., Xie, L., Yao, S., & Gao, Y. (2025). Algorithm application and optimization in intellectual property management. In 2025 International Conference on Digital Analysis and Processing, Intelligent Computation (DAPIC) (pp. 458–463). IEEE. https://doi.org/10.1109/DAPIC66097.2025.00091.

[5] Y, A. F., Sundaram, A., & Ruby Helen, F. (2025). Analyzing social media data misuse and intellectual property rights: A dual legal-empirical analysis approach in the digital landscape. In 2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI) (pp. 1803–1810). IEEE. https://doi.org/10.1109/ICMCSI64620.2025.10883512.

[6] Wu, B., Zheng, S., & Han, M. (2024). Innovation efficiency and influencing factors of high-tech industries: An analysis from the intellectual property perspective. In 2024 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 1480–1484). IEEE. https://doi.org/10.1109/IEEM62345.2024.10857198.

[7] Xiang, D., & Wu, Y. (2022). Analysis and research of internet user behaviors under the context of big data. In 2022 International Conference on Big Data, Information and Computer Network (BDICN) (pp. 243–247). IEEE. https://doi.org/10.1109/BDICN55575.2022.00054.

[8] Qiu, B., Liu, D., Cao, S., Mu, C., Yan, S., & Liu, Y. (2024). Risk analysis and protection suggestions for artificial intelligence data security. In 2024 IEEE 9th International Conference on Data Science in Cyberspace (DSC) (pp. 392–398). IEEE. https://doi.org/10.1109/DSC63484.2024.00059.

[9] Wang, Y., Sun, J., Lu, X., Chen, C., & Yang, F. (2024). Research on data privacy calculation and data traceability technology for power monitoring system. In 2024 Asia-Pacific Conference on Software Engineering, Social Network Analysis and Intelligent Computing (SSAIC) (pp. 867–871). IEEE. https://doi.org/10.1109/SSAIC61213.2024.00175.

[10] Zhang, L., Li, Y., Qiu, B., Zhang, J., & Liang, W. (2021). Design of communication power centralized remote monitoring system based on big data technology. In 2021 International Conference on Electronics, Circuits and Information Engineering (ECIE) (pp. 46–49). IEEE. https://doi.org/10.1109/ECIE52353.2021.00017.

[11] Lai, S., Pan, Z., Ren, Q., Wang, P., Zhao, J., & Chen, H. (2024, October). IoT-Based Site Safety Monitoring and Early Warning System. In 2024 3rd International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI) (pp. 870-874). IEEE. https://doi.org/10.1109/ICDACAI65086.2024.00164.

ДОСЛІДЖЕННЯ ІНТЕЛЕКТУАЛЬНИХ МЕТОДІВ АНАЛІЗУ ДАНИХ У ПРОЦЕСАХ МОНІТОРИНГУ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ СИСТЕМ

***Андрій Сеник*[1]*, *Володимир Коцун*[2], *Богдан Пенюх*[3], *Богдан Цибуляк*[4]**

[1] Національний університет «Львівська політехніка»,
вул. Степана Бандери, 12, Львів, 79013, Україна
[2] Приватний вищий навчальний заклад «Європейський університет»,
бульв. Академіка Вернадського, 16в, Київ, 03115, Україна
[3] Кафедра фізичної та біомедичної електроніки,
[4] Кафедра системного проектування,
Львівський національний університет імені Івана Франка,
вул. Драгоманова 50, Львів, 79005, Україна

АНОТАЦІЯ

Вступ. У сучасних системах моніторингу інформаційно-комунікаційних систем ключовим завданням є своєчасне виявлення аномалій при збереженні низького рівня хибнопозитивних спрацювань. Традиційні методи машинного та глибинного навчання часто виявляють компроміс між високою точністю та здатністю виявляти більшість відхилень, що знижує їхню ефективність у динамічних мережевих умовах.

Матеріали та методи. Запропоновано гібридний адаптивний метод моніторингу з багаторівневою перевіркою аномалій (HAM-MAV), який поєднує автокодер для первинного виявлення відхилень із класифікатором на основі алгоритму випадкового лісу для їх уточнення та адаптивним механізмом коригування порогу. Автокодер визначає підозрілі зразки на основі похибки реконструкції, після чого їх аналізує випадковий ліс, що зменшує кількість помилкових спрацювань. Поріг оновлюється автоматично залежно від статистичних характеристик останнього вікна спостережень. Для перевірки використано відкритий набір даних виявлення знань у базах даних, до якого застосовано попередню обробку (нормалізацію, кодування категоріальних ознак та відбір ознак за кореляційними критеріями).

Результати. Отримані результати демонструють, що запропонований метод забезпечує високу точність (96,92%), прийнятну повноту (62,67%), збалансований F1-показник (76,12%) та значення ROC-AUC 0,8003. У порівнянні з відомими методами (автокодер, випадковий ліс, ізоляційний ліс) запропонований підхід зменшує кількість хибнопозитивних сигналів і водночас покращує здатність виявляти реальні відхилення при збереженні швидкості обробки даних (1.52 с).

Висновки. Метод HAM-MAV демонструє ефективне поєднання точності, повноти та швидкодії, перевершуючи класичні підходи в умовах реального часу. Його архітектура дозволяє адаптуватися до змін у мережевому трафіку, що робить цей метод перспективним для застосування у сфері кібербезпеки, зокрема в автоматизованих системах виявлення вторгнень. Водночас проведені експерименти підкреслюють баланс, досягнутий запропонованим методом між чутливістю до виявлення та стабільністю роботи. Цей баланс особливо важливий для критично важливих інфраструктур, де як пропущені загрози, так і надмірні хибні спрацювання можуть призвести до серйозних наслідків.

Ключові слова: виявлення аномалій, глибоке навчання, випадковий ліс, адаптивний поріг, виявлення вторгнень, відкритий набір даних

Received / Одержано	Revised / Доопрацьовано	Accepted / Прийнято	Published / Опубліковано
31 August, 2025	22 September, 2025	26 September, 2025	31 October, 2025