Intrusion Detection System Traffic Classification Based on Machine Learning with Correlation-Based Filtering and a Genetic Algorithm-Inspired Feature Selection Method for IoT Networks

Alaa A. Almelibari
Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia
aamelibari@uqu.edu.sa (corresponding author)

Received: 18 July 2025 | Revised: 29 July 2025 and 3 August 2025 | Accepted: 15 August 2025

 DOI: https://doi.org/10.48084/etasr.13511

ABSTRACT

Securing Internet of Things (IoT) networks against diverse cyber-attacks remains a critical challenge due to their constrained resources and complex traffic patterns. This paper proposes a lightweight, multiclass Intrusion Detection System (IDS) that addresses the limitations of prior binary models by classifying five types of network traffic: Normal, DoS, Mirai, Man-in-the-Middle (MITM), and Scan. A key contribution of this work is the application of a Genetic Algorithm (GA)-inspired feature selection method, which significantly enhances model accuracy and efficiency by isolating the most relevant attributes. Combined with traditional machine learning models, the proposed approach was evaluated using a simulated dataset modeled after IoTID20. Among the classifiers, the Random Forest model, when integrated with GA-inspired feature selection, achieved the highest accuracy of 96.5%. The results highlight the effectiveness of combining lightweight feature optimization with robust classification techniques, making the system highly suitable for real-world IoT deployments.

Keywords-IDS; traffic classification; IoT; DoS; cyber security; network security

I. INTRODUCTION

The proliferation of the IoT has embedded smart technologies into the fabric of daily life, spanning domains from home automation and healthcare to industrial control and smart cities. This expansion of connectivity, however, has concurrently unveiled new security weaknesses, positioning IoT ecosystems as lucrative targets for cyber criminals. In these settings, traditional security protocols often prove inadequate, largely due to the resource limitations of IoT devices and the diverse array of communication standards in use. An IDS is an indispensable tool in any network security plan, engineered to identify and flag unauthorized or malicious activities. The application of machine learning has propelled the development of intelligent IDS that can autonomously recognize attack signatures within network traffic data. A significant portion of past research, however, concentrated on binary classifiers, which are restricted to differentiating normal network traffic from a single category of attack, such as a DoS event. This narrow focus fails to capture the intricate threat landscape of real-world IoT networks, which often face multiple simultaneous or sequential attacks. Authors in [1] tackled the issue of DoS attacks in high-dimensional data, a prevalent challenge for modern IDS. They introduce a novel hybrid IDS framework that improves detection performance by optimizing the dataset using a stacked feature selection method, which integrates RFS, Relief, and PCE to reduce dimensionality and select relevant features. After this optimization, stacked learning classifiers are used to identify DoS attacks. Tested on the CICDDoS-2019 dataset, the model outperformed traditional classifiers, achieving 96.5%. Authors in [2] focused on cybersecurity threats in connected intelligent vehicles, specifically vulnerabilities in the CAN bus system. They presented a hybrid approach combining LSTM neural networks with BFO to improve the detection of fuzzy and DoS attacks. BFO aids in adaptive feature selection to pinpoint relevant CAN bus parameters, while LSTM captures temporal anomalies in message sequences. Tested on the Car-Hacking dataset, the proposed LSTM-BFO model outperformed traditional methods, achieving a detection time of 0.0838 s with a precision of 94.6%,

Authors in [3] addressed the growing cybersecurity issues related to the increasing use of IoT technologies, which are susceptible to various complex attacks such as DoS, DDoS, probing, malware, and port scans. To improve real-time threat

detection, the authors introduce a hybrid deep learning model that combines CNN with Bi-LSTM networks. This CNN-BiLSTM model is thoroughly evaluated on three benchmark datasets—KDDCup99, NSL-KDD, and CIC_IDS_2017—achieving impressive accuracy rates of 99.9%, 99.8%, and 98.0%, respectively. Authors in [4] focused on addressing cybersecurity vulnerabilities in modern vehicles, specifically targeting the unsecured CAN that lacks encryption and authentication. The authors propose an advanced intrusion detection framework utilizing a TAN to detect attacks through three phases: feature extraction, feature classification, and final detection via adversarial learning. The model was optimized with the Greylag Goose Optimization algorithm for enhanced performance. Tested on a real-world CAN traffic dataset with injection attacks, this approach outperformed traditional machine learning models, achieving 96.3% accuracy. Authors in [5] addressed emerging security threats in the IoT, focusing on DoS attacks and botnets. They proposed a hybrid deep learning model that combines CNN for spatial feature extraction and GRUs for capturing temporal dependencies in network traffic. This model analyzes both static and dynamic patterns and is enhanced by the SUCMO algorithm for hyperparameter tuning. Experimental results on the UNSW-NB15 and BoT-IoT datasets showed that the proposed CNN-GRU-SUCMO model significantly outperformed others. Authors in [6] proposed an innovative technique to address class imbalance in intrusion detection datasets by integrating RBB with an LSTM ensemble. Their model achieved notable performance, with 91.04% accuracy.

Several recent studies have explored advanced methods for improving IDS in IoT and network environments. An IoT-focused IDS framework that integrates optimization algorithms to enhance detection accuracy was proposed in [7]. The study utilized swarm intelligence techniques to fine-tune model parameters and feature selection, achieving improved threat identification with reduced computational cost in constrained IoT environments. Authors in [8] developed an ensemble-based detection architecture combining a feature-augmented CNN with a deep autoencoder. Authors in [9] introduced a novel SHAP-based intrusion detection model employing QNN on IonQ hardware. Authors in [10] proposed a lightweight heterogeneous ensemble learning framework for network intrusion detection. The model integrates multiple base learners with different characteristics in a cascading structure, ensuring robustness and adaptability across varied attack scenarios. Authors in [11] proposed an intrusion detection model tailored for a Digital Twin-enabled IIoT environment, aiming to enhance industrial sustainability through secure and intelligent monitoring frameworks.

Recent research has increasingly explored hybrid approaches that combine feature selection or optimization techniques with machine learning and deep learning to enhance intrusion detection in IoT-related domains. For example, in [12] an integrated Genetic Algorithm (GA) and deep learning approach was proposed for effective cyber-attack detection and classification in Industrial IoT, demonstrating improved accuracy in complex industrial settings. Similarly, in [13] a GA-LR wrapper method was applied for efficient feature selection in network intrusion detection, yielding interpretable and computationally efficient models. In [14], a rule-based IDS was developed and evaluated using both the UNSW-NB15 dataset and real-time traffic, providing a robust framework for hybrid threat detection. Authors in [15] employed a combination of Particle Swarm Optimization (PSO) and deep learning to design a lightweight and accurate IDS for the Internet of Medical Things (IoMT), addressing the resource constraints and security challenges in healthcare environments. In the context of phishing detection, a recent study [16] demonstrated how BPSO can be integrated with deep learning models to significantly improve classification performance while minimizing false positives. These works collectively underscore the effectiveness of integrating optimization algorithms with intelligent models to enhance IDS performance across diverse IoT ecosystems.

This paper advances upon a prior binary IDS framework by introducing a multiclass classification model. The system is designed to identify a wider spectrum of cyber threats, including DoS, Mirai, MITM, and various scanning activities, in addition to benign traffic. By doing so, this work seeks to elevate the practical utility and realism of machine learning-driven IDS solutions for contemporary IoT infrastructures.

II. METHODOLOGY

To enhance anomaly detection capabilities in IoT networks, this study proposes a multiclass IDS founded on supervised machine learning. The work extends a prior binary model to a more comprehensive system that can identify and distinguish between Normal traffic and multiple attack categories, including DoS, Mirai, MITM, and Scan were considered. An anomaly-based detection approach was selected for its capacity to identify both known and novel threats based on behavioral deviations. The workflow of the proposed system is depicted in Figure 1.

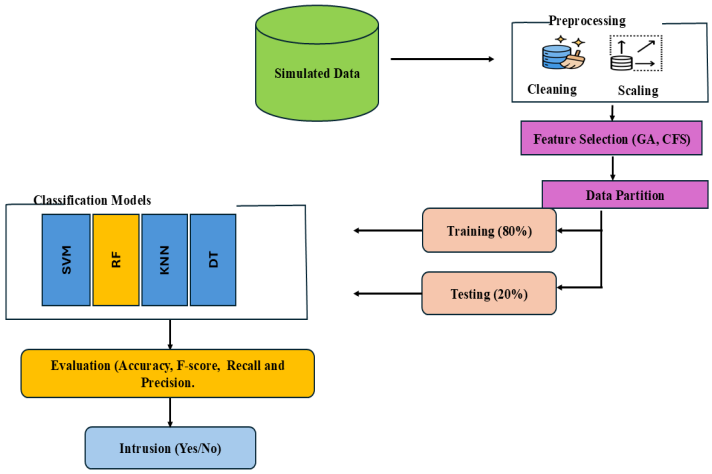

Fig. 1. Proposed methodology workflow.

A. Dataset and Simulation

Instead of using the IoTID20 dataset [17] in its original binary form, this research simulates a realistic multiclass dataset inspired by its structure and statistical distribution. The simulation process was performed using a combination of

Python-based traffic generators and statistical sampling methods. For each of the five traffic classes (Normal, DoS, Mirai, MITM, and Scan), traffic patterns were generated to reflect key attributes observed in IoTID20, including packet rates, protocol distributions (e.g. TCP/UDP/ICMP), flag combinations, and flow durations. Attack behaviors were modeled based on published threat signatures and existing metadata in the IoTID20 logs. For example, DoS traffic was emulated by generating high-frequency SYN packets, while MITM traffic included altered payloads with tampered checksums. Each class contained a comparable number of samples to prevent class imbalance and was validated by comparing feature distributions with those from the original dataset.

B. Data Preprocessing

To ensure the quality and consistency of the input data, a rigorous preprocessing stage was conducted. This involved removing any null values, applying feature scaling using StandardScaler to achieve a uniform distribution, and encoding string-based labels into integer values for classification. Noisy or irrelevant features, such as flow identifiers, were excluded to prevent model bias and improve generalizability. To ensure a fair evaluation of the classification models, the dataset was partitioned using a stratified 80/20 train-test split. Stratification was applied to maintain class distribution consistency across both training and testing sets, which is crucial for multiclass classification tasks. The training set was used for model fitting and hyperparameter tuning via cross-validation, while the test set was reserved strictly for final performance evaluation.

C. Feature Selection

To reduce dimensionality and isolate the most informative attributes, two distinct methods were evaluated:

- Correlation-Based Feature Selection (CFS): This is a filter-based approach that evaluates feature subsets by identifying attributes that are highly correlated with the output class but are uncorrelated with each other. The goal is to eliminate redundant information that can degrade detection accuracy and increase computation time.
- Genetic Algorithm (GA)-Inspired Selection: This method utilizes an optimization technique derived from the principles of natural selection. While not a full GA implementation, this study employed simplified logic inspired by GA, using a scoring system based on ANOVA F-tests to rank and select the top-k features. This approach balances performance with computational efficiency and interpretability.

In the GA-inspired feature selection process, we employed an F-test (ANOVA) statistical method to evaluate each feature's discriminatory power across the multiclass labels. The F-score measures the ratio of inter-class variance to intra-class variance, helping to identify features that contribute most to class separability. After computing the F-scores for all features, they were ranked in descending order. The top 20 features were selected based on empirical testing, which showed that this subset provided the best balance between model performance and computational complexity. This method, while inspired by GAs, avoids the overhead of full evolutionary optimization and offers an interpretable, filter-based approach.

III. EXPERIMENTAL SETUP

All experiments were conducted on a personal workstation equipped with an Intel® Core™ i7-11800H CPU, 32 GB RAM, and an NVIDIA GeForce RTX 3060 GPU, running Windows 11 Pro (64-bit). The implementation was carried out using Python 3.9, with key libraries including Scikit-learn 1.3.0 for machine learning, Pandas and NumPy for data processing, and Matplotlib and Seaborn for visualization. Jupyter Notebook and Visual Studio Code were used as development environments. As the study utilized traditional machine learning models (Random Forest, SVM, KNN), the experiments were lightweight and did not require extensive GPU usage, making the setup suitable for typical research-grade or edge-level systems.

IV. RESULTS AND DISCUSSION

The results were obtained from the evaluation of four machine learning classifiers using the two feature selection techniques described earlier: Correlation-based Feature Selection (CFS) and Genetic Algorithm-inspired feature ranking. Each model was tested on the same simulated dataset representing multiclass IoT traffic with five distinct classes. When using the top 20 features selected by the GA-inspired method, the Random Forest model consistently outperformed all other classifiers across all metrics. Table I summarizes the performance of each model.

TABLE I. CLASSIFIER PERFORMANCE SUMMARY (USING GA-INSPIRED FEATURE SELECTION)

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	**0.965**	**0.966**	**0.963**	**0.964**
Decision Tree	0.943	0.944	0.941	0.942
SVM (RBF Kernel)	0.956	0.958	0.951	0.954
KNN (k=5)	0.928	0.930	0.926	0.927

Among the evaluated models, Random Forest consistently outperformed the others in terms of accuracy and F1-score. Its ensemble approach and robustness against overfitting make it particularly effective for handling the diverse and complex distributions typical of IoT traffic. Support Vector Machine (SVM) also demonstrated strong performance, especially in terms of macro-averaged precision and F1-score. Its effectiveness in high-dimensional spaces and its resilience to class imbalance contribute to its reliability. The Decision Tree model produced commendable results with the added advantage of low computational cost, making it suitable for real-time or edge-based applications. In contrast, the K-Nearest Neighbors (KNN) algorithm with k=5, although easy to implement, showed relatively lower performance in terms of precision and recall, particularly when dealing with overlapping feature distributions.

Table II displays the comparison of CFS vs GA Selection. The comparison shows that GA-based selection produced a more discriminative subset of features, improving the models' ability to distinguish between subtle attack patterns.

TABLE II. ACCURACY COMPARISON

Classifier	Accuracy (CFS)	Accuracy (GA-inspired)
DT	0.928	0.943
RF	0.953	0.965
SVM	0.944	0.956
KNN	0.914	0.928

Figure 2 displays the confusion matrices, showcasing that Random Forest achieved the most accurate classifications across all categories, especially for "Normal," "DoS," "Mirai," and "MITM" classes, with minimal misclassifications. SVM (RBF kernel) followed closely, maintaining high accuracy and very few false positives, particularly in "Mirai" and "MITM" detection. KNN also performed well, especially in the "Mirai" and "MITM" categories, but lacked in predicting "Scan" and "Normal" classes. In contrast, the Decision Tree model demonstrated noticeably higher misclassifications, especially confusing "Normal" with "Mirai" and "Scan," and misclassifying a portion of "Scan" and "DoS" instances.

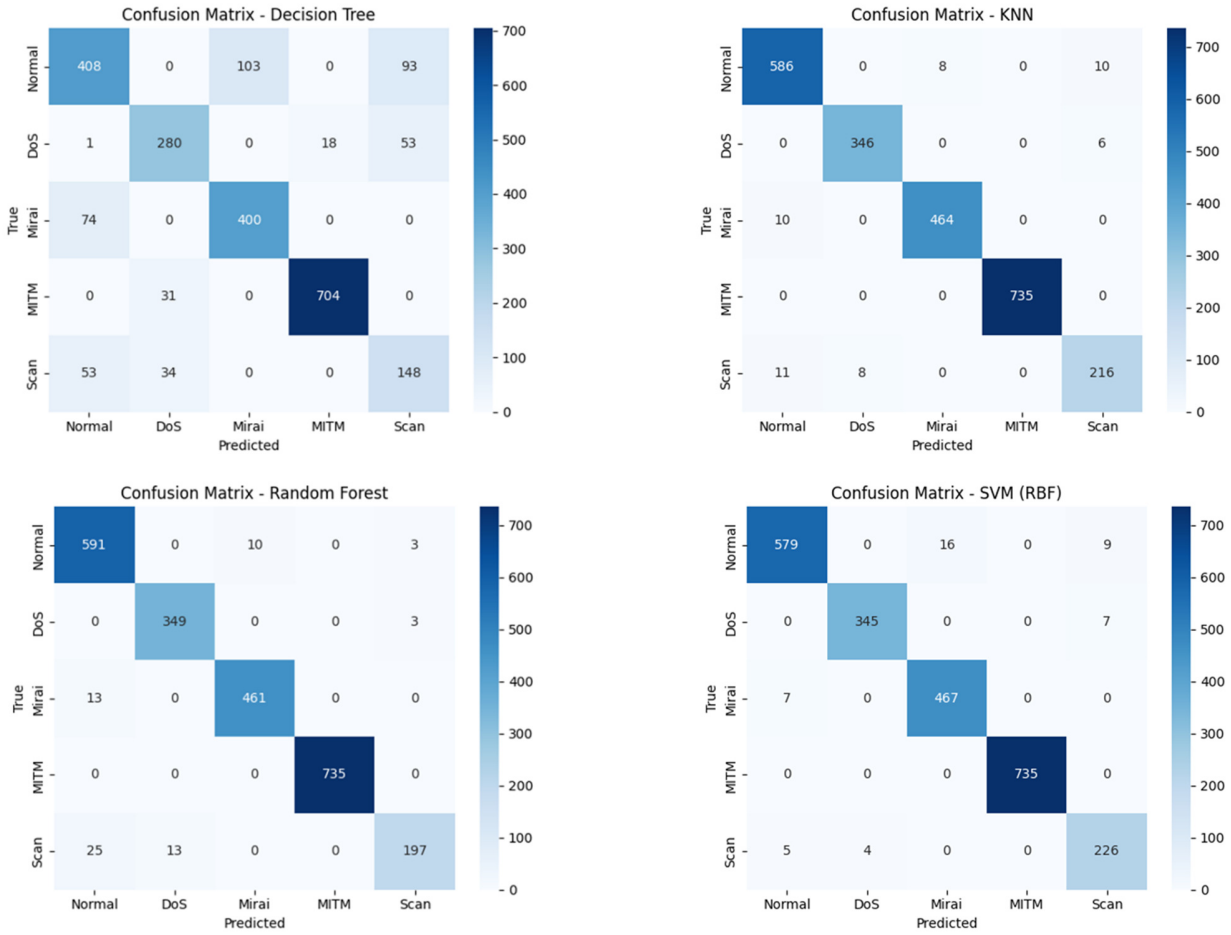

Fig. 2. Confusion matrices of the considered ML models.

Table III displays the hyperparameters selected using grid search with 5-fold cross-validation to optimize accuracy on the training set while avoiding overfitting.

TABLE III. HYPERPARAMETER TUNING CONFIGURATION OF CLASSIFIERS

Classifier	Tuned Hyperparameters	Selected Values
Random Forest	n_estimators, max_depth, min_samples_split	100, 20, 2
Decision Tree	criterion, max_depth, min_samples_split	"gini", 15, 2
SVM (RBF)	C, gamma, kernel	1.0, 0.01, "rbf"
KNN	n_neighbors, weights, metric	5, "uniform", "minkowski"

As shown in Table IV, the proposed model matches or exceeds the accuracy of several recent works while employing simpler and more computationally efficient models, such as Random Forest and SVM. Unlike deep models requiring large resources and complex tuning, our method achieves high precision and recall using lightweight ML algorithms combined with optimized feature subsets. This makes it suitable for deployment on edge or resource-constrained IoT environments.

TABLE IV. COMPARISON WITH RECENT STUDIES

Study	Dataset	Feature Selection	Model	Classes	Accuracy
[1]	CICDDoS-2019	RFS + Relief + PCE	Stacked Classifiers	Binary (DoS)	96.5%
[4]	CAN Dataset	GGO	Triple-Attention DL	Multiclass	96.3%
Proposed	Simulated IoTID20	CFS / GA-inspired	Random Forest	Multiclass	96.5%

V. CONCLUSIONS AND FUTURE WORK

In this study, we proposed a lightweight, multiclass IDS for IoT networks based on supervised machine learning. Building upon a prior binary classification model that focused solely on detecting DoS attacks, our work extends the detection scope to include multiple attack types — DoS, Mirai, MITM, and Scan — alongside normal traffic, resulting in a more comprehensive and realistic IDS solution. The novelty of this work lies in its shift from traditional binary IDS approaches to a lightweight multiclass intrusion detection framework tailored for IoT environments. Unlike deep learning-based methods that require high computational resources, our approach leverages GA-inspired feature selection combined with classical machine learning models, especially Random Forest, to achieve high detection accuracy (96.5%) across five distinct traffic classes. This demonstrates the potential of using interpretable and resource-efficient models for real-time IoT deployment. Furthermore, our simulation of a realistic multiclass dataset inspired by IoTID20 enhances the applicability and practical relevance of the system in constrained network environments.

Despite the promising results, this study has several limitations. First, the dataset used was simulated based on the IoTID20 dataset, which may not capture the full complexity of real-world network behavior. Second, the evaluation was conducted in an offline setting without testing real-time performance. Third, the model targets a limited number of attack types and may need adaptation for broader threat coverage. Fourth, the use of traditional machine learning models, while computationally efficient, may limit the detection of highly sophisticated or zero-day attacks. Lastly, the GA-inspired feature selection method, though effective, is a simplified version and does not fully utilize the search capabilities of advanced evolutionary algorithms. To further enhance the proposed model, several avenues will be pursued. First, the proposed IDS will be evaluated using real-world IoT traffic datasets or deployed in a live testbed to validate its performance under realistic conditions. Second, we plan to integrate real-time detection capabilities by adapting the system for streaming data environments. Third, future work will explore hybrid deep learning models, such as CNN-LSTM or Transformer-based architectures, to improve the detection of complex and zero-day attacks. Fourth, explainable AI (XAI) methods will be incorporated to increase model transparency and support human-in-the-loop analysis, as in recent studies in different domains [18]. Finally, we aim to deploy and benchmark the system on edge devices (e.g., Raspberry Pi, NVIDIA Jetson) to assess performance and feasibility in constrained IoT environments.

REFERENCES

[1] P. Mamatha, S. Balaji, and S. S. Anuraghav, "Development of Hybrid Intrusion Detection System Leveraging Ensemble Stacked Feature Selectors and Learning Classifiers to Mitigate the DoS Attacks," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, Feb. 2025, Art. no. 20, https://doi.org/10.1007/s44196-025-00750-6.

[2] W. B. Dennyson and C. Jothikumar, "Securing Automotive Networks from DoS and Fuzzy Attacks with Optimized LSTM Models," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, Apr. 2025, Art. no. 95, https://doi.org/10.1007/s44196-025-00782-y.

[3] M. Zahid and T. S. Bharati, "Enhancing cybersecurity in IoT systems: a hybrid deep learning approach for real-time attack detection," *Discover Internet of Things*, vol. 5, no. 1, Jul. 2025, Art. no. 73, https://doi.org/10.1007/s43926-025-00156-y.

[4] H. Yang and M. Effatparvar, "A deep learning based intrusion detection system for CAN vehicle based on combination of triple attention mechanism and GGO algorithm," *Scientific Reports*, vol. 15, no. 1, Jun. 2025, Art. no. 19462, https://doi.org/10.1038/s41598-025-04720-y.

[5] A. Sagu, N. S. Gill, P. Gulia, N. Alduaiji, P. K. Shukla, and M. A. Shah, "Advances to IoT security using a GRU-CNN deep learning model trained on SUCMO algorithm," *Scientific Reports*, vol. 15, no. 1, May 2025, Art. no. 16485, https://doi.org/10.1038/s41598-025-99574-9.

[6] O. Martins Onyekwelu, S. Yanxia, and D. Mashao, "Deep Learning-Based Intrusion Detection System: Embracing Long Short-Term Memory (LSTM) and Roughly Balanced Bagging Synergies," *Inteligencia Artificial*, vol. 28, no. 76, pp. 40–65, Jun. 2025, https://doi.org/10.4114/INTARTIF.VOL28ISS76PP40-65.

[7] L. Shan, "(IoT) Network intrusion detection system using optimization algorithms," *Scientific Reports*, vol. 15, no. 1, Jul. 2025, Art. no. 21706, https://doi.org/10.1038/s41598-025-04638-5.

[8] S. B, S. M, M. K, and L. B, "Ensemble of feature augmented convolutional neural network and deep autoencoder for efficient detection of network attacks," *Scientific Reports*, vol. 15, no. 1, Feb. 2025, Art. no. 4267, https://doi.org/10.1038/s41598-025-88243-6.

[9] K. Rajkumar and S. M. Shalinie, "SHAP-based intrusion detection in IoT networks using quantum neural networks on IonQ hardware," *Journal of Parallel and Distributed Computing*, vol. 204, Oct. 2025, Art. no. 105133, https://doi.org/10.1016/j.jpdc.2025.105133.

[10] Z. Zhang, A. Das, G. Huang, and S. Baskiyar, "CAT: A simple heterogeneous ensemble learning framework for network intrusion detection," *Peer-to-Peer Networking and Applications*, vol. 18, no. 4, Jun. 2025, Art. no. 213, https://doi.org/10.1007/s12083-025-02000-0.

[11] M. A. Ahmed and S. Alnatheer, "Intrusion Detection in a Digital Twin-Enabled Secure Industrial Internet of Things Environment for Industrial Sustainability," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 21263–21269, Apr. 2025, https://doi.org/10.48084/etasr.10128.

[12] N. Alkhafaji, T. Viana, and A. Al-Sherbaz, "Integrated Genetic Algorithm and Deep Learning Approach for Effective Cyber-Attack Detection and Classification in Industrial Internet of Things (IIoT) Environments," *Arabian Journal for Science and Engineering*, vol. 50, no. 15, pp. 12071–12095, Aug. 2025, https://doi.org/10.1007/s13369-024-09663-6.

[13] C. Khammassi and S. Krichen, "A GA-LR wrapper approach for feature selection in network intrusion detection," *Computers & Security*, vol. 70, pp. 255–277, Sep. 2017, https://doi.org/10.1016/j.cose.2017.06.005.

[14] V. Kumar, D. Sinha, A. K. Das, S. C. Pandey, and R. T. Goswami, "An integrated rule based intrusion detection system: analysis on UNSW-NB15 data set and the real time online dataset," *Cluster Computing*, vol. 23, no. 2, pp. 1397–1418, Jun. 2020, https://doi.org/10.1007/s10586-019-03008-x.

[15] R. Chaganti, A. Mourade, V. Ravi, N. Vemprala, A. Dua, and B. Bhushan, "A Particle Swarm Optimization and Deep Learning Approach

for Intrusion Detection System in Internet of Medical Things," *Sustainability*, vol. 14, no. 19, Jan. 2022, Art. no. 12828, https://doi.org/10.3390/su141912828.

[16] E.-S. M. El-Kenawy, M. M. Eid, H. L. Hussein, A. M. Osman, and A. M. Elshewey, "Optimized Deep Learning Model Using Binary Particle Swarm Optimization for Phishing Attack Detection: A Comparative Study," *Mesopotamian Journal of CyberSecurity*, vol. 5, no. 2, pp. 685–703, Jul. 2025, https://doi.org/10.58496/MJCS/2025/041.

[17] R. Amin Labid, "iotid20 dataset." [Online]. Available: https://www.kaggle.com/datasets/rohulaminlabid/iotid20-dataset.

[18] A. M. Elshewey, S. A. Z. Hassan, R. Y. Youssef, H. M. El-Bakry, and A. M. Osman, "Enhancing Hydrogen Energy Consumption Prediction Based on Stacked Machine Learning Model with Shapley Additive Explanations," *Process Integration and Optimization for Sustainability*, May 2025, https://doi.org/10.1007/s41660-025-00539-2.