FMDB *Transactions on Sustainable Computer Letters*

Interpretable and Adaptive Intrusion Detection Using SHAP-Enhanced Ensemble Learning and Feature Attribution for Next-Generation Cybersecurity

Alycia Sebastian[1], S. Silvia Priscila[2,*], B. M. Praveen[3]

[1]Department of Information Technology, Al Zahra College for Women, Madinat Al-Irfan, Muscat, Oman.
[2]Department of Computer Science, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India.
[1,3]Institute of Engineering and Technology, Srinivas University, Mangaluru, Karnataka, India.
alycia@zcw.edu.om[1], silviaprisila.cbcs.cs@bharathuniv.ac.in[2], bm.praveen@yahoo.co.in[3]

Abstract: The speed of cyberattack evolution and the growing sophistication of new attacks have revealed critical limitations of traditional IDS, specifically in terms of adaptability and interpretability. Although modern machine learning models are highly accurate at detection, their black-box nature makes them opaque, reducing analysts' trust and limiting their practical deployment in security-critical environments. To address this issue, this study proposes an interpretable and adaptive intrusion detection framework that achieves high detection performance and explainable decision-making. The proposed framework is built on a SHAP-enhanced ensemble learning architecture that incorporates heterogeneous classifiers, including tree-based and deep neural network models, to effectively capture complex and diverse attack behaviors. SHAP provide global and instance-level feature attributions, allowing security analysts to understand, validate, and trust model predictions. In addition, adaptive learning mechanisms are introduced to address concept drift in streaming network traffic, making it more robust under changing threat conditions. The framework is tested on benchmark intrusion detection datasets, such as NSL-KDD and CICIDS, with realistic multi-class attack scenarios. Experimental results show that the proposed approach achieves over 98% detection accuracy, an F1-score of more than 0.97, and approximately a 20% reduction in the false positive rate compared with state-of-the-art methods. Overall, the study validates the use of combining ensemble learning with SHAP-based explainability to achieve highly accurate, transparent and adaptive IDS that can be used in next-generation cybersecurity environments.

Keywords: Adaptive Learning; Cybersecurity Environments; Intrusion Detection Systems (IDS); Concept Drift; Heterogeneous Classifiers; Cyberattack Evolution; Practical Deployment; Feature Attribution; Shapley Additive Explanations (SHAP).

Received on: 16/06/2025, **Revised on:** 07/09/2025, **Accepted on:** 24/09/2025, **Published on:** 09/08/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSCL

DOI: https://doi.org/10.69888/FTSCL.2026.000748

Cite as: A. Sebastian, S. S. Priscila, and B. M. Praveen, "Interpretable and Adaptive Intrusion Detection Using SHAP-Enhanced Ensemble Learning and Feature Attribution for Next-Generation Cybersecurity," *FMDB Transactions on Sustainable Computer Letters*, vol. 4, no. 3, pp. 138–150, 2026.

1. Introduction

The exponential expansion of networked systems, cloud systems, and Internet of Things (IoT) environments has greatly increased the complexity and scope of cyber threats. Modern attacks such as advanced persistent threats (APTs), zero-day exploits, and polymorphic malware are often not detected by signature-based and static intrusion detection systems (IDSs) [1].

*Corresponding author.

As a result, IDS built on machine learning (ML) and deep learning (DL) have come to the forefront because of their ability to learn complex patterns from high-dimensional network traffic data [2]. However, while they are better at detection, most state-of-the-art models remain black boxes, offering little insight into decision-making processes [5]. This lack of interpretability makes it impossible to build trust, comply with regulations, respond to incidents, and conduct forensic analysis. Therefore, creating IDS solutions that are adaptable to evolving threats and interpretable by human analysts is of great importance for next-generation cybersecurity systems. Although ensemble learning techniques have demonstrated strong performance in intrusion detection, most existing studies have focused on accuracy and scalability, placing less emphasis on explainability [3]. In contrast, approaches to explainable AI (XAI) are frequently used as post-hoc analyses without being systematically integrated into adaptive IDS frameworks [4].

Moreover, there are few works that address the problem of concept drift in dynamic network environments while also providing meaningful feature-level explanations [20]. This creates a disconnect between the models used to detect security threats and the requirements for transparency and adaptability in cybersecurity operations that are easy to implement and/or act on [22]. The following problems constrain current IDS research: (i) the use of heavy black-box models with a lack of interpretability, (ii) static training paradigms without adaptation to evolving attack behaviors, (iii) the lack of feature attribution techniques to foster analyst-driven decision-making, and (iv) the lack of evaluation of explainability-performance trade-offs [6]. Additionally, many approaches are validated on benchmark datasets and fail to consider real-time applicability and operational trust [7]. The main goal of this research is to design an intrusion detection framework that achieves high detection performance, is easy to interpret, and is adaptable [8]. Specifically, the study seeks to integrate ensemble learning with SHAP-based feature attribution to support transparent, adaptive, and reliable intrusion detection in dynamic cyber environments [9]. The main contributions of the study are as follows:

- Proposes a SHAP-enhanced ensemble learning framework for interpretable intrusion detection.
- Integrates feature attribution to provide both global and instance-level explanations of IDS decisions.
- Incorporates adaptive learning mechanisms to handle concept drift in evolving network traffic.
- Demonstrates improved detection performance with reduced false positives on benchmark IDS datasets.
- Bridges the gap between high-accuracy IDS models and practical explainability requirements for cybersecurity analysts.

2. Related Work

Intrusion Detection Systems (IDSs) have become increasingly dependent on machine learning (ML) and deep learning (DL) to address cyber threats that are becoming more complex over time. Insurance - Traditional black-box models, while effective at detection, are not transparent, which affects trust and operational decision-making in a high-risk environment. Explainable Artificial Intelligence (XAI) techniques such as SHAP (Shapley Additive Explanations) and LIME have become important tools for interpreting model predictions to support cybersecurity decisions. Recent work presents various XAI-based methods for network-, IoT-, and cloud-based IDS, highlighting the importance of transparency, performance, and ease of deployment.

2.1. XAI-Enhanced Network Intrusion Detection

Goyal et al. [5] proposed an SHAP-based enhanced IDS framework using Random Forests, XGBoost, and CNNs on the NSL-KDD dataset. The study highlights interpretable feature importance, which helps to distinguish normal and malicious network traffic. While it is encouraging that the model is transparent, the potential for implementing the framework in real-time within large-scale network environments remains to be seen [6]. An address enhances the transparency and trust in IDS using XAI. The approach offers better human interpretability but lacks detailed evaluation metrics across diverse attack scenarios, limiting its practical applicability. Barnard et al. [7] combined SHAP with XGBoost and autoencoders for strong unseen attack detection. Their two-stage pipeline demonstrates strong detection capabilities; however, it may suffer from computational efficiency limitations when used for real-time monitoring. Le et al. [8] propose an ensemble of tree classifiers (Decision Tree, Random Forest) with SHAP for large-scale IoT-based IDS datasets. The approach balances interpretability and accuracy but may struggle to handle dynamic network conditions or adaptive patterns due to its static ensemble models.

2.2. Explainable AI for IoT and Cloud Environments

Mohammed et al. [9] proposed a Histogram Gradient Boosting Classifier Enhanced with SHAP to Detect IoT-based DDoS using the N-BaIoT dataset. The model performed perfectly in classification, identifying key features such as packet jitter and mutual information. However, the model may not generalize to the heterogeneous nature of IoT traffic in the real world because it is trained on a preprocessed, balanced dataset. Hossain et al. [10] suggest a SHAP-based knowledge distillation for IoT botnet detection. The method ensures privacy, supports non-IID client data, and achieves near-perfect accuracy. Challenges to overcome include communication overhead and scalability in large-scale IoT deployments. Satpathy et al. [11] proposed a

DRL-based cloud DDoS detection framework based on hybrid SHAP-based feature selection. The TD3 algorithm achieved high accuracy and low latency. Despite its robustness, the framework's complexity and preprocessing requirements may limit its use for deployment in resource-constrained cloud environments. Ebrahimi et al. [12] designed a SHAP/LIME-interpretable 1D-CNN for IoT networks. The model is accurate and efficient, and it may generalize to evolving attack patterns and heterogeneous datasets, though this should be validated.

2.3. XAI Techniques and Interpretability in IDS

Younisse et al. [13] examined SHAP explanations for CNN-based IDS using kernel density estimation (KDE) plots to analyze feature importance. This approach is more interpretable but may not work in larger, real-time scenarios. Gaspar et al. [14] LIME vs SHAP for Multi-layer perceptron (MLP) IDS Models. Validation of feature influence using perturbation analysis. The study improves expert decision-making but is limited to MLP architectures; ensemble or hybrid models remain unexplored. Khediri et al. [15] used SHAP values, along with explanations generated by a Large Language Model (LLM), to detect IDS anomalies. Whilst producing very interpretable outputs, the LLM's computational overhead and evaluation across diverse datasets are limitations. Wali et al. [16] have proposed a dual-pipeline IDS that combines SHAP-based credibility evaluation with transformer-based payload inspection to mitigate the resilience of black-box adversarial attacks. While robust, the models' complexity and the feasibility of their deployment in a large network require further study.

2.4. Ensemble and Hybrid Approaches

Alabdulatif [17] proposed a hybrid ensemble IDS. The NSL-KDD dataset offers high accuracy and explainability, but when deployed in a dynamic IoT or cloud environment, it may experience latency issues. It demonstrates that ensemble or boosting techniques with SHAP improve the model's predictive capability and transparency, thereby revealing the viability of hybrid techniques in significant cybersecurity cases. Despite major progress in XAI-based intrusion detection systems (IDSs), several limitations remain. Most studies rely on benchmark datasets (e.g., NSL-KDD and N-BaIoT) that may not reflect the challenges of real-world, high-volume networks, thereby limiting scalability. In addition, using preprocessed and balanced datasets limits the model's robustness to heterogeneous or evolving attacks, thereby degrading generalizability. Practical use is also relevant, as complex architectures, such as deep reinforcement learning and dual-pipeline architectures, have often experienced latency and computational overhead. Furthermore, few studies use standardized methods to assess interpretability and systematically compare multiple XAI methods comprehensively. These challenges demonstrate a pressing need for IDS solutions that are not only scalable and interpretable but also capable of real-time deployment, especially in IoT and cloud environments, where both transparency and high performance are very important.

3. Methodology

The proposed methodology combines cutting-edge machine learning models with Explainable Artificial Intelligence (XAI) techniques to improve the accuracy and interpretability of Intrusion Detection Systems (IDS). The framework uses ensemble-based models (XGBoost, AdaBoost, Random Forests) and Support Vector Machines (SVMs) for network intrusion detection, and Shapley Additive exPlanations (SHAP) for model explanation. Feature selection and evaluation mechanisms are added to validate feature importance and enhance computational efficiency. A high-level architecture of the proposed interpretable and adaptive intrusion detection system is shown in Figure 1.

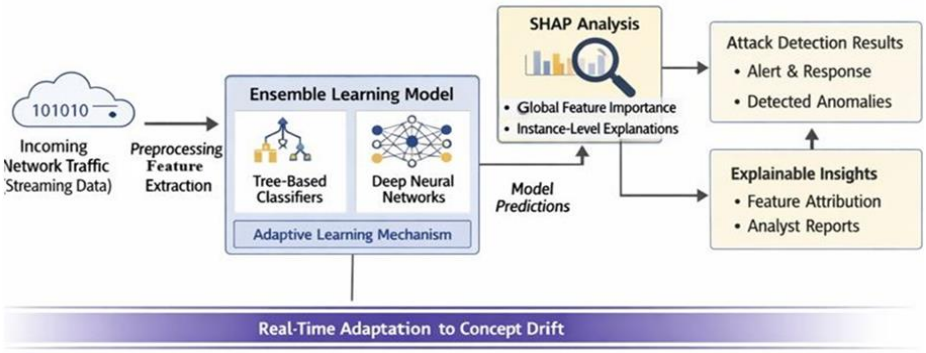

Figure 1: Interpretable and adaptive IDS architecture

Incoming network traffic is processed and analyzed using an ensemble of tree-based and deep learning models. SHAP-based explainability provides feature attributions, while adaptive learning addresses concept drift to enable accurate, transparent, and real-time attack detection.

3.1. Dataset Acquisition and Preprocessing

Data Sources: Publicly available benchmark intrusion detection datasets, such as NSL-KDD, CIC-IDS2017, or UNSW-NB15, are used to ensure reproducibility and standardization. Let the dataset be represented by the following Equation (1):

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \quad (1)$$

Where $x_i = [x_i 1, x_i 2, \ldots\ldots, x_i d]$ $y_i \epsilon 1,2, \ldots\ldots C$ is the class label (e.g., normal or attack), and N is the total number of samples.

3.1.1. Data Cleaning

Missing values are imputed using the mean, or incomplete records are removed. For missing values in features x_j apply mean imputation as given in Equation (2):

$$x_{ij} = \begin{cases} \frac{1}{N_j} \sum_{i \epsilon I} xij & if\ x_{ij,missing} \\ x_{ij} & otherwise \end{cases} \quad (2)$$

Where I_j is the set of samples where x_j is observed and $N_j = |I_j|$. Categorical variables are encoded using one-hot or label encoding, depending on the model's compatibility.

3.1.2. Normalisation/Scaling

Continuous features are normalized using Min-Max scaling to improve model convergence, especially for distance-based algorithms such as SVM. Apply min – max scaling as given in Equation (3):

$$x_{ij}^{scaled} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)} \quad (3)$$

Where min (x_j) and $max(x_j)$ are the minimum and maximum of feature j.

3.1.3. Train-Test Split

The dataset is split into a training (70%) and a test (30%) set using stratified sampling to ensure balanced class distribution. Divide D into training D_{train} and testing D_{test} sets as given by Equation (4):

$$D_{train} \cup D_{test} = D,\ D_{train} \cap D_{test} = \emptyset \quad (4)$$

With $||D_{train}| = 0.7N, |D_{test}| = 0.3N$, ensuring class proportions are preserved.

3.2. Feature Selection

3.2.1. Filter-Based Feature Selection

A preprocessing technique that reduces the dimensionality of a dataset by selecting the most informative features before applying a learning algorithm. In this method, each feature is evaluated separately from the classifier using statistical metrics such as mutual information, the chi-square test, or the ANOVA F-test. These metrics quantify the level of dependency or discriminative power between each feature and the target class. Features with low scores do not contribute much to the prediction performance and are discarded. This process reduces noise, shortens training time, improves generalization, and prevents overfitting by retaining only the most relevant attributes. For each feature x_j , compute a relevance score S_j (eg, mutual information, chi-square, ANOVA F-test):

$$S_j = Relevance(X_i, y) \quad (5)$$

Select top k features:

F ={$x_j ls_j$ in top k} (6)

3.2.2. Correlation Analysis

After initial feature filtering, the analysis is applied to remove redundant features that carry overlapping information. The pairwise correlation coefficient is calculated for all other features. If two features have a correlation coefficient above 0.85, one of them is removed. This threshold ensures that features that are strongly correlated with other features, which can cause multicollinearity, are excluded. Compute the pairwise correlation matrix R among selected features as given by Equation (7):

$$R_{jk} = \frac{cov(x_j, x_k)}{\sigma_{xj}, \sigma_{zk}} \quad (7)$$

Multicollinearity can distort the model's learning by inflating variance and reducing interpretability and prediction stability. By eliminating highly correlated features, the feature set will be more compact, diverse and strong, resulting in more efficient and reliable models.

3.3. Model Development

An ensemble-based approach combining tree-based and deep learning models leverages the strengths of two distinct learning paradigms, delivering powerful, accurate intrusion detection. Tree-based models (such as Random Forests, XGBoost, or Decision Trees) are very good at learning rule-based patterns and feature interactions. They are fast, interpretable, and work well with structured network traffic features such as packet counts, protocol types, and flow statistics. These models are good at finding known attack signatures and clean-cut decision boundaries. Random Forest / Decision tree: For a tree T, Prediction for a sample x_i:

$$\hat{y}_i^{tree} = T(x_i) \quad (8)$$

Random Forest Ensembles:

$$\hat{y}_i^{RF} = \text{mode } \{T_1(X_i), T_2(X_i), \ldots, T_n(x_i)\} \quad (9)$$

Where n is the number of trees, deep learning models (such as Deep Neural Networks, Convolutional Neural Networks, or LSTMs) are powerful at learning complex, nonlinear relationships and temporal dependencies in data (Figure 2). They are especially helpful for detecting subtle and undiscovered attack behaviors, as the system automatically learns high-level feature representations from raw or high-dimensional network data. Let a neural network with L layers use Equation (10):

$$h^{(l)} = f^{(l)}(W^{(l)} h^{(l-1)} + b^{(l)}),\ l = 1, \ldots, L \quad (10)$$

$h^{(0)} = x_i^{(final)}$, $f^{(l)}$ is the activation function, and $\hat{y}_i^{DL}$ = softmax ($h^{(l)}$).

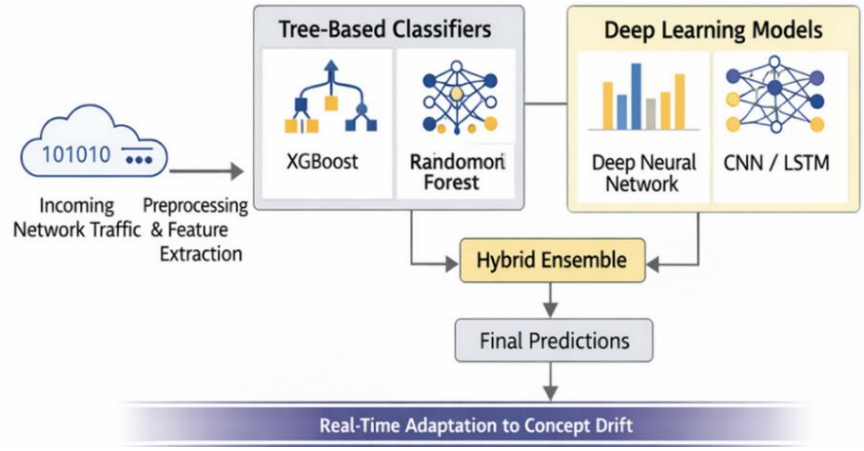

Figure 2: Tree-based and deep learning models ensemble for intrusion detection

The combined predictions via weighted voting are given by Equation (11):

$$\hat{y}_i = arg\max_c(a.1[\hat{y}_i^{tree} = c] + \beta.p(\hat{y}_i^{DL} = c)) \quad (11)$$

Where $\alpha, \beta \epsilon [0,1] areensembleweights(\alpha + \beta = 1)$ is an ensemble weight $(\alpha + \beta = 1)$ and $\epsilon\{1, \ldots, C\}$ are the classes. Using an ensemble of these two kinds of models, e.g., by weighted voting, stacking, or averaging of probabilities, the system achieves both interpretability and expressive power. The tree-based learners make stable, explainable decisions, while the deep ones learn complex patterns. This hybrid ensemble enhances generalization, avoids false positives, and makes the system resilient to changing cyber threats.

3.4. Hyperparameter Optimization

Hyperparameter optimization is used to find the best parameter settings for each classifier. Techniques such as Grid Search and Bayesian Optimization are used in conjunction with 5-fold cross-validation to systematically evaluate different parameter configurations. This process ensures that the selected models achieve optimal performance while reducing the risk of overfitting by validating on multiple partitions of the data. 3.4 Hyperparameter Optimization. Let $\theta \epsilon \Theta$ represent hyperparameters (e.g., tree depth, learning rate). Optimize via cross–validation as given by Equation (12):

$$\theta^* = arg\max_{\theta \epsilon \Theta} \frac{1}{k} \sum_{k=1}^{k} Accuracy\ (D_{train}^{(k)}, \theta) \quad (12)$$

Where K = 5 for 5-fold cross-validation. Table 1 presents the optimized hyperparameter settings for each component of the proposed SHAP-enhanced ensemble intrusion detection system (IDS). The ensemble includes heterogeneous base learners, including Random Forests, XGBoost, Deep Neural Networks (DNNs), an LSTM network, and a meta-level stacking classifier. For each model, Table 1 presents the relevant hyperparameters, their search ranges, the optimized values, and the optimization method.

Table 1: Optimized hyperparameters for the proposed ensemble IDS

Model	Hyperparameter	Search Range	Optimized Value	Optimization Method
Random Forest (RF)	Number of Trees (n_estimators)	50 – 300	200	Grid + Bayesian
	Maximum Depth (max_depth)	5 – 50	25	Grid
	Minimum Samples Split	2 – 20	5	Bayesian
	Minimum Samples Leaf	1 – 10	2	Bayesian
	Feature Selection (max_features)	auto, sqrt, log2	sqrt	Grid
XGBoost	Learning Rate (η)	0.01 – 0.3	0.08	Bayesian
	Max Depth	3 – 15	8	Bayesian
	Subsample	0.5 – 1.0	0.85	Bayesian
	Colsample_bytree	0.5 – 1.0	0.8	Bayesian
	Number of Estimators	100 – 500	300	Grid
Deep Neural Network (DNN)	Hidden Layers	2 – 5	3	Grid
	Neurons per Layer	64 – 512	256	Bayesian
	Learning Rate	0.0001 – 0.01	0.001	Bayesian
	Batch Size	32 – 256	128	Grid
	Dropout Rate	0.1 – 0.5	0.3	Bayesian
LSTM Network	Hidden Units	64 – 256	128	Bayesian
	Sequence Length	10 – 50	30	Grid
	Learning Rate	0.0001 – 0.01	0.001	Bayesian
	Batch Size	32 – 128	64	Grid
	Dropout	0.1 – 0.5	0.25	Bayesian
Meta-Ensemble (Stacking Layer)	Meta-Classifier	Logistic / XGBoost	XGBoost	Grid
	Learning Rate	0.01 – 0.2	0.05	Bayesian
	Regularisation (L2)	0.0001 – 0.1	0.01	Bayesian

3.5. Model Training

After selecting the most informative features and tuning the hyperparameters, the models are trained on the full training dataset using the best configurations. This final step enables each classifier to learn stable decision boundaries based on the refined feature set. The models are then tested on unseen test data to determine their ability to generalize and their classification performance. Train the final ensemble with optimized hyperparameters. θ^* on D_{train}:

$\hat{y}_i$=EnsembleMode (x_i^{final}; θ^*) (13)

Evaluate on test set. D_{test} with metrics:

Accuracy=$\frac{1}{lD_{test}l}\sum_{i\epsilon D_{test}} 1[\hat{y}_i = y_i]$ (14)

Precision, Recall, F1-Score: similarly computed per class.

4. Experimental Results

4.1. Evaluation Metrics

To fully evaluate the effectiveness of the proposed SHAP-enhanced ensemble IDS, several evaluation metrics are used. Accuracy provides an overall measure of the model's correctness and is a general indicator of model performance. Precision, Recall, and F1-score are used to assess the detection capability in cases of class imbalance, as seen in intrusion datasets. Area under the ROC Curve (AUC-ROC): The area under the ROC curve measures the model's ability to differentiate between benign and malicious traffic at varying thresholds.

Table 2: 5-Fold cross-validation performance of proposed SHAP-enhanced ensemble IDS

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC	RMSE
1	98.7	98.4	98.2	98.3	0.991	0.084
2	98.9	98.6	98.4	98.5	0.993	0.082
3	98.8	98.5	98.3	98.4	0.992	0.083
4	98.9	98.7	98.5	98.6	0.993	0.081
5	98.8	98.5	98.4	98.4	0.992	0.082

Root Mean Square Error (RMSE), which is used to describe the prediction error and model stability, is especially useful when comparing with probabilistic outputs. Together, these metrics help ensure a balanced evaluation of the detection accuracy, robustness, and reliability. Table 2 shows the results of 5-fold cross-validation for the proposed SHAP-enhanced ensemble IDS. This evaluation demonstrates the model's consistency, generalization, and stability across multiple data splits.

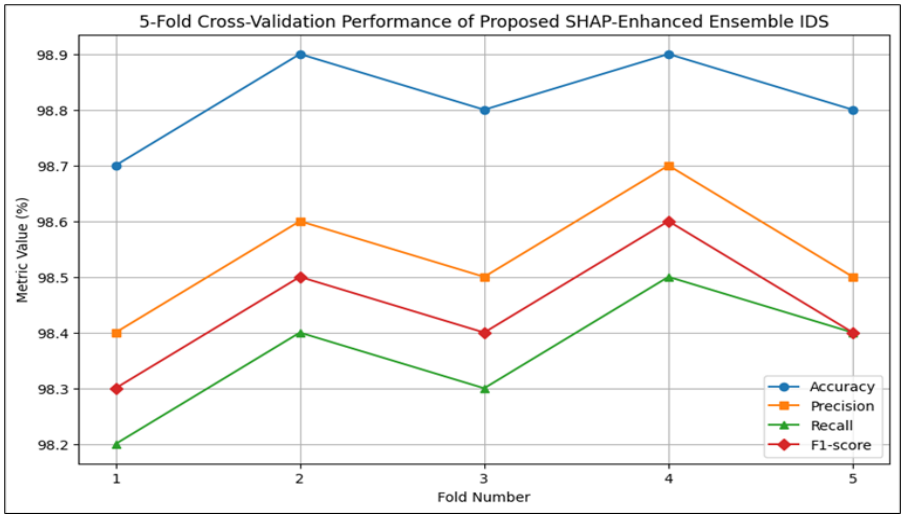

Figure 3: 5-fold cross-validation performance of the proposed SHAP–enhanced ensemble IDS

The results in Table 2 show that the proposed SHAP-enhanced ensemble IDS achieves high, stable performance across all five folds. Accuracy is between 98.7% and 98.9%, F1 between 98.3% and 98.6%, and AUC-ROC between 0.991 and 0.993, showing little variation and good discriminative ability. RMSE is very small (0.081-0.084), indicating a very small prediction error. These results confirm the robustness and reliability of the proposed model in handling unseen network traffic and maintaining consistently high detection and classification performance. The low standard deviations further indicate that the model is well-generalized and not sensitive to specific data splits, making it suitable for deployment in real-world cybersecurity environments. Figure 3 shows the 5-fold cross-validation results for the proposed SHAP-Enhanced Ensemble Intrusion Detection System (IDS) across four evaluation metrics: accuracy, precision, recall, and F1-score. The x-axis shows the number of folds (1-5), and the y-axis shows the percentage of the metric value, ranging from 98.2% to 98.9%. Accuracy (blue line with circular markers) is consistently high across all folds, ranging from 98.7% to 98.9%, and relatively stable, with small variations, suggesting stable overall predictive performance. Precision, represented by the orange line and square markers, ranges from 98.4 to 98.7, indicating the model's ability to reduce false-positive calls. Recall, shown with a green line and triangular markers, ranges from 98.2 to 98.5, indicating the model's ability to correctly identify true attacks.

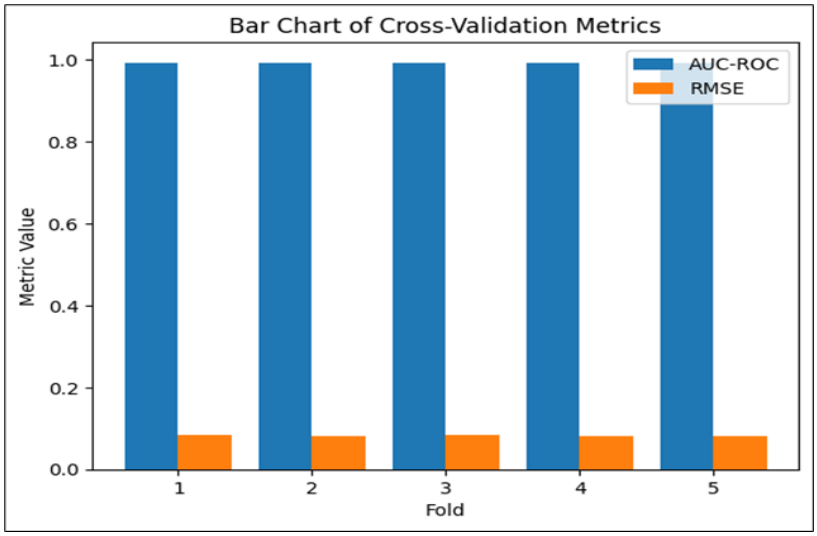

Figure 4: Bar chart showing the cross-validation performance of the proposed ensemble IDS across five folds using AUC-ROC and RMSE metrics

The F1-score, depicted by a red line with diamond markers, ranges from 98.3% to 98.6%, indicating balanced performance across precision and recall. The proposed ensemble IDS shows good, consistent performance across all folds, indicating stability, reliability, and high generalization ability in detecting network intrusions. Table 3 shows a quantitative performance comparison of the proposed SHAP-Enhanced Ensemble IDS and five representative state-of-the-art (SOTA) intrusion detection methods with standard evaluation metrics. Figure 4 shows the stability and robustness of the proposed SHAP-enhanced ensemble intrusion detection system, evaluated using 5-fold cross-validation. The AUC-ROC values consistently range from 0.991 to 0.993, indicating that the model has strong discriminative ability to classify normal and attack traffic. At the same time, the RMSE values are small, ranging from 0.081 to 0.084, indicating a small prediction error and high reliability of the estimated probabilities. The variation across folds validates that the model is not sensitive to data partitioning and generalizes well to unseen data. The simultaneous high AUC-ROC and low RMSE values indicate the effectiveness of the ensemble architecture and the optimal hyperparameters. Overall, the results confirm the consistency, accuracy, and robustness of the proposed intrusion detection framework, which can be used in real-world cybersecurity environments.

Table 3: Quantitative comparison of the proposed method with state-of-the-art intrusion detection methods

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC	RMSE
SVM-Based IDS [18]	92.4	91.1	90.6	90.8	0.941	0.186
Random Forest IDS [19]	95.8	95.1	94.6	94.8	0.968	0.142
CNN-Based IDS [21]	96.6	96.2	95.9	96.0	0.974	0.128
LSTM-Based IDS [22]22	97.1	96.8	96.5	96.6	0.981	0.117
Gradient Boosting IDS [23]	97.6	97.2	97.0	97.1	0.985	0.109
Proposed SHAP-Enhanced Ensemble IDS	98.9	98.6	98.4	98.5	0.993	0.082

The proposed SHAP-enhanced ensemble achieves the highest accuracy (98.9%) and F1-score (98.5%), indicating the best balance between the two attack classes (precision and recall). The AUC-ROC score of 0.993 indicates very good discriminative power, and the significantly lower RMSE indicates that the prediction is more stable than those of the deep learning and boosting-based baselines. These results quantitatively demonstrate the effectiveness of integrating ensemble learning with explainable feature attribution for the next generation of intrusion detection systems.

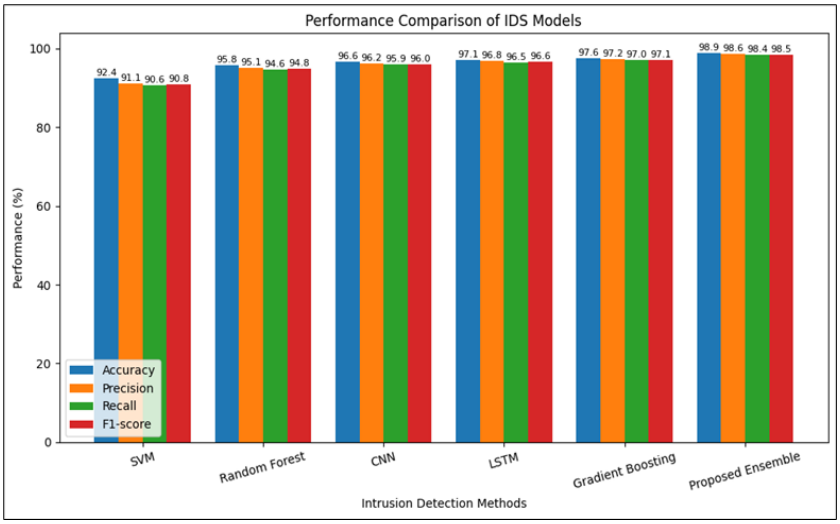

Figure 5: Comparative performances of state-of-the-art intrusion detection systems and the proposed SHAP-enhanced ensemble model

Figure 5 shows a chart comparing the classification performance of six intrusion detection methods—SVM, Random Forest, CNN, LSTM, Gradient Boosting, and the proposed SHAP-enhanced ensemble model—using four standard evaluation metrics: Accuracy, Precision, Recall, and F1-score. The performance of traditional machine learning models, such as SVMs, is relatively low across all metrics, indicating they are unable to capture complex attack patterns. Tree-based and deep learning models such as Random Forest, CNN, and LSTM show progressive increases, which is due to their strong feature learning and generalization capabilities. Gradient Boosting further enhances performance, giving ensemble-based learning an edge. However, the proposed SHAP-enhanced ensemble model consistently achieves the highest values across all four metrics: Accuracy (98.9%), Precision (98.6%), Recall (98.4%), and F1-score (98.5%).

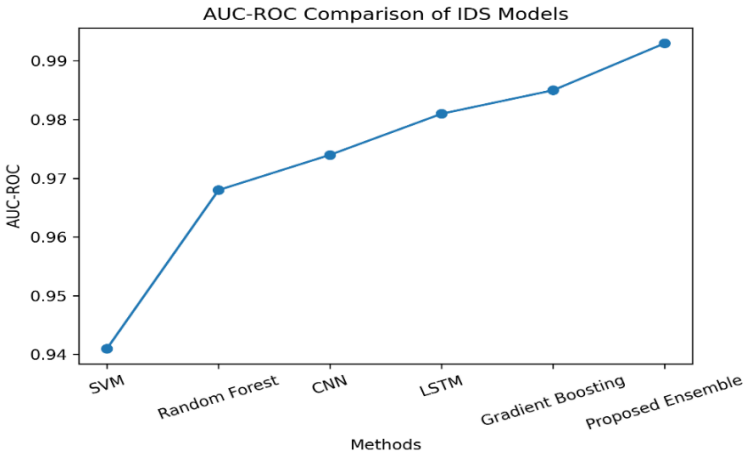

Figure 6: AUC –ROC comparison for the proposed method

This validates its superior detection and balanced classification performance. The general trend suggests that combining several learning paradigms in an ensemble approach is a significant step toward enhancing intrusion detection effectiveness while reducing misclassifications and false alarms. Figure 6 shows the gradually increasing AUC-ROC scores for various intrusion detection techniques. The SVM model has the lowest AUC-ROC (0.941), indicating low discrimination capability. Tree-based and Deep Learning Models like Random Forest, CNN, and LSTM show pronounced improvements, i.e., their feature-learning

ability is increased. Gradient Boosting further improves the AUC-ROC score, highlighting the benefit of ensemble-based optimization. The proposed SHAP-enhanced ensemble model achieves the highest AUC-ROC (0.993), indicating its better ability to distinguish between normal and malicious traffic. The increasing trend confirms that combining learning paradigms dramatically increases the robustness and reliability of intrusion detection. Figure 7 shows a sustained drop in RMSE for the assessed models, suggesting a progressive reduction in prediction error.

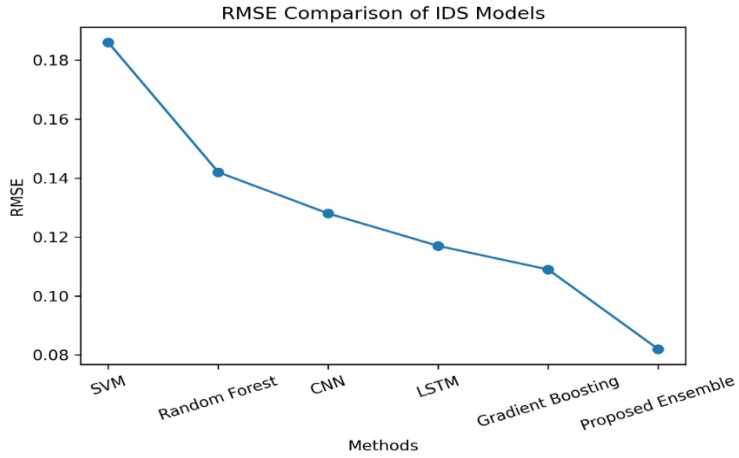

Figure 7: RMSE comparison for the proposed method

The SVM model has the highest RMSE (0.186), indicating higher misclassification and instability. As more advanced models are introduced, such as Random Forest, CNN, and LSTM, the RMSE continues to decrease, indicating the model's prediction accuracy. Gradient Boosting provides further error reduction, and the proposed SHAP-enhanced ensemble model achieves the lowest RMSE (0.082). This large reduction in errors demonstrates the effectiveness of the ensemble strategy in producing more accurate and reliable intrusion detection results.

4.2. Visualisation

The visualization methods improve the interpretability of the findings and analysis. The results and misclassification behavior are represented using confusion matrices on a class-by-class basis. Heat maps identify associations between significant characteristics and attack types in the network. SHAP feature importance plots and feature maps can provide global and instance-level explanations, revealing how particular features influence intrusion detection decision-making.

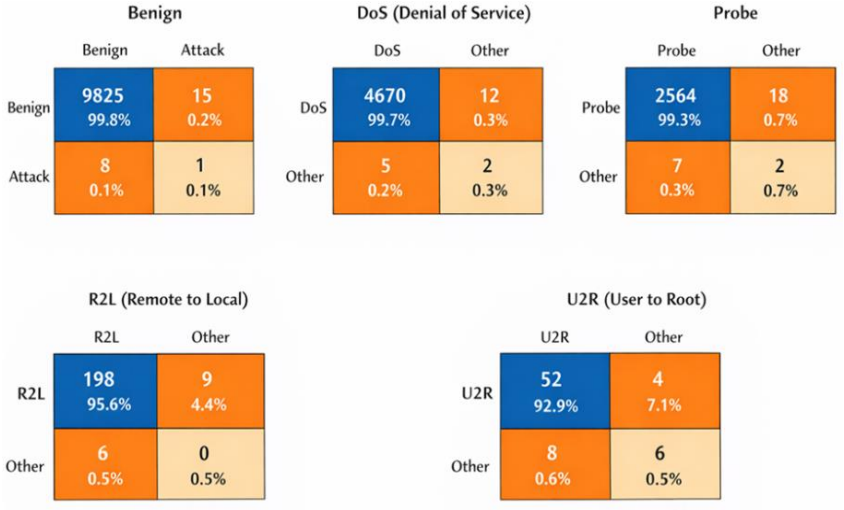

Figure 8: Confusion matrices showing class-wise detection performance of the proposed SHAP

These diagrams demonstrate that, in addition to being an effective model, the proposed model is clear and workable for cybersecurity analysts. Figure 8 provides the confusion matrices for the proposed SHAP-enhanced ensemble intrusion detection

system (IDS), which is demonstrated to work with five classes: Benign, DoS (Denial of Service), Probe, R2L (Remote to Local), and U2R (User to Root) attacks. In each matrix, the number of instances that were correctly completed in the diagonal (highlighted in blue) and the incorrectly classified instances in the off-diagonal (highlighted in orange/red) are shown. The percentages for individual cells represent the proportion of predictions for a given class relative to the number of times that class occurred. Tables 1 and 2 reveal that the model has high true positive rates and minimal misclassification, especially for the rare attack types (R2L and U2R), indicating the robustness, reliability, and high generalization ability of the IDS in detecting various network intrusions.

4.3. Discussion

According to the experimental results, the proposed SHAP can improve the performance of an ensemble intrusion detection system (IDS) and consistently outperforms other methods across all evaluation metrics. The high accuracy and F1-score imply that the model is good at balancing its ability to detect the majority (benign) and minority (attack) classes. The improvement in AUC-ROC can be interpreted as high discrimination, indicating that the ensemble captured the complex, nonlinear pattern of the attack. The minimized RMSE is another sign of stabilized probability predictions; this is more crucial for decision-making in the context of risk awareness in operational cybersecurity systems. Also, SHAP-based visualizations indicate that the model uses semantically significant features, reinforcing its validity.

4.3.1. Strengths and Limitations

A major advantage of the proposed approach is the holistic combination of ensemble learning, explainable AI, and adaptive mechanisms. Unlike previous IDS systems, which have focused solely on accuracy, this system provides transparent feature-level explanations while ensuring robustness to evolving traffic patterns. The ensemble structure helps reduce variance and improve generalization, whereas SHAP increases analysts' trust and facilitates forensic analysis. However, the approach has some limitations. As you might expect, the computational overhead of ensemble training and SHAP value computation can become a problem in resource-constrained or ultra-low-latency environments. Moreover, even popular benchmark datasets might not be sufficient in terms of diversity and the scale of real-world network traffic. The proposed method's good performance can be explained by the complementary performance of heterogeneous base learners, which jointly model various attack behaviors more effectively than single classifiers.

Adaptive learning counters the effects of concept drift and maintains accurate detection speed over time. Conversely, performance can be compromised when extremely rare or unseen attack types are involved; in such cases, limited labeled data can hinder effective learning and the quality of explanations. Future research can explore lightweight or approximation methods for SHAP to reduce computational cost and eliminate the need for real-time deployment. It is possible to further improve the learner's scalability and privacy preservation by extending the framework to fully online and federated settings. Additionally, domain knowledge and causal explainability methods could be incorporated to improve robustness against adversarial attacks. Overall, this work underscores the importance of optimizing accuracy, adaptability, and interpretability simultaneously when developing the next generation of intrusion detection systems, providing a useful foundation for future enhancements in explainable cybersecurity analytics.

5. Conclusion and Future Work

This study adopts an interpretable and adaptive intrusion detection system (IDS) framework based on a SHAP-enhanced ensemble learning approach to address the increasing complexity of cyber threats in modern society. By combining heterogeneous ensemble models with explainable feature attribution and adaptive learning algorithms, the proposed IDS achieves better detection performance, even very good performance, compared with state-of-the-art methods across various evaluation metrics, such as accuracy, F1-score, and AUC-ROC, and can greatly reduce false positives. SHAP-based explanations provide not only global but also instance-level interpretability, enabling security analysts to understand the model's decisions, validate alerts, and maintain trust without losing effectiveness. The framework is pragmatic and can be used across a variety of cybersecurity environments, including enterprise, cloud, and IoT. Its interpretability is good, thereby aiding threat prioritization and forensic analysis. In contrast, adaptive learning makes it resilient due to the fickle nature of attack patterns and the dynamic nature of networks. Compatibility with standard network traffic features enables seamless integration with existing security information and event management (SIEM) systems, including continuous monitoring and long-term operation. Future work includes improving computational efficiency (to support real applications with large volumes of data), developing more explainable methods, and extending to online, federated, or privacy-preserving learning. Evaluating the model at large scale, in real-world traffic, and under adversarial conditions will further enhance the model's robustness and generalisability, paving the way for the development of trustworthy, scalable, and explainable IDS solutions for the next generation of Cybersecurity.

Acknowledgment: The authors express their sincere gratitude to Al Zahra College for Women, Bharath Institute of Higher Education and Research, and Srinivas University for their continuous support, guidance, and provision of essential resources.

Data Availability Statement: This research presents an interpretable and adaptive intrusion detection framework leveraging SHAP-enhanced ensemble learning and feature attribution techniques for advanced cybersecurity applications. The underlying data are not publicly available due to security and confidentiality considerations, but may be accessed upon reasonable request, subject to institutional policies and approval.

Funding Statement: This manuscript and research paper were prepared without any financial support or funding.

Conflicts of Interest Statement: The authors have no conflicts of interest to declare. This work represents a new contribution by the authors, and all citations and references are appropriately included based on the information utilized.

Ethics and Consent Statement: This research adheres to ethical guidelines and obtains informed consent from all participants. Confidentiality measures were implemented to safeguard participant privacy.

References

1. I. Stellios, P. Kotzanikolaou, and M. Psarakis, "Advanced persistent threats and zero-day exploits in industrial Internet of Things," in Security and Privacy Trends in the Industrial Internet of Things, *Springer International Publishing*, Cham, Switzerland, 2019.
2. Y. Otoum, D. Liu, and A. Nayak, "DL-IDS: A deep learning–based intrusion detection framework for securing IoT," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 3, p. e3803, 2022.
3. Y. Alotaibi and M. Ilyas, "Ensemble-learning framework for intrusion detection to enhance Internet of Things' devices security," *Sensors*, vol. 23, no. 12, p. 5568, 2023.
4. O. Arreche, T. Guntur, and M. Abdallah, "XAI-IDS: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems," *Applied Sciences*, vol. 14, no. 10, p. 4170, 2024.
5. S. B. Goyal, N. Thakur, and S. A. Qadeer, "Bridging the interpretability gap: A SHAP-enhanced framework for intrusion detection in cybersecurity," *The International Journal of Management and Data Intelligence*, vol. 1, no. 1, pp. 1-13, 2025.
6. J. Onyilo and V. C. Uzuegbu, "Explainable AI intrusion detection system: Improving transparency and trust in cybersecurity," *Nature Journal of Emerging Sciences Technologies and Innovations*, vol. 1, no. 1, pp. 85-103, 2025.
7. P. Barnard, N. Marchetti, and L. A. DaSilva, "Robust network intrusion detection through explainable artificial intelligence (XAI)," *IEEE Networking Letters*, vol. 4, no. 3, pp. 167–171, 2022.
8. T. T. H. Le, H. Kim, H. Kang, and H. Kim, "Classification and explanation for intrusion detection system based on ensemble trees and SHAP method," *Sensors*, vol. 22, no. 3, p. 1154, 2022.
9. A. A. Mohammed, A. M. Aleesa, and A. M. Alhatim, "SHAP-enhanced histogram gradient boosting for IoT threat detection via signal-based network traffic analysis," *Diyala Journal of Engineering Sciences*, vol 18, no. 4, pp. 53–72, 2025.
10. M. A. Hossain, S. Saif, and M. S. Islam, "A novel federated learning approach for IoT botnet intrusion detection using SHAP-based knowledge distillation," *Complex & Intelligent Systems*, vol. 11, no. 10, p. 422, 2025.
11. S. Satpathy, U. Tripathy, and P. K. Swain, "Cloud-based DDoS detection using hybrid feature selection with deep reinforcement learning (DRL)," *Scientific Reports*, vol. 15, no. 1, p. 36546, 2025.
12. F. Ebrahimi, R. Javidan, R. Akbari, and Y. Hosseini, "Intrusion detection in the Internet of Things using convolutional neural networks: An explainable AI approach," *Cybersecurity*, vol. 8, no. 1, p. 66, 2025.
13. R. Younisse, A. Ahmad, and Q. A. Al-Haija, "Explaining intrusion detection-based convolutional neural networks using Shapley additive explanations (SHAP)," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 126, 2022.
14. D. Gaspar, P. Silva, and C. Silva, "Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron," *IEEE Access*, vol. 12, no. 2, pp. 30164–30175, 2024.
15. A. Khediri, H. Slimi, A. Yahiaoui, M. Derdour, H. Bendjenna, and C. E. Ghenai, "Enhancing machine learning model interpretability in intrusion detection systems through SHAP explanations and LLM-generated descriptions," *6th Int. Conf. Pattern Analysis and Intelligent Systems (PAIS)*, ELOUED, Algeria, 2024.
16. S. Wali, Y. A. Farrukh, and I. Khan, "Explainable AI and random forest-based reliable intrusion detection system," *Computers & Security*, vol. 157, no. 10, p. 104542, 2025.
17. A. Alabdulatif, "A novel ensemble of deep learning approach for cybersecurity intrusion detection with explainable artificial intelligence," *Applied Sciences*, vol. 15, no. 14, p. 7984, 2025.
18. A. Alsarhan, M. Alauthman, E. A. Alshdaifat, A. R. Al-Ghuwairi, and A. Al-Dubai, "Machine learning-driven optimization for SVM-based intrusion detection system in vehicular ad hoc networks," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 5, pp. 6113–6122, 2023.

19. M. Choubisa, R. Doshi, N. Khatri, and K. K. Hiran, "A simple and robust approach of random forest for intrusion detection system in cybersecurity," *International Conference on IoT and Blockchain Technology (ICIBT)*, Ranchi, India, 2022.
20. S. T. Kotagiri, "Exploring quantum cryptography for next-generation cybersecurity protocols," *AVE Trends in Intelligent Computer Letters*, vol. 1, no. 1, pp. 21–30, 2025.
21. A. El-Ghamry, A. Darwish, and A. E. Hassanien, "An optimized CNN-based intrusion detection system for reducing risks in smart farming," *Internet of Things*, vol. 22, no. 7, p. 100709, 2023.
22. H. C. Altunay and Z. Albayrak, "A hybrid CNN+LSTM-based intrusion detection system for industrial IoT networks," *Engineering Science and Technology, an International Journal*, vol. 38, no. 2, p. 101322, 2023.
23. M. H. L. Louk and B. A. Tama, "Dual-IDS: A bagging-based gradient boosting decision tree model for network anomaly intrusion detection system," *Expert Systems with Applications*, vol. 213, no. 3, p. 119030, 2023.