ADVANCED INTRUSION DETECTION AND CLASSIFICATION USING TRANSFER LEARNING WITH SQUEEZE-AND-EXCITATION NETWORK AND ADAPTIVE OPTIMIZATION IN BIG DATA

Anoop Mohana Kumar[1,2], Joseph Emerson Raja [3], Chinnaiyan Senthilpari [4]

[1] Postdoctoral Researcher, Department of Engineering and Technology, Multimedia University, Malaysia.
[2] Associate Professor, Department of Computer Science and Engineering, Saveetha School of Engineering, SIMATS, India
[3] Assistant Professor, Centre for Advanced Analytics, COE of Artificial Intelligence, Faculty of Engineering and Technology, Multimedia University, Malaysia
[4] Associate Professor, Centre for Advanced Devices and Systems, Multimedia University (MMU), Persiaran Multimedia, Malaysia

ABSTRACT

The rapidly growing number of inter-networked computer devices and upsurge of big data in cyber domain have made cyberattacks, particularly Denial-of-Service (DoS) attacks, a very serious threat. Traditional intrusion detection systems face scaling issues with increasing complexity of big data, while extracting local and global features causes redundancy. This research bridges this gap by integrating transfer learning and Squeeze-and-Excitation Network (SENet) within the proposed eXplainable Artificial Intelligence-driven Intrusion Detection Model (XIIDM). The feature extraction process is through a correlated univariate-elimination-based autoencoder, to preserve local and global features of input data and eliminate all redundant information. SENet further enhances representational power of proposed model by recalibrating channel-wise feature responses, leading to improved DoS classification accuracy. An adaptive partial reinforcement optimizer dynamically adjusts model parameters during training, thus optimizing precision and reducing time complexity. Moreover, incorporation of explainable artificial intelligence units makes the outcome of XIIDM transparent and accountable. The proposed XIIDM is then rigorously evaluated on five benchmark datasets: CSE-CIC-IDS2018, CIC-DDoS2019, NSL-KDD, KDD Cup-99, and UNSW-NB15, achieving 99.988% accuracy, 99.934% precision, and 99.932% recall, with 0.00014% error rate. This research further justifies the robustness and generalization capacity of the proposed model by performing k-fold cross-validation and ablation experiments, confirming its high performance and reliability.

KEYWORDS

Intrusion detection systems, Denial-of-service attacks, Imbalanced dataset, Transfer learning, Cyber threats, Explainable artificial intelligence

1. INTRODUCTION

Cybersecurity has become one of the significant challenges of the digital age in past few years. With the growth of internet and rapid proliferation of network-connected devices, securing online systems and infrastructures is paramount [1-2]. Large-scale expansion of big data, combined with ever-increasing volume, variety, and velocity of the network traffic, is excessively making critical

systems excessively susceptible to cyber-attacks. Sophistication of cyber threats keeps evolving, making conventional security systems inadequate against these emergent risks. DoS attacks stand out as one of the most severe and prevalent cyber threats, disrupting data of legitimate users, hence rendering services inaccessible. This results in higher period of downtime and huge financial implications along with future complications for organizations [3]. Intrusion Detection Systems (IDS) have long since played a vital role in defending systems from such threats. IDS detect any unauthorized access or unusual behavior, indicating potential threat, thus constituting important procedure related to detecting and mitigating security breaches. Although the signature-based detection approach has achieved effective response against well-documented threats, identified critical limitation is that it is not capable of detecting new or previously unseen attacks [45].

In response to these limitations, researchers began exploring Machine Learning (ML) or Artificial Intelligence (AI) strategies to enhance intrusion detection [6-8]. Since ML has great potential to learn from data and evolve with fresh patterns, it is promisingly paving the way against complex cybersecurity challenges. Early techniques in field of intrusion detection, based on ML, mainly relied on supervised learning, wherein classifiers used for creating model are trained on labeled dataset comprising both normal and attack instances [34, 35]. These classifiers include, among others, Decision Trees (DTs), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) [9]. While these techniques show promise, several drawbacks exist, particularly with big data. Further advancements led to the exploration of Deep Learning (DL)-based strategies, more capable of further complex attribute extraction and classification precision in attacks [36-38]. Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) emerge as efficient architectural types of choice for intrusion recognition with the ability to automatically learn an acceptable hierarchical representation of raw data-based features, thereby alleviating the requirement for the manual process of feature engineering. However, it is characterized by the fact that DL models always need significant amounts of labeled data for the training process, and also the demand for computational power increases dramatically [10]. In summary, traditional intrusion detection and early ML approaches made an important contribution to network security, but the operational modeling itself was limited by its inability to scale with added complexity in data and the changing nature of the cyber threat. Hence, an explainable Artificial Intelligence (XAI)-driven model is designed for the classification of DoS attacks. The contributions of the proposed XIIDM are:

- The feature extraction process involves selecting key features based on known domain criteria and exploratory data analysis. The correlated univariate-elimination-based autoencoder has two main components: a feature network, built on a convolutional block, captures the local features from attack data, while the relation network captures global relationships.
- This research integrates transfer learning with SENet to enhance DoS attack detection and multi-level classification. With transfer learning, local and global features are integrated most effectively and comparatively with efficiency, scalability, and accuracy.
- Adaptive Partial Reinforcement Optimizer (APRO), an innovative optimization technique, is introduced to enhance classification accuracy with minimal time complexity. By adjusting model parameters, APRO dynamically adjusts itself to the evolving landscape of big data while training, thus increasing model efficiency and accuracy.

The research is structured with following sections. Section 2 presents a review of the previously existing IDS for intrusion detection and classification; Section3 describes the processes involved in the proposed XIIDM applicable for DoS attack detection and classification; Section4 presents the experimentation findings, and Section 5 completes the research and particularizes on future research areas.

2. RELATED WORKS

Prior methodologies used to detect various cyberattacks with IDS to enforce defense mechanisms against intrusions are discussed. Xie and Chen [11] employed the Convolutional Spiking Neural Network (CSNN) to provide a compact yet impactful solution for intrusion recognition. It functioned reliably in environments with restricted computational and energy resources. By combining eXtreme Gradient Boosting (XGBoost) with Class-Wise Focal Loss (CWFL) and Variational AutoEncoder (VAE), Abdulganiyu et al. [12] developed an XGBoost-based Intrusion Detection of Imbalance Network Traffic via class-wise Focal Loss Variational AutoEncoder (XIDINTFL-VAE) framework to advance the recognition of minority class intrusions while a strong performance overall. Arafah et al. [13] developed an Enhanced Bidirectional Generative Adversarial Network (E-BiGAN) merging an Auto Encoder (AE) and a Wasserstein Generative Adversarial Network (WGAN) to resolve high-dimensional network traffic data with imbalanced classes.

Imtiaz et al. [14] presented the eXplainable Internet of Things (XIoT) model to address the challenges of real-time processing, scalability, and interpretability. XIoT, using CNNs, analyzed network traffic spectrograms to distinguish complex attack patterns while integrating XAI for transparency. Mahmoud et al. [15] introduced the eXplainable Intelligent 2-Stage Intrusion Detection System (XI2S-IDS), combining binary and multi-class classification with explainability using SHapley Additive exPlanations (SHAP) to advance detection and understandability of low-frequency attacks. Berguiga et al. [16] established a Hybrid Intrusion Detection System for Routing Protocol for Low-power and lossy networks (HIDS-RPL). It merged CNN for attribute extraction and Long Short-Term Memory neural network (LSTM) for sequence prediction.

Tserenkhuuet al. [27] suggested an intrusion detection framework for SDN-based IoT networks using DL models enhanced with hyperparameter tuning and XAI-based feature selection. Domain-constrained features were employed to improve detection accuracy and reduce computational complexity. Kilincer [28] developed a Comprehensive Layer 2 - IDS (CL2-IDS) dataset and a hybrid DL model to enhance intrusion detection in organizational local networks. The hybrid model combined CNN and Bidirectional-LSTM architectures to detect attack vectors. Albashayreh et al. [29] established a detection framework designed to automatically detect and mitigate DoS and distributed DoS attacks in next-generation wireless networks. It employed ML and DL techniques to enhance precision and optimize resource utilization for detecting attacks on 5G network's application and transport layers.

Hermosilla et al. [30] addressed challenges of transparency and explainability in DL models used for digital forensics and incident response. XAI methods were applied with XGBoost and TabNet models. Almadhor et al. [31] implemented a DL model to classify obfuscated malware. Transfer learning was used to familiarise model for intrusion detection to create more accurate and less time-consuming detection across different cybersecurity domains. Assadhan et al. [32] classified DDoS attacks from actual network traffic in a more accurate way by using LSTM, an increased interpretability method which pointed to some of the packet traffic features that drove LSTM's predictions.

Compared against recent state-of-the-art models [11-16, 27-32], XIIDM proposes many methodological improvements that strengthen its practical significance. Other frameworks stand alone like CL2-IDS and SDN-IoT IDS, while transfer learning is also integrated with SENet in a way that channel-wise feature responses are dynamically recalibrated to provide more discriminative and scalable feature representations across heterogeneous network environments. Inclusion of APRO further addresses the optimization stagnation and convergence instability commonly associated with deep IDS architectures, facilitating adaptive parameter tuning and

efficient model learning. In addition, correlated univariate-elimination-based autoencoding paired with explainable AI modules effectively overcomes long-standing interpretability challenges of deep models, such as those supporting LSTM and BiLSTM frameworks. Collectively, these design decisions enable XIIDM to attain architectural efficacy, interpretability, and adaptability, which are key practical characteristics for deploying intrusion detection systems in large-scale cybersecurity contexts.

2.1. Problem Statement

Escalating complexity of cyber threats and limitations of traditional IDS demand a robust and adaptive approach to intrusion detection. Conventional methods are often incapable of handling fast-paced and dynamic nature of network traffic. Motivation for this research lies in addressing these challenges by using DL and transfer learning strategies. The Proposed approach aims to enhance detection of DoS attacks while ensuring that false positives and computational complexity are kept to a minimum. Furthermore, integration of XAI approaches, such as SHAP and Local Interpretable Model-agnostic Explanations (LIME), confirms that decision logic is transparent and accountable, which is crucial for building trust in any automated security system. By improving detection accuracy and interpretability, this research contributes to furtherance of security and resilience development of network infrastructures.

3. PROPOSED DENIAL-OF-SERVICE CLASSIFICATION METHODOLOGY

Proposed DoS attack classification methodology, shown in Figure 1, utilizes a comprehensive AI-driven approach to effectively detect and classify DoS attacks. Initially, various datasets are used to collect both benign and attack data. To overcome class imbalance commonly found in large datasets, Deep Convolutional Generative Adversarial Network (DCGAN) is employed to produce artificial data for minority groups, while Synthetic Minority Oversampling TEchnique (SMOTE) further balances the dataset. Data is preprocessed through normalization of numerical columns and one-hot encoding of categorical features to optimize dataset for DL models. Feature extraction, based on a correlated univariate-elimination-based autoencoder, captures fine local features from narrower perspective and broad global relationships from within data. XIIDM also integrates transfer learning with SENet to further improve feature representation and enhance performance classification. Additionally, APRO is introduced to enhance model optimization, thus improving classification accuracy and minimizing training time. To ensure transparency and liability in model's decision logic, SHAP and LIME techniques are used to provide readable explanations of feature importance and the model predictions. This all-inclusive approach results in robust, efficient, and explicable DoS detection and classification system.

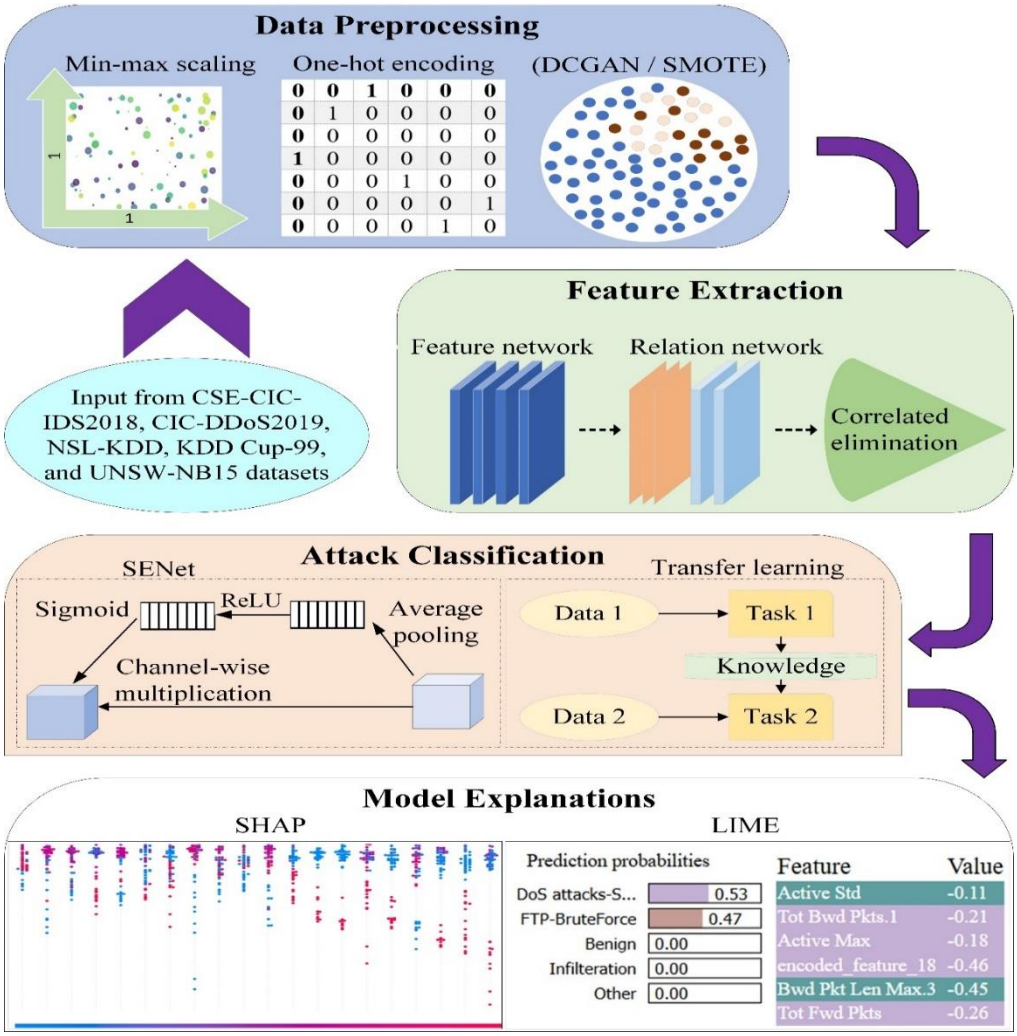


Figure 1. Proposed architecture of XIIDM for detection and classification of DoS attacks

3.1. Datasets

Data samples of DoS and benign instances are collected from CSE-CIC-IDS2018 [17], CIC-DDoS2019 [18], NSL-KDD [19], KDD Cup-99 [20], and UNSW-NB15 [21] datasets.

a) CSE-CIC-IDS2018: It contains seven discrete attack scenarios, like Brute Force, online attack, DoS, infiltration, and heartleech, with 83 samples each with unique features.
b) CIC-DDoS2019: It includes 56,000 assault rows and 56,000 benign traffic rows, each with 88 features.
c) NSL-KDD: It has 67,342 benign and 58,630 malicious records, with attacks categorized as DoS (45,927 records), Remote to Local (R2L, 955 records), User to Root (U2R, 52 records), and Probe (11,656 records).
d) KDD Cup-99: It includes 494,020 samples and 41 attributes, covering R2L, U2R, DoS, and Probe attacks.
e) UNSW-NB15: It includes 2,540,044 entries separated into four CSV archives.

Datasets used in this research were from Kaggle due to open accessibility, standardized formatting, and reliable preprocessing to allow reproducibility of research. Kaggle stores several benchmark

datasets for intrusion detection studies, and this ensures that evaluations are consistent and thus allow comparison with existing models on fair terms. Coupled with other forms of attack behaviors and network traffic patterns, these datasets, therefore, make for a proper setting for validating the robustness and generalization of XIIDM. More specifically, CSE-CIC-IDS2018 comprises over 80 network flow features, for instance, source/destination IP, flow duration, and packet length statistics; CIC-DDoS2019 includes 88 features associated with statistical traffic features such as flow bytes per second and packet inter-arrival times; NSL-KDD and KDD Cup-99 hold 41 and 42 attributes respectively, covering basic TCP/IP connection features (protocol type, service, flag), content-based features (failed logins, file creation), and traffic-based features (count, srv_count); while UNSW-NB15 contains 49 features classified into nine categories such as flow, basic, content, time, plus pertinent attributes related to attack. These datasets as a whole ensure equilibrium amidst representations of classical and modern network attacks, which assist both in the scalability and credibility of model evaluation.

3.2. Data Preprocessing

Preprocessing and augmentation methods are applied to improve robustness and generalization of XIIDM, resulting in better recognition of minority attack classes.

a) Data normalization: Min-max scaling is used to normalize the numeric features. It scales feature values to fit within a specific range between 0 and 1. Min-max scaling is expressed through equation (1).

$$A' = \frac{A - A_{mn}}{A_{mx} - A_{mn}} \tag{1}$$

where, A' is scaled data point, A is original data point, A_{mn} is minimum value, and A_{mx} is maximum data point in feature column.

b) Handling categorical features: One-hot encoding transforms categorical variables into binary columns.

c) Handling class imbalance: XIIDM uses DCGAN [22] to generate synthetic data for underrepresented DoS attacks and SMOTE to create new samples by interpolating between existing minority class instances, thus balancing datasets [33].

3.3. Feature Extraction

Proposed correlated univariate-elimination-based autoencoder [23] integrates a hybrid architecture for local feature extraction with global relationship modeling making it capable of processing network traffic data with effectiveness. These compromises modeling borderline patterns (local features) with more encompassing patterns (global relationships) while concurrently relieving itself from any redundancy by eliminating correlated features. This proposed method for feature extraction consists of two main components: feature network is responsible for capturing local features from the data, whereas relation network provides global relationships between the features. All steps of correlated univariate elimination strongly help to curb redundancy in features, thereby ensuring that only information remaining is relevant for effective learning and analysis.

3.3.1. Feature Network

Feature network extracts local features from input data by utilizing convolutional layers. The local features reflect small-scale patterns regarding certain attack signatures or anomalies in traffic appearing at isolated parts of the data. The network thus applies convolutional operations to capture fine-grained patterns in raw data which signal a possible DoS attack or benign behavior.

3.3.2. Relation Network

Relation network extracts global relationships from local features produced by the feature network. It holds dependencies and interactions between features across the entire dataset. This matters for understanding broader patterns, like large-scale strategies for attack or network-wide behaviors.

3.3.3. Correlated Univariate Elimination

Through correlated univariate elimination, redundant or highly correlated features are removed to improve the efficiency and robustness of feature extraction process. The Features which are too similar or correlated with each other don't contribute new information, therefore these are eliminated, subsequently reducing dimensionality of data. Correlated univariate elimination step is executed by estimating pairwise correlations between features and dismissing the features which the surpassed a certain threshold, thereby guaranteeing that only non-redundant and informative features are preserving for later analysis.

After feature network, relation network, and correlated univariate elimination process, the model merges the output from two networks (local and global features) into a final feature representation. This combined feature vector is then employed for the ensuing classification task. The final feature representation is defined as concatenation of local features and global relationships. It enables the model to extract local and global patterns for improved detection performance toward DoS attacks, featuring local anomalies (traffic spikes) and global attributes (distributed attack behavior).

3.4. Denial-of-Service Attack Detection and Multi-Level Classification

Transfer learning is integrated with SENet [24] to enhance attack detection and multi-level classification. Extracted features from the correlated univariate-elimination-based autoencoder are passed on to a transfer learning pre-trained CNN model for classifying network traffic data. After fine-tuning the pre-trained model, the transfer learning model generates a collection of advanced features (feature maps) that represent critical patterns related to network traffic and attack detection. This includes both local features (packet-level data) and global features (traffic patterns or long-term dependencies).

SENet adaptively recalibrates channel-wise feature responses by learning which features are more important in each layer of the network. This channel-wise attention enhances the representational power of CNN. The feature maps from transfer learning are input into the SENet, where the squeeze operation computes the global information by applying Global Average Pooling (GAP) to the convolutional layers' feature maps. GAP pools the spatial information from all positions in a channel into a single scalar value. The result is a squeezed representation, S, of the feature maps for each channel, as depicted in equation (2).

$$S = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j-1}^{W} D_{i,j,c} \tag{2}$$

where, $D_{i,j,c}$ represents the input feature representation, and H and W are the altitude and breadth of the feature representation. This gives a scalar value for each channel c.

After squeezing the spatial information into a global context, during the excitation operation, the network learns to recalibrate each channel's (feature map) importance using a fully connected layer followed by a sigmoid activation. This allows the network to decide which channels are more significant and to be emphasized or de-emphasized during the processing of features. The output from SENet stands, ultimately, as recalibrated feature maps on which the model learned to focus more while considering features that help distinguish between attack and benign traffic. Thus, critical attack patterns are emphasized, while irrelevant or noisy data is down-weighted. Hence, transfer learning with SENet proves extremely effective in building a robust and efficient intrusion detection model that improves accuracy in attack classification.

3.5. Model Parameter Optimization

The standard Partial Reinforcement Optimizer (PRO) [25] is an evolutionary algorithm that balances exploration and exploitation through partial reinforcement, improving DoS attack classification speed and effectiveness. Candidate solutions are evaluated based on the learner's performance (XIIDM), with reinforcement timing defined by equation (3).

$$T = \frac{E_v}{Max_{E_v}} \tag{3}$$

where, E_v is the number of evaluations during reinforcement, Max_{E_v} is the total permissible evaluations, and T represents the normalized time between two consecutive reinforcements. Selection rate, stimulation factor, and learner updates are driven by fitness improvements and priority scores. Reinforcement scheduling adapts based on fitness and variability to maintain exploration. APRO enhances PRO by dynamically adjusting reinforcement parameters to avoid premature convergence and improve optimization.

3.6. Explanations with Explainable Artificial Intelligence

XAI helps researchers understand AI decisions, making the process more transparent and easier to follow. The proposed DoS classification model consists of two post-hoc local XAI blocks to explain and visualize feature importance so that the model decisions become understandable. The SHAP values calculated on cooperative game theory concepts give a global perspective on the influence of each feature on model forecasts. LIME explains XIIDM's performance in the context of particular predictions. This helps to comprehend XIIDM's decision logic for particular instances, offering valuable perceptions of the local decision boundaries [26]. Together, SHAP and LIME improve model transparency and detect potential vulnerabilities and biases present in the DoS classification model, confirming better trust and accountability in automated decision-making.

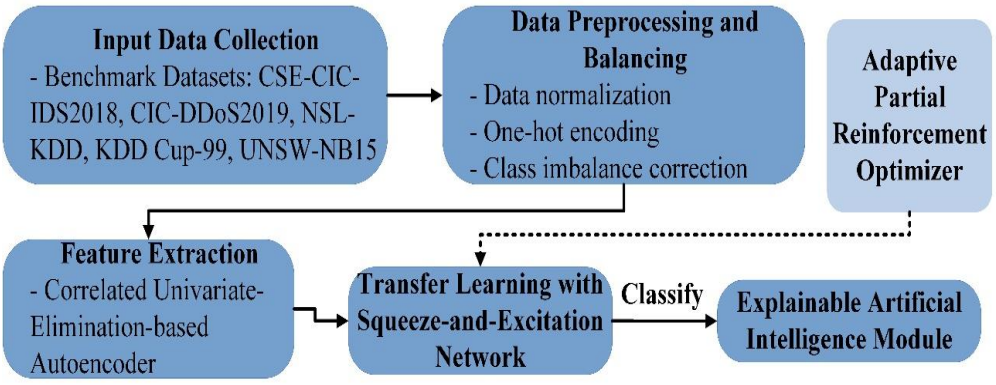


Figure 2. Workflow of XIIDM for intrusion detection and classification

Thus, XIIDM unifies advanced mechanisms into an explainable intrusion detection pipeline. It begins with multi-source data preprocessing, normalization, one-hot encoding, and class rebalancing via DCGAN and SMOTE, followed by feature extraction through a correlated univariate-elimination-based autoencoder to retain key dependencies while reducing redundancy. These features feed into SENet for refined feature recalibration, optimized using APRO for faster, stable convergence. Finally, SHAP and LIME ensure global and local interpretability, making the model's decisions transparent. A Complete end-to-end workflow is illustrated in Figure 2.

4. EXPERIMENTAL RESULTS

Performance of XIIDM is evaluated on the five datasets, focusing on DoS attack detection and classification. Efficiency of XIIDM is assessed in various assessment metrics, like accuracy (%), precision (%), specificity (%), processing time (seconds (sec)), and error rate (%). It is implemented in Python using PyTorch library, and all experiments are conducted on CPU with base speed of 3.40 GHz, featuring4coresand8logical processors.

Table 1. Experimental setup of XIIDM

Parameters	**Values**
Optimizer	APRO
Learning rate (LR)	0.001 with decay factor of 0.9
Batch size	128
Epochs	100
Dropout rate	0.3
Activation function	ReLU

Table 1 summarizes experimental configuration adopted to evaluate XIIDM across all benchmark datasets.

4.1. Proposed XIIDM's Performance Analysis

Performance evaluation of XIIDM is presented in Table 2, achieving near-perfect scores for accuracy, precision, recall, and F-measure across all five datasets. Accuracy scores of 99.96% to 100% indicate that XIIDM is reliable with classification error ranging from 0.0000% to 0.0004%, thus attesting its reliability in DoS attack detection. High specificity scores (99.98%-100%) indicate that XIIDM efficiently separates normal from attack traffic and thus minimizes false

positives. The results show that XIIDM outperforms traditional systems in precision and robustness.

Table 2. Performance assessment of XIIDM

Dataset/Metrics	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	Specificity (%)	Error rate (%)	Computationaltime(sec)
CSE-CIC-IDS2018	99.99	99.95	99.94	99.94	100	0.0001	61.1141
CIC-DDoS2019	99.99	99.95	99.95	99.95	99.99	0.0001	50.2936
NSL-KDD	100	99.98	99.98	99.98	100	0.0001	143.9598
KDD Cup-99	100	99.99	99.99	99.99	100	0.0000	136.1562
UNSW-NB15	99.96	99.80	99.80	99.80	99.98	0.0004	43.1213

Table 3 shows quantitative evaluation of XIIDM across different attack types and demonstrates its strong generalization and detection beyond conventional DoS attacks. In NSL-KDD, XIIDM detectsProbe and R2L intrusions, achieving 99.82% and 98.94% accuracy, while achieving 97.36% in KDD Cup-99 for U2R and as high as 99.93% accuracy with effective F-measure of 99.90% and detection rate of 99.94% for different DoS subtypes.

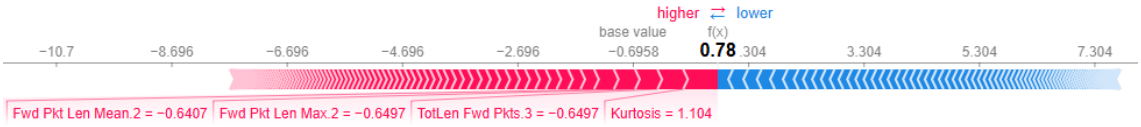
Table 3. Performance assessment of XIIDM on additional attack categories

Datasets	Attack types	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	Detection rate (%)
NSL-KDD	Probe	99.82	99.75	99.77	99.76	99.77
	R2L	98.94	98.85	98.72	98.78	98.72
KDD Cup-99	U2R	97.36	97.12	96.85	96.98	96.85
	DoS (Neptune, Smurf, Teardrop, Back)	99.93	99.89	99.91	99.90	99.91
CSE-CIC-IDS2018	Brute Force	99.71	99.65	99.63	99.64	99.63
	Infiltration	99.61	99.49	99.57	99.53	99.57
CIC-DDoS2019	DDoS	99.95	99.91	99.94	99.92	99.94
UNSW-NB15	Generic / Fuzzers	99.45	99.30	99.34	99.32	99.34
	Backdoor	98.87	98.73	98.69	98.71	98.69

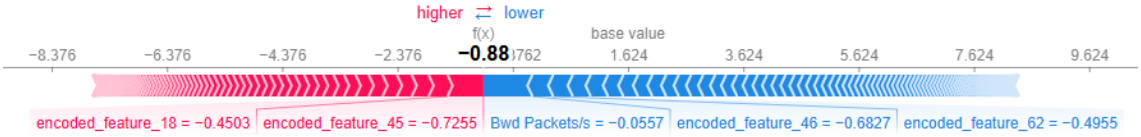
Similarly, in CSE-CIC-IDS2018, Brute Force and Infiltration attacks are detected with 99.71% and 99.61% accuracy, and in CIC-DDoS2019 and UNSW-NB15, it yields 99.95% and 99.45%/98.87% for DDoS, Generic/Fuzzer, and Backdoor attacks, respectively. These results (97.3%-99.9%) highlight XIIDM's robustness and adaptability to heterogeneous intrusion behaviors, supported by DCGAN and SMOTE for class balance and correlated univariate-elimination autoencoder for preserving key feature dependencies, validating model's stability and unbiased learning across datasets.

4.2. Explanation Illustrations

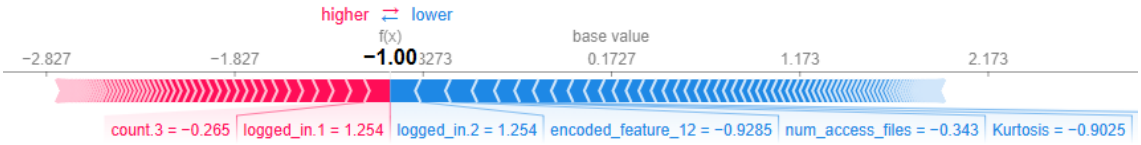
Descriptive instances of explanations with XAI for classification tasks are provided for CSE-CIC-IDS2018, CIC-DDoS2019, NSL-KDD, KDD Cup-99, and UNSW-NB15 datasets.



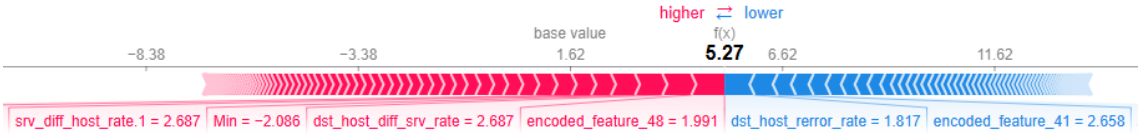
(a)



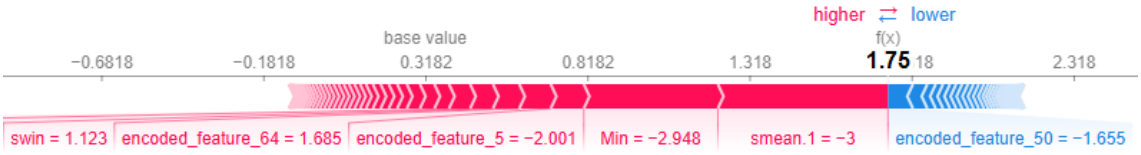
(b)



(c)



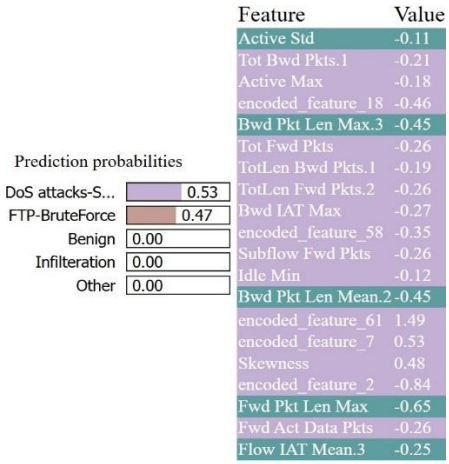
(d)



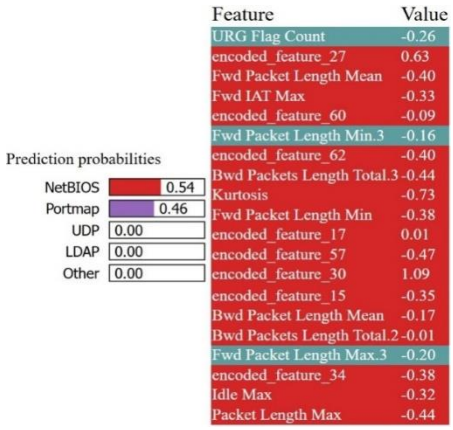
(e)

Figure 3. SHAP force plot for local explanation of single prediction for (a) CSE-CIC-IDS2018, (b) CIC-DDoS2019, (c) NSL-KDD, (d) KDD Cup-99, and (e) UNSW-NB15

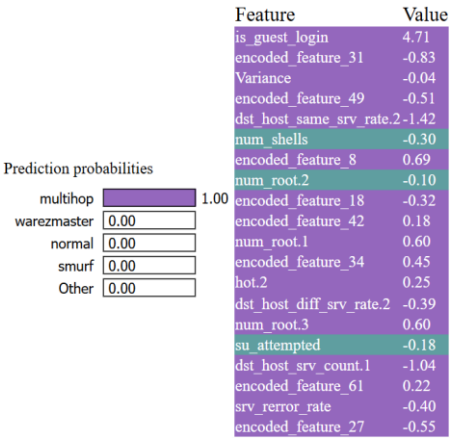
Figure3 shows a local explanation for a single forecast, with contributions of each feature displayed to explain the shift in XIIDM's output from the base value (average prediction) to the final forecast. The red bars denote features that push the prediction in the direction of the predicted class, while the blue bars show features that oppose it. Each of these bars is annotated with the name and value of the variable, indicating its influence on the prediction.



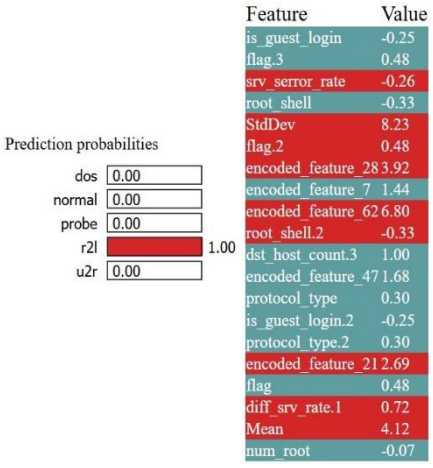
(a)



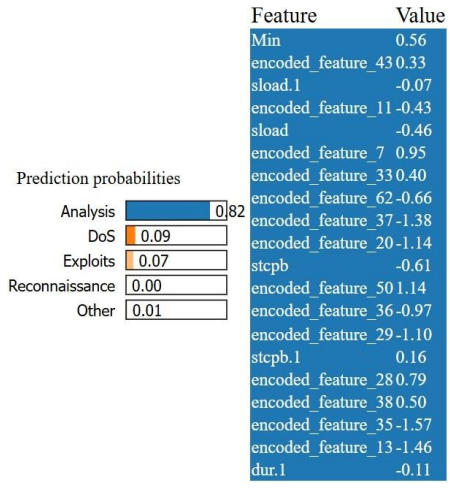
(b)



(c)



(d)



(e)

Figure 4. Local explanation for classification tasks (attack sample) with LIME for (a) CSE-CIC-IDS2018, (b) CIC-DDoS2019, (c) NSL-KDD, (d) KDD Cup-99, and (e) UNSW-NB15

An explanation of XIIDM's prediction for an instance is given in Figure4, which approximates locally the prediction with an interpretable one. The left side shows the model predicting the different classes' probabilities, with the most likely class predicted with a very high confidence score. The right side indicates the most influential features and the values for this particular prediction, sorted according to the features' contribution to the final prediction.

Integration of SHAP and LIME enhances interpretability and offers actionable insights for cybersecurity practitioners. Case-level analyses show that features like flow duration, packet length mean, and destination port are key indicators of DoS attacks in SHAP results, helping prioritize monitoring and adjust detection thresholds. LIME explanations allow analysts to trace anomalies and validate predictions, improving response. Table 4presents examples of these interpretability tools supporting network defense decisions.

Table 4. Case-level interpretability outcomes using SHAP and LIME for operational decision-making

Datasets	**Top contributing features (SHAP/LIME)**	**Interpretation insight**	**Practical decision**
CSE-CIC-IDS2018	Flow Duration, Packet Size Mean, Destination Port	High flow duration and large packet size indicate abnormal sustained traffic patterns	Adjust IDS threshold and trigger DoS-specific firewall rule
CIC-DDoS2019	Source IP Count, Flow Bytes/s, SYN Flag Count	Excessive SYN requests and high byte rate suggest SYN flood behavior	Block suspicious IPs; deploy rate-limiting mechanisms
NSL-KDD	Service Type, Logged-in, Count	Repeated login attempts over same service	Implement account lockout and enhanced authentication
KDD Cup-99	Protocol Type, Src Bytes, Diff_srv_rate	Abnormal traffic distribution across services	Enable dynamic service-based anomaly filters
UNSW-NB15	Duration, Inbound Packets, TCP Window Size	Unusual inbound packet growth and unstable TCP window	Reconfigure flow monitoring for early-stage DoS alerts

4.3. Comparative Experiments

Figures 5-9 present XIIDM's comparison with existing models. Figure5 indicates that XIIDM achieves superior accuracy of 99.99% and best recall and F-measure scores on CSE-CIC-IDS2018, compared to CSNN [11], XIDINTFL-VAE [12], Hybrid IDS [14], and spikformer [11].

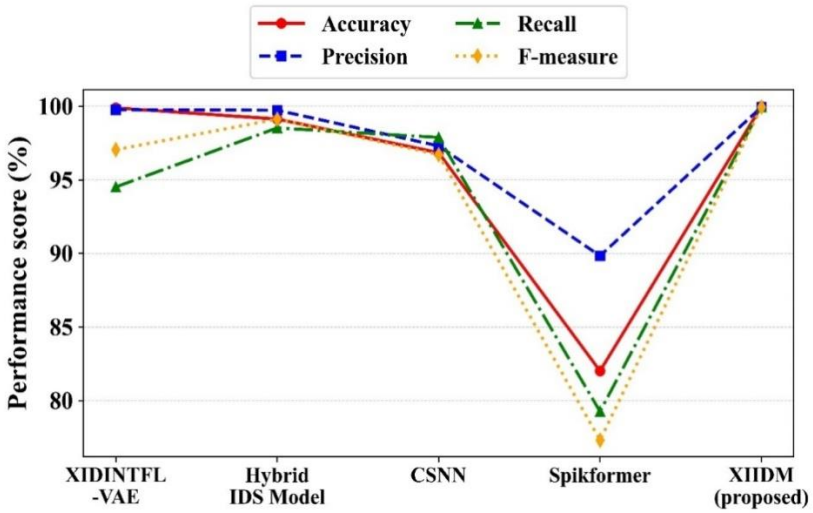


Figure 5. Evaluation of XIIDM with prevailing methods on CSE-CIC-IDS2018

Figure 6 strengthens the validity of XIIDM over CSNN, HIDS-RPL [16], and spikformer because it surpasses these models based on key assessment metrics on CIC-DDoS2019. In particular, XIIDM achieved an accuracy rate of 99.99%, while CSNN achieved 98.76%, HIDS-RPL 99.87%, and spikformer 93.68%.

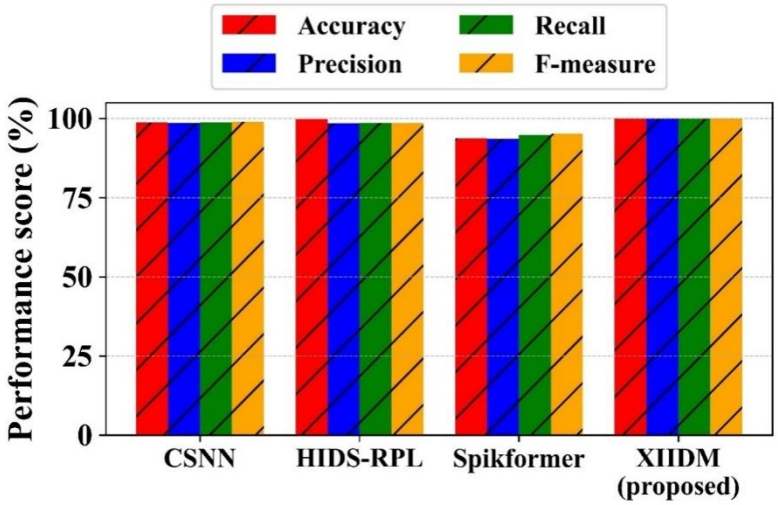


Figure 6. Evaluation of XIIDM with prevailing methods on CIC-DDoS2019

Figure 7 indicates XIIDM surpassing E-BiGAN [13], Random Forest (RF) [14], Fuzzy [14], and XIDINTFL-VAE on NSL-KDD with an accuracy of 100%. E-BiGAN comes close with 98%, RF has a lead with 99.66%, Fuzzy earned 95.32%, and XIDINTFL-VAE is at 99.79%, demonstrating its better capability in intrusion detection.

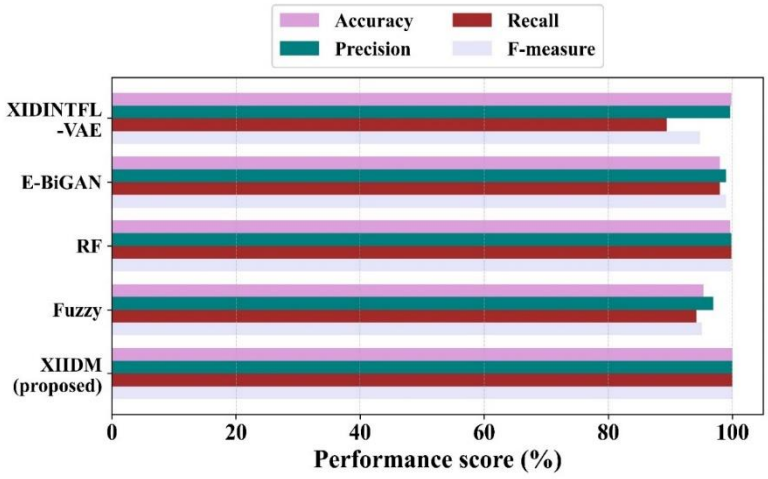


Figure 7. Evaluation of XIIDM with prevailing methods on NSL-KDD

Figure 8 shows XIIDM performing better than Capsule neural Network (CapsNet) and BiLSTM network [16] on KDD Cup-99 with outstanding 100% accuracy. In comparison, CapsNet achieved 96%, and BiLSTM achieved 97%, demonstrating XIIDM's potential to handle complex attack patterns and network traffic.

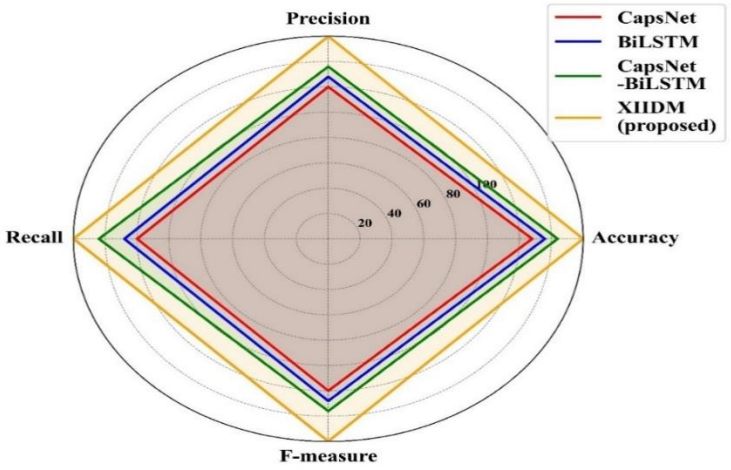


Figure 8. Evaluation of XIIDM with prevailing methods on KDD Cup-99

Figure 9 shows XIIDM, which achieves an accuracy of 99.96% on UNSW-NB15, performing better than Hybrid IDS, XIoT [14], XI2S-IDS [15], and VAE [12]. In comparison, Hybrid IDS produced an accuracy of 99.5%. XIoT achieved 98.34%, while XI2S-IDS reached 97%, whereas VAE got 93%. This shows XIIDM's improved detection capability across various types of network attacks.

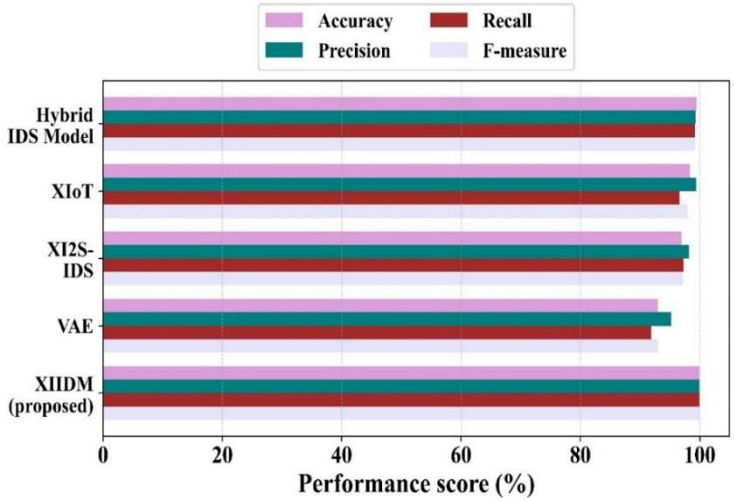


Figure 9. Evaluation of XIIDM with prevailing methods on UNSW-NB15

Near-perfect performance metrics of XIIDM are worth putting in perspective. First, the benchmark datasets chosen for this evaluation are well-structured and widely chosen for intrusion detection research. Secondly, XIIDM is compared against numerous state-of-the-art models (Figures5-9) and demonstrates superior performance in terms of accuracy, precision, and recall. This stellar performance is credited to a convergence of technological advancements: generation of balanced dataset through DCGAN and SMOTE, a powerful feature extraction mechanism via correlated univariate-elimination-based autoencoder, adaptive optimization through APRO, and a robust classification system through transfer learning using SENet.

4.4. K-Fold Cross-Validation

Key performance measures are recorded in 2, 4, 6, 8, and 12 folds to assess XIIDM's capability of generalization in detection of DoS attacks.

Table 5. K-fold test analysis of XIIDM on CIC-DDoS2019

Fold	Accuracy (%)	Precision (%)	Recall (%)
2	99.12	99.47	99.56
4	99.47	99.38	99.48
6	99.25	99.19	99.57
8	99.37	99.72	99.87
12	99.99	99.99	99.99
Average	99.44	99.55	99.69

Exceptionally, average scores for all folds are very high, which represents ability of XIIDM to perform highly on varying datasets. Specifically, on CIC-DDoS2019 (Table 5), it attained highest accuracy of 99.44%, precision of 99.55%, and recall of 99.69%.

4.5. Ablation Experiment

Ablation study serves to provide insights into impact of each component on performance capability of XIIDM.

Table 6. Hyperparameters-based ablation analysis of XIIDM

Configurations	LR	Batch size	Epochs	Accuracy (%)	Precision (%)	Recall (%)	Error rate (%)
A	0.001	128	100	99.99	99.93	99.93	0.0001
B	0.01	128	100	98.42	98.35	98.48	0.0021
C	0.0001	128	100	99.21	99.08	99.15	0.0009
D	0.001	64	100	99.53	99.47	99.49	0.0005
E	0.001	256	100	98.97	98.88	98.91	0.0010
F	0.001	128	50	98.65	98.57	98.60	0.0013
G	0.001	128	150	99.87	99.82	99.84	0.0002

The hyperparameters-based ablation study (Table 6) examines configurations A-G, where A (LR = 0.001, batch = 128, epochs = 100) represents the optimal baseline, B (0.01, 128, 100) uses an increased LR, C (0.0001, 128, 100) a decreased LR, D (0.001, 64, 100) a smaller batch size, E (0.001, 256, 100) a larger batch size, F (0.001, 128, 50) fewer epochs, and G (0.001, 128, 150) more epochs. Among these, Configuration A achieves highest accuracy (99.99%) and lowest error rate (0.0001%), confirming it as the effective training setup for XIIDM.

To further strengthen interpretability of ablation analysis and highlight individual effect of each architectural component, an extended ablation experiment was conducted. As shown in Table 7, each module independently contributes to improving detection performance. Base CNN starts at 94.85%, but improves as more modules are added: DCGAN (97.46%), SMOTE (96.98%), autoencoder (98.25%), SENet (98.74%) and APRO (99.02%). Complete configuration of XIIDM including all parts has almost perfect results of 99.99% in accuracy, 99.93% precision, recall, and F-measure, and low error rate of 0.0001%. This shows important contribution of all components to overall performance.

Table 7. Comprehensive ablation experiments showing independent contribution of each XIIDM component

Configurations	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	Error Rate (%)
Base CNN (without any enhancement)	94.85	95.12	94.67	94.89	0.51
Base + DCGAN	97.46	97.62	97.35	97.48	0.26
Base + SMOTE	96.98	97.03	96.79	96.91	0.31
Base + Autoencoder (Feature extraction)	98.25	98.51	98.13	98.32	0.18
Base + SENet (Transfer learning)	98.74	98.85	98.66	98.75	0.15
Base + APRO (Optimization)	99.02	99.12	99.06	99.09	0.10
Full XIIDM (All components integrated)	99.99	99.93	99.93	99.93	0.0001

4.6. Discussion

Recent advances in DL and XAI have enhanced IDS [11-16, 27-32], but such advancements are still limited by challenges of scalability, interpretability, and optimization efficiency. The proposed XIIDM addresses all these areas of concern by bringing together transfer learning with SENet, correlated univariate-elimination-based autoencoding, and APRO, paving way for enhanced detection precision, computation efficiency, and transparency. SHAP and LIME are further

embedded within XIIDM for global and local insights on its decision-making (Table 4, Figures 3-4). Its autoencoder efficiently eliminates the redundancy of features while at the same time preserving dependencies in data, resulting in powerful performance through heterogeneous datasets without needing retraining, as reflected in Tables 2 and 3. Through APRO, convergence is accelerated, and with the computation cost reduced, XIIDM becomes suitable for large-scale, resource-constrained environments, as shown in Tables 5-7.

Comparative experiments using state-of-the-art frameworks, like CL2-IDS and XIoT, show XIIDM achieving up to 99.99% accuracy with minimal error rates, setting new benchmark in both precision and explainability (Figures 5-9). Limitations still remain with regard to the fact that validation was only done using offline benchmark datasets, which do not account for performance in real-time streaming; this causes a potential latency in live conditions of network operation.

5. CONCLUSION

The proposed XIIDM for the classification of DoS attacks enhances research in cybersecurity significantly by overcoming limitations of conventional systems in classifying and detecting DoS attacks. Proposed classification model uses transfer learning with SENet and APRO to achieve unparalleled accuracy (99.988%), precision (99.934%), recall (99.932%), and specificity (99.994%) on multiple benchmark datasets. XAI techniques, including SHAP and LIME, lend further transparency and liability, making decision logic of XIIDM more interpretable and reliable. The Thoroughness of k-fold cross-validation and ablation experiments proves the resilience and generalization ability of XIIDM. Thus, this research outlines effective means to detect complex cyber threats and serves as a basis for further development of intelligent and adaptive cybersecurity systems. Future work will focus on designing a hybrid IDS with optimization algorithms on an open-source, unified stream and batch processing framework for accurate attack filtering and data storage. This ensures real-time analysis and scalability in big data environments.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENT

None

DATA AVAILABILITY STATEMENTS

The data supporting the findings of this study are cited and the link is listed in the references.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] M.M. Rahman, S.A. Shakil, and M.R. Mustakim, (2025), “A survey on intrusion detection system in IoT Networks,” *Cyber Security and Applications*, Vol. 3, p. 100082.

[2] L. Diana, P. Dini, and D. Paolini, (2025), “Overview on intrusion detection systems for computer networking security,” *Computers*, Vol. 14, No. 3, p. 87.

[3] C. Hamroun, A. Fladenmuller, M. Pariente, and G. Pujolle, (2025), "Intrusion detection in 5G and Wi-Fi Networks: A survey of current methods, challenges, and perspectives," *IEEE Access*, Vol. 13, pp. 40950–40976.

[4] A.H. Janabi, T. Kanakis, and M. Johnson, (2024), "Survey: Intrusion detection system in software-defined networking," *IEEE Access*, Vol. 12, pp. 164097–164120.

[5] V. Sharma, (2025), "Comparative analysis of machine learning models for intrusion detection systems," *Panamerican Mathematical Journal*, Vol. 35, No. 3s, pp. 273–285.

[6] S.K. Mallidi and R.R. Ramisetty, (2025), "Advancements in training and deployment strategies for AI-based Intrusion Detection Systems in IoT: A systematic literature review," *Discover Internet of Things*, Vol. 5, No. 1.

[7] V.Z. Mohale and I.C. Obagbuwa, (2025), "A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhance transparency and interpretability in cybersecurity," *Frontiers in Artificial Intelligence*, Vol. 8, p. 1526221.

[8] B.R. Kikissagbe and M. Adda, (2024), "Machine learning-based Intrusion Detection Methods in IoT Systems: A comprehensive review," *Electronics*, Vol. 13, No. 18, p. 3601.

[9] F. Nabi and X. Zhou, (2024), "Enhancing intrusion detection systems through dimensionality reduction: A Comparative Study of Machine Learning Techniques for Cyber Security," *Cyber Security and Applications*, Vol. 2, p. 100033.

[10] Y. Zhang, R.C. Muniyandi, and F. Qamar, (2025), "A review of deep learning applications in intrusion detection systems: Overcoming challenges in spatiotemporal feature extraction and data imbalance," *Applied Sciences*, Vol. 15, No. 3, p. 1552.

[11] Y. Xie and H. Chen, (2024), "A novel method for effective intrusion detection based on convolutional speaking neural networks," *Journal of King Saud University - Computer and Information Sciences*, Vol. 36, No. 2, p. 101975.

[12] O.H. Abdulganiyu, T.A. Tchakoucht, Y.K. Saheed, and H.A. Ahmed, (2025), "XIDINTFL-VAE: XGBoost-based intrusion detection of imbalance network traffic via class-wise focal loss variational autoencoder," *The Journal of Supercomputing*, Vol. 81, No. 1, pp. 1–38.

[13] M. Arafah, I. Phillips, A. Adnane, M. Alauthman, and N. Aslam, (2025), "An enhanced Bigan architecture for network intrusion detection," *Knowledge-Based Systems*, Vol. 314, p. 113178.

[14] N. Imtiaz, A. Wahid, S. Z. Ul Abideen, M. M. Kamal, N. Sehito, S. Khan, B. S. Virdee, L. Kouhalvandi, and M. Alibakhshikenari, (2025), "A deep learning-based approach for the detection of various internet of things intrusion attacks through optical networks," *Photonics*, Vol. 12, No. 1, p. 35.

[15] M.M. Mahmoud, Y.O. Youssef, and A.A. Abdel-Hamid, (2025), "Xi2s-ids: An explainable intelligent 2-stage intrusion detection system," *Future Internet*, Vol. 17, No. 1, p. 25.

[16] A. Berguiga, A. Harchay, and A. Massaoudi, (2025), "Hids-RPL: A hybrid deep learning-based intrusion detection system for RPL in Internet of Medical things network," *IEEE Access*, Vol. 13, pp. 38404–38429.

[17] *https://www.kaggle.com/datasets/solarmainframe/ids-intrusion-csv (accessed on 3rd April 2025).*

[18] *https://www.kaggle.com/datasets/dhoogla/cicddos2019 (accessed on 3rd April 2025).*

[19] *https://www.kaggle.com/datasets/hassan06/nslkdd (accessed on 3rd April 2025).*

[20] *https://www.kaggle.com/code/girishji/kdd-cup-99-intrusion-detection-lda-randomforests (accessed on 3rd April 2025).*

[21] *https://www.kaggle.com/datasets/alextamboli/unsw-nb15 (accessed on 3rd April 2025).*

[22] M. Shoman, T. Ghoul, G. Lanzaro, T. Alsharif, S. Gargoum, and T. Sayed, (2024), "Enforcing Traffic Safety: A deep learning approach for detecting motorcyclists' helmet violations using Yolov8 and deep convolutional generative adversarial network-generated images," *Algorithms*, Vol. 17, No. 5, p. 202.

[23] S. Bijari, S. Sayfollahi, S. Mardokh-Rouhani, S. Bijari, S. Moradian, Z. Zahiri, and S. M. Rezaeijo, (2024), "Radiomics and deep features: Robust classification of brain hemorrhages and reproducibility analysis using a 3D Autoencoder neural network," *Bioengineering*, Vol. 11, No. 7, p. 643.

[24] S. Surati, H. Trivedi, B. Shrimali, C. Bhatt, and C.M. Travieso-González, (2023), "An enhanced diagnosis of monkeypox disease using deep learning and a novel attention model set on a diversified dataset," *Multimodal Technologies and Interaction*, Vol. 7, No. 8, p. 75.

[25] A. Taheri, K. RahimiZadeh, A. Beheshti, J. Baumbach, R. V. Rao, S. Mirjalili, and A. H. Gandomi, (2024), "Partial reinforcement optimizer: An evolutionary optimization algorithm," *Expert Systems with Applications*, Vol. 238, p. 122070.

[26] O. Arreche, T.R. Guntur, J.W. Roberts, and M. Abdallah, (2024), "E-xai: Evaluating black-box explainable AI frameworks for network intrusion detection," *IEEE Access*, Vol. 12, pp. 23954–23988.

[27] M. Tserenkhuu, M.D. Hossain, Y. Taenaka, and Y. Kadobayashi, (2025), "Intrusion detection system framework for SDN-based IoT networks using deep learning approaches with XAI-based feature selection techniques and domain-constrained features," *IEEE Access*, vol. 13, pp. 136864 – 136880.

[28] I.F. Kilincer, (2025), "Explainable AI supported hybrid deep learning method for layer 2 intrusion detection," *Egyptian Informatics Journal*, vol. 30, No. 100669.

[29] A. Albashayreh, S. Al-Sharaeh, Y. Tashtoush, and P. Zahariev, (2025), "Enhancing 5G network security: A deep learning framework for real-time DDoS detection and explainable threat analysis," *IEEE Access*, vol. 13, pp. 168217 – 168235.

[30] P. Hermosilla, M. Díaz, S. Berríos, and H. Allende-Cid, (2025), "Use of explainable artificial intelligence for analyzing and explaining intrusion detection systems," *Computers*, vol. 14, No. 5, p. 160.

[31] A. Almadhor, S. Alsubai, N. Kryvinska, A. Al Hejaili, B. Bouallegue, M. Ayari, and S. Abbas, (2025), "Transfer learning with XAI for robust malware and IoT network security," *Scientific Reports*, vol. 15, No. 1, p. 26971.

[32] B. Assadhan, A. Bashaiwth, and H. Binsalleeh, (2024), "Enhancing explanation of LSTM-based DDoS attack classification using SHAP with pattern dependency," *IEEE Access*, vol. 12, pp. 90707–90725.

[33] Alshamy, Reem, and MUHAMMET AKCAYOL. "Intrusion Detection Model Using Machine Learning Algorithms on NSL-KDD Dataset." *International Journal of Computer Networks and Communications* 16, no. 6 (2024).

[34] Touzene, Abderezak, Ahmed Al Farsi, and Nasser Al Zeidi. "High performance Nmf based intrusion detection system for big data IoT traffic." *International Journal of Computer Networks and Communications* 16, no. 2 (2024): 43-58.

[35] Chimphlee, W., and S. Chimphlee. "Intrusion detection system (IDS) development using tree-based machine learning algorithms." *International Journal of Computer Networks and Communications* 15, no. 4 (2023): 93-109.

[36] Sultan, Mohamad T., Hesham El Sayed, and Manzoor Ahmed Khan. "An intrusion detection mechanism for MANETs based on deep learning artificial neural networks (ANNs)." *arXiv preprint arXiv:2303.08248* (2023).

[37] Songma, Surasit, Watcharakorn Netharn, and Siriluck Lorpunmanee. "Extending Network Intrusion Detection with Enhanced Particle Swarm Optimization Techniques." *arXiv preprint arXiv:2408.07729* (2024).

[38] Trong, Thua Huynh, and Thanh Nguyen Hoang. "Effective Multi-Stage Training Model For Edge Computing Devices In Intrusion Detection." *arXiv preprint arXiv:2401.17546* (2024).

AUTHORS

Dr. Anoop Mohana Kumar earned an M.Sc (Integrated) degree in software engineering from Noorul Islam College of Engineering in Kumaracoil, Tamil Nadu, India, as well as a ME degree in computer science and engineering from PET Engineering College in Vallioor, Tamil Nadu, India.

Dr. Joseph Emerson Raja received his PhD in Engineering from Multimedia University in 2014. He has been working at the Centre for Advanced Analytics, COE of Artificial Intelligence, Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia, since 2004.

Dr. Chinnaiyan Senthilpari holds postgraduate degrees in M.Sc. (Applied Electronics) and M.E. (Material Science) from the National Institute of Technology, Tiruchirappalli. He began his academic career as a lecturer at Multimedia University, Malaysia, in 2005 and earned his Ph.D. from the same institution in 2009.