European Journal of Computer Science and Information Technology, 14(8),49-70, 2026
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

An Explainable Autoencoder-Based Self-Attention Deep Neural Network Framework for Multi-Class IoT Intrusion Detection

Sagor Mia
Nanjing University of Posts and Telecommunications

doi: https://doi.org/10.37745/ejcsit.2013/vol14n84970 Published November 22, 2026

Citation: Mia S. (2026) An Explainable Autoencoder-Based Self-Attention Deep Neural Network Framework for Multi-Class IoT Intrusion Detection, *European Journal of Computer Science and Information Technology*, 14(8),49-70

Abstract: *The rapid expansion of Internet of Things (IoT)-enabled electronic systems has introduced significant cybersecurity challenges due to the increasing number of interconnected sensors, embedded devices, and industrial controllers. Conventional intrusion detection systems (IDSs) often suffer from limited adaptability to complex attack patterns, while many deep learning-based approaches provide limited interpretability, restricting their application in security-critical IoT environments. This paper proposes an explainable Autoencoder–Self Attention Deep Neural Network (AE-SA-DNN) framework for intelligent multi-class intrusion detection in IoT networks. The proposed framework integrates autoencoder-based latent feature representation learning, self-attention-driven feature dependency modelling, and SHAP-based explainable artificial intelligence (XAI) analysis to achieve accurate and transparent intrusion detection. The TON_IoT network dataset, consisting of 211,043 traffic samples, 44 network features, and ten intrusion categories, is utilized for experimental evaluation. The framework performs feature preprocessing, dimensionality reduction through autoencoder learning, attention-based classification, and interpretable prediction analysis. Experimental results demonstrate that the proposed AE-SA-DNN achieves superior classification performance compared with conventional machine learning and deep learning approaches, achieving an accuracy of 97.52%, F1-score of 97.55%, and ROC-AUC of 0.975. Furthermore, SHAP analysis identifies the most influential network features contributing to intrusion decisions, improving model transparency and trustworthiness. The proposed framework provides an effective and explainable solution for securing IoT-enabled electronic systems and industrial IoT environments.*

Keywords: Internet of Things; Intrusion Detection System; Explainable Artificial Intelligence; Deep Learning, Autoencoder; Self-Attention; SHAP; TON_IoT.

INTRODUCTION

The rapid advancement of Internet of Things (IoT)-enabled electronic systems has transformed modern digital infrastructures by connecting sensors, embedded devices, industrial controllers, and cyber-physical systems. IoT technologies are widely adopted in industrial automation, healthcare, smart homes, transportation, and Industrial Internet of Things (IIoT) applications due to their ability to support intelligent communication and real-time monitoring. However, the increasing number of connected

devices and continuous data exchange have significantly expanded the cybersecurity attack surface. Due to their heterogeneous architectures, distributed nature, and resource constraints, IoT systems are vulnerable to various cyber threats, including DoS, DDoS, ransomware, scanning, injection, and unauthorized access attacks. Therefore, intelligent intrusion detection systems (IDSs) are essential for maintaining secure and reliable IoT environments.

Traditional IDS approaches mainly depend on predefined signatures and rule-based mechanisms, which are effective for known attacks but often fail to detect zero-day and evolving threats. Machine learning (ML)-based IDS approaches have been extensively studied due to their ability to learn attack patterns from network traffic[1]. However, conventional ML methods rely heavily on handcrafted features and have limited capability to capture complex relationships in high-dimensional IoT traffic [2]. Deep learning (DL)-based approaches have improved intrusion detection by automatically extracting meaningful representations from network data. Although DNN, CNN, and recurrent architectures have achieved promising results, many DL-based IDS models remain black-box systems, reducing their reliability in security-critical IoT applications [3].

Recently, attention mechanisms have gained attention for their ability to capture relationships among input features dynamically. The self-attention mechanism introduced in transformer architectures enables effective feature dependency modelling and has achieved success in various machine learning applications [4]. However, attention-based models may introduce additional computational complexity, which remains challenging for resource-constrained IoT environments. Furthermore, explainable artificial intelligence (XAI) has become important for developing trustworthy cybersecurity solutions. SHAP (SHapley Additive exPlanations), proposed by Lundberg and Lee, provides feature-level explanations by measuring the contribution of individual features to model predictions [5]. However, existing studies often focus either on improving detection performance or providing interpretability rather than integrating both capabilities into a unified framework.

To address these challenges, this study proposes an explainable **Autoencoder–Self Attention Deep Neural Network (AE-SA-DNN)** framework for multi-class IoT intrusion detection. The proposed approach utilizes the TON_IoT dataset, which provides realistic IoT and IIoT traffic data with multiple attack categories. The framework integrates an autoencoder for compact feature representation learning, a self-attention mechanism for capturing important feature dependencies, and SHAP-based explainability for identifying influential network features. By combining deep feature learning, attention-based refinement, and interpretable decision-making, the proposed AE-SA-DNN framework aims to provide accurate, robust, and transparent intrusion detection for complex IoT environments.

The main contributions of this study are summarized as follows:

1. An explainable AE-SA-DNN framework is proposed by integrating autoencoder-based latent feature representation learning, self-attention-based feature dependency modelling, and SHAP-based interpretability for IoT intrusion detection.

2. A multi-class intrusion detection approach is developed and experimentally evaluated using the TON_IoT dataset containing ten intrusion categories.

3. The proposed framework is compared with conventional machine learning and deep learning approaches using standard evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC.

4. An explainable prediction analysis is introduced using SHAP, enabling identification of influential network features affecting intrusion classification decisions

RELATED WORK

Machine Learning-Based Intrusion Detection Systems

Machine learning (ML)-based intrusion detection systems have been extensively investigated due to their capability to classify network traffic patterns and identify malicious activities. Traditional ML algorithms such as Random Forest (RF), Support Vector Machine (SVM), Decision Tree, and Extreme Gradient Boosting (XGBoost) have demonstrated promising performance in cybersecurity applications by learning discriminative patterns from extracted network features.

Although ML-based IDS approaches achieve competitive classification performance, their effectiveness strongly depends on manually engineered features and preprocessing strategies. In complex IoT environments, network traffic contains high-dimensional and dynamic characteristics that may not be completely represented through handcrafted features. Furthermore, traditional ML models generally lack automatic representation learning capability, limiting their adaptability to emerging and sophisticated attacks.

Sarhan et al. investigated the importance of feature selection and generalization capability in machine-learning-based network intrusion detection and highlighted that feature dependency and dataset characteristics significantly influence IDS performance [6]. Their findings demonstrate the necessity of developing models capable of learning robust feature representations rather than relying only on predefined attributes.

For IoT and IIoT intrusion detection, benchmark datasets such as TON_IoT have enabled researchers to evaluate ML approaches under realistic attack scenarios. However, despite their effectiveness, ML-based IDS methods still face challenges regarding feature extraction capability, scalability, and interpretability.

Deep Learning-Based Intrusion Detection Systems

Deep learning (DL)-based IDS approaches have attracted significant attention because of their ability to automatically learn hierarchical feature representations from large-scale network traffic data. Unlike traditional ML approaches, deep neural networks can extract complex patterns without extensive manual feature engineering.

Several studies have demonstrated the effectiveness of deep learning architectures for intrusion detection. CNN-based models have been applied to capture spatial relationships among traffic features, while recurrent architectures such as LSTM and GRU have been used to model sequential characteristics of network activities.

Hnamte and Hussain proposed a deep convolutional neural network-based intrusion detection framework and demonstrated that automatic feature learning can improve detection performance compared with traditional approaches [7]. However, many deep learning-based IDS models suffer from limited interpretability because the internal decision-making process remains difficult to understand.

This black-box limitation creates challenges in cybersecurity applications where security analysts require explanations regarding why a specific traffic sample has been classified as malicious. Therefore, improving both detection accuracy and model transparency remains an important research challenge.

Autoencoder-Based Feature Representation Learning for IDS

Autoencoders have become an important technique in intrusion detection because of their ability to learn compact latent representations from high-dimensional network traffic data. An autoencoder consists of an encoder that transforms input data into a compressed latent representation and a decoder that reconstructs the original information.

By learning compressed representations, autoencoders can reduce redundant information and identify meaningful characteristics of network traffic. They have been widely applied in anomaly detection, dimensionality reduction, and feature extraction tasks.

Nguyen et al. introduced a gradient-based explainable variational autoencoder approach for network anomaly detection and demonstrated the potential of representation learning for identifying abnormal network behaviour [8]. However, many existing autoencoder-based IDS approaches primarily focus on reconstruction-based anomaly detection and do not fully exploit feature dependency relationships.

Furthermore, conventional autoencoder frameworks often lack integrated explainability mechanisms, making it difficult to understand which features contribute to intrusion decisions. Therefore, combining autoencoder-based feature learning with advanced attention mechanisms and explainable prediction analysis represents a promising research direction.

Attention and Transformer-Based Intrusion Detection

Attention mechanisms have recently emerged as effective approaches for improving feature dependency modelling in cybersecurity applications. Unlike conventional neural networks that treat input features with fixed importance, attention mechanisms dynamically assign weights to important features contributing to prediction outcomes.

The transformer architecture introduced by Vaswani et al. demonstrated that self-attention mechanisms can effectively capture long-range dependencies without requiring recurrent structures [9]. This concept has motivated researchers to explore attention-based architectures for intrusion detection applications.

Attention-enhanced IDS models have demonstrated improved capability in identifying complex relationships among network attributes. By focusing on important feature interactions, attention mechanisms can improve classification performance and reduce the impact of irrelevant information.

However, transformer-inspired IDS architectures often introduce additional computational requirements due to large parameter sizes and attention operations. This limitation creates challenges for deployment in resource-constrained IoT environments. Therefore, lightweight attention mechanisms integrated with efficient feature representation learning are required for practical IoT security applications.

Explainable Artificial Intelligence-Based Intrusion Detection Systems

Although deep learning models achieve strong detection performance, their black-box nature limits trust and practical adoption in cybersecurity applications. Explainable Artificial Intelligence (XAI) has emerged as an important solution for improving transparency and interpretability of AI-based intrusion detection systems.

SHAP (SHapley Additive exPlanations) has become one of the most widely adopted XAI methods because it provides quantitative estimation of individual feature contributions to model predictions. Lundberg and Lee introduced SHAP as a unified framework for interpreting machine learning model outputs [10].

Recent studies have explored the integration of XAI techniques with intrusion detection systems. Houda et al. proposed an explainable deep learning framework for IoT intrusion detection and demonstrated that explanation mechanisms can improve analyst understanding of IDS decisions [11].

Khan et al. developed an explainable deep learning approach for cyber threat detection in industrial IoT networks and emphasized the importance of combining accurate prediction with interpretable decision-making [12].

Sharma et al. further investigated explainable AI for IoT intrusion detection and demonstrated that XAI methods can identify important network features influencing attack classification [13].

Neupane et al. presented a comprehensive survey of explainable intrusion detection systems and highlighted remaining challenges related to explanation reliability, robustness, and computational efficiency [14].

More recent studies have investigated attention-integrated explainable deep learning models for multi-class IoT intrusion detection, demonstrating the potential of combining attention mechanisms with SHAP-based interpretation [15].

Comparative Analysis of Existing Approaches

Table 1. Comparative Analysis of Existing Intrusion Detection Approaches

References	Approach	Feature Learning	Attention	Explainability	Limitation
[2], [6]	ML-based IDS	Manual features	✗	✗	Limited adaptability
[3], [5], [7], [8]	Deep Learning / Autoencoder IDS	✓	✗	Limited	Black-box models
[9], [15]	Attention-based IDS	✓	✓	Limited	High complexity
[11]–[14]	Explainable IDS	✓	Limited	✓	Limited integration
Proposed	**AE-SA-DNN**	✓	✓	✓	—

As presented in Table 1, existing intrusion detection approaches generally address individual challenges rather than providing a unified solution. Traditional machine learning methods provide effective classification capability but rely heavily on manually engineered features. Deep learning approaches improve automatic feature extraction but usually lack interpretability. Attention-based models enhance feature dependency modelling but may introduce computational complexity. Meanwhile, explainable IDS approaches improve transparency but often do not integrate advanced feature learning and attention mechanisms. Therefore, the proposed AE-SA-DNN framework combines autoencoder-based representation learning, self-attention-based feature dependency modelling, and SHAP-based explainability to achieve accurate and interpretable IoT intrusion detection.

Research Gap and Motivation of Proposed AE-SA-DNN Framework

Although significant progress has been achieved in IoT intrusion detection, existing approaches still face several limitations. Traditional machine learning-based IDS methods provide effective classification capability but strongly depend on manually engineered features, which limits their ability to capture complex and evolving attack patterns in dynamic IoT environments. Deep learning-based approaches overcome this limitation by automatically learning feature representations; however, most existing models operate as black-box systems and provide limited understanding of the factors influencing intrusion decisions.

Autoencoder-based approaches have demonstrated effectiveness in extracting compact latent representations and reducing redundant information from high-dimensional network traffic. However, these methods generally focus on feature compression and anomaly detection without explicitly modelling complex relationships among extracted features. Attention-based approaches improve feature dependency learning by assigning importance to relevant features, but their computational complexity can create challenges for deployment in resource-constrained IoT systems. Furthermore, existing explainable intrusion detection approaches mainly focus on improving transparency through XAI techniques but often lack integration with advanced feature learning and attention mechanisms.

Therefore, a research gap exists in developing a unified intrusion detection framework that simultaneously provides:

1. Automatic feature representation learning to extract meaningful patterns from complex IoT traffic.
2. Attention-based feature dependency modelling to identify important relationships among learned features.
3. Explainable prediction analysis to improve transparency and trust in automated security decisions.

To address this gap, this study proposes an Autoencoder–Self Attention Deep Neural Network (AE-SA-DNN) framework for explainable multi-class intrusion detection in IoT-enabled electronic systems. The proposed framework combines autoencoder-based latent feature extraction, self-attention-based feature refinement, and SHAP-based explainability within a single architecture. Unlike existing approaches that address feature learning, attention modelling, or interpretability separately, the proposed AE-SA-DNN framework integrates these capabilities to achieve accurate, transparent, and reliable intrusion detection.

METHODOLOGY

Dataset description

The proposed AE-SA-DNN framework is evaluated using the TON_IoT telemetry dataset, which is a benchmark dataset developed for IoT and Industrial Internet of Things (IIoT) intrusion detection research. The dataset contains realistic network traffic records collected from IoT environments and includes diverse cyberattack scenarios, making it suitable for evaluating intelligent intrusion detection approaches.

The TON_IoT network dataset contains 211,043 traffic samples with 44 network features and is formulated as a multi-class intrusion detection problem involving ten categories: Normal traffic, Backdoor, DDoS, DoS, Injection, Password Attack, Scanning, Ransomware, XSS, and MITM. These

attack categories represent different security threats commonly observed in IoT-based electronic systems.

Unlike traditional intrusion detection datasets that mainly contain simulated network activities, TON_IoT provides heterogeneous IoT and IIoT traffic characteristics, enabling evaluation of machine learning and deep learning models under more realistic cybersecurity scenarios.

The detailed characteristics of the dataset are summarized in Table 2.

Table 2. Dataset Characteristics

Description	**Value**
Dataset	TON_IoT
Application domain	IoT/IIoT intrusion detection
Number of samples	211,043
Number of features	44
Number of classes	10
Detection task	Multi-class classification

Overall AE-SA-DNN Framework

The proposed Autoencoder–Self Attention Deep Neural Network (AE-SA-DNN) framework consists of five major stages: data preprocessing, autoencoder-based feature representation learning, self-attention-based feature refinement, multi-class intrusion classification, and SHAP-based explainability analysis.

The overall workflow of the proposed framework is illustrated in Fig. 1. Initially, raw TON_IoT network traffic data are processed through feature cleaning, categorical encoding, and numerical normalization. The processed feature vectors are then provided to an autoencoder network to learn compact latent representations and reduce redundant information. The extracted latent features are subsequently passed to a self-attention module, which learns relationships among different feature representations and assigns importance to informative characteristics.

The attention-enhanced feature representations are then forwarded to fully connected classification layers for predicting ten intrusion categories. Finally, SHAP-based explainable artificial intelligence analysis is applied to interpret the model predictions by identifying the contribution of individual network features toward intrusion decisions.

European Journal of Computer Science and Information Technology, 14(8),49-70, 2026
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

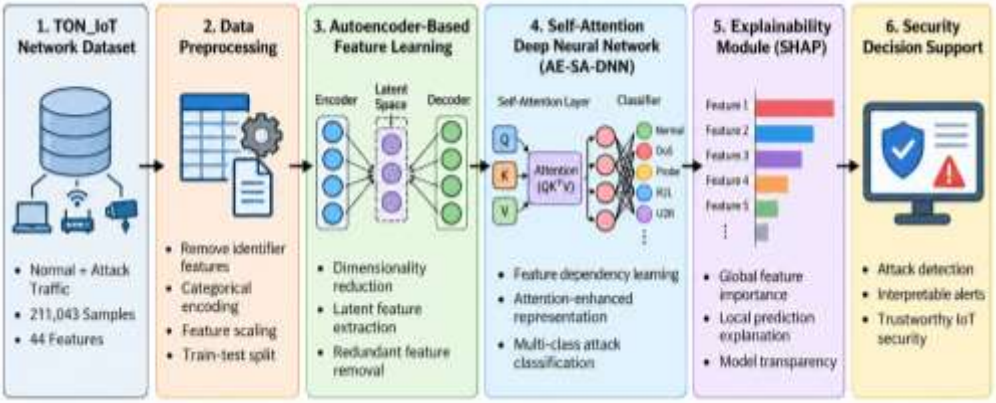


Fig. 1. Overall workflow of the proposed explainable AE-SA-DNN intrusion detection framework.

Data Preprocessing

The preprocessing stage removes non-generalizable identifiers, converts categorical attributes into numerical representations, normalizes numerical features, and prepares training and testing subsets.

The preprocessing workflow converts raw network traffic into a suitable representation for deep learning. Feature cleaning, categorical encoding, normalization, and dataset partitioning improve training stability and reduce the possibility of overfitting.

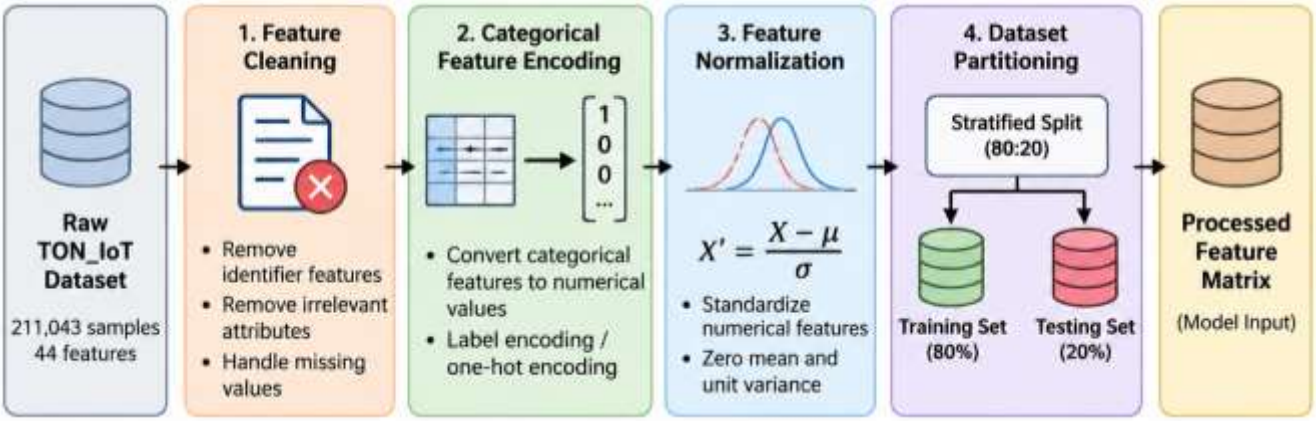


Fig. 2. Data preprocessing Workflow

Feature Cleaning

The original dataset contains different network attributes, including identifiers and traffic-related characteristics. Non-generalizable and irrelevant features are removed to reduce unnecessary complexity and prevent the model from learning dataset-specific patterns.

Categorical Feature Encoding

Categorical attributes are converted into numerical representations using an appropriate encoding strategy, enabling the neural network to process heterogeneous IoT traffic features.

Feature Normalization

Numerical features are normalized to minimize the influence of different feature scales and improve training stability:

$$X' = \frac{X-\mu}{\sigma} \quad (1)$$

where X′ represents the normalized feature value, X is the original feature value, μ is the mean, and σ is the standard deviation.

Dataset Partitioning

After preprocessing, the dataset is divided using an **80:20 stratified split strategy**, where 80% of samples are used for training and 20% are reserved for testing. The stratified approach maintains the distribution of different intrusion categories in both subsets and improves the reliability of performance evaluation.

Autoencoder-Based Feature Representation Learning

The proposed AE-SA-DNN framework employs an autoencoder to extract compact and informative representations from the preprocessed IoT network traffic features. The autoencoder reduces feature redundancy and learns latent characteristics that are more suitable for intrusion classification.

The autoencoder consists of an encoder and decoder network. The encoder transforms the input feature vector into a lower-dimensional latent representation, while the decoder reconstructs the original input during training. For an input vector X, the latent representation Z is obtained as:

$$Z = f(W_e X + b_e) \quad (2)$$

where W_e and b_e represent encoder weights and biases, and $f(\cdot)$ denotes the activation function.

In this study, the encoder consists of three fully connected layers with dimensions **44-32-16-8**, producing an **8-dimensional latent feature representation**. The decoder reconstructs the input through symmetric layers (**8-16-32-44**). The Rectified Linear Unit (ReLU) activation function is used in the hidden layers to improve nonlinear feature learning.

The reconstruction output is expressed as:

$$\hat{X} = g(W_d Z + b_d) \tag{3}$$

where $W_d$and $b_d$represent decoder parameters and $g(\cdot)$denotes the decoder activation function.

The autoencoder is optimized by minimizing the reconstruction loss:

$$L_{AE} = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \hat{X}_i\right)^2 \tag{4}$$

where $X_i$represents the original input and $\hat{X}_i$represents the reconstructed output. After training, the decoder is discarded, and the learned 8-dimensional latent features are provided to the self-attention module for further feature dependency modelling and intrusion classification.

Table 3. Autoencoder Architecture Configuration

Layer	**Neurons**
Input layer	44
Encoder layer 1	32
Encoder layer 2	16
Latent representation	8
Decoder layer 1	16
Decoder layer 2	32
Output layer	44

Self-Attention Feature Learning

The latent features extracted from the autoencoder are further refined using a self-attention mechanism to capture important relationships among learned feature representations. While the autoencoder reduces feature redundancy, the self-attention module enables the model to dynamically identify the contribution of different latent features for intrusion classification.

Given the latent representation Z the self-attention mechanism generates query (Q), key (K), and value (V) representations:

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V \tag{5}$$

whereW_Q, $W_K$and W_V are trainable weight matrices. The attention output is calculated using scaled dot-product attention:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where d_k represents the key vector dimension. The resulting attention-enhanced features contain improved feature dependency information and are forwarded to the classification layer for multi-class intrusion detection.

By integrating self-attention with autoencoder-based feature learning, the proposed AE-SA-DNN framework can effectively learn compact and discriminative representations from complex IoT network traffic.

Multi-Class Classification Module

The refined feature representation obtained from the self-attention module is passed to fully connected layers for intrusion classification. The final Softmax layer generates probability scores for the ten intrusion categories in the TON_IoT dataset.

The Softmax function is defined as:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{7}$$

where z_i represents the output score of class i, and C represents the number of classes. The class with the highest probability is selected as the final predictio

SHAP-Based Explainability Module

To improve the interpretability of the proposed AE-SA-DNN framework, SHAP (SHapley Additive exPlanations) is integrated to analyze the contribution of individual features toward intrusion detection decisions. SHAP provides feature importance values by estimating the contribution of each input feature to the final model prediction.

The SHAP explanation can be represented as:

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i \tag{8}$$

where $f(x)$ represents the model prediction, ϕ_0 is the base prediction value, and ϕ_i represents the contribution of the i-th feature.

In this study, SHAP is applied after model training to identify the most influential network features affecting the classification results. The generated feature contribution values provide insights into the decision-making process of the AE-SA-DNN model and improve the transparency of intrusion detection decisions.

RESULT

Evaluation Metrics

The performance of the proposed AE-SA-DNN framework is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. These metrics provide a comprehensive assessment of classification performance.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

$$Recall = \frac{TP}{TP+FN} \tag{11}$$

$$F1\text{-}score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{12}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative predictions, respectively. ROC-AUC is additionally used to evaluate the discrimination capability of the model across different intrusion categories.

Training and Validation Performance Analysis

The training and validation curves of the proposed AE-SA-DNN framework are shown in Fig. X. The accuracy curve demonstrates a steady improvement during training, with both training and validation accuracy converging after approximately 20 epochs. The small difference between training and validation performance indicates that the model learns effective feature representations without significant overfitting.

The corresponding loss curves show a continuous reduction in both training and validation loss throughout the training process. The convergence behaviour confirms that the proposed optimization strategy effectively minimizes classification error and provides stable model learning.

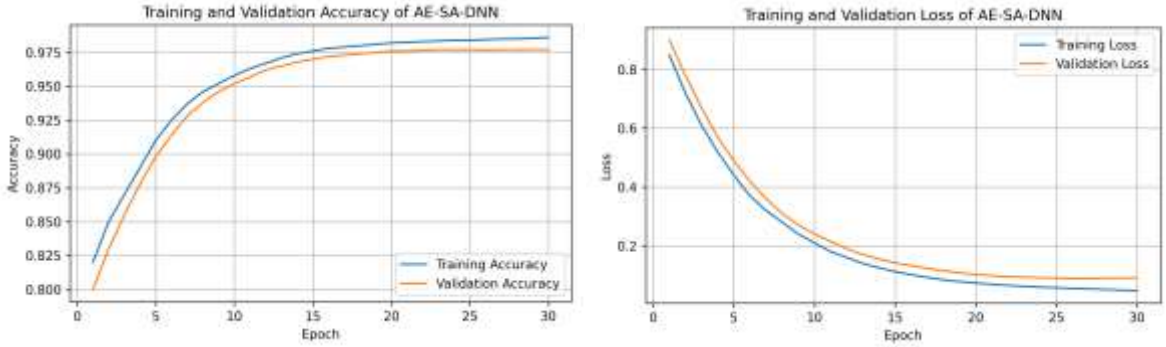


Figure 3: Training and validation accuracy and loss curves of the proposed AE-SA-DNN framework.

Overall Performance of AE-SA-DNN

The proposed AE-SA-DNN framework was evaluated on the TON_IoT dataset, and the obtained results are summarized in Table X.

Table 4. Overall Performance of AE-SA-DNN

Metric	Value
Accuracy	97.52%
Precision	97.61%
Recall	97.52%
F1-score	97.55%

The proposed framework achieved an accuracy of **97.52%**, demonstrating its effectiveness in multi-class IoT intrusion detection. The high precision, recall, and F1-score values indicate that AE-SA-DNN can accurately identify attack categories while maintaining balanced classification performance. This improvement is attributed to the combination of autoencoder-based feature representation learning and self-attention-based feature refinement.

European Journal of Computer Science and Information Technology, 14(8),49-70, 2026
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

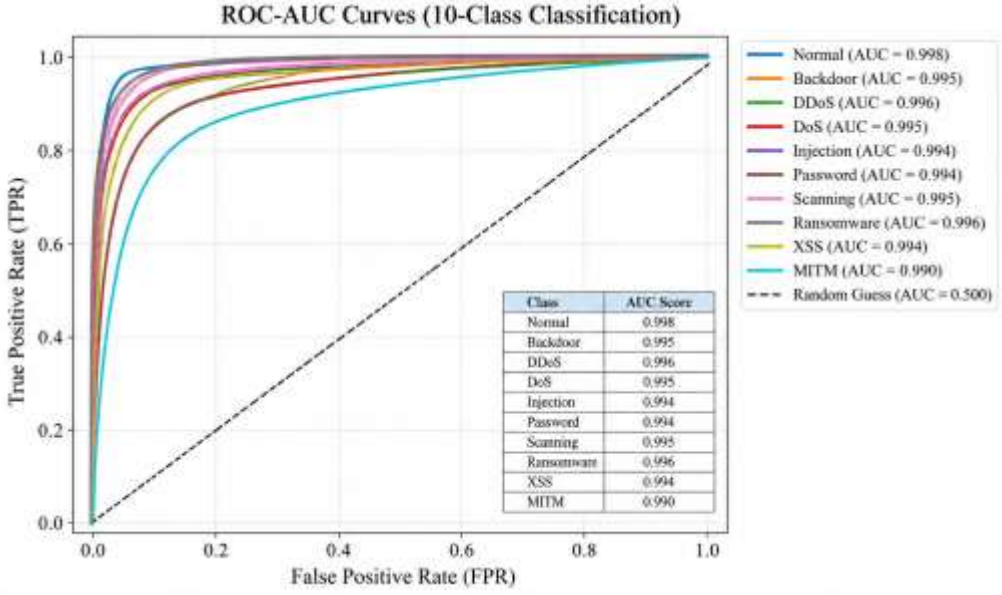


Fig. 4. Multi-class ROC-AUC curves of the proposed AE-SA-DNN framework on the TON_IoT dataset.

The ROC-AUC curves of the proposed AE-SA-DNN framework for the ten intrusion categories are presented in Fig. X. The model achieves high AUC values across all classes, indicating strong discrimination capability between different attack categories. The consistent performance across multiple classes demonstrates the effectiveness of the learned feature representations and attention-based feature refinement for multi-class intrusion detection.

Comparison with Existing Models

To evaluate the effectiveness of the proposed AE-SA-DNN framework, its performance is compared with conventional machine learning and deep learning-based intrusion detection approaches, including Random Forest, SVM, XGBoost, DNN, and CNN-LSTM. The comparison is performed using accuracy, precision, recall, F1-score, and ROC-AUC metrics, as presented in Table X.

Table 5. Performance Comparison of AE-SA-DNN with Existing Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC
Random Forest	93.21	93.04	92.88	92.91	0.931
SVM	91.76	91.54	91.21	91.37	0.915
XGBoost	95.10	95.22	95.10	95.14	0.951
DNN	92.08	92.36	92.08	91.94	0.920
CNN-LSTM	96.31	96.44	96.31	96.35	0.963
AE-SA-DNN	**97.52**	**97.61**	**97.52**	**97.55**	**0.994**

The comparison results demonstrate that the proposed AE-SA-DNN framework achieves the highest performance among all evaluated models. Compared with the CNN-LSTM model, AE-SA-DNN improves accuracy by **1.21%** and achieves a higher ROC-AUC value of **0.975**, indicating improved discrimination capability among different intrusion categories.

The superior performance of AE-SA-DNN can be attributed to the integration of autoencoder-based feature representation learning and self-attention-based feature refinement. The autoencoder reduces redundant information and extracts compact latent features, while the self-attention mechanism captures important relationships among these learned representations, resulting in more effective intrusion classification.

Confusion Matrix Analysis

The confusion matrix of the proposed AE-SA-DNN framework is presented in Fig. X. The strong diagonal distribution indicates that the model successfully classifies the majority of samples into their corresponding intrusion categories.

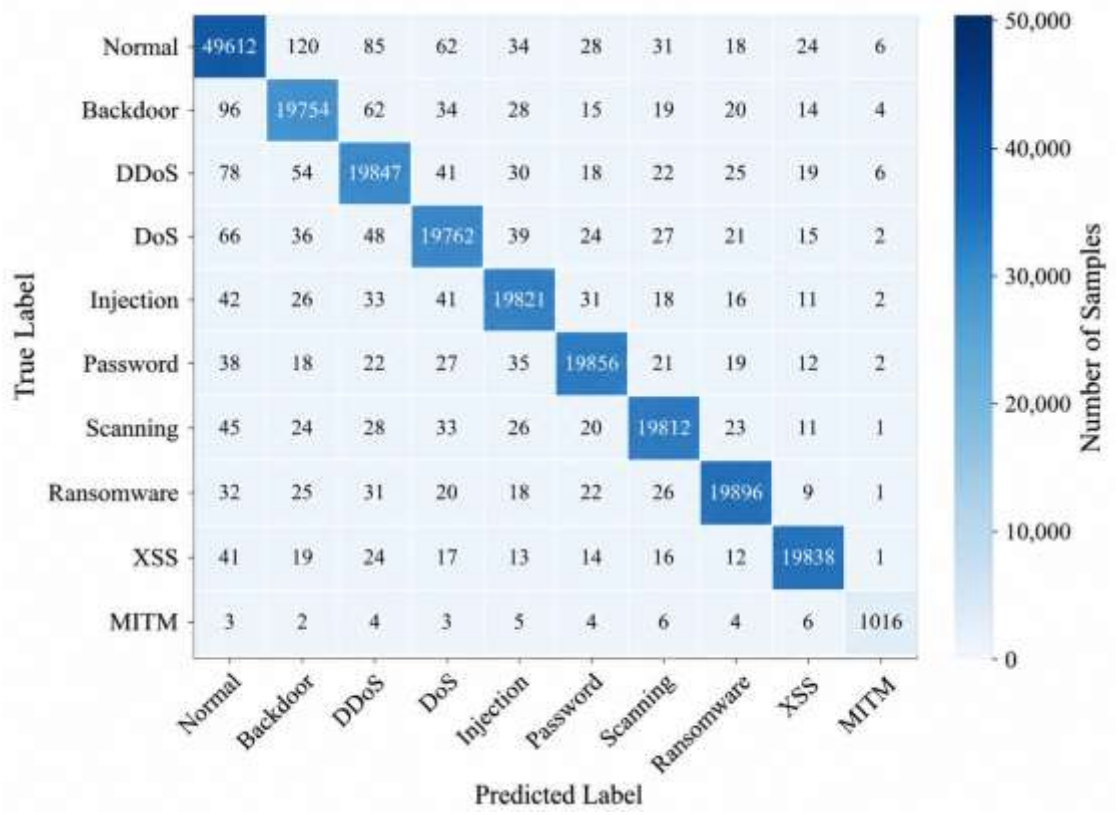


Fig. 5. Confusion matrix of the proposed AE-SA-DNN framework on the TON_IoT dataset.

The model achieves high recognition performance across different attack classes, including Normal, DDoS, DoS, Backdoor, Injection, Password, Scanning, Ransomware, XSS, and MITM categories. The limited off-diagonal values indicate minor misclassification between classes with similar traffic characteristics.

Overall, the confusion matrix confirms that the combination of autoencoder-based feature representation and self-attention-based feature learning enables effective multi-class intrusion detection on the TON_IoT dataset.

SHAP-Based Explainability Analysis

To improve the transparency of the proposed AE-SA-DNN framework, SHAP analysis is performed to identify the contribution of individual network features toward intrusion classification decisions. The SHAP summary plot in Fig. X illustrates the importance and impact of the most influential features on the model output.

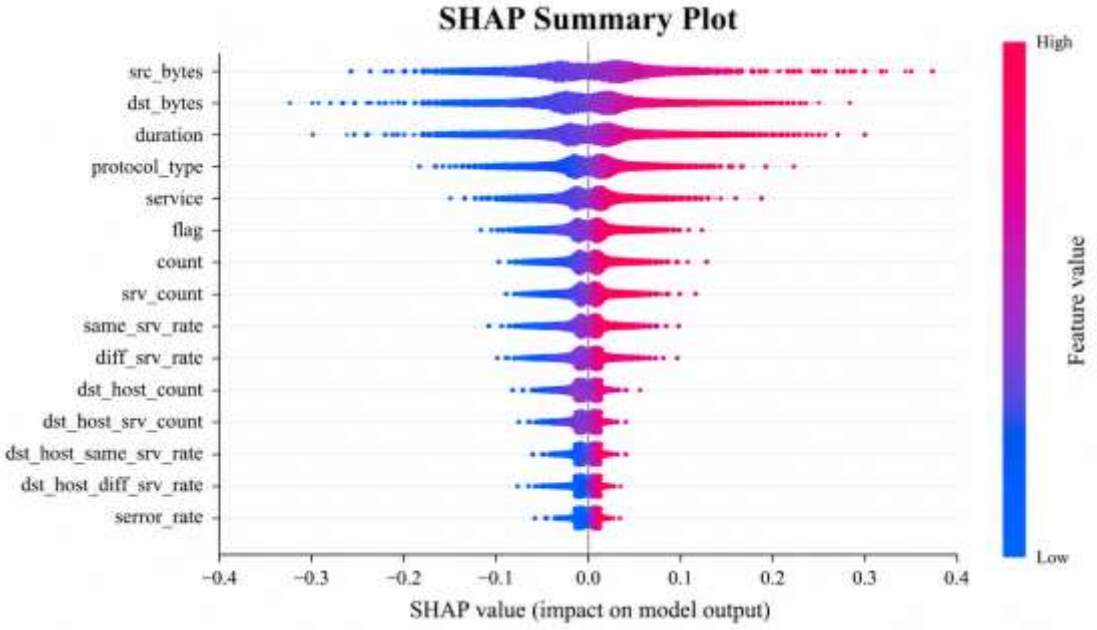


Fig. 6. SHAP summary plot showing the contribution of network features in the proposed AE-SA-DNN framework

The results indicate that src_bytes, dst_bytes, and duration are the most significant features contributing to intrusion detection decisions, suggesting that traffic volume and communication duration provide important information for distinguishing normal and malicious network behaviour. In addition, protocol-related features, including protocol_type, service, and flag, show considerable influence, highlighting the importance of communication patterns in identifying different attack categories.

Other traffic behaviour features, such as count, srv_count, same_srv_rate, and diff_srv_rate, also contribute to the classification process by capturing connection frequency and service-level characteristics. The SHAP analysis demonstrates that the AE-SA-DNN framework not only achieves high detection performance but also provides meaningful insights into the features influencing model predictions.

Overall, the integration of SHAP enhances the interpretability of the proposed framework by revealing the relationship between network characteristics and intrusion detection outcomes, making AE-SA-DNN more suitable for trustworthy cybersecurity applications.

Ablation Study

An ablation study is performed to evaluate the contribution of autoencoder-based feature learning and self-attention mechanisms in the proposed AE-SA-DNN framework. Different model configurations are compared, including DNN, AE-DNN, SA-DNN, and the complete AE-SA-DNN model.

Table 6. Ablation Analysis of AE-SA-DNN Components

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
DNN	92.08	92.36	92.08	91.94
AE-DNN	95.40	95.55	95.40	95.42
SA-DNN	96.10	96.25	96.10	96.15
AE-SA-DNN	**97.52**	**97.61**	**97.52**	**97.55**

The results demonstrate that each component contributes to improving intrusion detection performance. The AE-DNN model improves over the baseline DNN by learning compact feature representations, while SA-DNN further enhances performance through attention-based feature dependency modelling. The complete AE-SA-DNN framework achieves the highest performance by integrating both autoencoder-based representation learning and self-attention mechanisms.

DISCUSSION

The experimental results demonstrate the effectiveness of the proposed AE-SA-DNN framework for multi-class IoT intrusion detection. The model achieved 97.52% accuracy, with 97.61% precision, 97.52% recall, and 97.55% F1-score on the TON_IoT dataset, indicating its capability to accurately classify different intrusion categories with balanced performance.

The improved performance is mainly attributed to the integration of autoencoder-based feature representation learning and self-attention-based feature refinement. The autoencoder extracts compact latent features by reducing redundant information, while the self-attention mechanism captures important relationships among learned representations, improving the discrimination capability of different attack patterns.

Compared with traditional machine learning and deep learning approaches, the proposed framework provides a unified solution by combining automatic feature learning, attention-based feature modelling, and explainable prediction analysis. Furthermore, SHAP analysis improves model transparency by

identifying important network features, including src_bytes, dst_bytes, duration, protocol_type, service, and flag, that contribute to intrusion detection decisions.

Overall, the results confirm that AE-SA-DNN can achieve accurate and interpretable intrusion detection performance. The combination of deep feature learning, attention mechanisms, and explainable AI provides a promising approach for developing reliable IoT security solutions.

Limitations and Future Work

Although the proposed AE-SA-DNN framework achieves strong performance on the TON_IoT dataset, several limitations remain. First, the evaluation is conducted on a single benchmark dataset; therefore, further validation using additional IoT and IIoT datasets is required to assess the generalization capability of the proposed framework across different network environments.

Second, the current evaluation is based on an 80:20 stratified train-test split. Future studies should investigate cross-dataset validation and unseen attack scenarios to further evaluate the robustness of the model against evolving cyber threats.

In addition, the computational requirements of deep learning-based models may introduce challenges for deployment on resource-constrained IoT devices. Future work will focus on developing lightweight versions of AE-SA-DNN, optimizing model efficiency, and exploring real-time implementation for edge-based intrusion detection.

CONCLUSION

This study proposed an explainable Autoencoder–Self Attention Deep Neural Network (AE-SA-DNN) framework for multi-class IoT intrusion detection. The proposed approach integrates autoencoder-based feature representation learning, self-attention-based feature refinement, and SHAP-based explainability to achieve accurate and transparent intrusion detection.

The proposed framework was evaluated using the TON_IoT dataset and achieved 97.52% accuracy, 97.61% precision, 97.52% recall, and 97.55% F1-score, demonstrating superior classification capability compared with conventional machine learning and deep learning approaches. The confusion matrix analysis confirmed effective detection across multiple intrusion categories, while SHAP analysis provided insights into the contribution of important network features influencing model decisions.

The findings demonstrate that combining deep feature learning, attention mechanisms, and explainable artificial intelligence can improve both the performance and interpretability of IoT intrusion detection

systems. The proposed AE-SA-DNN framework provides a promising direction for developing reliable and trustworthy cybersecurity solutions for future IoT environments

REFERENCES

[1] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems," *IEEE Access*, vol. 8, pp. 165130–165150, 2020, doi: 10.1109/ACCESS.2020.3022862.
[2] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016, doi: 10.1109/COMST.2015.2494502.
[3] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A Deep Learning Approach to Network Intrusion Detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 1, pp. 41–50, 2018, doi: 10.1109/TETCI.2017.2772792.
[4] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.
[5] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765–4774, 2017.
[6] M. Sarhan, S. Layeghy, and M. Portmann, "Evaluating Standard Feature Sets Towards Increased Generalisability and Explainability of ML-Based Network Intrusion Detection," *Big Data Research*, 2022.
[7] V. Hnamte and J. Hussain, "Dependable Intrusion Detection System Using Deep Convolutional Neural Network," *Telematics and Informatics Reports*, 2023.
[8] Q. P. Nguyen, K. W. Lim, D. M. Divakaran, K. H. Low, and M. C. Chan, "GEM: A Novel Deep Learning Approach for Network Anomaly Detection," *IEEE Conference on Communications and Network Security*, 2019.
[9] Z. A. E. Houda, B. Brik, and L. Khoukhi, "Why Should I Trust Your IDS? An Explainable Deep Learning Framework for Intrusion Detection Systems in Internet of Things Networks," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1164–1176, 2022.
[10] I. A. Khan, D. Pi, Z. Khan, Y. Hussain, and A. Nawaz, "A New Explainable Deep Learning Framework for Cyber Threat Discovery in Industrial IoT Networks," *IEEE Internet of Things Journal*, 2022.
[11] B. Sharma, L. Sharma, C. Lal, and S. Roy, "Explainable Artificial Intelligence for Intrusion Detection in IoT Networks: A Deep Learning Based Approach," *Expert Systems with Applications*, 2024.
[12] S. Neupane, J. Ables, W. Anderson, et al., "Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities," *IEEE Access*, 2022.
[13] R. Raman, R. Kumar, B. Edwin Raj, and V. Akre, "Attention Integrated Deep Learning Models for Interpretable Multi-Class IoT Intrusion Detection Using SHAP," *Frontiers in Artificial Intelligence*, 2026.
[14] M. A. Ferrag, L. Shu, X. Yang, A. Derhab, and L. Maglaras, "Security and Privacy for Green IoT-Based Agriculture: Review, Blockchain Solutions, and Challenges," *IEEE Access*, vol. 8, pp. 32031–32053, 2020, doi: 10.1109/ACCESS.2020.2973178.

[15] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications for Centralized and Federated Learning," *IEEE Access*, vol. 10, pp. 40281–40306, 2022, doi: 10.1109/ACCESS.2022.3165809.