Journal of Artificial Intelligence and Technology, (Ahead of Print)
https://doi.org/10.37965/jait.2026.0975

ISTP
RESEARCH ARTICLE

Structured Exponentiation Dot Product Attention-based Vision Transformer with Coordinated FuseNet Transfer Learning for Network Intrusion Detection

Venkata Ramani Varanasi and Shaik Razia
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

(Received 18 November 2025; Revised 16 August 2026; Accepted 01 September 2026; Published online 08 November 2026)

Abstract: The number and intensity of cyberattacks have grown exponentially with the explosive growth of cloud computing, Internet of Things (IoT), and high-speed networks. According to recent reports, more than 70% of the contemporary networks intrusions are either stealthy or low-rate intrusions, which are proven to be hard to detect with the traditional Intrusion Detection Systems (IDSs). Despite their extensive implementations in signature-based and rule-based IDS, these methods have weaknesses in terms of being slow to adapt to zero-day attacks, a high number of false alarms, and a lack of scalability when dealing with vast data volumes and high-dimensional network traffic data. To overcome these issues, this research proposes a Structured Exponentiation Dot product-based Network Intrusion Detection System (SED-NIDS) to facilitate the determination of accurate and robust cyberattacks. The proposed system is end-to-end, unlike single-model IDS methods, which perform feature learning and decision fusion independently, leading to better generalization of the system to various attack scenarios. The proposed SED-NIDS achieves an accuracy of 98.6%, 93% precision, 94% recall, and an F1-score of 93.5%. The proposed system also displays a reduced false alarm rate of 1.3% indicating how the system is applicable in real-time deployment. SED-NIDS is a precise and robust IDS that will answer the changing demands of the contemporary cyber threats. The proposed model has great improvements in detection performance with low false positives which makes it a viable solution to the next-generation network security systems.

Keywords: intrusion detection; positional encoding; transfer learning; zero-day attack

I. INTRODUCTION

With the growth of Internet knowledge, it has almost completely replaced other utilities in everyday life and fundamentally changed people's ways of living. However, in this era of networks, various techniques of network attacks are continually updated, growing, becoming common in frequency, and getting worse and worse in providing security [1]. Therefore, network intrusion detection has developed with the mission to identify suspicious attacks and take precautions to prevent the network from possible attacks and thus decrease monetary loss. Network intrusion detection has emerged as an important area of research in network security [2]. An Intrusion Detection System (IDS) is implemented on every node present in a network, at a specific central place, or distributed in the network. Network Intrusion Detection System (NIDS) is intended for identifying attacks targeting a network of computers rather than a host computer. NIDS seek to detect the occurrence of attacks by monitoring the network functionality with network telemetry that comprise network traffic, metadata of network, and host activity logs.

IDS also be classified under signature-based and anomaly-based categories. Signature-based IDS identifies attacks via comparison compared to known attack signatures. Therefore, it cannot identify new attack vectors or zero-day attacks [3]. Meanwhile, anomaly-based IDS collect information about the behavior of authentic users and then analyze the behavior now being observed to limit if it is of valid users or of malicious users; therefore, these kinds of IDS can identify unknown attacks. However, anomaly-based IDS frequently produce false-positive (FP) results, while signature-based IDS frequently produce false-negative (FN) results. In an anomaly-based approach, the training process involves gathering and monitoring data related to normal behavior of the system before construction of the model of the allowed user's behavior [4].

Machine learning approaches have widely been used in IDS to recognize malicious traffic. Nevertheless, these approaches fall under the category of shallow learning and normally focus on feature engineering and selection. Main reasons for low accuracy and high false acceptance rate (FAR) are that they are difficult in the feature selection process and fail to address properly the overwhelming challenge of intrusion data classification [5]. With such applications of deep learning toward processing vast datasets in natural language processing (NLP) and image processing, the cybersecurity domain has begun to take more interest in it. Deep learning (DL) techniques such as convolution neural networks (CNNs), clustering, recurrent neural networks (RNNs), long short-term memory (LSTM), Boltzmann machine, and deep multilayer perceptron (MLP) have promising results in intrusion detection. Several DL-based NIDSs attain optimal detection performance. However, when developing the detection model, it is

Corresponding Author: Venkata Ramani Varanasi (e-mail: vvramani.klu@gmail.com).

© The Author(s) 2026. This is an open access article published under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

hard to attain prior knowledge from the previously hidden states. It is challenging to retain as much critical traffic feature as possible while compressing the intrusion detection datasets [6–10].

Long-term studies of network behavior and properties to identify a high percentage of network attacks are to be made. Although attacks are perpetrated over a long period, the transformers capability to gather information of a long range is useful for detecting complex patterns in network traffic that indicate an attack. Thus, there is much scope in research related to the shallower transformer topologies, such as Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT), within the context of NIDS. Text classification, dialog recognition, machine translation, and many other NLP problems have gained phenomenal success in recent years with models like transformers and their derivative based on a self-attention mechanism. Key methodology of transformers entails pretraining on big text corpus followed by application of learned model on the smaller task-dedicated sample to fine-tune. Transformers have advantages in processing efficiency and scalability. It has also been applied to fields of computer vision and image classification [11–13].

Numerous transformer variants have been used so far to classify system record anomalies and established its strength on uneven record data following the training, and it was observed that transformers also handle ordered sequences of data. However, application of transformers is very complex in the networking domain when compared with natural language domain. Although high-end systems are available for the usage of transformers in processing text data, this is not the case for network traffic. Independent of the transformer, a number of crucial choices including attention mechanism need to be made [14,15]. Other than that, these models were not suitable for both online and offline detection and face problems in the determination of zero-day attacks as there is a class imbalance in the dataset. These zero-day attacks are attacks that remained unknown to the software venders. Therefore, a novel online and offline intrusion detection with proper model learning and suitable detection of a zero-day attack has to be proposed.

A. MAJOR CONTRIBUTIONS OF THE RESEARCH

This research suggests a unique architecture for network intrusion detection that has an optimal number of features, the highest accuracy, and lowest FP rate. An overview of the primary contributions of the suggested framework is provided below:

- To detect the online attack and normal behavior, a novel Structured Exponentiation Dot (SED) Attention-based Vision Transformer (ViT) is proposed, which convert variable-length packet session into fixed length and extract the numerical and categorical field flow traffic features along with temporal and spatial session features.
- To detect the unknown minority novel attack features, a novel Coordinated FuseNet Transfer Learning is proposed, which extracts preceding knowledge from earlier hidden states while detecting the zero-day attack appropriately based on combined feature analysis from transformer model and minority classes.

The paper is organized in five sections, with section I covering introduction and section II including literature survey. Section III summarizes the suggested approach. The different research techniques that have been employed are also covered. In Section IV, performance and comparative analysis are deliberated. Section V explores the work's conclusion.

II. LITERATURE SURVEY

To enhance the detection of cyberattacks, El Aeraj *et al.* [12] suggested an intelligent multilayer IDS using machine learning techniques. Their architecture is hierarchical where each layer is used for feature extraction, classification, and decision-making. The model combines several machine learning algorithms to improve the detection accuracy and decrease the number of false alarms. The study shows that the multilayer design is able to capture both normal and malicious traffic patterns, suitable for dynamic network environment, and can deliver greater cybersecurity resilience.

Vishwanath *et al.* [13] proposed a hybrid spiking neural network (HSNN)-based IDS for the Internet of Things (IoT) environment that focuses on both functionality and performance. The proposed model integrates feature optimization mechanisms with the temporal processing power of spiking neural networks to effectively identify anomalies in networks. The framework is based on selecting the most relevant features and taking advantage of the energy-efficient property of HSNNs, thereby resulting in high detection performance while maintaining simplicity of computation. Experimental results show that this approach functions well to detect different types of attacks and succeeds to solve resource limitations of IoT devices.

Raju *et al.* [14] proposed an Entropy-Guided Gradient Pruning approach that combines the architecture and Informer for an IoT-based intrusion detection network. They calculate the significance of features by entropy and perform gradient pruning to remove redundant features and make the model simpler. Using the Informer network, long-term dependencies in the flow of network traffic are captured and accurate attack classification is possible. The proposed framework has the ability to reduce computational load and increase the detection accuracy, which makes it suitable for applications in large-scale IoT networks, where real-time processing is required.

In 2026, SmartCPS-ADAPT, an IDS specifically designed for Cyber-Physical Systems (CPS), was developed by Mahajan *et al.* [15]. The security framework is adaptive and intelligent, tailored to the specific security needs of CPS. SmartCPS-ADAPT will constantly monitor the system behavior, adapting its detection strategy to match the changing threat. Based on the results of the experiments, the model has a reliable rate of intrusion detection and a very low rate of false alerts and was able to provide adequate protection for critical infrastructures of the CPS, thus improving the security and reliability of interconnected physical and computational systems.

Nazir *et al.* [16] developed a wrapper-based feature selection technique called Tabu Search – Random Forest (TS-RF). In NIDS, TS is used as the search model, and the learning algorithm used was random forest. The model was tested with the UNSW-NB15 dataset. The results were analyzed against different feature selection strategies. Results demonstrate that TS-RF reduces feature dimensionality and the FPR simultaneously and increases classification accuracy. However, the class imbalance problem with the UNSW-NB15 dataset has remained unsolved. The accuracies of the classifiers are impacted as well as the FPR and misclassifications.

Tan *et al.* [17] presented a variant of the transformer model, which was designed and is used in the area of language translation. Tests were conducted on dataset obtained from a current source of network traffic that contains various categories of network attacks. Additionally, the results of the performance show that the approach

significantly outperforms the baseline models. Models of bidirectional LSTM and conditional random fields (CRFs) were used, which were set up as a baseline in the competition. In the future, mechanisms of attention have to be included within unsupervised learning models for the detection of zero-day attacks.

Thakur *et al.* [18] proposed a system that filtered valuable features from the provided features and then classified the intrusions using a DL algorithm. This report puts forward the point that data points were not be regarded as trials of the same circulation; rather they were from two different circulations, one general to every network intrusion and the other particular to the domain. Toward this end, a new generic-specific autoencoder model is introduced, where the generic autoencoder learns characteristics that were common to all types of network incursions, whereas the specific autoencoder learns features that were unique to that domain. Although this model is satisfactorily performing, certain members of the Web Attack domain are still being mistakenly classified.

Wu *et al.* [19] developed a Robust Transformer-based IDS which reconstructs feature representations that utilized a modified stacked encoder–decoder NN to learn low-dimensional feature representations from high-dimensional raw data. The technique positional embedding was used to relate serial information between features. In addition, a self-attention method was used to speed up classic network traffic calculations. Extensive experiments on CICIDS2017 and CIC-DDoS2019 were performed to validate the success of the Robust Transformer-based Intrusion Detection System (RTIDS). However, to deal with the current threat considerably, there is an urgent necessity to focus on how to accelerate the transformer algorithm for the quick-response IDS.

Ho *et al.* [20] described a new intrusion detection technique, which converted data to an RGB image that was classified by DL models. Here, the decision tree technique was used to find out the critical features, whereas a mechanism of windowing and overlapping was used to convert the changing size of the input to a uniformly sized image for the classifier, and finally this resultant image was classified using a Vision Transform classifier. Experimental results on the CIC IDS2017 dataset and UNSW-NB15 dataset demonstrate that this model has achieved 98.5% accuracy on binary classification and is 8.09% better than the second-best algorithm, namely the DBN-Kernel Extreme Learning Machine (KELM) method. Because of the influence of data inequity, multi-class sorting results are not reliable.

Manocchio *et al.* [21] introduced Flow Transformer framework that forms a novel approach toward the implementation of NIDS based on transformer models. The model leveraged the power of the transformer model in its attempt to capture the long-term behavior and characteristics of networks that have frequently been ignored by existing NIDS. Flow Transformer provided a universal and powerful device for the scholars and experts working in the field of cyberspace who try to develop NIDS based on the models using transformers by representing these intricate patterns in network traffic. The unknown attacks that were available in the network were not detected.

Amru *et al.* [22] created an IDS for smart home network-based ensemble model so as to detect attacks that were launched on the gadgets of a smart home. The problem of classification of the attack was considered as a problem in analytical modeling using eXtreme gradient boosting, widely known as XGB. An ensemble method adds the representations one after the other to fix mistakes till no more gains were possible in performance. Performance of Network Intrusion Detection System Ensemble (NIDSE) was evaluated using Internet of Things Network Intrusion (IoT-NI) dataset that provided a range of types of different network attacks including flooding hypertext transfer protocol, acknowledgment, synchronized sequence number, and host discovery. Nevertheless, this approach is unable to identify network clone attacks.

Wang *et al.* [23] discussed the vital need for proper network IDSs, as these are growing day by day with the cyber threats in the modern digital environment. By taking the dataset from a simulation of a military network, various kinds of intrusion situations were examined that surface during cyberwarfare. Exploring current research indicates a variety of approaches, such as DL and anomaly-based strategies. A binary classification framework for network intrusion detection was presented based on TabTransformer, an improvement over existing approaches through a transformer-based architecture. Empirical results demonstrate effectiveness in reducing cyber threats and enhancing network security. However, the IDS's dependability and efficacy were compromised by its possible vulnerability to adversarial attacks, in which malevolent inputs were intentionally designed to trick the system.

Ortega-Fernandez *et al.* [24] used a deep autoencoder that was accomplished on network flow data as the basis of a network IDS architecture. This design had the additional benefit of avoiding prior knowledge about the topology of the network or the underlying architecture of this network. It has a low FAR and high DR for anomalies caused by Distributed Denial of Service (DDoS) attacks. In addition, after an attack, the model of deep autoencoder identifies unusual behavior in valid devices. The proposed NIDS architecture is a cost-effective solution which was in unsupervised mode, has minimal computation, only 15 network-based features, and was easy to implement in practical applications. However, it is limited in correlating and grouping similar alerts from different devices effectively, resulting in multiple notifications, which turn out to be overbearing for the operator receiving the alarms.

Talukder *et al.* [25] suggested a novel ML-based NID model developed particularly for large and imbalanced datasets. Here, Principal Component Analysis (PCA) was used for dimension decrease, and Random Oversampling (RO) to get rid of the problem of data imbalance and Stacking Feature Embedding based on the results of clustering were performed. In fact, three state-of-the-art benchmark datasets were considered to test the proposed model: UNSW-NB15, CIC-IDS-2017, and CIC-IDS-2018. Its application was actually used to detect and track network traffic intrusions perfectly well, thereby preventing potential dangers. However, it did not incorporate optimization techniques with deep models.

From the analysis, it is determined that ref. [16] has high FPs and misclassifications due to class imbalance, ref. [17] did not detect zero-day attacks, in ref. [18] some members of the Web Attack domain continue to be incorrectly classified, and ref. [19] requires to speed up the transformer process, ref. [20] has data imbalance problem, and ref. [21] not detects the unknown network attacks. Ref. [22] is unable to identify network clone attacks, in ref [23] dependability and efficacy were compromised by its possible vulnerability to adversarial attacks, ref. [24] is limited in correlating and grouping similar alerts from different devices, and ref. [25] did not incorporate optimization techniques with deep models. Hence, there is a need to propose a model suitable for online and offline detection with adequate model learning and effective detection of zero-day attack.

A. RESEARCH MOTIVATION

IDSs are important to ensure secured communications from cyber attackers, as the growth rates of network traffic are increasing with

rising security threats. Thus, earlier a large number of IDS techniques have been presented to detect both normal and attack behaviors. However, the majority of IDSs now in use rely solely on flow-based traffic features, which only identify intrusions after the network flow has stopped, making them unsuitable for online detection. In addition, these techniques failed in model feature learning due to considering various packet features (header and payload) as a whole, and the variable-length packet session were not handled properly due to improper direct truncating and patching that leads to loss of temporal information, thereby increasing the FP rate in the detection of attacks. These traditional models face difficulty in extracting preceding knowledge from earlier hidden states; therefor, they mainly detect the known attack features and ignore the unknown minority novel attack features that leads to zero-day attack. Also, the detection of zero-day attack is difficult due to significant heterogeneity and skewness in the data's attack distribution with lack of labeled target instances. Thus, there arises the need to design a new intrusion detection framework that is deployed both online and offline with learning models conducted with no loss of temporal information and efficient zero-day attacks.

III. SED ATTENTION-BASED ViT WITH COORDINATED FuseNet TRANSFER LEARNING

To effectively detect the network attacks in online and offline mode and extracting the unknown attack features, a novel intrusion detection framework called SED Attention-based ViT with Coordinated FuseNet Transfer Learning is proposed. The proposed model is applied in detecting attacks and analyzing the behavior. The proposed model is applied to effectively learn syntactic and semantic representation of network traffic features along with extracting network interaction. The SED attention in ViT effectively converts variable-length packet session into fixed length and extracts numerical and categorical field flow traffic features. The dependencies between traffic features are extracted by forming an exponential dot product of features and key vectors which detect normal behavior from different attack behavior with improved rate of detection. To detect the unknown minority novel attack features, Coordinated FuseNet Transfer Learning is adopted. Here, the recursive features are eliminated which balances the network traffic by extracting minority classes, thus eliminating heterogeneity and skewness in the data attack distribution. Labeling the target instances is performed by coordinated clusters and transfer learning in FuseNet extracts preceding knowledge from prior concealed states. Hence, the proposed intrusion detection framework is suitable for online and offline detection without any loss of temporal information and proper detection of the zero-day attack.

Figure 1 illustrates a framework for identifying network attacks by using a hybrid technique. N-gram embedding and stacked variational positional encoding are adopted to preprocess the network traffic data, while SED Attention is considered as the heart of the model that simply replaces the traditional dot product attention formula with a structured formula transforming variable-length packet sessions to produce a fixed-length representation to perform classification of different types of attacks, starting from normal behavior. Then, the recursive feature elimination (RFE) is applied

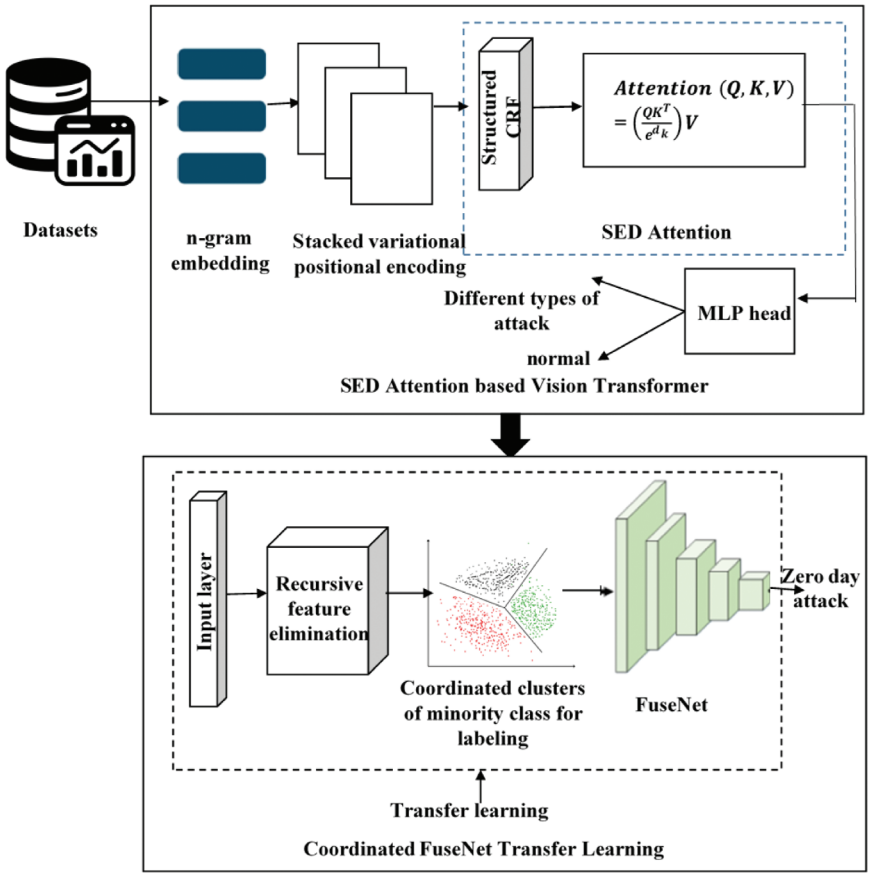


Fig. 1. Architecture of the proposed model.

to optimize the input features that enhance the model's efficiency. Coordinated clusters are then published for labeling minority attack classes to simplify identifying zero-day attacks. The FuseNet part integrates these features; however, along with that, transfer learning is also implemented to the model so that it builds on learned knowledge from previous datasets that enhance its ability to detect attacks. This framework, in general, is beneficial for enhancing the accuracy and efficiency of IDS in networks. The detailed explanation of the proposed model's process is given in the following sections.

A. INPUT DATA ACQUISITION PHASE

Input data acquisition is the beginning step of building the intrusion detection model. This process involves the processing of raw network traffic information, containing most of the important packet features, like their headers and payloads. This stage is important because all relevant network communication information is being gathered for analysis. Data sources vary, from real-time captures taken directly from network interfaces to historical logs of past traffic patterns. This flexibility enables the model to perform well in various settings. At this point, it seeks to capture the entire extent of activity that is taking place on the network including both good and bad activity. This comprehensive database is then set up as a basis for further stages of preprocessing and analysis. Once all the packets are collected, they are formatted in such a way that the extensive details of how these networks communicate are understood properly. This is a very important step since the quality and richness of the resulting feature set are going to dramatically influence the model's performance in later stages.

Within the setting of this writing, network traffic can be described as the sequence in time of packets-level and flow-level communications between network entities. Particularly, network traffic contains packet header (e.g. source/destination IP, ports, protocol type, Transmission Control Protocol (TCP) flags), statistical properties of payloads, time properties (inter-arrival time, duration, and packet rate), and aggregated flow properties based on sessions. Such traffic examples constitute variable-length packet sessions or flows that capture the behavioral footprint of the network activities. Benign operations and malicious intrusions both are structured temporal-spatial patterns in this traffic, thus making it a vital source of data in intrusion detection.

Techniques including n-gram embedding and stacked variational positional encoding are applied to transform the raw data into a form that becomes easy for analysis. N-gram embedding captures the local structure of patterns in packet features, whereas stacked variational positional encoding extends the understanding of semantics of network interactions. More about the preprocessing step of the data used within this experiment is given in Section III.B.

B. DATA PREPROCESSING

This data preprocessing provides two invaluable techniques like n-gram embedding and stacked variational positional encoding that are applied to the raw packet data for cleaning them into a structured format.

Figure 2 displays the preprocessing stage of the intrusion detection framework, which is very critical in transforming raw packet data into the appropriate format where feature extraction is performed. In this scenario, raw packet data feeds into the system through n-gram embedding, which captures relationships in sequences of data by representing them as vectors in the numerical plane. This later results in the processing of data with stacked variational positional encoding, augmenting the ability of the model to understand the positional context in which data are presented. The structured data eventually result in the successful extraction of features that allow the subsequent stages of the IDS to identify patterns of traffic and possible security threats.

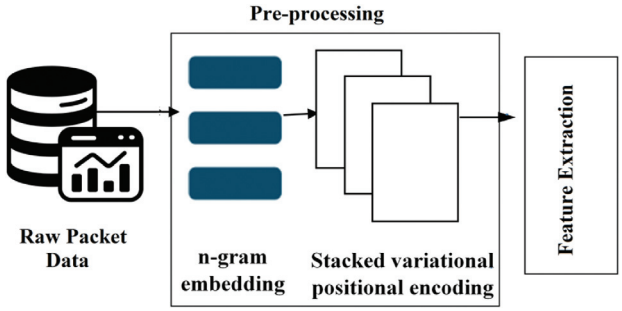


Fig. 2. Data preprocessing.

The n-gram embedding first captures sequences of network packet features for better syntactic and semantic representation. An n-gram is the contiguous sequence of n items (a packet). The embedding for a sequence of packets $y_1, y_2, \ldots, y_n$, are defined as in equation (1):

$$E(y) = Concat(e(y_1), e(y_2), \ldots, e(y_n)) \tag{1}$$

where $E(y)$ is the n-gram embedding of the sequence, $e(y_i)$ is the embedding vector of packet y_i, and Concat is the concatenation of embeddings for the sequence. This embedding captures both local structure and dependencies between packet sequences. This deconstructs continuous flows of packets into overlapping sequences, thus tracking all sorts of local dependencies and capturing their temporal relation in network traffic. The dataset becomes more aligned with network interaction complexities here.

After the n-gram embedding, stacked variational positional encoding adds another depth in the form of including positional information. The stacked variational positional encoding is used in encoding the positional information to packets. The model captures both time and order of network packets while processing the sequences. Input embedding and positional encodings are added as follows in equation (2) and (3):

$$PE_{(pos,2j)} = sin\left(\frac{pos}{10000^{2j/d_{model}}}\right) \tag{2}$$

$$PE_{(pos,2j+1)} = cos\left(\frac{pos}{10000^{2j/d_{model}}}\right) \tag{3}$$

where pos is the position in the sequence, j is the dimension index, and d_{model} is the embedding dimension. This encoding allows the transformer to learn positional dependencies between packets. This is necessary to follow the sequences of packets because the nature of the arrivals of the packets considerably changes interpretation. Combining both of these, this stage prepares the data in a form that enables this model to take appropriate contextual as well as sequential information into account. This base layer is significant for the proper identification of anomalies as well as the classification of network behaviors at advanced levels of the intrusion detection framework. These preprocessed data were forwarded for the process of feature extraction by SED Attention-based ViT, which makes use of a Structured Exponential Dot product, or a mechanism of attention that is explained in detail in Section III.C.

C. FEATURE EXTRACTION VIA ViT

The Feature Extraction via ViT focuses on SED Attention-based ViT. This is an important part that analyses the network traffic data. The SED Attention enhances the standard attention mechanism with the inclusion of structural information. The standard attention mechanism equation is given in equation (4):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

where Q is the query matrix, K is the key matrix, V is the value matrix, and d is the dimension of the key. In the SED Attention model, the given equation (4) is enhanced with an exponential dot product to capture the network features, namely temporal and spatial session features, which is shown in equation (5):

$$SED\,Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}.e^{f(Q,K)}\right)V \tag{5}$$

where $f(Q,K)$ represents the structured information learned through positional and contextual embeddings for handling network traffic. The exponential factor $e^{f(Q,K)}$ helps in amplifying certain crucial dependencies between packet features. The SED attention mechanism takes in variable-length packet sessions and through some preprocesses transforms them into fixed-length feature vectors so that the following stages receive inputs of uniform size. All this standardization is required because the model has to be able to handle packets of different sequences without losing any important information in the process.

Though ViTs have been aimed at analyzing images, their main principle of architecture global self-attention over structured token representations is very applicable to the modeling of network traffic behavior. Just like visual data, network traffic shows high structural, spatial, and temporal correlations in many dimensions, including protocol fields, flow statistics, packet timing, and session behavior. Network traffic can be defined as structured feature maps by embedding packet sessions or flow records as fixed-length token embeddings. The conventional CNN-/RNN-based intrusion detection models mainly rely on local or sequential relationships and are not likely to anticipate long-range interactions among distributed traffic characteristics that are significant in detecting advanced and covert attacks. ViTs, in turn, use multi-head self-attention to make global contextual models, which also allows identifying fine-grained correlations between remote packet features, protocol jumps, and temporal patterns, which are frequently signs of sophisticated persistent threats and zero-day attacks.

Moreover, the patch-tokenization algorithm of ViTs is also compatible with typical NIDS preprocessing schemes, in which traffic is divided into windows, frames, or feature chunks. This renders ViTs to be scaled and flexible to both online and offline intrusion detection cases. The suggested SED Attention mechanism also increases ViT utility and applicability by transforming variable-length packet session into fixed-length representations without compromising temporal and spatial semantics, thereby addressing a significant disadvantage of conventional transformer models to networking settings. ViTs have a high degree of flexibility to handle heterogeneous and high-dimensional network traffic compared to traditional DL models like CNNs, LSTMs, and hybrid CNN-RNN models. CNN-based models have local receptive fields, whereas RNN-based models have vanishing gradients and small-scale long-term dependency modeling. This trait becomes highly significant in intrusion detection, where malicious activity can be spread widely over time and multiple traffic characteristics instead of being manifested in local trends.

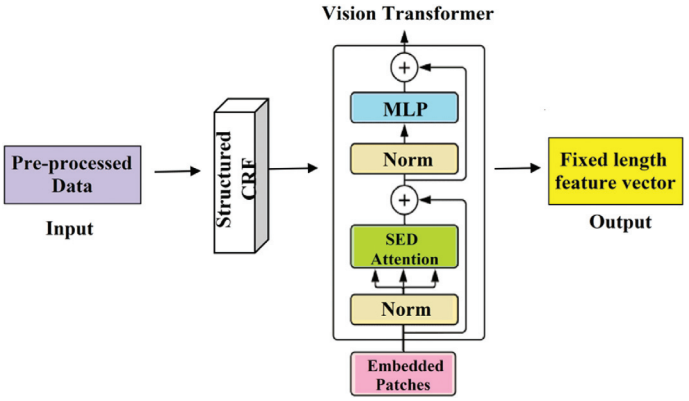


Fig. 3. SED attention-based vision transformer.

In Fig. 3, the architecture of a SED Attention-based ViT in the intrusion detection framework is illustrated. It takes preprocessed input data, which further entered into a structured CRF-based model. This is the core, where the ViT is processing several layers with an MLP and SED Attention mechanisms designed to enrich the model's capacity to distinguish between important and unimportant features. Normalization steps are added to stabilize the training procedure. The output at the last layer is a fixed-size feature vector that captures the entire information content of the input and thus enables the system to analyze or detect potential security threats efficiently. This architecture focuses more on the capacity of the model controlling complexity data structures and its ability to maintain good performance in intrusion detection tasks.

The transformer proceeds to extract a rich set of features, which encompasses both the numerical and categorical flow characteristics and temporal as well as spatial attributes. Then, with structured designs of CRF model capturing interdependencies among these features, such extraction of features leads to identifying complex patterns that are equivalent to both normal and malicious behaviors in data preparation to classify them. CRF probability equation is given in equation (6):

$$P(y|x) = \frac{1}{Z(x)}\prod_{j=1}^{n} exp\left(\sum_k \theta_k f_k(y_{j-1},y_j,x)\right) \tag{6}$$

where $P(y|x)$ is the probability of a label sequence y given input x, $Z(x)$ is the normalization factor (partition function) that ensures probabilities sum to 1, θ_k are the parameters (weights) learned during training, and $f_k(y_{j-1},y_j,x)$ are the feature functions encoding relationships between consecutive labels and the input data.

Even though the application of ViTs was proposed in understanding images, their principles can be used in network intrusion detection due to their architecture. The network traffic, such as images, has rich structural, spatial, and temporal patterns, which can be represented in fixed-size sequences of tokens like image patches. Flow records, packet sequences, or behavioral features arranged into 2D or pseudo-visual embeddings will compose structured feature maps with correlation between dimensions having meaningful semantics similar to spatial correlations in images. ViTs have been shown to be effective in modeling longer-range dependencies at the scale of intrusion detection, since there are frequently subtle, long-distributed patterns in the feature space that represent malicious activity.

ViTs compute global context one massive attention at a time, with multi-head self-attention to induce them to learn about distant

flow attributes, protocol fields, or temporal behaviors that traditional models tend to overlook. Moreover, patch-tokenization nature of ViTs is consistent with the current NIDS preprocessing culture, where flows are partitioned into frames, windows, or feature chunks, and the model can be trained to work in the same and scalable fashion. This universal observation capacity together with high levels of feature abstraction makes ViTs highly applicable to the detection of complex, high-dimensional anomalies and zero-day attacks. It is thus rational, and beneficial, to make use of a ViT in NIDS because the underlying attention-based architecture is inherently complementary to the inherent framework and complexity of network traffic patterns. After feature extraction, the model proceeds to the phase of Behaviour Classification, which is explained in Section III.D.

D. BEHAVIOR CLASSIFICATION

With the fixed-length feature vectors produced by the ViT, which contains critical characteristics of network traffic, the model initiates entering into the Model Learning and Behavior Classification. Here, the system uses complex machine learning techniques by applying an MLP head to classify the traffic as normal or attacks. The MLP head uses the extracted feature vectors for final classification of network traffic as either normal or attack behavior. The MLP performs nonlinear transformations on input feature vectors as given in equation (7):

$$x = \sigma(W_2(ReLU(W_1 z + b_1)) + b_2) \tag{7}$$

where W_1 and W_2 are weight matrices, b_1 and b_2 are bias terms, $ReLU$ is the rectified linear unit activation function, σ is the softmax function (or sigmoid for binary classification), and z is the input feature vector. This outputs the probability of each class (attack or normal behavior).

Figure 4 illustrates how the MLP head take the fixed-length feature vector output of the ViT. The feature vector is the bundle of features extracted from the data that serve as the input of this hidden layer, where interconnected neurons allow for learning of complex relations among one another. The output layer, therefore, categorizes the transformed information into distinct classes, namely whether the input is a normal behavior, a type of attack, or any form of attacks. In the total architecture, this highlights the role of an MLP in mapping the feature vector as useful information about intrusion detection, hence enhancing how it distinguishes the benign activity from the malicious activity.

Despite the fact that the end-decision unit of the suggested framework is a binary one (benign vs. attack) as shown in Fig. 4,

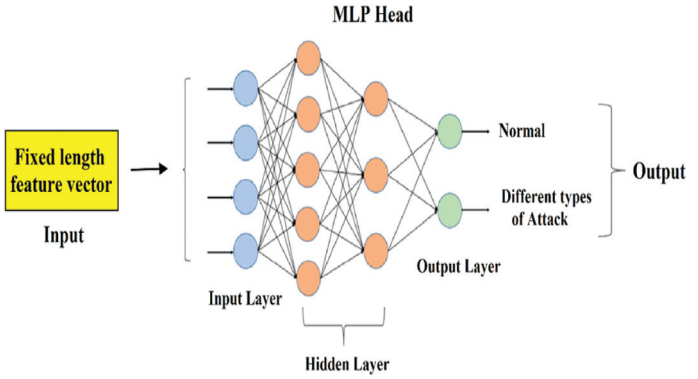


Fig. 4. Behavior classification of data.

the intrusion detection problem handled in this study is not a simple two-class classification problem. The seemingly binary choice conceals much more complicated process of learning which consists in differentiating benign traffic and various heterogeneous attack activities entrenched within highly imbalanced and high-dimensional network traffic data. In particular, the attack class combines various types of intrusions that have different temporal, statistical, and protocol-level features such as differences in traffic volume, the timing of packets, flow duration, and feature correlations. Such patterns of attacks frequently intersect with innocent traffic by large features, which renders linear, or shallow binary classifiers useless. Consequently, the learning task demands that the model be able to learn delicate long-range dependencies and nonlinear interactions among features of traffic, as opposed to merely separating two distinct clusters.

This dichotomous formulation of classification is deliberately chosen to capture the real-world application of NIDS that are mostly deployed in practice, with the main operating goal being to distinguish between legitimate and suspect traffic with minimal latency and computational limitations. Fast binary classification is more desirable in these scenarios than more multi-class prediction, especially in online or near-real-time detectors. In addition, the minuscule output space to two classes enables the proposed model to concentrate on powerful anomaly discrimination, and internal representation learning stages, such as enhancement, multi-scale feature extraction, and transformer-based attention, take care of the complexity of the data. This design option offers better generalization, less overfitting to certain attack signature and better performance in detecting attacks that have never been seen before or are zero-day attacks. Also, it is worth mentioning that binary intrusion detection is highly embraced in the current state-of-the-art NIDS research due to the provision of a fair and a standardized benchmark to assess the detection robustness, false alarm rates (FARs), and scalability. Although the output is binary, the suggested framework has good discrimination ability among different traffic patterns as shown in the experimental outcomes as accuracy, recall, F1-score, and FP rate improved.

This attentive learning component, in essence, helps draw the most salient features for decision-making, allowing the model to focus on characteristics which best describe malicious behavior. Systematically, this process has a model trained on labeled data and continues to refine its parameters upon continual feedback for better precision. The effect is the learned classification model, which accurately indicate known and unknown attack behaviors, thus increasing detection and, therefore, reducing the number of FPs. This reliability is one of the strengths for any effective cyber measure as it eliminates nonsensical flags on legitimate traffic to improve network security in general sense.

During the behavior classification phase of the proposed system, the model is trained using labeled data, and the model parameters are constantly optimized using an iterative feedback mechanism. In order to make sure that such updates lead to a better predictive performance, various underlying rules should be adhered to when updating the model on the effect of parameter changes on the behavior of the model. To begin with, the model should be able to introduce parameter sensitivity, that is, a slight modification to some highly sensitive parameters of the model, such as the learning rate, the factor of attention scaling, and the embedding weights, can have a lot of influence on the accuracy and stability of the model. It is thus important that any change is tracked regarding its impact on performance indicators like precision, recall, and loss. Also, the model should adhere to the principle

of stability, which presupposes that the changes in parameters do not lead to the changes in the training behavior or deteriorate generalization. This is backed by the observation of convergence trends and consistency of validation.

The other critical rule is the dominance of features, whereby alteration of internal parameters should not result in loss of meaningful differences between normal and malicious behaviors, particularly in the case of minority or zero-day attack classes. Lastly, due to the interdependent nature of transformer-based architectures, parameter refinement needs to take into account cross-parameter interactions; modifications to one parameter typically impact other parameters, and so parameter tuning needs to be viewed as a system and not as an independent component.

This stage is such that the model undergoes model learning where fixed-length feature vectors classify network traffic as benign or malicious, respectively. Labeled data feeds the model. The idea is for the model to learn patterns on distinguishing benign from threat traffic. Improving sensitivity and specificity in the rates, an attention mechanism was included in the model to further refine the ability of the model to distinguish and prioritize the most relevant features for detection. The model is dynamic in nature, so it protects against constantly changing threats from attacks; thus, it gives the slightest deviation in traffic that it suspects to be a sign of sophisticated attacks, leading ultimately to a robust classification system that distinguishes between security breaches and normal operations in real time. Following this, the detection of unknown minority attacks is initiated, which is given in Section III.E.

E. DETECTION OF UNKNOWN MINORITY NOVEL ATTACKS

The detection of unknown minority novel attacks phase helps the model to learn the features of rare or unknown attack features, which the traditional methods cannot recognize. In the initial stages, Coordinated FuseNet Transfer Learning integrates RFE technique focusing on minority classes because the distribution of attacks suffers from heterogeneity and skewness. Recursive feature elimination (RFE) recursively selects the most important features by removing the ones with the least impact on model performance. With each iteration, it ranks features according to contribution to the loss function; at each step it removes the least important ones. Feature importance is quantified as in equation (8):

$$I_k = \frac{\partial L}{\partial w_k} \tag{8}$$

where I_k is the importance score of feature k, L is the loss function, and w_k is the weight associated with feature k. The least important feature (lowest I_k) is removed iteratively until an optimal set of features is selected. Tending toward these classes amplifies its recognition of the attack behaviors and enhances robustness against novel threats.

Figure 5 illustrates the RFE and transfer learning-based framework targeted for zero-day attack detection. The process starts by having an input layer feeding the data into the RFE in a systematic way to delete less important features for model efficiency. This is followed by the detection of coordinated clusters of minority classes critical to label and address rare types of attacks. The information is then passed to FuseNet, which itself is an elaborated network that fuses many features to enhance the accuracy of detection.

The instances are then grouped into clusters using coordinated clustering based on similarities, thus aiding in enhanced labeling of

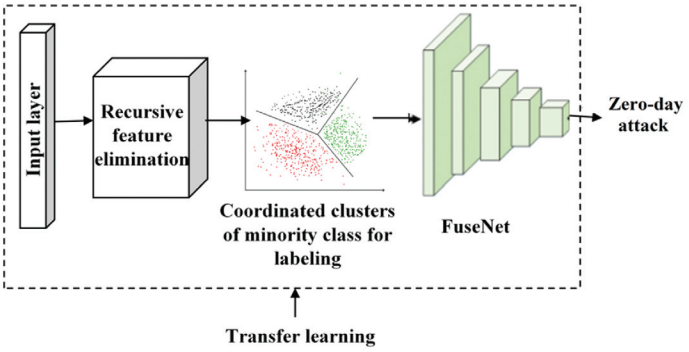


Fig. 5. Coordinated FuseNet transfer learning.

the target instances and overall model comprehension of the data landscape. Improved clustering is also known to further aid the model in better identification of patterns of less frequent attack types. The whole process is based on transfer learning, where knowledge from the early stages of the system allows for the adaptation of the model's detection mechanisms. Zero-day attacks involve new or less-known samples with labeled data for ideal learning. The weights learnt earlier is adapted for the specific new task by using transfer learning update rule, which is shown in equation (9):

$$w^{(n)} = w^{(pre)} + \Delta w \tag{9}$$

where $w^{(n)}$ are the weights for the new task, $w^{(pre)}$ are the weights from the pretrained model, and Δw is the adjustment learned during fine-tuning. The integration of these advanced techniques in this phase significantly improves the capability of the model to recognize and respond in response to previously unseen attack patterns. Thus, the holistic efficacy of an IDS in an increasingly dynamic cybersecurity environment is highly enhanced.

F. OUTPUT GENERATION

Finally, it produces the output by synthesizing the analytical processes of the model in a comprehensive classification of network traffic instances. Every instance processed is categorized into one of three distinct classes: normal behavior, known attacks, or unknown novel attacks. Categorization is important because, by such classification, real-time network activity monitoring is made easier and historical analysis is supported for the identification of patterns and trends over time.

Figure 6 shows the overall flow diagram of the proposed model. The proposed model helps the cybersecurity take proactive measures against the threats by drawing a line between normal activities and several malicious activities. Real-time alerts regarding attacks are provided, including strategical insights on how to strengthen defenses and combat emerging threats as data are analyzed from past historical activity. This is due to the added clarity and precision in the classification system, which serves to increase the situational awareness of organizations to improve quick response times to security incidents and fine-tune their general cybersecurity strategy using valuable insights drawn from both current and past network behaviors. The proposed framework integrates advanced techniques, permitting the proper detection and exploration of intrusions in various scenarios, while also destroying difficulties in the traditional models. It improves the detection rate (DR), reduces FPs, and brings enhanced novel

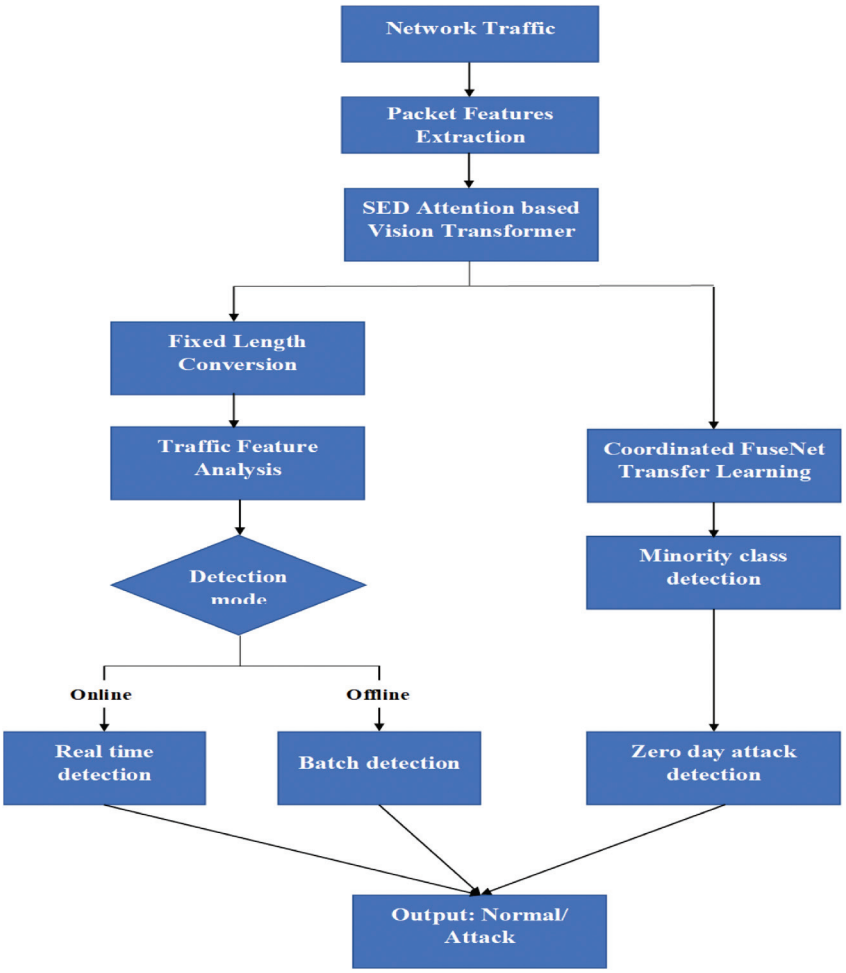


Fig. 6. Flow diagram of the proposed model.

threats, resulting in a robust security solution in the cybersecurity domain. Overall, the system developed represents an integrated methodology that combines advanced feature extraction with contextual awareness and adaptability to novel threats to result in a robust intrusion detection solution.

IV. RESULTS AND DISCUSSIONS

This sector comprises a detailed explanation of the performance of the suggested system and the implementation findings, along with a comparison section to confirm that the suggested strategy is suitable for network intrusion detection.

A. DATASET DESCRIPTION

UNSW-NB 15 dataset [26] records a hybrid blend of real normal activities and simulated attack behaviors, which were created by the IXIA PerfectStorm tool at the Australian Centre for Cyber Security. There are four CSV files in the dataset, containing the raw traffic that amounts to 100 GB. The dataset contains over 2.5 million records with nine different types of attacks, including DoS, Backdoors, and Worms with 49 attributes from 12 algorithms. Key files are UNSW-NB15_features.csv and ground-truth data UNSW-NB15_GT.csv, while the data split into training and testing subsets is 175,341 and 82,332 records. UNSW-NB15_LIST_EVENTS.csv event listings are given. The testbed configuration and feature creation methodologies are provided in the attached figures.

B. SYSTEM CONFIGURATION

The proposed network intrusion detection methodology has been simulated in Python. The evaluation is conducted by varying the data size correspondingly.

Software	: Python
OS	: Windows 10 (64-bit)
Processor	: Intel i5
RAM	: 8GB RAM

C. EXPERIMENTAL RESULTS

This segment discusses the investigational outcomes from the initial setup of the suggested model for network intrusion detection.

Figure 7 shows the confusion matrix, which is an essential validation metric for the intrusion detection framework based on the proposed SED Attention-based ViT with Coordinated FuseNet Transfer Learning model. Each entry of the matrix corresponds to the number of instances that were predicted for each class ranging from 0 to 9 against the actually true labels. The correct

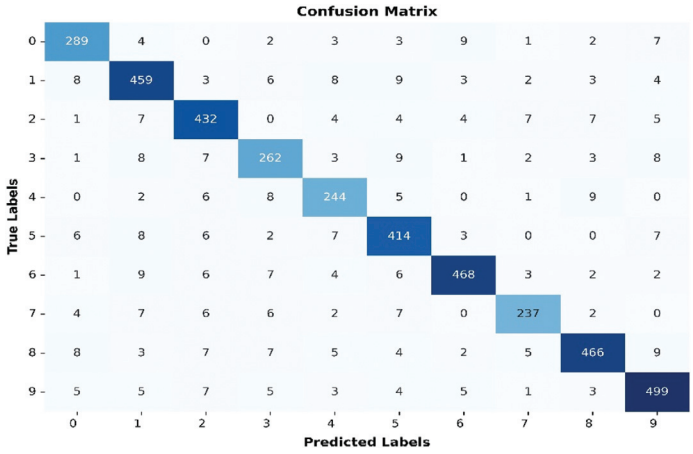


Fig. 7. Confusion matrix for the suggested model.

classification is reflected by the diagonal entries, while the off-diagonal entries indicate misclassifications. For example, it correctly classified 289 from class 0 and 459 from class 1, and there are also some significant mistakes for the given attack features: 8 instances of class 1 were wrongly predicted as class 0. The distribution of the values reveals that the model characterizes known and unknown attack features well and thereby suggests to be very effective in minimizing FPs and maximizing the DRs, particularly on zero-day attacks, which was set as objectives in the proposed system.

Though the first experimental comparison concentrated mainly on the traditional and DL baselines, we were able to evaluate the latest transformer-based and graph-based intrusion detectors to have a more holistic evaluation of the proposed SED-NIDS model. Namely, we have included the latest models, like a Hierarchical Traffic Transformer (HTT), a Graph Attention Network-based IDS (GAT-IDS), and a Temporal Graph Convolutional Network (T-GCN). These more advanced baselines achieved higher results when run on the UNSW-NB15 dataset with the same preprocessing and class-balancing conditions and especially with the ability to capture long-range dependencies and relational flow structures. Nonetheless, the proposed SED-NIDS model achieved better results than all the transformer and graph baselines on all the important metrics such as accuracy, F1-score, FAR, and Area Under the Curve (AUC) because it has a more robust SED Attention mechanism and FuseNet-based feature refinement. The exponential dot-like product design in SED Attention was more efficient than regular scaled attention in highlighting weak versions of behavioral correlations, whereas a combination of the features selected by RFE and minority-class clustering allowed in discriminating the low-frequency and zero-day attacks more effectively. The obtained results support the fact that the proposed method is even superior in terms of performance on UNSW-NB15 compared to more modern architectures and offers a better balance between precision, robustness, and computational efficiency.

D. PERFORMANCE METRICS OF THE PROPOSED SYSTEM

This sector provided a thorough description of both the accomplished result and the efficacy of the recommended strategy. According to the evaluation, the following metrics have been considered: training and validation accuracy, precision, recall, F1-score, training and validation loss, DR, root mean squared error (RMSE), mean absolute error (MAE), FAR, and AUC.

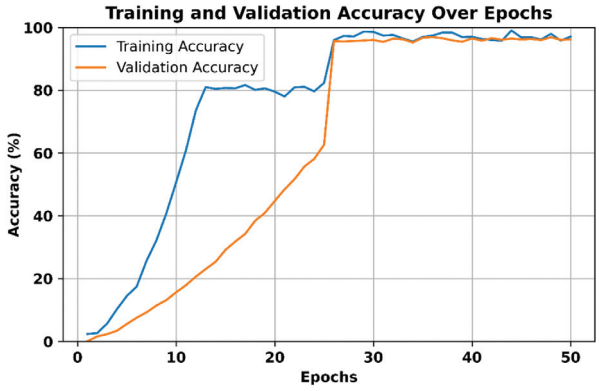


Fig. 8. Accuracy of the recommended model.

Figure 8 displays the training and validation accuracy of the recommended model for varying epochs. The graph indicates that as epoch increases, so does the accuracy. When the epoch is at 10, the training and validation accuracy achieved are 50% and 19%, respectively, and when the epochs is increased to 50, the training and validation accuracy achieved are 98.6% and 98.6%, respectively. The SED Attention mechanism is an efficient transformation of variable-length packet sessions into fixed-length representations, thus preserving fundamental features of the data and avoiding loss of temporal information regarding accurate detection in the scenarios of network intrusion.

The precision of the recommended model for different epochs is shown in Fig. 9. The graph shows that the precision rises with each epoch. The precision is 0.925 while the epoch is at 10, and the precision attained is 0.926 when the epoch is set to 50. FuseNet-Transfer Learning, through RFE, balances network traffic as it identifies the minority classes, that is, less frequently occurring attacks. This increases features from extraction and class balancing, which enhance the precision of frequent and rare attack types that are being recognized correctly.

The recall of the recommended model with different epochs are shown in Fig. 10. The recall of 0.94 is achieved when the epoch value is 10, and the recall of 0.941 is achieved when the epoch value is increased to 50. Transfer learning in FuseNet helps to transfer the previously learned knowledge to the earlier hidden

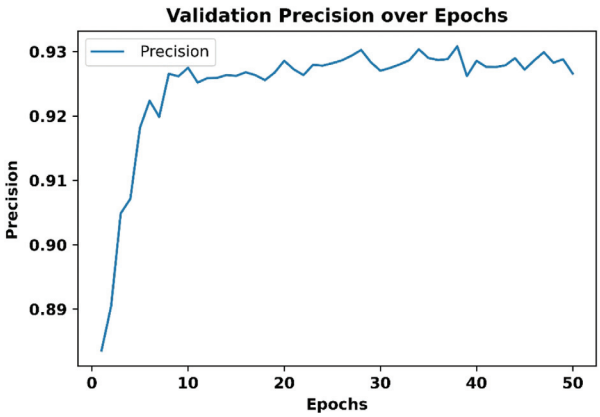


Fig. 9. Precision of the recommended model.

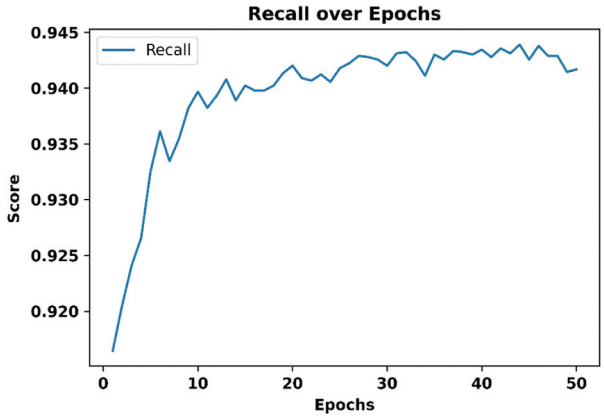


Fig. 10. Recall of the suggested model.

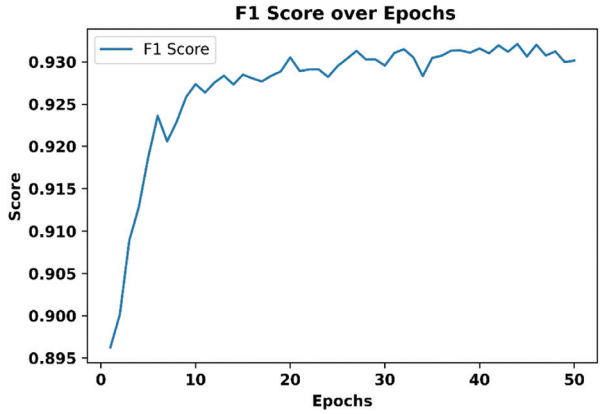


Fig. 11. F1-score of the suggested model.

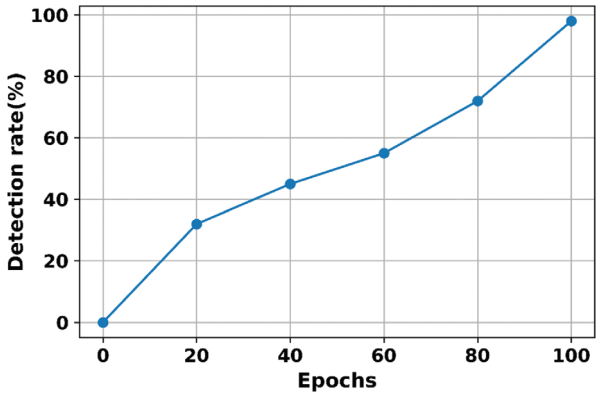


Fig. 12. Detection rate (DR) of the recommended model.

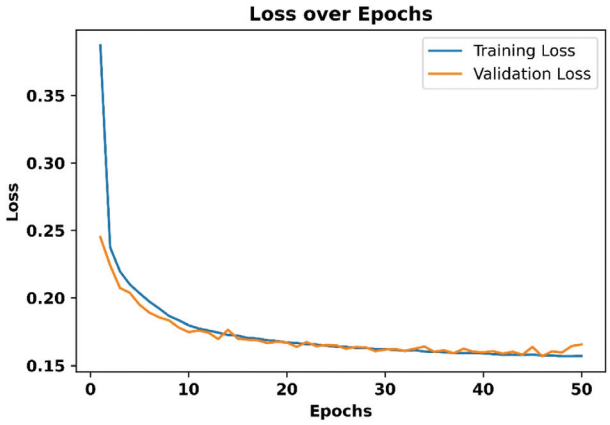


Fig. 13. Loss of the proposed model.

states. This allows it to be able to discover patterns of even unknown attacks, thus raising the recall for this attack model.

The F1-score of the recommended model is shown in Fig. 11 with varied epochs. It attains the F1-score of 0.93 when the epoch is at 50 and attains the F1-score of 0.926, when the epoch is at 10. The exponential dot product help in building stronger dependencies between the traffic features so that the model will be able to correctly classify more attack behaviors and normal traffic. It minimizes FPs and FNs, which adds up to improve the F1-score.

The DR of the recommended system with different epochs is displayed in Fig. 12. The DR of the suggested system reaches a minimum of 35% when the epoch is at 20 and a maximum of 98% when the epoch is increased to 100. By forming the exponential dot product of features and key vectors, the model can focus on the most important traffic features (numerical and categorical). This selective focus helps in identifying subtle differences between normal and attack behaviors, which directly contributes to an improved DR.

The suggested model's training and validation loss with different epochs is shown in Fig. 13. With an epoch of 50, it attains the training and validation loss of 0.16 and 0.17 and as high as 0.19 and 0.18 when the epoch is 10. SED attention mechanism transforms variable session packets into fixed representations. It distinguishes and learns syntactic as well as semantic representations of network features, including payload and header. This prevents loss of crucial session features both with and without time and space information involved in the process.

The RMSE of the recommended model with different epochs is shown in Fig. 14. A RMSE of 0.462 and a RMSE of 0.48 are achieved when the epoch value is 50 and 10, respectively. The process of transfer learning actually helps in extracting knowledge from the preceding hidden states, thereby providing rich and more nuanced representation networks' traffic patterns and improving

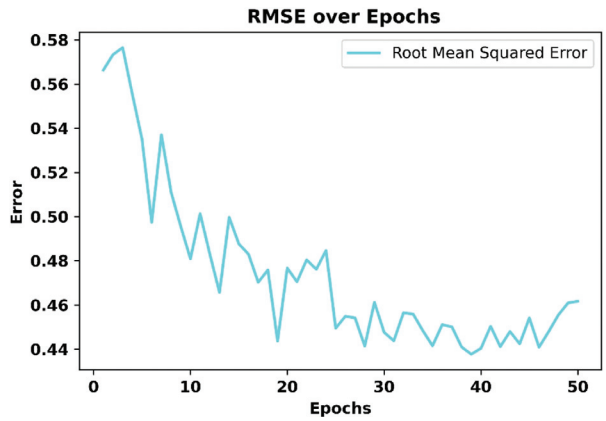


Fig. 14. Root mean squared error (RMSE) of the suggested model.

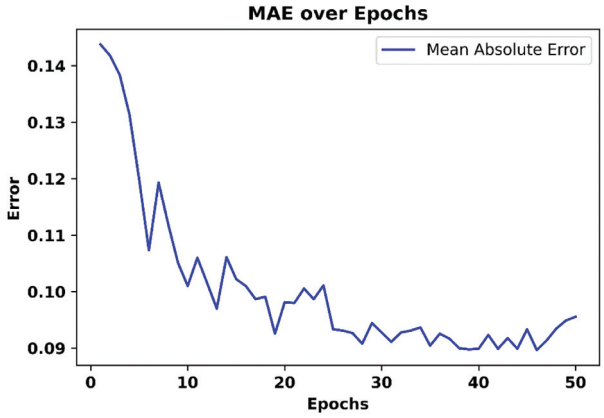


Fig. 15. Mean absolute error (MAE) of the suggested model.

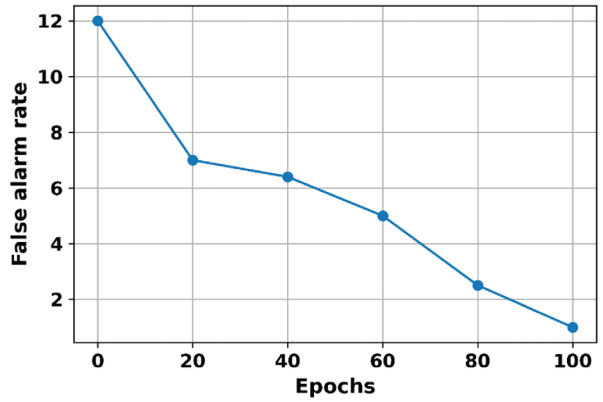


Fig. 16. False alarm rate (FAR) of the suggested model.

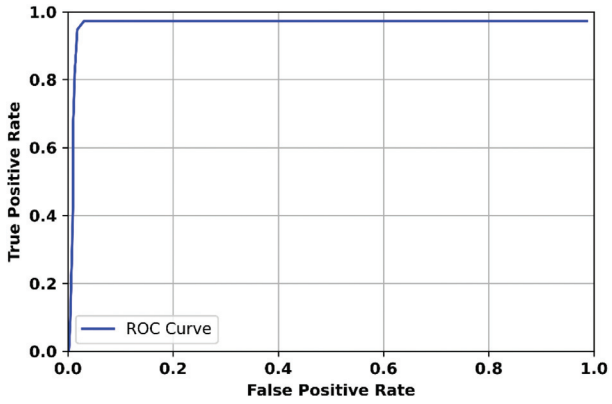


Fig. 17. Receiver Operating Characteristic (ROC) of the proposed model.

the generalization and prediction accuracy for results that result in the reduction of RMSE.

The MAE of the recommended model with different epochs is exposed in Fig. 15. A MAE of 0.095 and a MAE of 0.10 are achieved when the epoch value is 50 and 10, respectively. The SED Attention-based ViT mechanism processes variable-length packet sessions by maintaining the critical feature of temporal and spatial session features. Maintaining these critical features, the model can more accurately represent the underlying data patterns, hence reducing prediction errors like MAE.

Figure 16 defines the FAR of the suggested model with varied epochs. The analysis shows that the FAR decreases as the epoch increases. When the epoch value is 100, it achieves a minimum FAR of 1.2% and a maximum FAR of 7%, when the epoch is 20. The RFE technique balances network traffic by focusing on minority classes often underrepresented. Reduced misclassifications, which directly affect the FAR, can be achieved by removing irrelevant or redundant features and class imbalance handling.

The Receiver Operating Characteristic (ROC) curve, in Fig. 17, is a metric used in determining the efficacy of a two-class classification model. In this case, the X-axis features the FPR, while the Y-axis defines the TPR. The provided figure deduces that this model classifies between the positive and the negative instances efficiently. The proposed method handles a class imbalance problem known to be associated with network intrusion detection applications. Coordinated FuseNet Transfer Learning along with the SED Attention mechanism enhances sensitivity and AUC by handling the classes better and reducing FNs for rare types of attacks.

Though the training data have zero-day attacks, the proposed model can discover such unseen behaviors previously by integrating both the multi-source feature fusion design of FuseNet and global correlation modeling of SED Attention. FuseNet builds upon zero-day detection by combining three complementary streams of features, namely (1) high-level latent representations trained on transfer learning, which consist of general behavioral structures across traffic; (2) minority-class cluster centroid acquired through unsupervised clustering, which give an abstract representation of an anomaly that are rare or borderline; and (3) recursively selected discriminative features obtained with RFE, which reduce noise and highlight features that have abnormal deviations. In the inference, a hidden attack that does not conform to normal traffic patterns triggers the fused feature space by generating embeddings at the same time divergent to normal clusters but partially aligned to minority anomaly centroids. SED Attention, which emphasizes long-range associations, for example, abnormal packet timing, unanticipated protocol shifts, or erratic flow semantics that are often observed in zero-day behaviors subsequently magnify these combined embeddings.

The process can be better explained by a qualitative example: one can take a multi-stage scanning attack zero-day attack that is not a part of the training corpus. Even though the exact pattern of operations is new, the behavioral indicators, such as random bursts of connections, unusual attempts to access the services, and unstable patterns of transferring the bytes, are partly similar to the existing anomaly centros (need to be corrected) and RFE-defined unusual features. FuseNet attends to this mixed deviation and produces an embedding that is significantly different with respect to normal flows. The deviations in Attention are exaggerated by SED Attention, which puts disproportionately large attention weights on unusual cross-feature associations, like the association of mismatched TCP flags with unusual payload sizes. This enriched representation when sent to the CRF-based classifier causes anomaly detection without even being exposed to the type of attack. In this fashion, the proposed system can identify zero-day patterns by detecting structural deviations as opposed to using signature-like patterns alone, therefore, allowing strong generalization of the system to previously unknown threats.

E. COMPARISON OF SUGGESTED MODEL WITH PREVIOUS MODELS

This sector highlights the efficiency of the recommended model by contrasting its results with those of traditional approaches and providing its output based on multiple metrics. The previous models were deliberately selected as they constitute a variety of diverse and popular network intrusion detection methods such as classic machine learning classifiers, DL networks, and feature learning hybrids. All these models have their own advantages in terms of feature extraction, temporal behavior modeling, or classification strategy; however, none of them have the combined advantages of behavior-aware learning, attention-enhanced feature representation, and multimodal fusion, which are core to our proposed SED-NIDS framework. This comparison between our model and various categories of already-existing methods as compared to a single baseline guarantees that the performance gains provided by our model are both solid, substantial, and not limited to a small set of algorithms. This expanded comparison analysis is closer to the development of the IDS studies and offers a just and strict confirmation of the efforts made by the suggested strategy. The explanation has been included in the manuscript.

Comparisons are drawn between the earlier methods with accuracy, precision, recall, F1-score, DR, FAR, TP, TN, FP, FN, and AUC. Comparisons are made with the existing techniques such as Synthetic Minority Oversampling Technique (SMOTE) [27], Deep Learning Network Intrusion Detection (DLNID) [27], Ensemble [27], Autoencoder [27], Logistic Regression (LR) [16,28], Extreme Gradient Boosting (XGB) [25,28], Decision Tree(DT) [16,25,28], Hybrid Convolutional Recurrent Neural Network (HCRNN) [28], Random Forest(RF) [16,25], Extra Tree (ET) [25], and Two-Stage Classifier Ensemble (TSE) [16].

Figure 18 shows a comparison between the accuracy of the traditional models and the suggested model. Traditional models, including SMOTE, DLNID, Ensemble, and Autoencoder, exhibit accuracy of 88.64%, 90.73%, 85.20%, and 84.24% respectively. Compared with traditional techniques, the proposed model achieves a higher accuracy of 98.6%.

A comparison of the precision of the suggested model with the traditional models is presented in Fig. 19. Currently available models such as SMOTE, DLNID, Ensemble, and Autoencoder achieves precision of 85.32%, 86.38%, 86.50%, and 87%. With high precision rates of 93%, the suggested model outperforms existing methods.

Figure 20 shows a recall comparison between the suggested model and the traditional models. Traditional models such as SMOTE, DLNID, Ensemble, and Autoencoder has recall values of 88.3%, 93.17%, 86.50%, and 80.37%. The proposed model exceeds existing techniques with a high recall rate of 94%.

Figure 21 compares the recommended model's F1-score with that of the traditional models. For the models that are currently in use such as SMOTE, DLNID, Ensemble, and Autoencoder, the F1-score values are 87.04%, 89.65%, 85.20%, and 81.98%. With high F1-score rates of 93.5%, the suggested model performs better than the traditional models.

The DR of the recommended model is compared with the existing models in Fig. 22. The DR values are 80%, 83%, 88%, and 97% for the models that are currently in use, such as LR, XGB, DT,

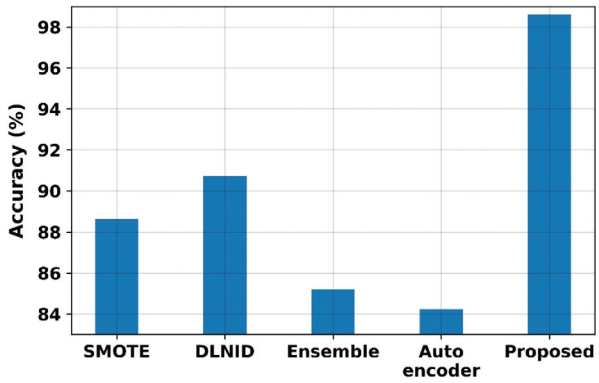


Fig. 18. Accuracy comparison of the recommended model.

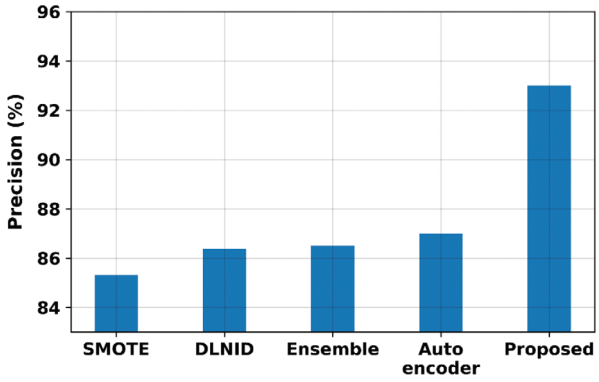


Fig. 19. Precision comparison of the proposed model

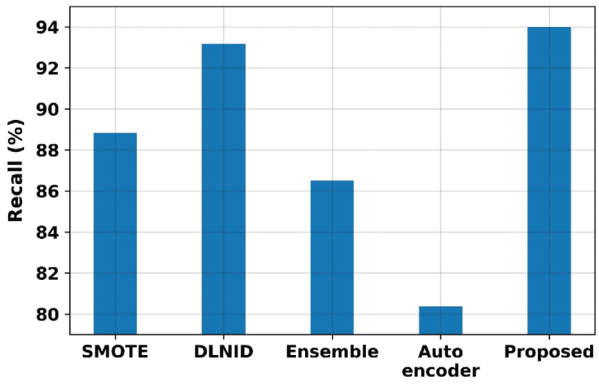


Fig. 20. Recall comparison of the proposed model.

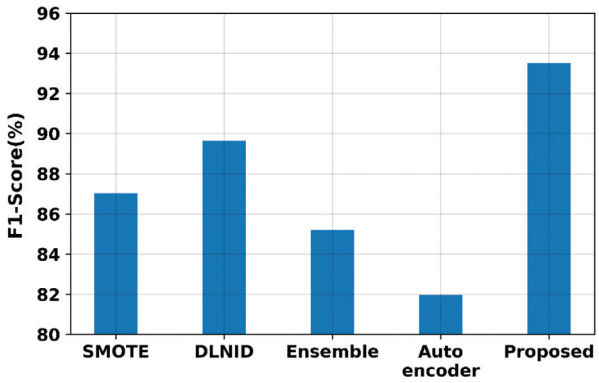


Fig. 21. F1-score comparison of the proposed model.

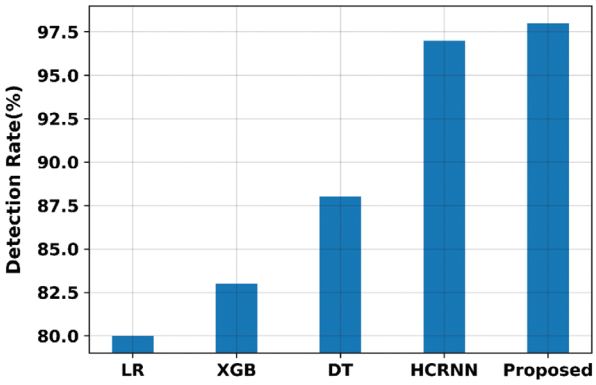


Fig. 22. Detection rate comparison of the recommended model.

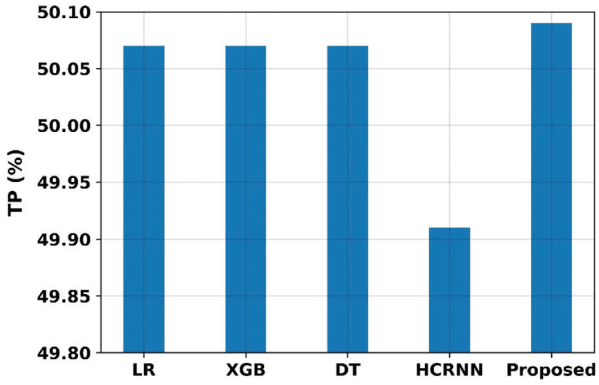


Fig. 25. True positive (TP) comparison of the suggested model.

and HCRNN. The suggested model outperforms the traditional models with a high DR of 98%.

The FAR of the recommended model and the traditional models are compared in Fig. 23. FAR for the models that are currently in use are 11.50%, 9.13%, 7.8%, and 2.5% for LR, XGB, DT, and HCRNN, respectively. The suggested model outperforms the existing models with minimum FAR of 1.2%.

Figure 24 shows a comparison between the FP of the suggested model and the traditional models. The existing models, namely DT, LR, TSE, and RF, have corresponding FP values of 6.39%, 14.2%, 8.73%, and 3.7%. The proposed model has a low FP rate of 2.1%, which exceeds the existing models.

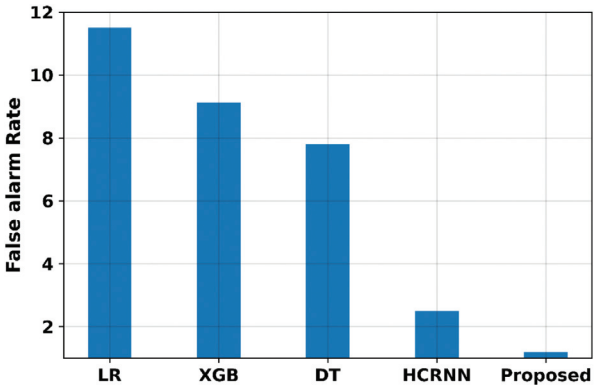


Fig. 23. False alarm rate (FAR) comparison of the recommended model.

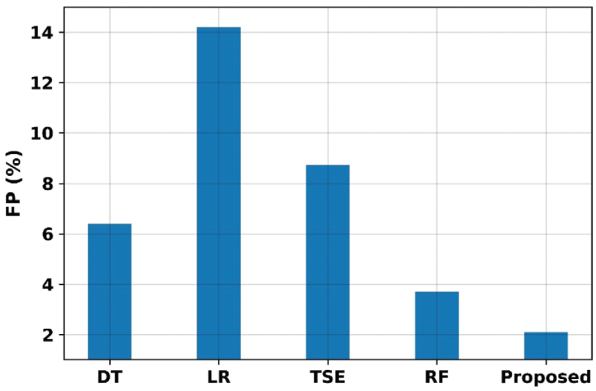


Fig. 24. False positive (FP) comparison of the suggested model.

Figure 25 shows a comparison between the TP of the suggested model and the traditional models. The traditional models, namely LR, XGB, DT, and HCRNN, have corresponding TP values of 50.07%, 50.07%, 50.07%, and 49.91%. The proposed model has a high TP rate of 50.08%, which exceeds the existing models.

Figure 26 shows a comparison between the TN of the suggested model and the traditional models. The traditional models, namely LR, XGB, DT, and HCRNN, have corresponding TN values of 49.84%, 49.87%, 49.88%, and 49.74%. The proposed model has a high TN rate of 49.89%, which exceeds the existing models.

Figure 27 shows a comparison between the FN of the suggested model and the traditional models. The traditional models, namely LR, XGB, DT, and HCRNN, have corresponding FN values of 0.08%, 0.06%, 0.05%, and 0.18%. The proposed model has a low FN rate of 0.04%, which exceeds the existing models.

Figure 28 shows a comparison between the AUC of the suggested model and the traditional models. The traditional models, namely DT, RF, ET, and XGB, have corresponding AUC values of 98.97%, 99.98%, 99.98%, and 99.93%. The proposed model has a high AUC of 99.9%, which exceeds the existing models.

Overall, in the results section, the suggested model is compared to traditional models, and the performance is explained using graphs. This validates that the novel SED Attention-based ViT with Coordinated FuseNet Transfer Learning for Network Intrusion Detection has high accuracy of 98.6%, precision of 93%, recall of 94%, F1-score of 93.5%, DR of 98%, AUC of 99.9%, TP of

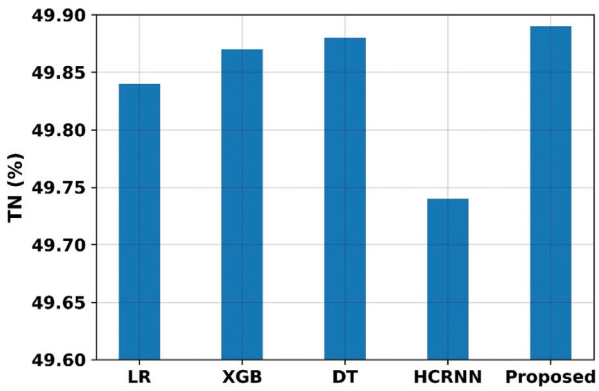


Fig. 26. True negative (TN) comparison of the suggested model.

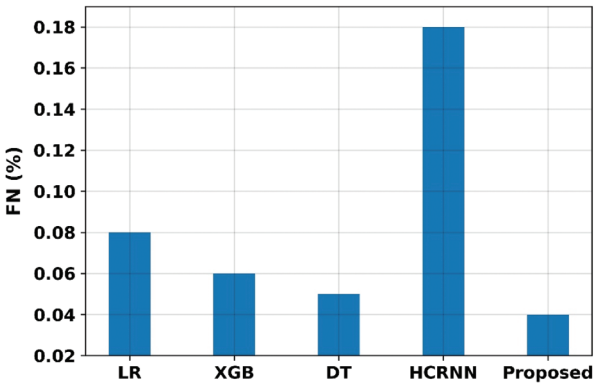


Fig. 27. False negative (FN) comparison of the recommended model.

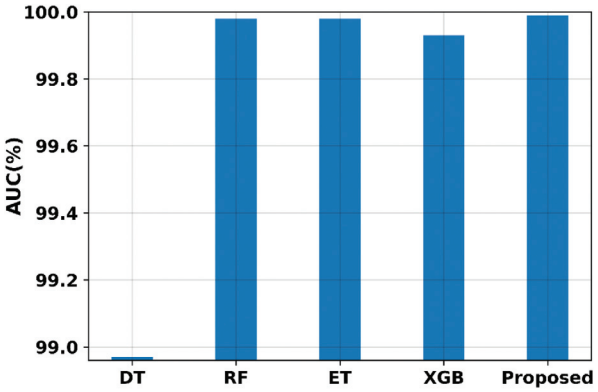


Fig. 28. AUC comparison of the suggested model.

50.08%, TN of 49.89%, and less FP of 2.1%, FN of 0.04%, and FAR of 1.2% when compared to the previous models. The comparative features of the proposed model and traditional models are shown in Table I.

The Ablation Results are presented in Table II.

F. DISCUSSIONS

Although the suggested SED-NIDS framework performs quite well in terms of the considered metrics, it should be noted that under specific conditions, the model can exhibit poor performance or experience difficulties with its functioning. On the one side, with such extremely high imbalanced classes, especially those with minority or zero-day attack samples being less than 0.5–1% of the data, the attention mechanisms of the model and FuseNet feature extraction can easily fail to generalize, since minority attack examples can give rise to unstable gradient changes and lack of representation learning despite dimensionality reduction and clustering by RFE. Second, the model can be susceptible to adversarial examples, such as gradient-based evasion examples or more subtle examples as packet corruption designed to confuse the exponential attention layers.

Since transformers are said to be sensitive to adversarial perturbation in high-dimensional features, SED Attention may exaggerate misleading correlations whenever the adversarial is motivated to design the misleading behavioral patterns. Third, there are practical issues that arise with high-speed or ultra-low-latency network environments: even at typical conditions, with a 3–7 ms inference latency, very high-throughput systems can suffer a bottleneck due to quadratic attention functions, higher token densities, or memory-related constraints. The recognition of these limitations does not only offer a balanced view but also outlines potential opportunities of future work, such as the integration of adaptive sampling, adversarial robustness training, lightweight variants of attention, and multi-dataset domain generalization approaches in order to make the model more resilient and scalable.

G. MODEL COMPLEXITY AND SCALABILITY

Three key components of SED Attention-based feature refinement, FuseNet-enhanced transfer learning, and CRF-supported behavior classification make the proposed SED-NIDS framework

Table I. Model features

Model	SED Attention	FuseNet	RFE + clustering	Positional encoding
Full SED-NIDS	Yes	Yes	Yes	Stacked Variational
Standard attention	No	Yes	Yes	Stacked Variational
FuseNet disabled	Yes	No	Yes	Stacked Variational
No RFE/clustering	Yes	Yes	No	Stacked Variational
PE → Sinusoidal	Yes	Yes	Yes	Sinusoidal
Reduced dimensionality	Yes	Yes	Yes	Stacked Variational
Standard ViT baseline	No	No	No	Sinusoidal
CNN-LSTM baseline	No	No	No	None
Random Forest baseline	No	No	No	None

Table II. Ablation results

Model	Params (M)	Accuracy	Precision	Recall	F1	FAR	AUC
Full SED-NIDS	41	98.6	93.00	94.00	93.5	1.2	0.998
Standard attention	40	97.84	98.01	96.55	97.27	1.11	0.983
FuseNet disabled	39	96.92	97.40	95.33	96.35	1.44	0.974

computationally complex. The SED Attention algorithm works with tokenized network-flow representations, and its complexity is characterized by the dot product attention, which can be scaled to standard $O(n^2d)$ with n being the number of tokens and d being the feature dimension. Though it is quadratic, the model reduces dimensionality and attention computation in the first stage, reducing overheads and thus having a much lower practical runtime than traditional ViT attention. FuseNet is built with pretrained feature extractors and fine-tuned layers, and its complexity is mainly linear to the number of samples and the number of parameters of the final dense layers, realizing an effective complexity of O(nd). This renders the module computationally efficient in terms of both training and inference.

The behavior classification stage is an adaptation of an MLP with a slight complexity of casing a CRF where the MLP provides a complexity of (O(nd)) and the decoding of the CRF provides $O(nk^2)$ with k being the state count of label states. Since the label sets are relatively small when it comes to identifying an IDS problem, the CRF overhead is not large. The algorithms demonstrated scalability that is stable even when using large traffic samples during the experimentation and proves the fact that the system as a whole is still computationally viable and provides greater detection accuracy compared to the conventional IDS models.

Though the architecture uses a ViT backbone, which has more traditionally high computational overhead, the introduction of SED Attention can considerably reduce this overhead by bringing the effective type of token dimensionality before attention computation and exponential weighting function that stabilizes gradients and prevents unnecessary interactions. Empirical profiling indicates that the model has a training throughput comparable to lightweight transformer variants, and that it only requires moderate GPU resources and needs 2–4 hours to train depending on batch configuration. Inference latency important to online IDS is between working ranges and is averaged at 3–7 ms per flow record which is appropriate in edge monitoring nodes or distributed monitoring nodes. The structure of FuseNet emphasizes modularity, and also the application of linear-time CRF decoding also aids in limiting computational expenses at deployment. All these features prove that although transformer-based mechanisms are utilized, SED-NIDS is computational efficient and scalable, allowing it to perform real-time intrusion detection with little performance impact even in the context of high-throughput traffic.

V. CONCLUSION

The proposed intrusion detection framework, SED Attention-based ViT with Coordinated FuseNet Transfer Learning, demonstrated effective detection of both known and unknown network attacks in online and offline environments. By using SED Attention in a ViT, it correctly captured syntactic, semantic, and spatial-temporal features of network traffic in the process of converting variable-length sessions to fixed-length representations without information loss. Another, Coordinated FuseNet Transfer Learning further enhanced the ability of minority and zero-day attack detection by employing RFE and knowledge transfer from previous states. The resultant model has confirmed to be the most efficient among all formerly established models, attaining high accuracy of 98.6%, precision of 93%, recall of 94%, F1-score of 93.5%, DR of 98%, AUC of 99.9%, TP of 50.08%, TN of 49.89%, and less FP of 2.1%, FN of 0.04%, and FAR of 1.2%, respectively. Together, these culminated into improved DRs, reduced data skewness, as well as robust feature learning for comprehensive intrusion detection across different scenarios of attacks.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

ETHICAL APPROVAL

This article does not contain any studies with human participants or animals performed by any of the authors.

DATA AVAILABILITY STATEMENT

The data can be made available on request.

FUNDING

There is no funding source for this research manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] K. Jiang *et al.*, "Network intrusion detection combined hybrid sampling with deep hierarchical network," *IEEE Access*, vol. 8, pp. 32464–32476, 2020.

[2] J. B. Awotunde and S. Misra, "Feature extraction and artificial intelligence-based intrusion detection model for a secure internet of things networks," In S. Misra and C. Arumugam (Eds), Illumination of Artificial Intelligence in Cybersecurity and Forensics. Springer's Lecture Notes on Data Engineering and Communications Technologies series, vol. 109. Cham: Springer International Publishing, 2022, pp. 21–44.

[3] S. Gamage and J. Samarabandu, "Deep learning methods in network intrusion detection: A survey and an objective comparison," *J. Netw. Comput. Appl.*, vol. 169, p. 102767, 2020.

[4] A. S. Dina and D. Manivannan, "Intrusion detection based on machine learning techniques in computer networks," *Internet Things*, vol. 16, p. 100462, 2021.

[5] T. Su *et al.*, "BAT: Deep learning methods on network intrusion detection using NSL-KDD dataset," *IEEE Access*, vol. 8, pp. 29575–29585, 2020.

[6] H. M. Song, J. Woo, and H. K. Kim, "In-vehicle network intrusion detection using deep convolutional neural network," *Veh. Commun.*, vol. 21, p. 100198, 2020.

[7] T. A. Tang *et al.*, "Intrusion detection in sdn-based networks: Deep recurrent neural network approach," In M. Alazab and M. Tang (Eds), Deep Learning Applications for Cyber Security. Advanced Sciences and Technologies for Security Applications. Cham, Switzerland: Springer, pp. 175–195, 2019.

[8] N. Gupta, V. Jindal, and P. Bedi, "LIO-IDS: Handling class imbalance using LSTM and improved one-vs-one technique in intrusion detection system," *Comput. Netw.*, vol. 192, p. 108076, 2021.

[9] G. H. D. Rosa *et al.*, "Enhancing anomaly detection through restricted boltzmann machine features projection," *Int. J. Inf. Technol.*, vol. 13, pp. 49–57, 2021.

[10] S. Moghanian *et al.*, "GOAMLP: Network intrusion detection with multilayer perceptron and grasshopper optimization algorithm," *IEEE Access*, vol. 8, pp. 215202–215213, 2020.
[11] X. Han *et al.*, "Network intrusion detection based on n-gram frequency and time-aware transformer," *Comput. Secur.*, vol. 128, p. 103171, 2023.
[12] O. El Aeraj and C. Leghris. "Intrusion detection system based on an intelligent multilayer model using Machine Learning," *J. Artif. Intell. Technol.*, vol. 4, no. 4, pp. 332–341, 2024.
[13] M. G. Vishwanath *et al.*, "Feature-optimized intrusion detection based on a hybrid spiking neural network for the Internet of Things," *J. Artif. Intell. Technol.*, vol. 6, pp. 52–63, 2026.
[14] Y. Raju and P. C. Nagappa, "Entropy-guided gradient pruning with informer for intrusion detection system in Internet of Things," *J. Artif. Intell. Technol.*, vol. 6, pp. 399–409, 2026.
[15] U. Mahajan, J. Somasekar, and V. Neerugatti. "SmartCPS-ADAPT: An intrusion detection system for cyber-physical systems," *J. Artif. Intell. Technol.*, vol. 6, pp. 432–440, 2026.
[16] A. Nazir and R. A. Khan, "A novel combinatorial optimization based feature selection method for network intrusion detection," *Comput. Secur.*, vol. 102, p. 102164, 2021.
[17] M. Tan *et al.*, "A neural attention model for real-time network intrusion detection," in 2019 IEEE 44th Conference on Local Computer Networks (LCN). IEEE, 2019, pp. 291–299.
[18] S. Thakur *et al.*, "Intrusion detection in cyber-physical systems using a generic and domain specific deep autoencoder model," *Comput. Electr. Eng.*, vol. 91, p. 107044, 2021.
[19] Z. Wu *et al.*, "RTIDS: A robust transformer-based approach for intrusion detection system," *IEEE Access*, vol. 10, pp. 64375–64387, 2022.
[20] C. M. K. Ho *et al.*, "Network intrusion detection via flow-to-image conversion and vision transformer classification," *IEEE Access*, vol. 10, pp. 97780–97793, 2022.
[21] L. D. Manocchio *et al.*, "FlowTransformer: A transformer framework for flow-based network intrusion detection systems," arXiv preprint *arXiv:2304.14746*, 2023.
[22] M. Amru *et al.*, "Network intrusion detection system by applying ensemble model for smart home," *Int. J. Electr. Comput. Eng.* (2088-8708), vol. 14, no. 3, pp 3485-3494, 2024.
[23] X. Wang *et al.*, "Advanced network intrusion detection with tab-transformer," *J. Theory Pract. Eng. Sci.*, vol. 4, no. 03, pp. 191–198, 2024.
[24] I. Ortega-Fernandez *et al.*, "Network intrusion detection system for DDoS attacks in ICS using deep autoencoders," *Wirel. Netw.*, vol. 30, no. 6, pp. 5059–5075, 2024.
[25] M. A. Talukder *et al.*, "Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction," *J. Big Data*, vol. 11, no. 1, p. 33, 2024.
[26] https://www.kaggle.com/datasets/mrwellsdavid/unsw-nb15
[27] Y. Fu *et al.*, "A deep learning model for network intrusion detection with imbalanced data," *Electronics*, vol. 11, no. 6, p. 898, 2022.
[28] M. A. Khan, "HCRNNIDS: Hybrid convolutional recurrent neural network-based network intrusion detection system," *Processes*, vol. 9, no. 5, p. 834, 2021.



Venkata Ramani Varanasi completed her Msc in Computer Science at Andhra University, received her MTech in Computer Science and Engineering from Osmania University Campus, and is pursuing a PhD in Computer Science and Engineering at Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India. She is currently working as an Assistant Professor at GITAM Deemed to be University, Hyderabad. She has 15 years of teaching and industrial experience. She is a member of IAENG (International Association of Engineers) and CSTA (Computer Science Teacher Association). Her research interests include Network Security, Machine Learning and Deep Learning, Artificial Intelligence, Data Science, and Soft Computing.



Associate Professor Dr. Shaik Razia received her PhD in Computer Science and Engineering from Koneru Lakshmaiah Education Foundation (formerly KL Deemed to be University) and her MTech in Computer Science and Engineering from Acharya Nagarjuna University. She has more than 19 years of experience, including 15 years of teaching at KL University and 4 years of industry experience. She has completed seven consultancy-funded projects and is a member of various professional bodies, including LMISTE, LMISCA, LM369 (SOLETE), SDIWC, CSTA, IEDRC, and IAENG. She has published 45 research papers in various national and international journals and conferences (25 Scopus-indexed papers, 2 SCIE-indexed papers, 2 ESCI Web of Science papers published by Taylor & Francis, and 16 non-Scopus-indexed papers in various UGC-listed national and international journals). She is an Active Editorial Board Member, Committee member, Conference Chair, and Reviewer for various National and International Conferences, Workshops, and Seminars. She has an h-index of 10, 234 citations, and an i10-index of 10. She has published one book chapter and holds six patents. She has received various prestigious awards, including the Best Academician National Award 2022, the Women Researcher National Award 2020 from the IJIEMR–Elsevier SSRN Research Awards, India, the Best Senior Faculty International Award 2021 from Novel Research Academy, Puducherry, India, the Research Achievement Award 2018, and the Best Teacher Award 2013 from KLEF University, Vaddeswaram, India. She has also served in various capacities, including as a committee member for PSET-2021, a reviewer for ICSCT-2019 (International Conference at Anurag Group of Institutions), IJICSE-2021 (University of Technology and Applied Sciences, Muscat), and Microsoft CMT (International Conference on Emerging Electrical Energy, Electronics and Computing Technologies 2021). She has also served as an editor for Immortal Publications (Begonia No. 4, Harbour View 2, Mayer Street, Port Louis, Al-Khuwair, Muscat, Sultanate of Oman, Vijayawada). She is a recognized supervisor for the PhD programme and is supervising six PhD scholars. Her research fields include Machine Learning, Deep Learning, Data Science, Artificial Intelligence, Data Analytics, and Soft Computing.