IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 January 2024, accepted 23 January 2024, date of publication 29 January 2024, date of current version 7 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3359595

RESEARCH ARTICLE

Intrusion Detection With Deep Learning Classifiers: A Synergistic Approach of Probabilistic Clustering and Human Expertise to Reduce False Alarms

ABDOUL-AZIZ MAIGA[1], (Graduate Student Member, IEEE), EDWIN ATARO[2], AND STANLEY GITHINJI[3]

[1]Pan African University, Institute for Basic Sciences Technology and Innovation (PAUSTI), Nairobi, Kenya
[2]Department of Electrical and Electronic Engineering, Technical University of Kenya, Nairobi, Kenya
[3]Department of Computing, United States International University-Africa (USIU-A), Nairobi, Kenya

Corresponding author: Abdoul-Aziz Maiga (abdoul.maiga@students.jkuat.ac.ke)

This work was supported by the African Union (AU).

ABSTRACT Intrusion detection systems (**IDS**) have seen an increasing number of proposals by researchers utilizing deep learning (DL) to safeguard critical networks. However, they often suffer from high false alarm rates, posing a significant challenge to their deployment in critical networks. This paper presents a comprehensive human-machine framework for mitigating false alarms in DL-based intrusion detection systems. The proposed approach uses probabilistic clustering to enable human-machine collaboration in a synergistic manner. Probabilistic clustering involves regrouping network traffic into clusters based on their probabilities (computed using the DL model). Clusters with high false alarms (H-FAR) are detected, and all traffic falling within them is considered uncertain for efficient classification by the DL model as malicious or benign. They are redirected to human experts to analyze and make a final decision. The proposed framework incorporates a next-generation firewall (NGFW) to help human experts handle the processed traffic efficiently. The proposed framework enhances the performance of DL-based intrusion detection classifiers by reducing false alarms. To validate the proposed concept, assessments were conducted using a customized high-performance convolutional neural network (CNN) and a hybrid recurrent neural network (RNN) model with three open-access benchmark datasets (CICDDoS2019, UNSW-NB15, and CICIDS2017). The evaluation through simulation demonstrated that combining human expertise with deep learning technology can significantly reduce the number of false positives (FPs) and false negatives (FNs) by up to 79.61% and 86.99%, respectively.

INDEX TERMS Deep learning, IDS, false alarms mitigation, human-in-the-loop expertise, human-machine.

I. INTRODUCTION

Intrusion Detection Systems (IDS) continuously evolve within the realm of network security to safeguard critical information assets owing to the increasing sophistication of cyber threats [1], [2], such as advanced persistent threats (APTs) [3], zero-day exploits, large-scale DDoS attacks, slow-rate DDoS attacks, malicious botnets, and many others [4], [5].

Over the years, deep learning (DL) classifiers have emerged as formidable tools to bolster the efficacy of intrusion detection. The DL-based IDS classifier utilizes deep learning to analyze network traffic for intrusion detection by classifying incoming traffic as malicious or benign based on a threshold [6]. When the DL classifier receives a packet as its input, it computes the probability that the packet is malicious or benign. It is then compared with the threshold to make the

The associate editor coordinating the review of this manuscript and approving it for publication was Vicente Alarcon-Aquino.

final decision. For example, when the output is greater than the threshold, the packet is classified as malicious. Otherwise, it is classified as benign. Their ability to discern complex patterns in network traffic data has revolutionized the detection of both known and novel threats [7]. However, like any technological advancement, deep-learning-based IDS classifiers are not devoid of challenges, most notably, the issue of false alarms [8], [9]. A false alarm occurs when the DL model classifies a benign packet as malicious, or vice versa. Although, in some cases, the model can identify previously unknown malicious traffic, it is susceptible to misclassifying well-known malicious traffic as benign, often in conjunction with some normal traffic, which is also misclassified. The persistence of false alarms within an IDS (referring to a situation in which an IDS, despite its efforts to identify threats, repeatedly misclassifies regular traffic as malicious or vice versa) presents a significant challenge [10]. These erroneous alerts not only deplete valuable resources but also impose an increased burden on security personnel, frequently leading to ''false alarm fatigue.'' Addressing these inherent weaknesses of the DL-based classifiers is crucial. The research questions were as follows:

- How can we detect the weaknesses of the DL-based IDS classifiers in their early stages and proactively mitigate false alarms?
- How effectively can the proposed system mitigate false alarms?
- How can the proposed human-machine system be deployed effectively in a live network environment?

The approach outlined in this study aims to address these questions and enhance the performance of deep learning (DL) IDS classifiers. By leveraging the capabilities of deep-learning classifiers, an innovative synergistic model that amalgamates probabilistic clustering, human expertise, and next-generation firewalls (NGFW) is introduced. The study is firmly grounded in the belief that while deep learning algorithms excel at recognizing intricate patterns, human insight and contextual awareness remain of added value in the battle against cyber intrusions.

The core of the proposed approach is the utilization of probabilistic clustering techniques to identify the critical prediction capability clusters in the DL model during the testing process. Grouping the model's probabilistic predictions into clusters allows for the processing and determination of the False Alarm Rate (FAR) of each cluster based on the model's best threshold for classifying traffic as normal or malicious. Clusters exhibiting a high false-alarm rate (H-FAR) are identified as critical, indicating a high probability of the DL model misclassifying traffic within these ranges. Consequently, the integration of human expertise within a DL-based IDS was proposed to enhance intrusion detection capabilities and reduce false alarms. This approach suggests that once the critical H-FAR clusters of the DL model are identified, all traffic within these clusters should be redirected to the Security Operations Center (SOC) for human experts' intervention for the final classification. The SOC is a centralized unit within an organization responsible for monitoring and managing security-related issues on an ongoing basis. Human experts have long played a central role in cybersecurity, and despite advancements in Artificial Intelligence (AI), they can continue to offer more efficiency, where AI models may fall short [11]. A pertinent question arises: How can traffic correctly classified by a human expert be managed effectively to avoid redundant tasks? Here, the use of an existing solution, NGFW [12], is advocated. An NGFW is a unified security system that enables advanced security features such as deep packet inspection (DPI), intrusion prevention system (IPS), threat intelligence, application control, and inherent security features from traditional firewalls, such as access control, security zone boundaries, and network traffic control through security policies. This allows the administrator to create their own exception security rules. A security rule is an instruction that tells the firewall to block or allow traffic based on its characteristics. A predefined rule comes with the firewall, whereas an exception rule is created by the administrator. Exception rules can be implemented to block newly discovered threats that do not have updated patches. The role of the NGFW in this study was to assist human experts in defining exception rules to accurately classify new traffic. Subsequently, similar traffic that has undergone human expert classification is automatically managed by the NGFW, continually reducing the workload on human experts by eliminating the need to reprocess the previously categorized traffic.

The proposed framework is highly flexible for enabling human-machine-based network traffic classification. Even very well-performing DL models that still generate small false alarms can be used to redirect uncertain traffic for human-expert analysis. To validate the approach, a high-performance Convolutional Neural Network (CNN) and a Hybrid Recurrent Neural Network (H-RNN) model were designed and used with three benchmark datasets (CICDDoS2019, UNSW-NB15, and CICIDS2017) in a simulation environment, which yielded promising results.

The contributions of this study to the field are summarized as follows:

1. A probabilistic clustering technique is proposed to identify the probability ranges in which DL models perform poorly with H-FAR in their classification.
2. Human-in-the-loop expertise is suggested for H-FAR reduction with the help of NGFW for enhanced management of previously categorized traffic.
3. We designed two high-performance DL models (CNN and H-RNN) that can effectively classify malicious traffic from normal traffic, but can still be improved by human experts in the loop using the proposed probabilistic clustering technique.

This section introduces the background of the study and its contributions to the field. Section II discusses previous studies that used deep learning for intrusion detection. Section III describes the proposed methodology in detail. In Section IV, we describe the datasets used and the simulation process.

Section V presents and discusses the results of this study. Finally, Section VI concludes the paper.

II. PREVIOUS WORKS

The use of human-in-the-loop in cyber security is not a new concept. Previously, Santhanam et al. [13] used a human-in-the-loop technique to monitor and detect software side-channel vulnerabilities. For deep learning-based network intrusion detection, it is very rare to obtain a DL model that achieves 100% efficiency for all data in network traffic classification. A combination of human expertise and DL can result in greater efficiency. In this section, we discuss the performance of existing deep-learning-based IDS in various areas of application.

Awajan [14] proposed a four-layer deep fully connected network architecture that can detect malicious traffic and attacks on connected internet of things (IoT) devices without requiring system attributes, communication protocol adjustment, or network structure virtualization. The study evaluated the performance of the proposed system on five common attacks: blackhole, distributed denial of service, opportunistic service, sinkhole, and wormhole. The model displayed an acceptable accuracy of 93.74%, with false alarms ranging from 3.5% to 11.6% based on the type of attack.

Sai Chaitanya Kumar et al. [15] proposed in their study a deep residual convolutional neural network (DRCNN) for network intrusion detection. They used an Improved Gazelle Optimization Algorithm (IGOA) for model optimization and a Novel Binary Grasshopper Optimization Algorithm (NBGOA) for feature selection. They evaluated the model using the UNSW-NB-15, CICDDoS2019, and CIC-IDS2017 datasets. The authors claimed that their model outperformed the previous methods in terms of accuracy and processing time.

Shah et al. [16] developed deep learning models combined with a blockchain for intrusion detection in the IoT. The proposed deep neural network (DNN) models are used to classify IoT smart contracts as malicious or benign. The best method achieved an accuracy of 99.27%. However, some false alarms reaching 12% were observed (based on the confusion matrixes presented).

Praveen et al. [17] combined a CNN with a Bidirectional Long Short-Term Memory (BiLSTM) for intrusion detection. The authors proposed a high-performance model to address the increasing number of cyber-attacks in a growing interconnected world. The hybrid model achieved a good accuracy of 99.308% but still had an FPR rate of 0.20%.

Li et al. [18] proposed a multi-CNN fusion model for intrusion detection to address the limitations of traditional machine learning models in meeting the current network environment. The model exhibited an accuracy of 76.67% for binary classification, indicating a high misclassification rate of the proposed model.

For botnet detection and mitigation, Filho et al. [19] proposed a federated learning-based CNN. The idea of the authors is to allow each device in the network to participate in malicious traffic detection. The model exhibited an accuracy of 89.753%. The achieved accuracy allowed us to conclude that it exposes a high misclassification rate compared to other studies.

Neto et al. [20] proposed a feedforward neural network model to detect DDoS attacks in a multi-tenant IoT network. The authors used federated learning to maintain the privacy of the tenants' device data while training a deep learning model. By adopting a simulation method using the CICDDoS2019 dataset, the model achieved an accuracy of 84.2% in DDOS attack detection, justifying its high false alarms generation.

The detection and mitigation of low-rate DDoS attacks have been a research focus of Ali et al. [21]. A low-rate DDoS can be very difficult to detect because it behaves like normal traffic. The authors developed an ANN-based weighted federated learning (WFL) method for detection and mitigation. The model exhibited an FPR value of 2.215% at the top of an acceptable accuracy.

Bousalem et al. [22] proposed a deep learning model capable of detecting DDOS attacks in 5G and future networks. The authors created a demo 5G core network to simulate a real-world scenario and tested their system, which achieved an accuracy of 97% with a false-positive rate of less than 4%.

Wu and Guo [23] developed a CNN +RNN model for intrusion detection. The model was designed to capture both spatial and temporal dependencies in network traffic for enhanced detection accuracy. The model tested through simulation exhibited an FPR of 3.96% using the UNSW-NB15 dataset.

Al-Haija and Zein-Sabatto [24] designed a CNN model to detect and classify cyber-attacks in IoT communication networks. The authors used feature engineering (data normalization and encoding) to enhance model performance. The simulation method showed a false alarm rate (FAR) of 1.28%.

Xu et al. [25] developed in their study a CNN-BiLSTM-Attention classifier for intrusion detection. They chose deep learning because of its advantages in processing large-scale and complex data. The hybrid designed model achieved a classification accuracy of 93.26% and FPR of 7.53% in a simulated environment. The performance of the model is quite good, but the false alarms generated are also not negligible.

Ravi et al. [26] suggested in their study the use of deep RNN model for network attack classification. The proposed model identifies relevant features from concealed layers in recurrent models using a kernel-based principal component analysis (KPCA) method, which enables the selection of the most impactful features. These optimal features from the RNN layers were combined, and a meta-classifier ensemble was used for classification. The classification accuracy achieved by the model using the SDN-IoT dataset was 97%.

Gurung et al. [27] proposed in their article a deep learning-based IDS using the NSL-KDD Dataset for evaluation. The model achieved an accuracy of 87.2%, which was not competitive with other studies. Consequently, it can be deduced that the model misclassifies a large amount of traffic, thereby generating false alarms.

Some researchers have achieved very good intrusion detection performance using deep-learning-based IDS. This is the case for Hnamte and Hussain [28], who developed a deep convolutional neural network (DCNN) model that showed a range of accuracies between 99.79% and 100% using four benchmark datasets (SCX-IDS 2012, DDoS dataset from Kaggle, CICIDS2017, and CICIDS2018). The authors used a GPU to boost the performance of their model during the training step. The same authors (Hnamte and Hussain) proposed a high-performance hybrid deep learning model by combining CNN and BiLSTM [29]. The authors claimed that they achieved accuracies of 100% and 99.64% using the CICIDS2018 and Edge_IIoT datasets, respectively.

Hnamte et al. [30] designed a two-stage deep learning model using LSTM and Auto-Encoders (AE) for network intrusion detection. The hybrid model showed promising results using open-access datasets. It exhibited accuracies of 99.99% and 99.10% for the CICIDS2017 and CSE-CICIDS2018 datasets, respectively.

The common challenge they encounter is the generation of false alarms in various application areas of DL-based IDSs. Although some models achieved 100% accuracy in some datasets, some false alarms were observed in others datasets, which can still be mitigated by the human-in-the-loop technique proposed in this study. A lower FAR value is not always optimal. When the dataset is extremely large, a low FAR can result in thousands of traffic events that are likely to be misclassified. The solution proposed in this study will help to effectively detect the type of traffic for which the DL model is inefficient in classifying and redirecting traffic for cyber security expert analysis for false alarms mitigation.

III. PROPOSED METHODOLOGY

In this methodology, probabilistic clustering is proposed for detecting network traffic with a high probability of misclassification using a deep learning model. This is proposed for domain experts to further analyze uncertain traffic, and the framework is completed with a next-generation firewall (NGFW) to assist human experts in handling the analyzed traffic. Two customized deep learning models (CNN and BiLSTM+LSTM) were designed to assess the proposed framework. These techniques are described in the following subsections. This study employs two prevalent terms that require thorough comprehension: **false alarms** within deep learning represent the misclassification of typical instances as malicious or benign, while the **False Alarm Rate** (FAR) measures the frequency of false alarms relative to the total anticipated detections.

A. PROBABILISTIC CLUSTERING: A NEW APPROACH

DL-based intrusion detection classifier methods often rely on deterministic classification techniques. However, these conventional approaches struggle to accommodate the innate probabilistic essence of the network data. The consequence of this deterministic nature is evident in the form of elevated false-alarm rates. This is because these traditional methods make binary decisions and classify network traffic as malicious or benign based on the prediction value of the model and a fixed threshold, thereby restricting their practical applicability in certain cases. The proposed probabilistic clustering method enables the DL model to determine the optimal conditions for accurately classifying network traffic, discerning instances where it performs reliably, and identifying scenarios that may be prone to misclassification.

This section outlines the theoretical foundations and mathematical underpinnings of the proposed Probabilistic Clustering method for the DL-based intrusion detection. We demonstrate how this approach effectively identifies clusters associated with high false-alarm rates, thereby providing a crucial link to expert analysis.

To address the challenge of high false alarm rates, an innovative approach known as Probabilistic Clustering, tailored for DL-based intrusion detection classification models, was suggested. The proposed methodology utilizes probabilistic models to partition network traffic into clusters. Each cluster was associated with a probability interval. For traffic to be part of a given cluster, it must have a probability value (the value predicted by the DL model) within that interval.

Probabilistic computing was performed to determine the FAR of each cluster during DL model testing after training. When the model is deployed, it can state which traffic falls in a cluster with H-FAR and redirect it for huma-in-the-loop expertise. By embracing uncertainty, this approach offers a robust framework for human-machine collaboration, enhancing intrusion detection performance. To understand this concept mathematically, we describe its operation.

In the Probabilistic Clustering approach, we seek to identify clusters with high false alarms from the DL model by using the testing dataset $\mathbf{D}=\mathbf{X_0},\mathbf{X_1},\ldots,\mathbf{X_{n-1}}$ with n data points. The model processes each data point X_i, and the output is a probability value $p_i^{C_k} \epsilon [a_k, b_k]$ associated with a cluster C_k, where $k = 0, 1, 2, \ldots, m-1$ represents the cluster number, and m is the maximum number of clusters. Assuming that the predicted model probability values range from zero to one, $[a_k, b_k]$ is such that $0 \leq a_k < b_k \leq 1$. $p_i^{C_k}$ can be defined by:

$$p_i^{C_k} = p(X_i \epsilon C_k) \tag{1}$$

It is important to note that the probability interval of a cluster should be independent and should not overlap with other cluster intervals. This means that data point X_i belongs to only one cluster. The probability intervals for k clusters can be represented by $C = [[a_0, a_1], [a_1, a_2], \ldots, [a_{k-2}, a_{k-1}]] \epsilon [0, 1]$.

To identify clusters with high false alarms, we evaluated the FPR and FNR for each cluster C_k. The FPR for cluster C_k is defined as

$$\mathrm{FPR}(C_k) = \frac{\mathrm{FP}\,(C_k)}{\text{Total data points in } C_k} \tag{2}$$

Similarly, the FNR for cluster C_k is calculated as

$$\mathrm{FNR}(C_k) = \frac{\mathrm{FN}\,(C_k)}{\text{Total data points in } C_k} \tag{3}$$

We set a predefined H-FAR threshold θ and focused our attention on the clusters in which FPR(Ck) $\geq \theta$ and/or FNR(C_k) $\geq \theta$. These clusters were identified as high false-alarm clusters (H-FAC), and further investigation is required. Network traffic within H-FAC is not directly classified by the DL model, but is instead redirected for expert analysis. This approach can help to mitigate the generation of false alarms. Conversely, clusters that do not meet the criteria, where FPR(C_k) $< \theta$ and/or FNR(C_k) $< \theta$, are considered Low False Alarm Clusters and network traffic within is directly classified by the DL model. In summary, ''an intelligent deep learning model'' is obtained that can dissociate network traffic, which can effectively be classified by the DL model, from traffic that is likely to be misclassified and redirect it for human expert analysis.

B. HUMAN-IN-THE-LOOP TECHNIQUE

As described in the previous section, the Probabilistic Clustering approach enables the identification of highly false-alarm clusters, which are indicative of the inherent uncertainty in the classification process. To further enhance the quality of intrusion detection and ensure false alarm mitigation, the ''Human-in-the-Loop'' mechanism is introduced. This subsection describes the human-in-the-loop approach and its significance for real-world applications.

A Human-in-the-loop is associated with human experts who possess domain-specific knowledge and expertise. In the context of this study, they are identified as ''cyber security experts.'' When the deep learning model detects uncertain traffic based on its cluster false-alarm rate, it redirects it to a cyber security specialist portal for analysis and the final decision regarding whether the traffic is malicious or normal. The remaining traffic belonging to clusters with low or null false-alarm rates are directly handled by the DL model. This human-machine hybrid technique can help in the early deployment of DL-based intrusion-detection models in critical networks. However, there is one more step in completing the framework to make it more efficient for real-world applications. The next subsection introduces a next-generation firewall (NGFW) that can help domain experts handle traffic categorization and similar traffic in the future.

C. NEXT GENERATION FIREWALL (NGFW)

There is a limitation in the use of human experts in this loop. After categorizing the traffic as malicious or normal, human experts will still be alerted to the same traffic in the future if it is handled incorrectly. There is a need for a solution to avoid redundancy in the alarms and to reduce the expert's work. To address this challenge, next-generation firewalls (NGFW) are a low-cost option. Low cost because it is an existing solution and there is no need to reinvent a new one.

NGFWs are advanced security devices that combine traditional firewall capabilities with intrusion detection and prevention systems (IDPS) and other advanced security features [31]. NGFW are designed to protect networks from a wide range of threats, including denial-of-service attacks, malware attacks, and unauthorized access attempts. It is currently used by a wide range of institutions, including large companies and governments [32], [33] for enhanced secured networks. By default, NGFW integrates a threat signature database for known attack detection. However, administrators can create their own signatures or exception rules to handle N-day threats (new threats discovered). This functionality is required by a human expert in the proposed method to complete the categorization of network traffic redirected by the DL model. Let us describe how it works in a step-by-step manner.

- **Step 1:** The deep learning model processes an incoming network packet to determine whether it is malicious. The output is a probability value assigned to a cluster (clusters have already been defined during model training and testing, and those with H-FAR are known). If a packet belongs to a low-FAR cluster, DL directly classifies it as malicious or normal based on its best threshold (also defined during model training and testing). If a packet belongs to a cluster with H-FAR, it is redirected to an NGFW for security filtering and rule matching, as defined by a human expert.
- **Step 2:** The NGFW matches the packet to previous exception rules, actions, and security policies configured by experts (or administrators). If a match is found, the corresponding action is executed. The NGFW can implement several actions to manage traffic effectively. 'Blocking' prevents unauthorized or potentially harmful traffic from reaching its destination. Conversely, 'Allowing' permits traffic to adhere to established rules, enabling its passage. 'Blacklisting' specifically denies traffic from known threats or suspicious sources based on identified signatures or addresses. These actions are implemented through predefined rules and policies established by the security administrators. These rules dictate which types of traffic are permitted, blocked, or blacklisted, considering factors such as IP addresses, protocols, and application signatures. If no match is found, the packet is directed to a human expert.
- **Step 3:** Human experts receive uncertain packet alarms and must use their expertise to determine whether a packet is normal or malicious. After categorizing the packet, the expert creates an exception signature or exception security policy rule with the corresponding action to be taken by the NGFW firewall when it receives the next same traffic type.

By setting up this technique, a strong human-machine security monitoring framework is obtained. The overall architecture of the framework is illustrated in Figure 1.

D. DESIGNED MODELS: CNN AND BiLSTM+LSTM

This section introduces the two deep learning models designed and employed in this study, namely Convolutional Neural Networks (CNN) and the combination of Bidirectional Long Short-Term Memory (BiLSTM) and Long

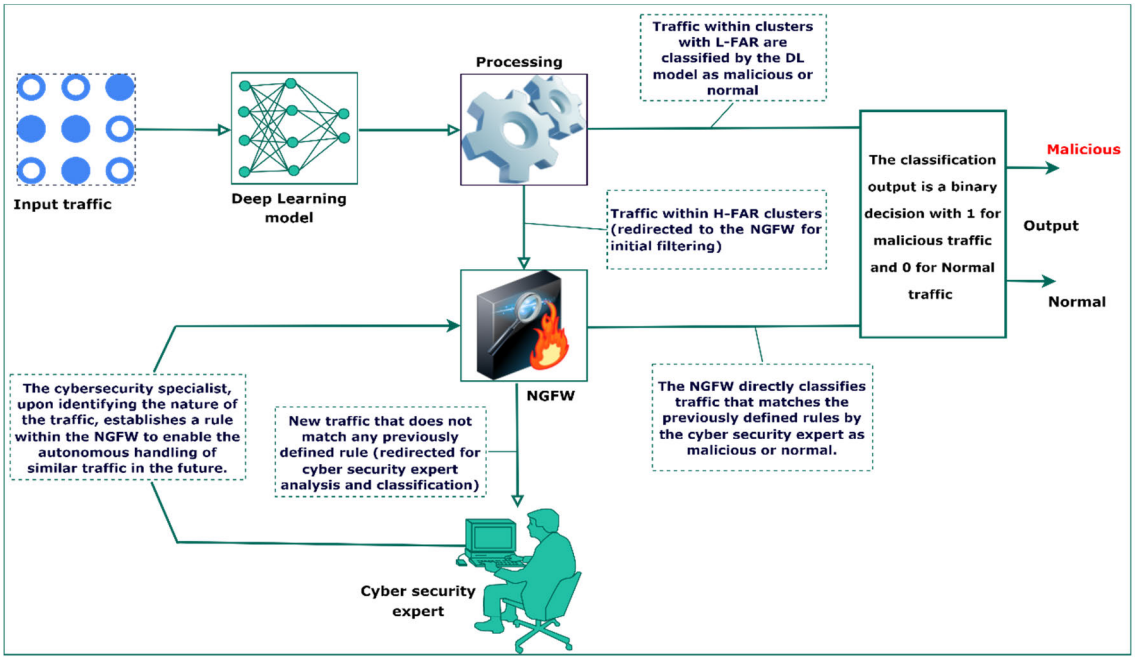

FIGURE 1. The proposed system workflow for human-machine IDS.

Short-Term Memory (LSTM). These models were carefully designed to play a pivotal role in processing data and extracting valuable features for classification tasks. The selection of CNN, BiLSTM, and LSTM models for intrusion detection has been motivated by their wide use in the literature for the same purpose. CNNs have proven their efficiency in capturing spatial dependencies within network traffic data using filters to detect patterns and features. They excel in extracting high-level representations from raw data, thereby offering robustness in feature extraction for intrusion detection. Conversely, the BiLSTM and LSTM models are efficient in capturing both forward and backward temporal dependencies within sequences and excel in learning intricate patterns in time-series data. By leveraging the bidirectional nature of BiLSTM and the memory retention of LSTM, the proposed architecture comprehensively captures nuanced patterns and long-term dependencies within sequential network traffic, thereby enhancing the model's ability to discern complex intrusion patterns. The choice of these models is strategic, harnessing the CNNs' spatial feature extraction capabilities and the proficiency of BiLSTM+LSTM in handling sequential data to offer a comprehensive and adaptable approach for effective intrusion detection. The last motivation is to prove that the proposed probabilistic clustering is applicable to various deep learning technologies.

1) DESIGNED CNN

Convolutional Neural Networks, commonly referred to as CNNs, are a class of deep neural networks predominantly used in computer vision and image processing tasks. However, their application extends to various domains, including natural language processing and time-series analysis. CNNs are characterized by their proficiency in learning hierarchical patterns and features from structured data. In this study, CNNs served as the initial data processing and feature extraction stages. The model comprises convolutional layers that perform local feature extraction, pooling layers to reduce spatial dimensions, and fully connected layers for classification tasks. The use of 1D CNNs is particularly effective for analyzing sequential data, making them well suited for data-processing objectives. For more information on CNN, readers can explore the article by Alzubaidi et al. [34]. Figure 2 illustrates the architecture of the proposed model.

2) DESIGNED HYBRID MODEL: BiLSTM+LSTM

BiLSTM and LSTM networks are recurrent neural networks (RNNs) designed to capture sequential dependencies in the data. LSTM cells are equipped with memory units that enable them to process and remember information over long sequences. In contrast, BiLSTM is an extension of LSTM that processes sequences in both the forward and backward directions, allowing them to capture contextual information effectively. In this study, a combination of BiLSTM and LSTM was employed to exploit the sequential characteristics of the data. This hybrid architecture leverages the benefits of LSTM memory retention and the ability of BiLSTM to simultaneously consider past and future information. By employing this architecture, we aimed to harness the inherent temporal dependencies present in the data, ultimately enhancing the performance of classification tasks. Further details of BiLSTM and LSTM can be found in [35]. The designed high-performance model is shown in Figure 3.

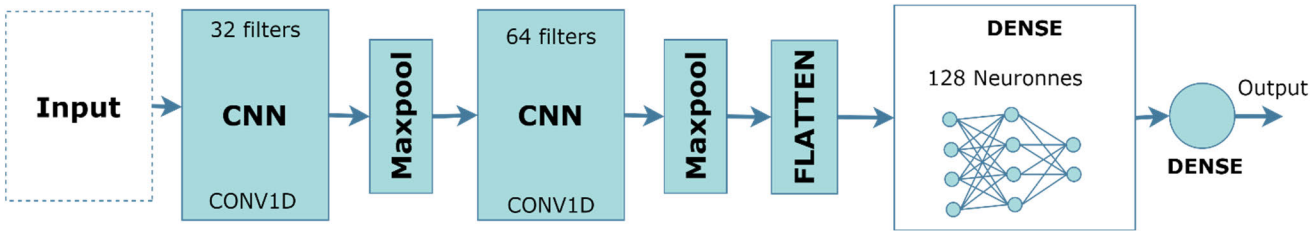

FIGURE 2. Designed CNN model.

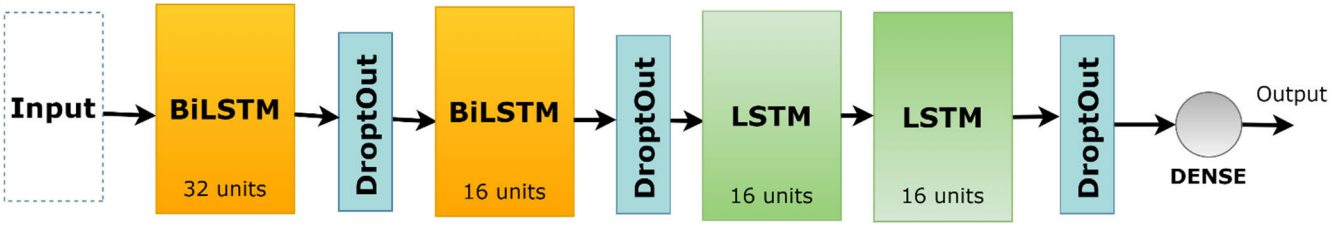

FIGURE 3. Designed BiLSTM+LSTM model.

IV. DATASETS AND SIMULATION PROCESS

In this section, we elucidate the fundamental components of the study, encompassing the datasets under scrutiny and the simulation environments. The selection of datasets and the simulation process are pivotal, as they form the bedrock upon which the investigation was conducted.

A. DATASETS

Three well-known benchmark datasets were used to evaluate the proposed DL models using the human-in-the-loop expertise techniques: CICDDoS2019 [36], UNSW-NB15 [37], and CICIDS2017 [38].

- **CICDDoS2019 dataset**: The First benchmark dataset used to evaluate the proposed models was the CICDDoS2019 dataset, which was developed by the Canadian Institute for Cybersecurity (CIC) and is one of the latest updated benchmark datasets used for evaluating anti-DDoS systems. This dataset contains different modern reflective DDOS attacks. The dataset was initially organized into two sets: training and testing. The size of the test set is significantly larger. The dataset was normalized, and the labels were binary-encoded with 0 for normal samples and 1 for malicious samples. Subsequently, the Nans (not a number) values were replaced by the mean value for each feature (or eliminated when excessive), and the features were scaled to have zero mean and unit variance for better performance during training. Subsequently, feature selection was performed to curate the dataset by eliminating highly correlated attributes. A correlation function with a stringent threshold of 0.80 was employed for this purpose.
- **UNSW-NB15 dataset:** In conjunction with the CICDDoS2019 dataset, this study leveraged the UNSW-NB15 dataset, which is a widely acknowledged resource in the domain of network security. The UNSW-NB15 dataset is well-known for its diverse network traffic scenarios and attack profiles, making it an ideal choice for evaluating and validating the proposed models. It was created by the Cyber Range Lab of the Australian Center for Cyber Security and includes nine types of attacks: Backdoors, Reconnaissance, DoS, Generic, Worms, Exploits, Fuzzers, Analysis, and Shellcode. For dataset normalization, we used the same techniques as for CICDDOS2019, except for the correlation function threshold, which was set to 0.95, yielding 34 features selected from 49 initial features of the dataset.
- **CICIDS2017 dataset:** The last dataset used to assess the proposed framework was the CICIDS2017 dataset. It has been widely used by researchers for IDS evaluation. It is newer than the UNSW-NB15 dataset and covers many types of attacks, including DDoS, DoS, Sql injection, web attacks, PortScan, and many others explorable from the source paper or website (IDS 2017 | Datasets | Research | Canadian Institute for Cybersecurity | UNB). Similar to the two datasets above, it was normalized using the same technique. All the attack samples were kept, but the normal samples were undersampled randomly for data balance and to avoid normal sample bias during the training. For features selection, the correlation function was set to 0.95, yielding to 43 final features used for the models training and testing. The resulting refined datasets following the preprocessing steps are presented in Table 1. Notably, 20% of the training set was used for validation during the training process.

B. SIMULATION PROCESS

Experiments and model development were conducted in both controlled and simulated environments. The computational infrastructure used for these activities included a Windows

TABLE 1. Datasets distributions used in the simulation.

Dataset	Total samples	Selected Features	Train samples	Test samples
CICDDoS2019	431,371	55	125,170	306,201
UNSW-NB15	257,673	34	82,332	175,341
CICIDS2017	851,750	43	596,225	255,525

11 operating system with 24 gigabytes of Random Access Memory (RAM) and an NVIDIA graphics card, specifically GeForce RTX 2060. TensorFlow Keras, Pandas, and Python were used to implement the proposed deep learning models for data processing and build a probabilistic clustering algorithm. To simulate human expert analysis, the process was readily configured, given the availability of a known testing dataset. A Python script was developed to autonomously mimic the role of a human expert. This script processes the redirected traffic and makes classification decisions using the original dataset labels as a reference. To simulate the creation of exception rules in the NGFW, we employed a Python script that utilized redirected traffic IDs from the testing datasets to filter traffic that had been categorized by human experts. In a real-world scenario, rule creation can be based on protocols, packet size, port number, signature, source/destination IP addresses, and so on.

For each dataset and the proposed DL model, we systematically varied the cluster size (N) within a set of values {10, 20, 30, 40, 50, 60}. For each cluster size, we explored different levels of High False Alarm Rate (H-FAR) with thresholds (θ) of {10%, 20%, 30%, 40%, 50%, and 60%}. This resulted in an extensive parameter grid for a comprehensive assessment of the performance of the proposed approach. The optimal hyperparameters of the proposed DL models for each dataset are presented in Tables 2, 3, and 4.

C. EVALUATION METRICS

In this study, the performance of the proposed system model was assessed before and after incorporating human expertise into the traffic redirection. Common evaluation metrics [39] were employed to quantify the effectiveness of the system, including accuracy (ACC), precision (PR), recall (R), F1-score, false positive rate (FPR), and false negative rate (FNR). FPR_a and FNR_a are used to represent FPR and FNR without human expertise (HE), and FPR_b and FNR_b, respectively, when the human expertise technique is enabled via probabilistic clustering. These metrics aid in assessing the classification performance of the system in both scenarios. The equations corresponding to each metric are as follows:

$$\text{ACC} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (4)$$

$$\text{PR} = \text{TP}/(\text{TP} + \text{FP}) \quad (5)$$

$$R = \text{TP}/(\text{TP} + \text{FN}) \quad (6)$$

$$\text{F1\_score} = 2 \times (PR \times R)/(\text{PR}+R) \quad (7)$$

$$\text{FPR\_a} = \text{FP}/(\text{FP} + \text{TN}) \quad (8)$$

$$\text{FNR\_a} = \text{FN}/(\text{FN} + \text{TP}) \quad (9)$$

$$\text{FPR\_b} = \text{FP}'/(\text{FP}' + \text{TN}') \quad (10)$$

$$\text{FNR\_b} = \text{FN}'/(\text{FN}' + \text{TP}') \quad (11)$$

where:

TP = True Positives without HE; TN = True Negatives without HE; FP = False Positives without HE; FN = False Negatives without HE.

TP′ = True Positives with HE; TN′ = True Negatives with HE; FP′ = False Positives with HE; FN′ = False Negatives with HE.

These metrics allow the theoretical quantification of the improvement in false-alarm mitigation after leveraging human expertise through traffic redirection.

V. RESULTS AND DISCUSSIONS

This section presents and discusses the simulation results. The proposed DL models were evaluated using the testing sets.

A. EVALUATION RESULTS: WITHOUT HUMAN EXPERTISE

In this section, we present the initial evaluation results which reflect the performance of the system before incorporating human expertise. Table 5 summarizes the performance of the models evaluated using the testing sets of the corresponding datasets. The outcomes in this section lay the foundation for a comparative analysis that delves into the transformative impact of human expertise and its associated improvements in later subsections. The initial simulation demonstrated excellent results for the proposed DL models alone. The CNN model achieved FPR and FNR values of 0.114% and 0.506%, 1.918% and 0.771%, and 0.122% and 0.541% for the CICDDoS2019, UNSW-NB15, and CICIDS2017 testing sets, respectively. The FPR and FNR values with BiLSTM+LSTM were 0.068% and 0.399%, 0.663% and 0.035%, and 0.133% and 0.551%, respectively, for the same testing sets. The hybrid BiLSTM+LSTM model yielded better results with fewer false alarms than the CNN model. This can be explained by the hybrid model, which is based on the LSTM architecture and can learn the dependencies between network packets. The confusion matrix of the three models is presented in Figure 4 for the CICDDoS2019 dataset, Figure 5 for the UNSW-NB15 dataset, and Figure 6 for the CICIDS2017 dataset.

TABLE 2. Optimal hyperparameters with CICDDoS2019 dataset.

Hyperparameter\Model	CNN	BiLSTM+LSTM
Learning Rate	-	0.005
Number of Epochs	15	15
Batch Size	42	42
Activation function	sigmoid	sigmoid
Loss function	binary_crossentropy	binary_crossentropy
optimizer	Adam	Adam

TABLE 3. Optimal hyperparameters with UNSW-NB15 dataset.

Hyperparameter\Model	CNN	BiLSTM+LSTM
Learning Rate	0.0015	-
Number of Epochs	10	15
Batch Size	32	32
Activation function	sigmoid	sigmoid
Loss function	binary_crossentropy	binary_crossentropy
optimizer	Adam	Adam

TABLE 4. Optimal hyperparameters with CICIDS2017 dataset.

Hyperparameter\Model	CNN	BiLSTM+LSTM
Learning Rate	-	-
Number of Epochs	40	40
Batch Size	42	42
Activation function	sigmoid	sigmoid
Loss function	binary_crossentropy	binary_crossentropy
optimizer	Adam	Adam

TABLE 5. The results without human-in-the-loop technique.

Designed Model	Metric\dataset	CICDDoS2019	UNSW-NB15	CICIDS2017
CNN	**ACC**	**99.55%**	**99.67%**	**98.76%**
	PR	99.97%	99.63%	98.10%
	R	99.49%	99.87%	99.45%
	F1_score	99.73%	99.75%	98.77%
	FPR_a	**0.114%**	**0.771%**	**1.918%**
	FNR_a	**0.506%**	**0.122%**	**0.541%**
BiLSTM+LSTM	**ACC**	**99.65%**	**99.89%**	**99.39%**
	PR	99.98%	99.98%	99.33%
	R	99.60%	99.86%	99.44%
	F1_score	99.79%	99.92%	99.39%
	FPR_a	**0.068%**	**0.035%**	**0.663%**
	FNR_a	**0.399%**	**0.133%**	**0.551%**

The designed DL models performed well without applying a human-in-the-loop technique using probabilistic clustering. However, they still generate false alarms, similar to many deep learning classifiers in the literature. The following subsection assesses how human expert intervention can contribute to the mitigation of these alarms.

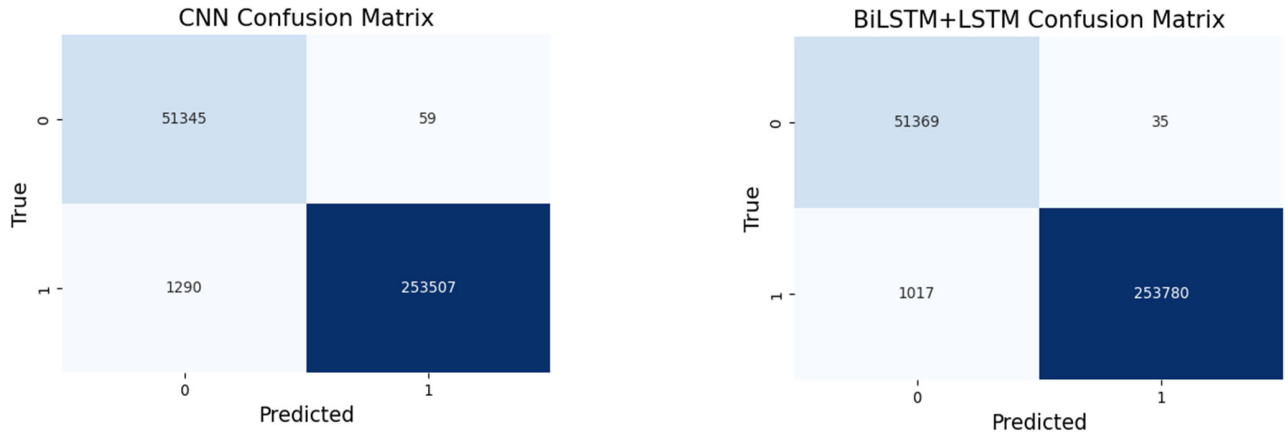

FIGURE 4. Confusion matrixes of the proposed CNN and BiLSTM+LSTM models with the CICDDoS2019 testing set.

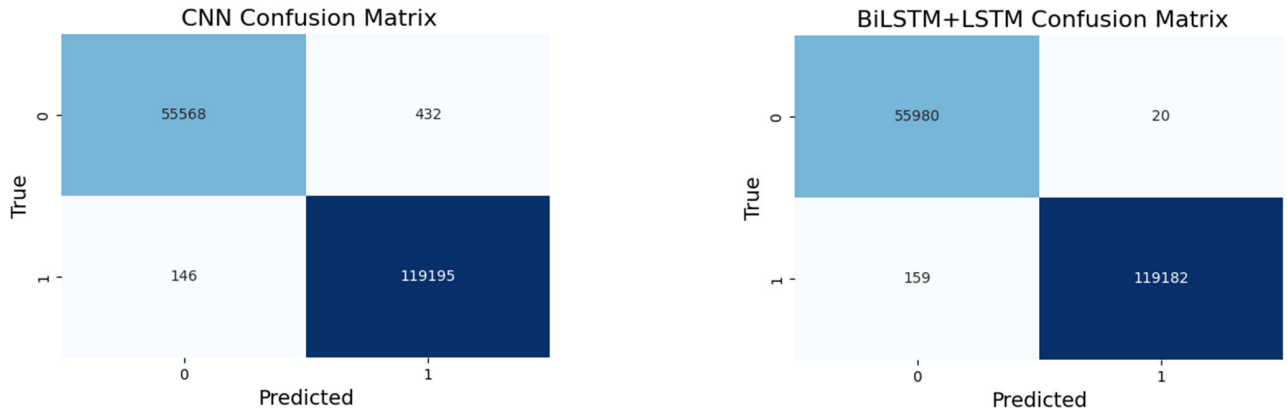

FIGURE 5. Confusion matrixes of the proposed CNN and BiLSTM+LSTM models with UNSW-NB15 testing set.

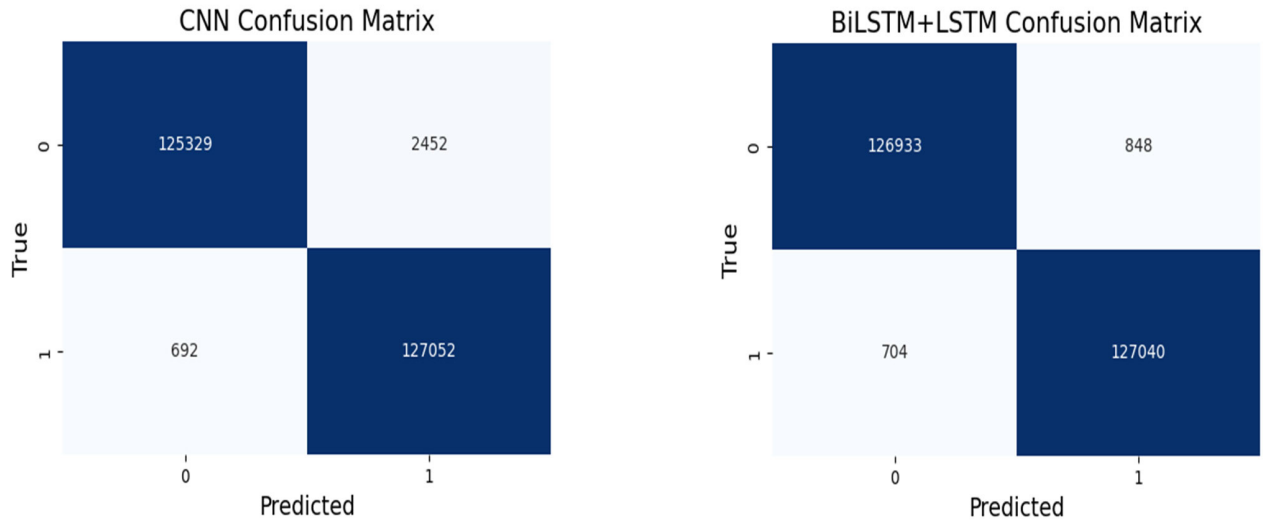

FIGURE 6. Confusion matrixes of the proposed CNN and BiLSTM+LSTM models with CICIDS2017 testing set.

B. EVALUATION RESULTS: WITH HUMAN-IN-THE-LOOP TECHNIQUE

We moved from evaluating the autonomous capabilities of deep learning models to examining the impact of human expertise on the proposed traffic-categorization framework. The focus shifted to a comparative analysis that highlights the crucial role of human expert intervention in supporting the application of deep learning classifiers (in the context of

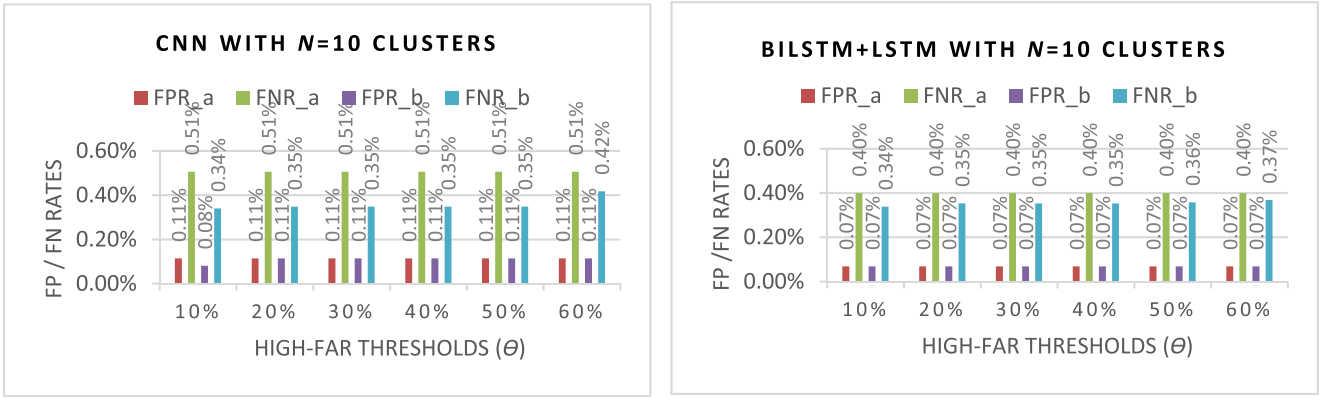

FIGURE 7. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the CICDDoS2019 Testing Set, considering H-FAR threshold (θ) variations.

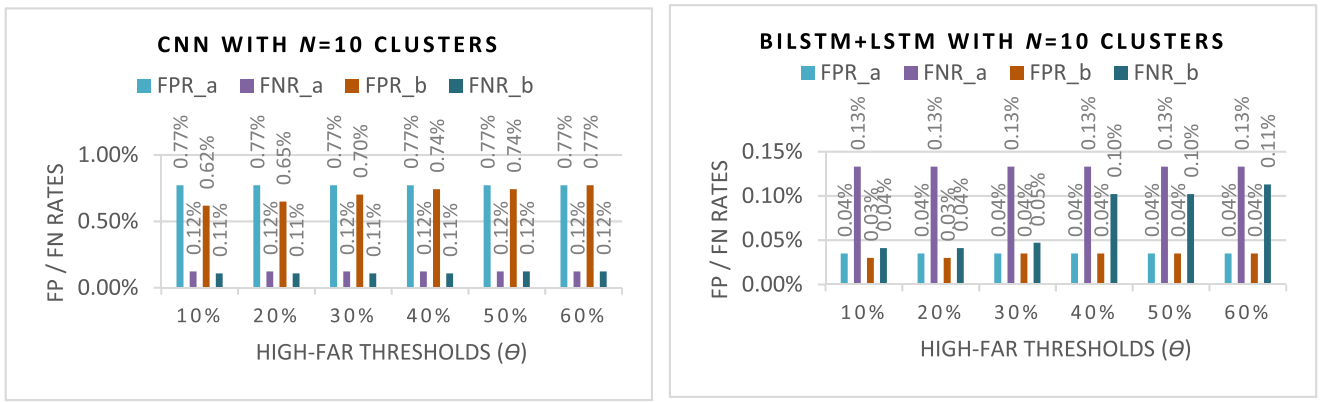

FIGURE 8. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the UNSW-NB15 Testing Set, considering H-FAR threshold (θ) variations.

intrusion detection) in real-world network scenarios. The assessment involved a comprehensive set of parameters, including varying the cluster count ($N = [10, 20, 30, 40, 50, 60]$) and adjusting the H-FAR thresholds ($\theta = [10\%, 20\%, 30\%, 40\%, 50\%, 60\%]$) for each cluster to ensure a thorough and robust evaluation.

1) RESULTS FOR $N = 10$ CLUSTERS

For $N = 10$ clusters, H-FAR clusters were determined, and a human domain expert intervention technique was employed for the identified H-FAR clusters, leading to the establishment of new metrics pertaining to false-positive (FPR_b) and false-negative (FNR_b) rates. Quantitative assessments of the reduction in False Positives (FPs) and False Negatives (FNs) were performed.

The comparative results derived from the CICDDoS2019, UNSW-NB15, and CICIDS2017 datasets when the human expertise technique was applied and when it was not are comprehensively presented in Figures 7, 8, and 9, respectively. Specifically, significant reductions were observed when the CNN model was used in conjunction with human-in-the-loop expertise on the CICDDOS2019 dataset. At a threshold of $\theta = 10\%$, a substantial 28.81% reduction in FPs and an impressive 32.79% reduction in FNs were achieved, yielding global reductions in FPR and FNR. However, elevating the threshold beyond 20% predominantly affected FNs reduction, ranging from 17.51% to 31.24%. Interestingly, the FPR remained consistently low, indicating a limited impact on mitigating it with higher thresholds.

Furthermore, the BiLSTM+LSTM model with human expertise demonstrated a more constrained impact. Specifically, the FNs reduction was up to 15.14%, whereas the FPs remained notably the same owing to a very low FPR.

In the evaluation using the UNSW-NB15 dataset, a significant reduction in FPs and FNs was observed when the human expertise technique was applied. When the CNN model was employed in tandem with human expertise at $\theta = 10\%$, a commendable reduction in FPs of 19.90% and FNs of 12.32% was observed. However, with a higher threshold of $\theta = 50\%$, only False Positives (FPs) demonstrated a modest reduction of 3.70%, whereas no clusters with high false alarm rates (H-FAR) were detected beyond this threshold. In contrast, the application of the BiLSTM+LSTM model with human expertise yielded distinctive outcomes. At $\theta = 10\%$, an impressive reduction of 68.55% in FNs was achieved, yielding a high reduction in the FNR. In Addition, a notable

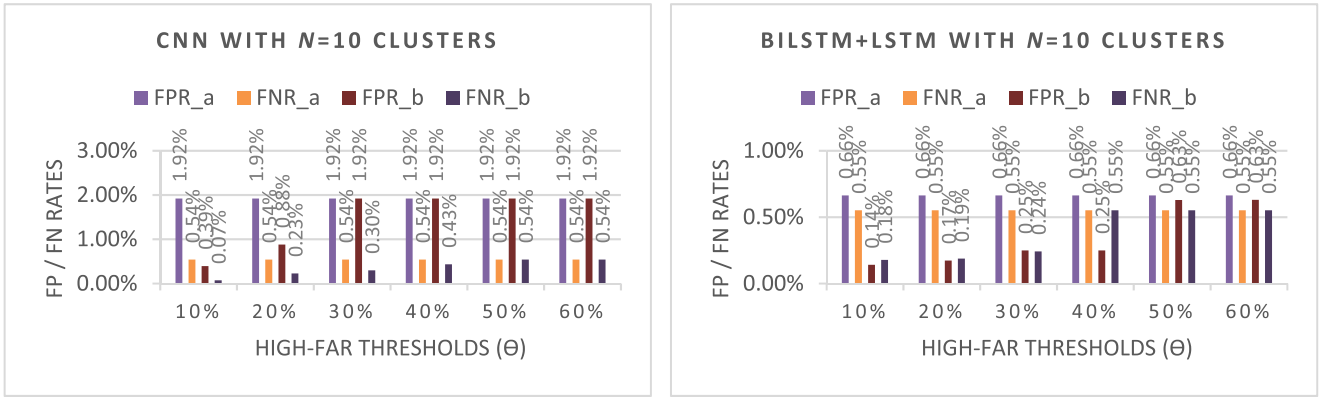

FIGURE 9. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the CICIDS2017 Testing Set, considering H-FAR threshold (θ) variations.

FNs reduction of 15.09% was observed at $\theta = 60\%$. FPs were reduced by 15% for $\theta = 10\%$ and 20% but no reduction was observed beyond $\theta = 20\%$.

The assessment using the CICIDS2017 dataset also yielded good results for 10 clusters. The CNN model with the human-in-the-loop technique yielded a 79.61% reduction in FPs and 86.99% reduction in FNs for an H-FAR threshold of 10%. For a threshold of 20%, the reduction in FPs was 54.16%, and that of FNs was 57.95%. No reduction in FPs was observed for and above a threshold of 30%. The FNs were reduced by 45.23% and 20.09% for thresholds of 30% and 40%, respectively. However, for and above 50%, no reduction was observed. The hybrid BiLSTM+LSTM model coupled with human expertise achieved a reduction in FPs between 5.07% (for a threshold of 60%) and 78.77% (for a threshold of 10%). That of the FNs was reduced between 55.97% (for a threshold of 30%) and 67.76% (for a threshold of 10%). For a threshold equal to or greater than 40%, no reduction in FNs was observed..

2) RESULTS FOR $N = 20$ CLUSTERS

In the evaluation using 20 clusters, a comprehensive reduction in false alarms was observed in the simulation results derived from the CICDDoS2019, UNSW-NB15, and CICIDS2017 datasets, as presented in Figures 10, 11, and 12, respectively. Using the CNN model in conjunction with human-in-the-loop expertise, substantial reductions in FPs and FNs were achieved at various thresholds. Specifically, at $\theta = 10 = 60\%$, only False Negatives (FNs) demonstrated a reduction of 17.51%. Above $\theta = 50\%$, no clusters exhibiting H-FAR-containing FPs were detected. Furthermore, the hybrid model with human expertise showed reductions ranging between 7.47% and 23.79% for the FNs, indicating a moderate impact on FNs mitigation. However, the reduction in FPs was comparatively limited, with a reduction of only 5.71%. These findings underscore the nuanced performance differences between the models and highlight the varying effectiveness of addressing false alarms in relation to different threshold values.

The evaluation using the UNSW-NB15 dataset yielded a reduction in FPs between 1.3% and 24.30% when the CNN model was coupled with human expertise. The FNs reduction ranged from 12.32% to 26.71% when the H-FAR threshold was less than 50%. The hybrid BiLSTM+LSTM allowed a reduction in FNs from 20.12% to 68.55% with human expertise; however, only 15% of the FPs was reduced for $\theta = 10\%$.

The evaluation using the CICIDS2017 dataset yielded the following results. With the CNN model, the FPs were reduced between 42.49% and 79.61% for an H-FAR threshold less than or equal to 30%, and no reduction otherwise. The reduction in FNs for the same model was between 20.37% and 79.47%, with a threshold of less than or equal to 40%, and no reduction was observed otherwise. By using the BiLSTM+LSTM model, the FPs were reduced by between 5.07% and 78.77%. The number of FNs showed a reduction between 35.51% and 83.52% for a threshold less than or equal to50%; otherwise, no reduction was observed.

3) RESULTS FOR $N = 30$ CLUSTERS

The simulation results in terms of false alarms reduction for $N = 30$ clusters for the three datasets are presented in Figures 13, 14, and 15. For the CNN model coupled with human expertise for the CICDD0S2019 dataset, the FPs were reduced by 25.42% and 13.55% for thresholds of 10% and 20%, respectively. Beyond a threshold of 20%, there was no reduction in FPs owing to the absence of H-FAR with FPs at this threshold. Concerning FNs, the reduction was between 24.34% and 51.86% based on the threshold. The BiLSTM+LSTM model with human expertise achieved FPs reduction of 22.85% only for a threshold of 10%. Beyond this threshold, there was no reduction in the FPs because of the very low FPR. The FNs values were reduced by 8.65% and 32.05% based on the threshold.

For the UNSW-NB15 dataset utilizing the CNN model, the integration of human expertise led to FPs reductions from 3.245% to 27.77% across thresholds ranging from 10% to 50%. Correspondingly, the reduction in FNs ranged

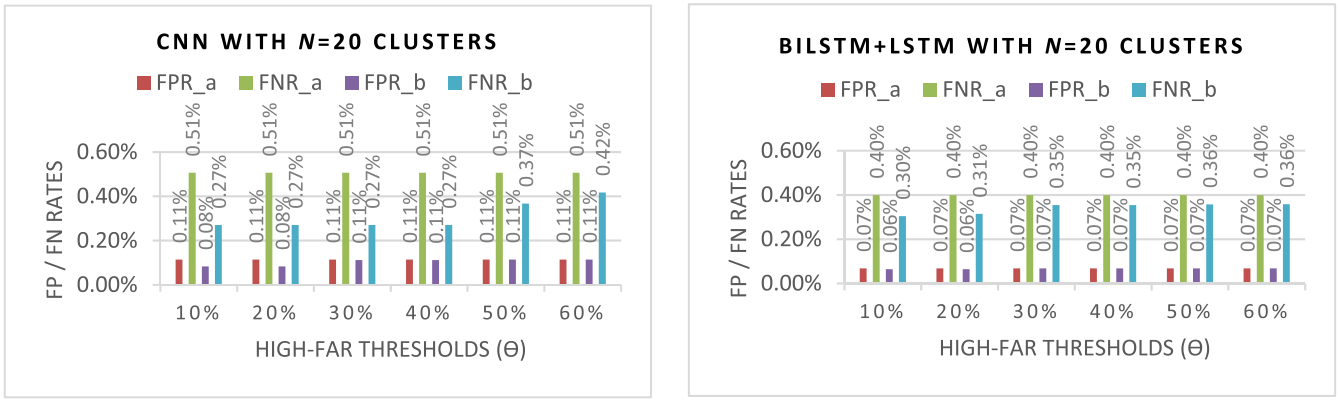

FIGURE 10. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the CICDDoS2019 Testing Set, considering H-FAR threshold (θ) variations.

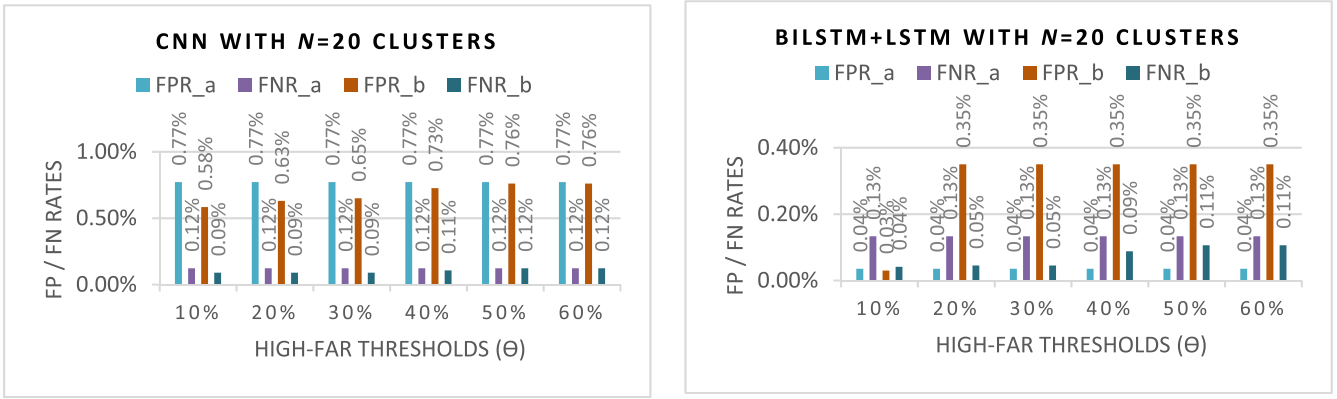

FIGURE 11. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the UNSW-NB15 Testing Set, considering H-FAR threshold (θ) variations.

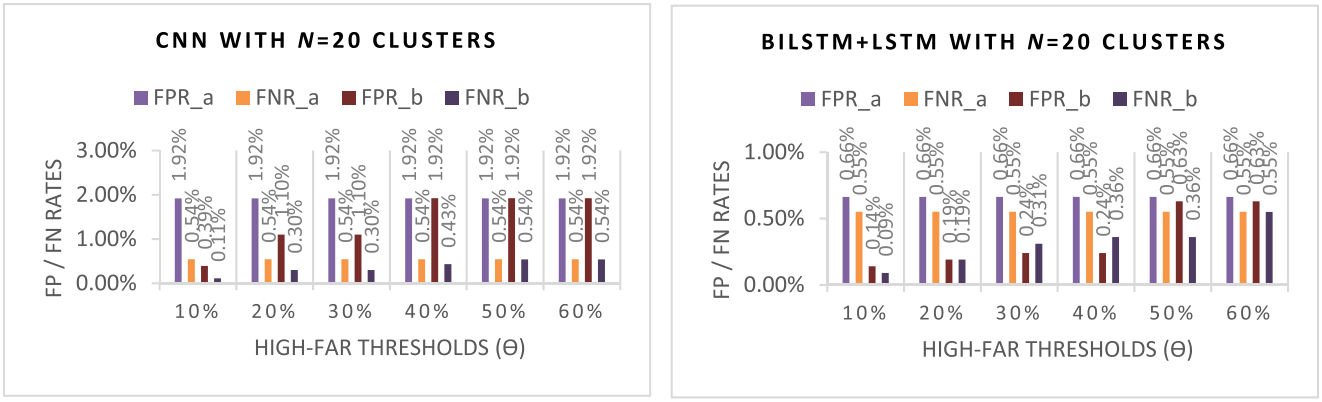

FIGURE 12. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the CICIDS2017 Testing Set, considering H-FAR threshold (θ) variations.

between 8.21% and 30.82%, correlating with the specific threshold settings. In contrast, employing the hybrid model (BiLSTM+LSTM) with human expertise resulted in notable reductions in FPs by 40% at $\theta = 10\%$ and 15% at $\theta = 20\%$. However, beyond the threshold of $\theta = 20\%$, no further reductions in FPs were observed, owing to a pre-existing low False Positive Rate. In contrast, FNs displayed reductions of up to 74.84% at $\theta = 10\%$ and 21.38% at$\theta = 60\%$.

The assessment using the CICIDS2017 dataset under the human-expertise technique gave the following results. The CNN model coupled with human expertise achieved FPs reduction between 1.67% and 79.61% for a threshold

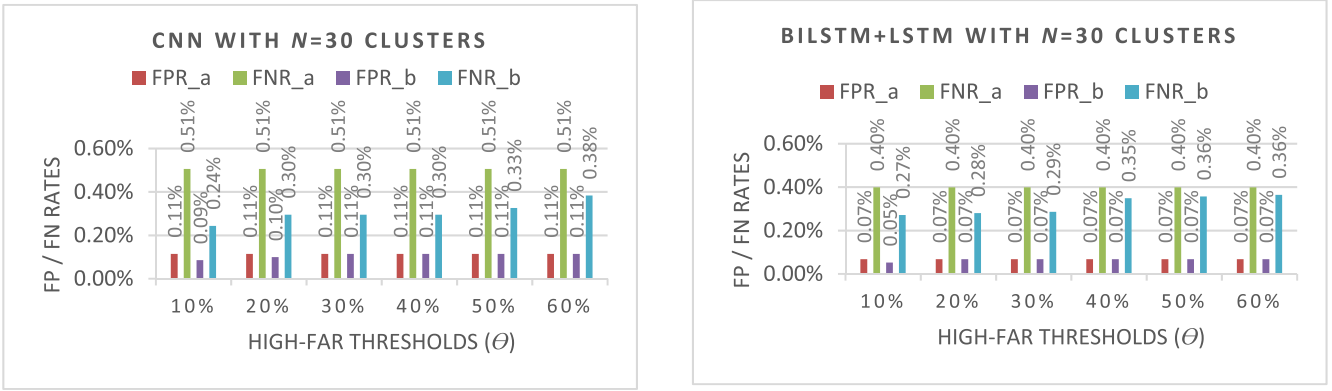

FIGURE 13. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the CICDDoS2019 Testing Set, considering H-FAR threshold (θ) variations.

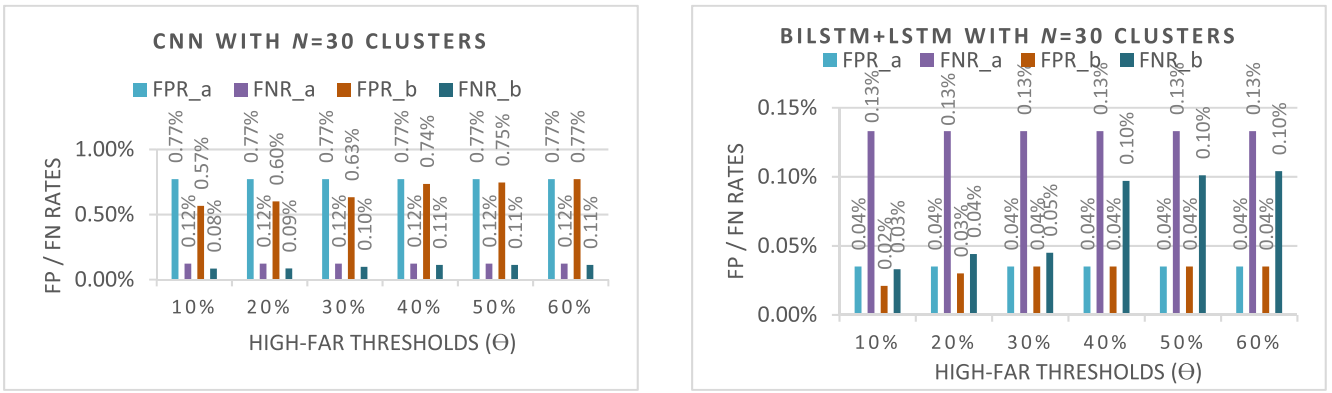

FIGURE 14. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the UNSW-NB15 Testing Set, considering H-FAR threshold (θ) variations.

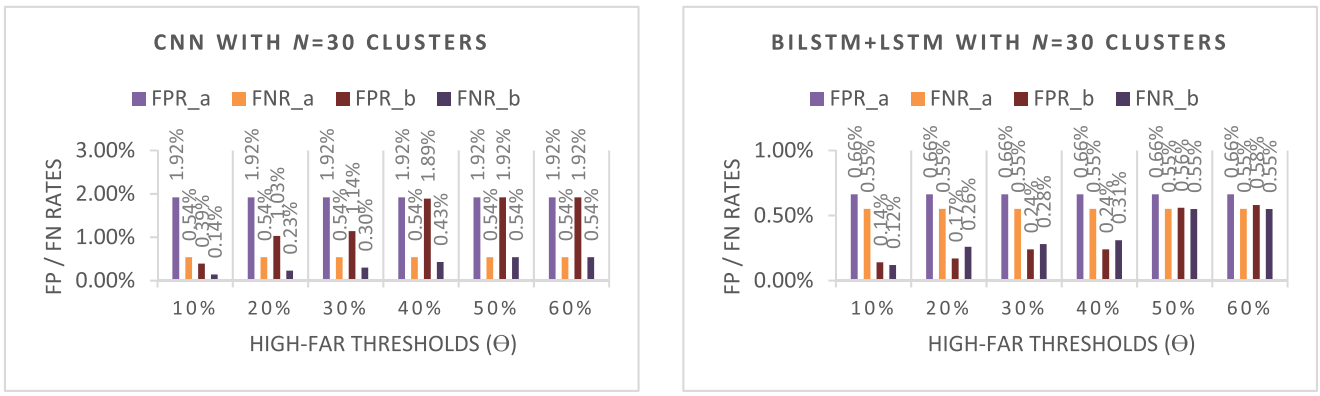

FIGURE 15. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the CICIDS2017 Testing Set, considering H-FAR threshold (θ) variations.

(?) less than or equal to 40%; otherwise, no reduction in FPs was observed. The reduction of FNs with the same model was between 21.24% and 74.86% for the same threshold range, and no reduction was observed above it. The BiLSTM+LSTM model with human expertise technique achieved FPs reduction between 12.5% and 78.30% based on the used threshold. The reduction in FNs was possible when the threshold was less than or equal to 40%. The reduction in FNs was between 43.04% and 77.70%.

4) RESULTS FOR $N = 40$ CLUSTERS

The false alarm rate reductions for 40 clusters are outlined in Figures 16, 17, and 18 for each dataset. Employing the

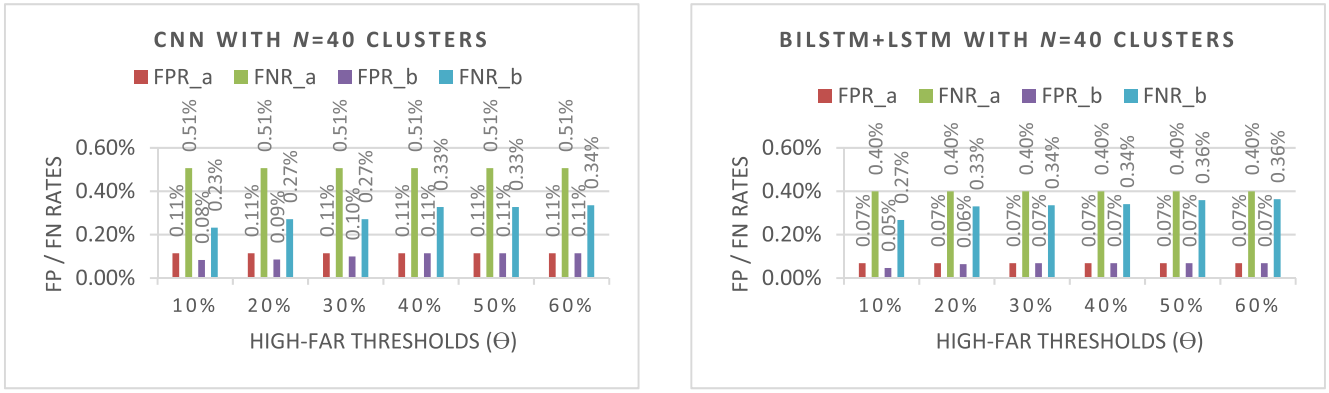

FIGURE 16. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the CICDDOS2019 Testing Set, considering H-FAR threshold (θ) variations.

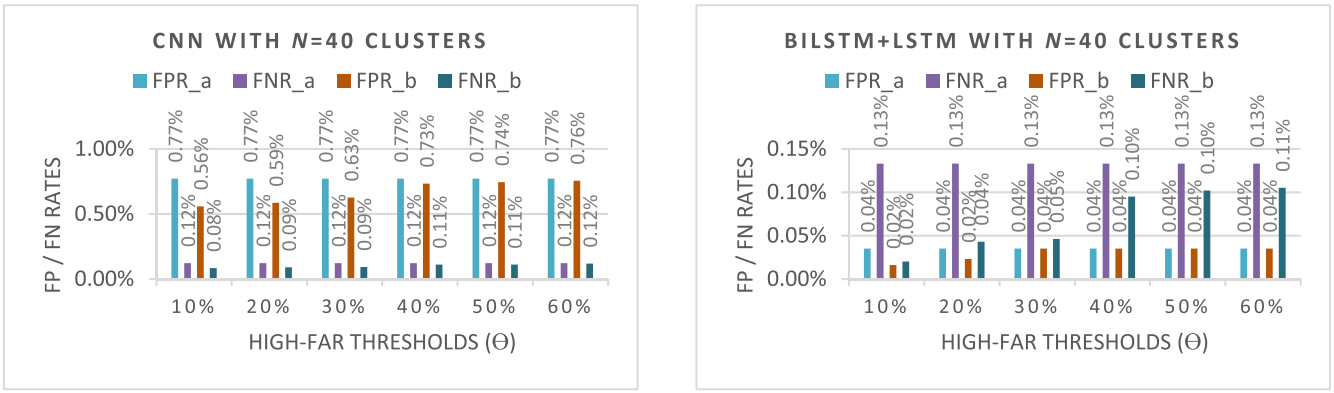

FIGURE 17. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the UNSW-NB15 Testing Set, considering H-FAR threshold (θ) variations.

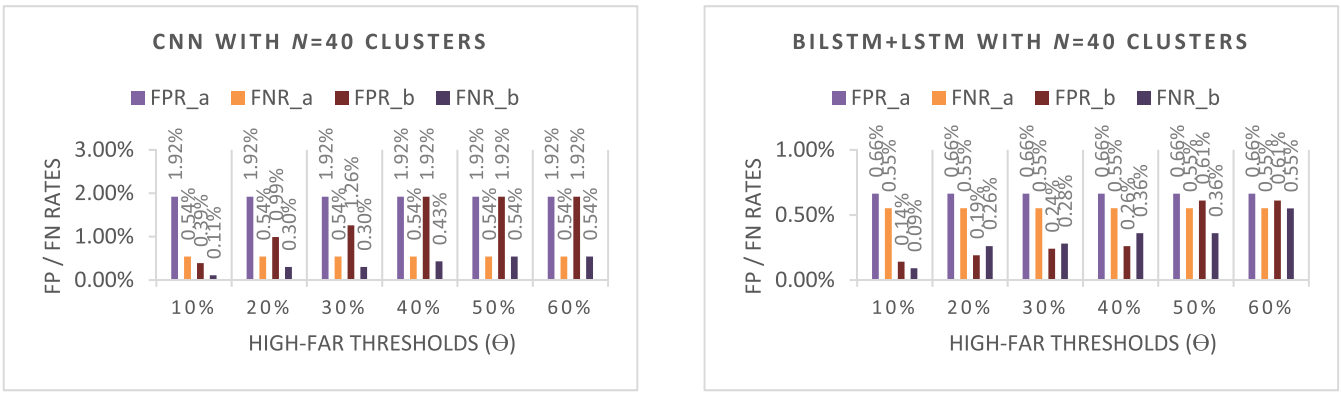

FIGURE 18. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the CICIDS2017 Testing Set, considering H-FAR threshold (θ) variations.

CNN model in conjunction with human expertise on the CICDDoS2019 dataset resulted in substantial reductions in FPs spanning from 13.55% to 27.11%, specifically within a H-FAR threshold equal to or less than 30%. In addition, the reduction in FNs ranged between 33.64% and 54.10%, contingent on the threshold settings between 10% and 60%. In contrast, utilizing the hybrid model on the same dataset demonstrated a reduction in FPs of 31.42% at $\theta = 10\%$ and 5.71% at $\theta = 20\%$. Correspondingly, reductions in FNs varied between 8.84% and 32.94% based on the threshold values employed.

Using the CNN model with human expertise resulted in a reduction in FPs, ranging from 2.08% to 27.54% within the UNSW-NB15 dataset. Correspondingly, the reductions in FNs ranged between 3.42% and 31.50%, illustrating the impact of the specific threshold θ utilized. Employing

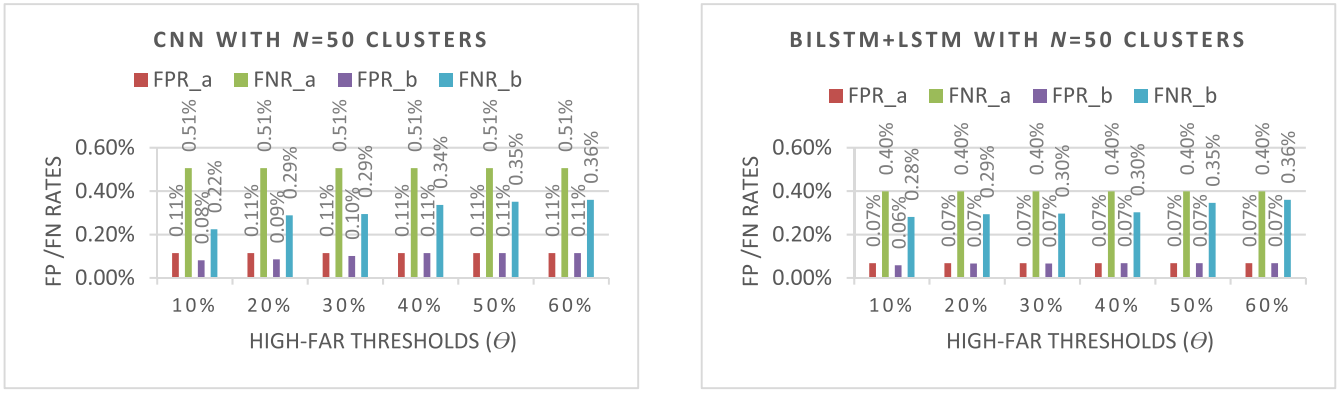

FIGURE 19. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the CICDD0S2019 Testing Set, considering H-FAR threshold (θ) variations.

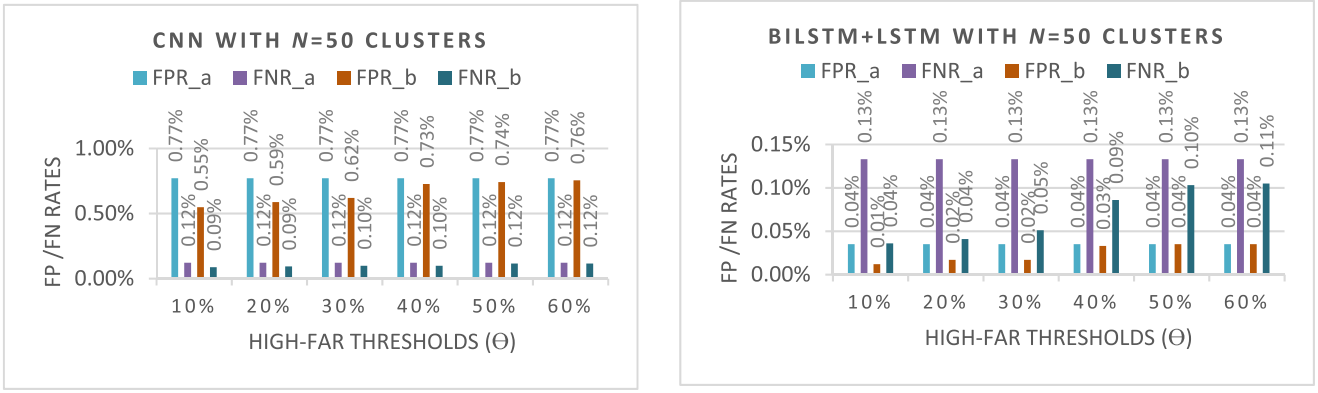

FIGURE 20. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the UNSW-NB15 Testing Set, considering H-FAR threshold (θ) variations.

the hybrid model (BiLSTM+LSTM) with human expertise resulted in substantial reductions in the FPs by 55.00% at $\theta = 10\%$ and 35.00% at $\theta = 20\%$. The reductions in FNs ranged from 20.75% to 84.90%, contingent on the threshold settings employed.

The results with the CICIDS2017 dataset using the human-in-the-loop technique are as follows. The CNN model with human expertise achieved a reduction in FPs between 34.54% and 79.61% for an H-FAR threshold less than or equal to 30%; otherwise, no reduction in FPs was observed. The reduction in FNs was between 19.94% and 79.48% when the H-FAR threshold was not greater than 40%; otherwise, no reduction in FNs was observed. The hybrid model coupled with human expertise allowed for the reduction of FPs between 7.78% and 79.25%, and FNs reduction was possible when the H-FAR threshold was not greater than 50%. In this case, FNs were reduced between 35.51% and 83.81%. The impact of human-in-the-loop expertise enabled by the proposed probabilistic clustering can be observed in the results graphs.

5) RESULTS FOR $N = 50$ CLUSTERS

The experimental results for $N = 50$ clusters are presented in Figures 19, 20, and 21. When employing the CNN model alongside human expertise with the CICDDoS2019 dataset, significant reductions in FPs ranging from 11.86% to 28.81% were achieved, notably within H-FAR thresholds equal to or less than 30%. Correspondingly, the reductions in FNs ranged between 28.83% and 55.58% based on the threshold, showing substantial improvements. Employing the hybrid model on the same dataset resulted in a reduction in FPs from 2.85% to 14.28%, specifically within the threshold θ less than or equal to 30%. Simultaneously, the reductions in FNs ranged from 9.63% to 29.40% based on the chosen threshold values. These findings distinctly outline the varying impacts of different models paired with human expertise, emphasizing their effectiveness in reducing both FPs and FNs within the dataset across distinct threshold configurations.

Using the UNSW-NB15 dataset, the CNN model with human expertise reduced the FPs by 2.08% to 28.93% and the FNs by 4.79% to 28.08% based on the threshold θ. The hybrid model (BiLSTM+LSTM) achieved FPs reductions from 5.00% to 65.00% and FNs reductions from 20.75% to 72.32% for $\theta \leq 40\%$.

The assessment with the CICIDS2017 dataset under human expertise yielded the following results. For an H-FAR threshold less than or equal to 40%, the CNN coupled with human expertise allowed the reduction of FPs between 1.14% and 79.61% and that of FNs between 23.99% and 76.88%. With the BiLSTM+LSTM coupled with human expertise, the

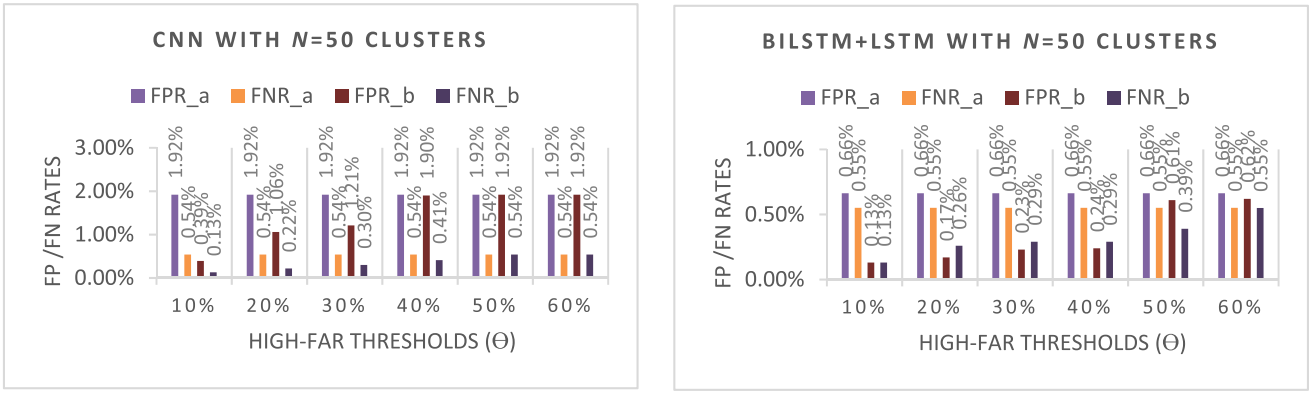

FIGURE 21. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the CICIDS2017 Testing Set, considering H-FAR threshold (θ) variations.

FPs reduction was between 7.08% and 80.31%. The reduction of FNs was only possible for an H-FAR threshold less than or equal to 50% and yielded a reduction of FNs between 28.41% and 76.30%.

6) RESULTS FOR $N = 60$ CLUSTERS

In Figures 22, 23, and 24, the results for $N = 50$ clusters are shown. Using the CICDDoS2019 dataset, the CNN model coupled with human expertise achieved reductions in FPs ranging from 13.55% to 25.42%, notably within H-FAR thresholds equal to or less than 40%. Correspondingly, the reductions in FNs ranged between 29.92% and 55.58%, showing substantial improvements based on the chosen threshold settings. In contrast, the hybrid model on the same dataset achieved FPs reduction of 17.14% at θ =10% and 20%, respectively. Additionally, the reductions in FNs ranged from 10.52% to 31.85%, based on the threshold values employed.

Using the UNSW-NB15 dataset, the CNN model with human expertise reduced FPs by 1.38% to 28.70%, and FNs by 8.21% to 35.61% based on the threshold θ. The hybrid model (BiLSTM+LSTM) achieved FPs reductions of 65.00% and 40.00% at θ = 10% and 20%, respectively. The FNs reductions ranged from 18.23% to 72.95% based on the threshold. Always note that the performance difference between the two models depends on their prediction accuracy and H-FAR clusters detection efficiency for each dataset.

The evaluation results using the CICIDS2017 dataset under the human-in-the-loop technique are as follows. By using the CNN model coupled with human expertise, the reduction of FPs was possible when the H-FAR threshold was less than or equal to 40% and 50% for FNs. The reduction in FPs was between 1.10% and 78.67%, and that of FNs was between 4.04% and 78.32%. The BiLSTM+LSTM model coupled with human expertise yielded a reduction in FPs between 13.68% and 78.89%, given the threshold. FNs reduction was possible when the threshold was less than or equal to 50%, with a reduction ranging from 25.85% to 80.40%.

Through all the evaluations under the human expertise technique enabled by the proposed probabilistic clustering, a few or large reductions in false positives or false negatives were observed, yielding a reduction in FPR and/or FNR for each model and for each testing dataset used.

C. RESULTS COMPARISON

This section first compares the best results obtained between the performance of the proposed models without the human-in-the-loop technique and when the human-in-the-loop technique is applied through probabilistic clustering. The results obtained using the human-in-the-loop technique were then compared with the performances of previous studies. As the models coupled with the human-in-the-loop technique were evaluated using different H-FAR thresholds under different cluster sizes, only the best outcomes are used for the comparison. Table 6 shows for each dataset and for each model the optimal cluster size and H-FAR value that achieved the best results in terms of false alarm reduction, yielding performance improvement.

1) HUMAN-IN-THE-LOOP IMPACT ON THE PROPOSED MODELS' PERFORMANCES

Previously, it was explained that human expertise was enabled by the proposed probabilistic clustering to assist deep learning models, where they can be inefficient in the traffic classification process. This subsection compares the simulation results of the models when the probabilistic clustering method is not applied and when it is applied for the human-in-the-loop technique. Table 7 presents the results of this comparison. It can be observed from the table the positive impact of human-in-the-loop intervention in terms of performance. Through all three datasets with different designed models, human expertise helped improve the accuracy, precision, recall, F1-score, FPR, and FNR.

2) RESULTS COMPARISON WITH PREVIOUS WORKS

In this subsection, the best performance of the proposed system when the human-in-the-loop technique is applied through

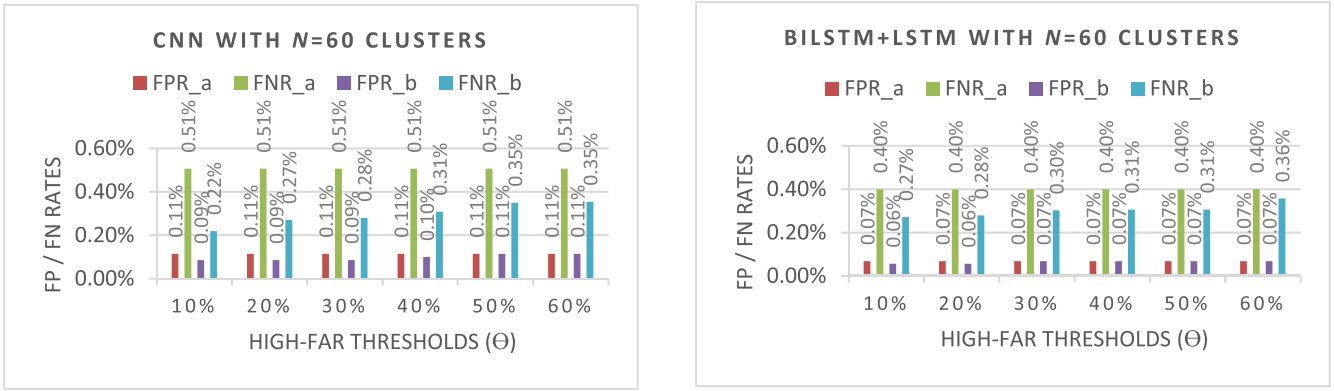

FIGURE 22. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the CICDDoS2019 Testing Set, considering H-FAR threshold (θ) variations.

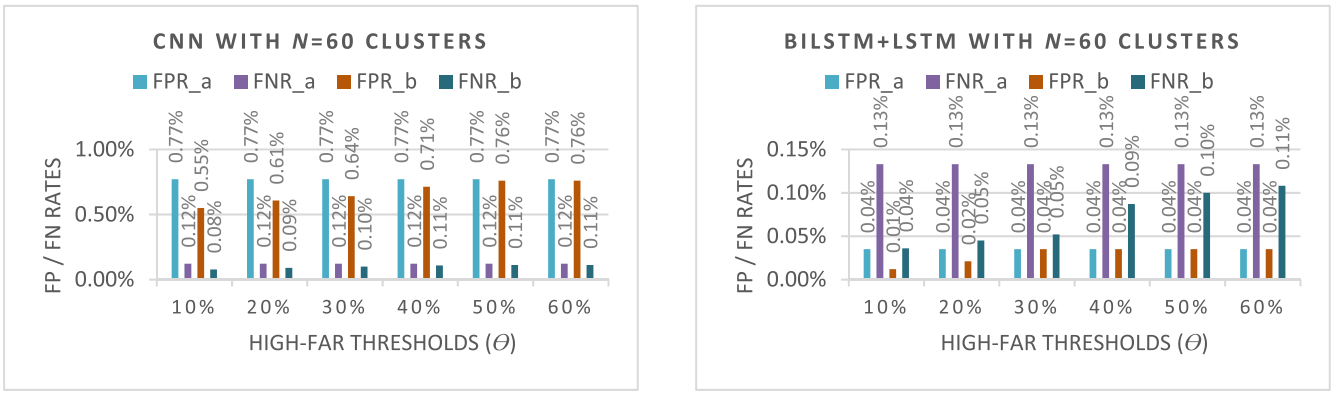

FIGURE 23. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the UNSW-NB15 Testing Set, considering H-FAR threshold (θ) variations.

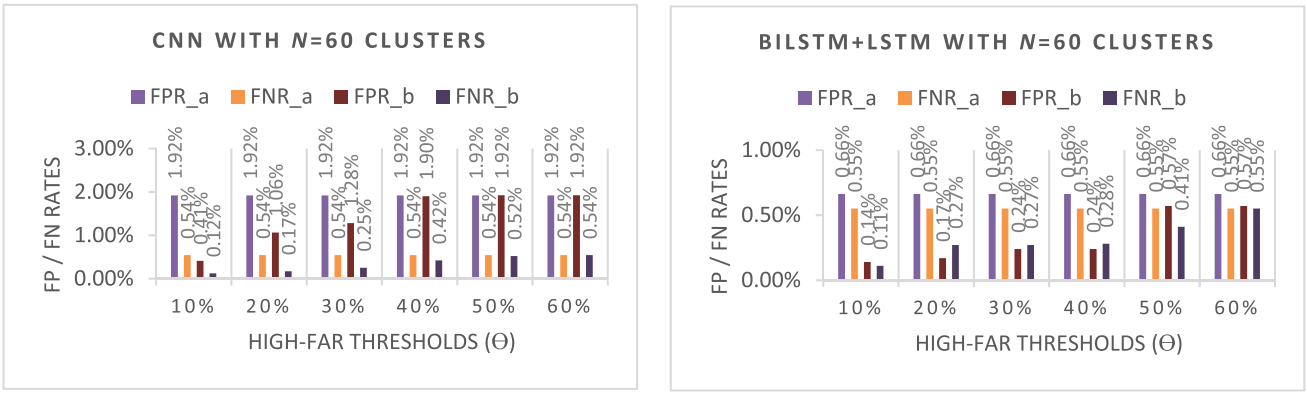

FIGURE 24. Comparison of FP and FN Rates (FPR_a, FNR_a) versus (FPR_b, FNR_b) without and with Human Expertise Intervention using Probabilistic Clustering for Each designed Model on the CICIDS2017 Testing Set, considering H-FAR threshold (θ) variations.

probabilistic clustering is compared with previous works. Table 8 provides the performance comparison in terms of accuracy, Precision, FPR, training time, and inference time of the proposed system with previous works. Only previous studies that utilized the most relevant or widely used public datasets were considered to maintain a more balanced comparison. The comparison tables show that the proposed system outperformed most of the previous works.

D. DISCUSSIONS

The primary objective of this study was to develop a novel approach that leverages human expertise within the framework of DL models to reduce false alarms in IDS. False alarms not only consume valuable resources but can also undermine the trustworthiness and effectiveness of intrusion detection systems. The proposed framework introduces probabilistic clustering to enable human-in-the-loop

TABLE 6. Optimal configuration that yielded the best results with the human-in-the-loop technique.

Dataset	Model	Number of Clusters (N)	H-FAR Threshold (θ)
CICDDoS2019	CNN	60	10%
	BiLSTM+LSTM	40	10%
UNSW-NB15	CNN	60	10%
	BiLSTM+LSTM	40	10%
CICIDS2017	CNN	10	10%
	BiLSTM+LSTM	40	10%

TABLE 7. Performance comparison when human-in-the-loop technique is not applied and when it is applied.

Dataset	Model	Human-in-the-loop	Accuracy	Precision	Recall	F1-score	FPR	FNR
CICDDoS2019	CNN	No	99.55%	99.97%	99.49%	99.73%	0.114%	0.506%
	BiLSTM+LSTM	No	99.65%	99.98%	99.60%	99.79%	0.068%	0.399%
	CNN	***YES***	***99.80%***	***99.98%***	***99.78%***	***99.88%***	***0.085%***	***0.219%***
	BiLSTM+LSTM	***YES***	***99.76%***	***99.99%***	***99.73%***	***99.86%***	***0.046%***	***0.267%***
UNSW-NB15	CNN	No	99.67%	99.63%	99.87%	99.75%	0.771%	0.122%
	BiLSTM+LSTM	No	99.89%	99.98%	99.86%	99.92%	0.035%	0.133%
	CNN	***YES***	***99.77%***	***99.74%***	***99.92%***	***99.83%***	***0.55%***	***0.078%***
	BiLSTM+LSTM	***YES***	***99.98%***	***99.99%***	***99.97%***	***99.98%***	***0.016%***	***0.020%***
CICIDS2017	CNN	No	98.76%	98.10%	99.45%	98.77%	1.918%	0.541%
	BiLSTM+LSTM	No	99.39%	99.33%	99.44%	99.39%	0.663%	0.551%
	CNN	***YES***	***99.76%***	***99.60%***	***99.92%***	***99.76%***	***0.391%***	***0.070%***
	BiLSTM+LSTM	***YES***	***99.88%***	***99.86%***	***99.91%***	***99.88%***	***0.137%***	***0.089%***

expertise integration, thereby bridging the gap between automated detection systems and human domain knowledge. The accuracy of the decision-making process can be enhanced by directing uncertain traffic to human experts for analysis. The Simulation results underscore the considerable reduction in false alarms through the implementation of human-in-the-loop techniques, as measured by metrics such as FPR_b and FNR_b. Across six clustering scenarios and six H-FAR thresholds, three benchmark datasets and two customized deep learning models were used.

The comparison with previous works, shown in Table 8, demonstrated the superiority of the proposed framework over most of them. In fact, with the UNSW-NB15 dataset, the results showed that the proposed DL models under the human-in-the-loop technique outperformed all previous works ([26], [45], [46], [47]) in terms of accuracy, precision, and FPR. A model proposed by Kasongo [46] (XGboost-Simple-RNN) demonstrated a lower training time of 68.37s on the same dataset, but its accuracy (87.07%) is much lower than our models, 99.77% and 99.98% for the proposed CNN and BiLSTM+LSTM under human expertise, respectively.

The results with the CICDDoS2019 dataset showed the superiority of the proposed models under human expertise over most previous works ([20], [25], [44]) in terms of accuracy, precision, and FPR with 99.80%, 99.98%, and 0.085% for the CNN and 99.76%, 99.99%, and 0.046% for the BiLSTM+LSTM, respectively. However, a DFNN model developed by the authors in [43] demonstrated a slight superiority in accuracy (99.94%) but a lower precision (99.95%).

The results with the CICIDS2017 dataset also proved the efficiency of the proposed framework in intrusion detection and false alarm mitigation. The proposed models under human expertise outperformed the recent works of [25], [26], and [42] in terms of accuracy and precision with 99.76% and 99.60% for the CNN and 99.88% and 99.86% for the BiLSTM+LSTM, respectively. However, the authors of [28] and [30] developed two models that slightly outperformed our models in terms of accuracy and precision, with 99.96% and 99.96% accuracy and precision for the first model and

TABLE 8. Comparison with previous works.

Reference (year)	Model	Dataset	Accuracy	Precision	FPR	Training time	Inference time
Parveen et al. [17] (2023)	CNN+BiLSTM	NSL-KDD	99.308%	-	0.23%	-	-
Li et al. [18] (2020)	Multi-CNN	NSL-KDD (Detest^{+}; and KDDTest^{-21})	86.95%; 76.67%	69.47%; 59.73%	13.448%; 52.695%	-	-
Neto et al. [20] (2022)	Feedforward NN	CICDDoS2019	84.8%	-	-	-	-
Ali et al. [21] (2023)	Weighted FL based ANN	CAIDA	98.85%	98.13%	2.2%	-	-
Al-Haija and Zein-sabatto [24] (2020)	CNN	NSL-KDD	99.3%	99.04%	1.28%	-	-
Xu et al. [25] (2023)	CNN-BiLSTM-Attention	NSL-KDD; CICIDS2018; CICIDS2017; CICDDoS2019	93.26%; 88.27%; 90.31%; 93.26%	95.17%; 91.54%; 93.28%; 94.17%	<= **7.53%**	-	-
Ravi et al. [26] (2022)	RNN-LSTM-GRU	UNSW-NB15; CICIDS2017; KDD-Cup-1999; WSN-DS	99%; 99%; 99%; 98%	99%; 99%; 97%; 96%	0.458%; 1.184%; 1.142%; 0.589%	-	-
Hnamte and Hussain [28] (2023)	DNN	CICIDS2017; CICIDS2018 (LIOS_HOIS); ISCX 2012	99.80%; 100%; 99.76%	99.80%; 100%; 99.76%	- ; 0.001%; 5.175% (based on confusion matrix)	33 s; 13 s; 23 s	17.41 s; 29.05 s; 9.04 s
	DCNN	CICIDS2017; CICIDS2018 (LIOS_HOIS); ISCX 2012	99.96%; 100%; 99.79%	99.96%; 100%; 99.79%	- ; 0.00%; 11.233% (based on confusion matrix)	40 s; 15 s; 26 s	19.50 s; 29.36 s; 9.91 s
Hnamte and Hussain [29] (2023)	DCNNBiLSTM	CICIDS2018; Edge_IIoT	100%; 99.62%	-	0.00%; -	3245 s; 8421 s	1712.03 s; 4177.53 s
Hnamte et al. [30] (2023)	LSTM-AE	CICIDS2017; CICIDS2018	99.99%; 99.10%	99.99%; 99.07%	-	184 s; 462 s	53.66 s; 128.24 s
Saud and Salim [40] (2023)	MLP-PB	*CSE-CICIDS2018*	98.97%	99.98%	0.13%	-	-
	MLP-PSO	*CSE-CICIDS2018*	96.25%	96.80%	6.48%	-	-
WU et al. [41] (2023)	DNN	CICIDS2018	97%	-	-	-	-
Chanu et al. [42] (2023)	MLP-GA	CICDDoS2017	98.8%	-	-	45.8 s	1.52 s
Garcia and	DFNN	CICDDoS2019	99.94%	99.95%	-	-	-

TABLE 8. *(Continued.)* **Comparison with previous works.**

Blandon [43] (2022)							
Kumar et al. [44] (2023)	LSTM	CICDDoS2019	98%	98%	-	-	-
Yin et al. [45] (2023)	MLP	UNSW-NB15	84.24%	83.60%	4.03%	-	-
Kasongo [46] (2022)	XGboost-Simple-RNN	UNSW-NB15; NSL-KDD	87.07%; 83.70%	-	-	68.37 s; 78.19 s	-
	XGboost-LSTM	UNSW-NB15; NSL-KDD	85.08%; 88.13%	-	-	380.46 s; 225.46 s	-
	XGboost-GRU	UNSW-NB15; NSL-KDD	88.42%; 84.66%	-	-	135.66 s; 161.00 s	-
Sharma et al. [47] (2024)	1D-CNN	UNSW-NB15	80%	48%	-	-	-
	2D-CNN	UNSW-NB15	81%	57%	-	-	-
This Paper	**CNN + Human-in-the-loop technique**	**CICDDoS2019; UNSW-NB15; CICIDS2017**	**99.80%; 99.77%; 99.76%**	**99.98%; 99.74%; 99.60%**	**0.085%; 0.55%; 0.391%**	**190.9 s; 105.75 s; 1260.25 s**	**26.91 s; 14.42 s; 15.75 s**
	BiLSTM+LSTM + Human-in-the-loop technique	**CICDDoS2019; UNSW-NB15; CICIDS2017**	**99.76%; 99.98%; 99.88%**	**99.99%; 99.99%; 99.86%**	**0.046%; 0.016 %; 0.137%**	**184.09 s; 437.06 s; 1386.87 s**	**26.21 s; 30.90 s; 19.14 s**

99.99% and 99.99%, respectively, for the second model. They also experienced a very fast training time using a GPU, but the inference time of the proposed models was lower than theirs. GPUs are known to be boosting tools for machine or deep learning model training processes compared to CPU.

The results demonstrate how well the proposed probabilistic clustering can enable human-in-the-loop expertise and help reduce the false alarms of DL classifiers, thus improving the overall performance. The comparison with previous works proved the high competitiveness of the proposed framework with previous studies.

The unique advantage of the proposed probabilistic clustering concept integrated into the framework is that it can be applied to any deep-learning-based IDS classifier that is not 100% accurate. This means that if the DL model generates false alarms, this technique can be used to detect traffic that has a high probability of being misclassified by the DL model and redirect it for domain expert analysis. Theoretically, the framework will always help in false alarm mitigation in an IDS if there is generation of false alarms and if the human expert is good enough to distinguish normal traffic from malicious traffic. This is a good start for human-machine collaboration in intrusion detection with the rise of deep learning technologies in cybersecurity.

E. LIMITATIONS

Although this study adopted a rigorous approach, it had some limitations. The research unfolds within a simulated environment wherein the role of a human expert is emulated through a Python script. The use of established benchmark datasets facilitated the precise classification of redirected traffic, creating an idealized scenario in which a human expert proficiently distinguished between malicious and normal network traffic. However, it is imperative to acknowledge that real-world circumstances may not guarantee the same level of efficiency and accuracy among experts. In practice, human judgment can exhibit variances that impact outcomes.

Despite the identified limitations, the proposed approach remains a valid and promising concept for enhancing intrusion detection systems through human-machine collaboration with deep learning.

VI. CONCLUSION

Deep learning-based intrusion detection system classifiers are emerging as a promising approach to network security. However, some still exhibit high false alarm rates (FARs) in some cases. Although some researchers have developed models with exceptionally low FARs, the validity of this metric is undermined by the dataset size used to evaluate the model. A very low FAR can correspond to thousands of traffic events that are misclassified as false alarms.

This study proposed a human-machine framework that utilizes a probabilistic clustering technique to identify network traffic that is highly likely to be misclassified by the deep learning IDS model and then route it to human experts for further analysis and accurate decision making.

In addition, we introduced a next-generation firewall (NGFW) as a complementary tool for human experts to effectively handle human-classified data types and prevent redundant traffic analysis, thereby creating a sustainable and evolving solution. The simulation results demonstrated the effectiveness of the concept across three datasets, namely CICDDoS2019, UNSW-NB15, and CICIDS2017, as well as two tailor-made deep learning models, CNN, and BiLSTM+LSTM. The results showed an accuracy ranging from 99.76% to 99.98%, a false-positive rate (FPR) ranging from 0.016% to 0.55%, and a false-negative rate (FNR) ranging from 0.020% to 0.267%. By incorporating the probabilistic clustering method for uncertain traffic redirection, we achieved a significant decrease in the FPR and FNR and an improvement in the performance of the deep learning model. The strength of our approach lies in its ability to enhance the efficiency of the most effective deep-learning-based intrusion detection systems (IDS) in the literature, which are not 100% accurate, by integrating human expertise for practical applications.

The probabilistic clustering technique was effective in this study when used with the designed models, but the results depended entirely on the deep learning classifiers with which it was paired. Some deep learning models may allow for more human–expert intervention through traffic redirection than others. Consequently, future research should investigate how the proposed probabilistic clustering can enhance the performance of different DL classifiers with H-FAR.

AUTHORS CONTRIBUTION

The authors have contributed to this work as follows: Abdoul-Aziz Maiga: Conceptualization, methodology, software, and writing of the original draft. Edwin Ataro: Validation, formal analysis, supervision, resources, writing—review and editing, and funding acquisition. Stanley Githinji: Validation, formal analysis, conceptualization improvement, supervision, and writing—Review and editing.

All Authors have read and agreed to the published version of the manuscript.

DECLARATION OF COMPETING INTEREST

The authors declare no competing financial interests or personal relationships that could influence the work reported in this study.

REFERENCES

[1] H. Mohammadian, A. A. Ghorbani, and A. H. Lashkari, "A gradient-based approach for adversarial attack on deep learning-based network intrusion detection systems," *Appl. Soft Comput.*, vol. 137, Apr. 2023, Art. no. 110173.

[2] A. Pinto, L.-C. Herrera, Y. Donoso, and J. A. Gutierrez, "Survey on intrusion detection systems based on machine learning techniques for the protection of critical infrastructure," *Sensors*, vol. 23, no. 5, p. 2415, Jan. 2023, doi: 10.3390/s23052415.

[3] A. Sharma, B. B. Gupta, A. K. Singh, and V. K. Saraswat, "Advanced persistent threats (APT): Evolution, anatomy, attribution and countermeasures," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 7, pp. 9355–9381, Jul. 2023, doi: 10.1007/s12652-023-04603-y.

[4] M. Imran, H. U. R. Siddiqui, A. Raza, M. A. Raza, F. Rustam, and I. Ashraf, "A performance overview of machine learning-based defense strategies for advanced persistent threats in industrial control systems," *Comput. Secur.*, vol. 134, Nov. 2023, Art. no. 103445, doi: 10.1016/j.cose.2023.103445.

[5] A. Chakraborty, A. Biswas, and A. K. Khan, "Artificial intelligence for cybersecurity: Threats, attacks and mitigation," in *Artificial Intelligence for Societal Issues*, A. Biswas, V. B. Semwal, and D. Singh, Eds. Cham, Switzerland: Springer, 2023, pp. 3–25, doi: 10.1007/978-3-031-12419-8_1.

[6] H. Hindy, R. Atkinson, C. Tachtatzis, J.-N. Colin, E. Bayne, and X. Bellekens, "Utilising deep learning techniques for effective zero-day attack detection," *Electronics*, vol. 9, no. 10, p. 1684, Oct. 2020, doi: 10.3390/electronics9101684.

[7] H. Mohammadian, A. A. Ghorbani, and A. H. Lashkari, "A gradient-based approach for adversarial attack on deep learning-based network intrusion detection systems," *Appl. Soft Comput.*, vol. 137, Apr. 2023, Art. no. 110173, doi: 10.1016/j.asoc.2023.110173.

[8] H. S. Milan and K. Singh, "Reducing false alarms in intrusion detection systems—A survey," *Int. Res. J. Eng. Technol.*, vol. 2395, p. 0056, Jan. 2018.

[9] M. Gamal, H. Abbas, and R. Sadek, "Hybrid approach for improving intrusion detection based on deep learning and machine learning techniques," in *Proc. Int. Conf. Artif. Intell. Comput. Vis. (AICV)* (Advances in Intelligent Systems and Computing), A.-E. Hassanien, A. T. Azar, T. Gaber, D. Oliva, and F. M. Tolba, Eds. Cham, Switzerland: Springer, 2020, pp. 225–236, doi: 10.1007/978-3-030-44289-7_22.

[10] W. Alhakami, A. ALharbi, S. Bourouis, R. Alroobaea, and N. Bouguila, "Network anomaly intrusion detection using a nonparametric Bayesian approach and feature selection," *IEEE Access*, vol. 7, pp. 52181–52190, 2019, doi: 10.1109/ACCESS.2019.2912115.

[11] Z. Zhang, H. Ning, F. Shi, F. Farha, Y. Xu, J. Xu, F. Zhang, and K.-K.-R. Choo, "Artificial intelligence in cyber security: Research advances, challenges, and opportunities," *Artif. Intell. Rev.*, vol. 55, no. 2, pp. 1029–1053, Feb. 2022, doi: 10.1007/s10462-021-09976-0.

[12] K. Neupane, R. Haddad, and L. Chen, "Next generation firewall for network security: A survey," in *Proc. SoutheastCon*, Apr. 2018, pp. 1–6, doi: 10.1109/SECON.2018.8478973.

[13] G. R. Santhanam, B. Holland, S. Kothari, and N. Ranade, "Human-on-the-loop automation for detecting software side-channel vulnerabilities," in *Information Systems Security* (Lecture Notes in Computer Science), R. K. Shyamasundar, V. Singh, and J. Vaidya, Eds. Cham, Switzerland: Springer, 2017, pp. 209–230, doi: 10.1007/978-3-319-72598-7_13.

[14] A. Awajan, "A novel deep learning-based intrusion detection system for IoT networks," *Computers*, vol. 12, no. 2, p. 34, Feb. 2023, doi: 10.3390/computers12020034.

[15] G. Sai Chaitanya Kumar, R. Kiran Kumar, K. Parish Venkata Kumar, N. Raghavendra Sai, and M. Brahmaiah, "Deep residual convolutional neural network: An efficient technique for intrusion detection system," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 121912, doi: 10.1016/j.eswa.2023.121912.

[16] H. Shah, D. Shah, N. K. Jadav, R. Gupta, S. Tanwar, O. Alfarraj, A. Tolba, M. S. Raboaca, and V. Marina, "Deep learning-based malicious smart contract and intrusion detection system for IoT environment," *Mathematics*, vol. 11, no. 2, p. 418, Jan. 2023, doi: 10.3390/math11020418.

[17] S. P. Praveen, S. Sindhura, P. N. Srinivasu, and S. Ahmed, "Combining CNNs and bi-LSTMs for enhanced network intrusion detection: A deep learning approach," in *Proc. 3rd Int. Conf. Comput. Inf. Technol. (ICCIT)*, Sep. 2023, pp. 261–268, doi: 10.1109/iccit58132.2023.10273871.

[18] Y. Li, Y. Xu, Z. Liu, H. Hou, Y. Zheng, Y. Xin, Y. Zhao, and L. Cui, "Robust detection for network intrusion of industrial IoT based on multi-CNN fusion," *Measurement*, vol. 154, Mar. 2020, Art. no. 107450, doi: 10.1016/j.measurement.2019.107450.

[19] F. L. de Caldas Filho, S. C. M. Soares, E. Oroski, R. de Oliveira Albuquerque, R. Z. A. da Mata, F. L. L. de Mendonça, and R. T. de Sousa Jùnior, "BotNet detection and mitigation model for IoT networks using federated learning," *Sensors*, vol. 23, no. 14, p. 6305, Jul. 2023, doi: 10.3390/s23146305.

[20] E. C. P. Neto, S. Dadkhah, and A. A. Ghorbani, "Collaborative DDoS detection in distributed multi-tenant IoT using federated learning," in *Proc. 19th Annu. Int. Conf. Privacy, Secur. Trust (PST)*, Aug. 2022, pp. 1–10, doi: 10.1109/PST55820.2022.9851984.

[21] M. N. Ali, M. Imran, M. S. ud Din, and B.-S. Kim, "Low rate DDoS detection using weighted federated learning in SDN control plane in IoT network," *Appl. Sci.*, vol. 13, no. 3, p. 1431, Jan. 2023, doi: 10.3390/app13031431.
[22] B. Bousalem, V. F. Silva, R. Langar, and S. Cherrier, "Deep learning-based approach for DDoS attacks detection and mitigation in 5G and beyond mobile networks," in *Proc. IEEE 8th Int. Conf. Netw. Softwarization (NetSoft)*, Jun. 2022, pp. 228–230, doi: 10.1109/NetSoft54395.2022.9844053.
[23] P. Wu and H. Guo, "LuNet: A deep neural network for network intrusion detection," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2019, pp. 617–624, doi: 10.1109/SSCI44817.2019.9003126.
[24] Q. Abu Al-Haija and S. Zein-Sabatto, "An efficient deep-learning-based detection and classification system for cyber-attacks in IoT communication networks," *Electronics*, vol. 9, no. 12, p. 2152, Dec. 2020, doi: 10.3390/electronics9122152.
[25] H. Xu, L. Sun, G. Fan, W. Li, and G. Kuang, "A hierarchical intrusion detection model combining multiple deep learning models with attention mechanism," *IEEE Access*, vol. 11, pp. 66212–66226, 2023, doi: 10.1109/ACCESS.2023.3290613.
[26] V. Ravi, R. Chaganti, and M. Alazab, "Recurrent deep learning-based feature fusion ensemble meta-classifier approach for intelligent network intrusion detection system," *Comput. Electr. Eng.*, vol. 102, Sep. 2022, Art. no. 108156, doi: 10.1016/j.compeleceng.2022.108156.
[27] S. Gurung, M. K. Ghose, and A. Subedi, "Deep learning approach on network intrusion detection system using NSL-KDD dataset," *Int. J. Comput. Netw. Inf. Secur.*, vol. 11, no. 3, pp. 8–14, Mar. 2019.
[28] V. Hnamte and J. Hussain, "Dependable intrusion detection system using deep convolutional neural network: A novel framework and performance evaluation approach," *Telematics Informat. Rep.*, vol. 11, Sep. 2023, Art. no. 100077, doi: 10.1016/j.teler.2023.100077.
[29] V. Hnamte and J. Hussain, "DCNNBiLSTM: An efficient hybrid deep learning-based intrusion detection system," *Telematics Informat. Rep.*, vol. 10, Jun. 2023, Art. no. 100053, doi: 10.1016/j.teler.2023.100053.
[30] V. Hnamte, H. Nhung-Nguyen, J. Hussain, and Y. Hwa-Kim, "A novel two-stage deep learning model for network intrusion detection: LSTM-AE," *IEEE Access*, vol. 11, pp. 37131–37148, 2023, doi: 10.1109/ACCESS.2023.3266979.
[31] M. S. Islam, M. A. Uddin, D. M. S. Ahmed, and G. Moazzam, "Analysis and evaluation of network and application security based on next generation firewall," *Int. J. Comput. Digit. Syst.*, vol. 13, no. 1, pp. 193–202, Jan. 2023.
[32] J. Liang and Y. Kim, "Evolution of firewalls: Toward securer network using next generation firewall," in *Proc. IEEE 12th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2022, pp. 0752–0759, doi: 10.1109/CCWC54503.2022.9720435.
[33] V. P. Sitorus and S. L. Siregar, "Nunukan state court's computer network security improvement using centralized next-generation firewall," *Budapest Int. Res. Critics Inst.-J.*, vol. 5, no. 2, pp. 10102–10113, 2022.
[34] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, and J. Santamaría, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.
[35] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of LSTM and BiLSTM in forecasting time series," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 3285–3292, doi: 10.1109/BigData47090.2019.9005997.
[36] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2019, pp. 1–8, doi: 10.1109/CCST.2019.8888419.
[37] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Proc. Mil. Commun. Inf. Syst. Conf. (MilCIS)*, Nov. 2015, pp. 1–6, doi: 10.1109/MilCIS.2015.7348942.
[38] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "A detailed analysis of the CICIDS2017 data set," in *Information Systems Security and Privacy* (Communications in Computer and Information Science), P. Mori, S. Furnell, and O. Camp, Eds. Cham, Switzerland: Springer, 2019, pp. 172–188, doi: 10.1007/978-3-030-25109-3_9.
[39] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*.
[40] S. Alzughaibi and S. El Khediri, "A cloud intrusion detection systems based on DNN using backpropagation and PSO on the CSE-CIC-IDS2018 dataset," *Appl. Sci.*, vol. 13, no. 4, p. 2276, Feb. 2023, doi: 10.3390/app13042276.
[41] C. Wu and S. Chen, "A heuristic intrusion detection approach using deep learning model," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2023, pp. 438–442, doi: 10.1109/ICOIN56518.2023.10049024.
[42] U. S. Chanu, K. J. Singh, and Y. J. Chanu, "A dynamic feature selection technique to detect DDoS attack," *J. Inf. Secur. Appl.*, vol. 74, May 2023, Art. no. 103445, doi: 10.1016/j.jisa.2023.103445.
[43] J. F. Cañola Garcia and G. E. T. Blandon, "A deep learning-based intrusion detection and preventation system for detecting and preventing Denial-of-Service attacks," *IEEE Access*, vol. 10, pp. 83043–83060, 2022, doi: 10.1109/ACCESS.2022.3196642.
[44] D. Kumar, R. K. Pateriya, R. K. Gupta, V. Dehalwar, and A. Sharma, "DDoS detection using deep learning," *Proc. Comput. Sci.*, vol. 218, pp. 2420–2429, Mar. 2023, doi: 10.1016/j.procs.2023.01.217.
[45] Y. Yin, J. Jang-Jaccard, W. Xu, A. Singh, J. Zhu, F. Sabrina, and J. Kwak, "IGRF-RFE: A hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset," *J. Big Data*, vol. 10, no. 1, p. 15, Feb. 2023, doi: 10.1186/s40537-023-00694-8.
[46] S. M. Kasongo, "A deep learning technique for intrusion detection system using a recurrent neural networks based framework," *Comput. Commun.*, vol. 199, pp. 113–125, Feb. 2023, doi: 10.1016/j.comcom.2022.12.010.
[47] B. Sharma, L. Sharma, C. Lal, and S. Roy, "Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 121751, doi: 10.1016/j.eswa.2023.121751.

ABDOUL-AZIZ MAIGA (Graduate Student Member, IEEE) received the bachelor's and master's degrees in networks and telecommunication from Saad Dahlab Blida 1 University, Algeria, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Pan African University Institute of Basic Sciences, Technology, and Innovation (PAUSTI), Nairobi, Kenya. Before pursuing the Ph.D. degree, he was a Mobile Core Network Engineer with HUAWEI Burkina Faso. His research interest includes the application of artificial intelligence to 5G virtual network function (VNF) security.

EDWIN ATARO received the bachelor's degree in electrical and communication engineering from Moi University, Eldoret, Kenya, in 1989, and the Master of Science (M.Sc.) and Doctor of Engineering (Dr.-Ing.) degrees in electrical communication engineering from the University of Kassel, Germany, in 2000 and 2005, respectively. He is currently the Executive Dean of the Faculty of Engineering and the Built Environment, Technical University of Kenya (TUK), Nairobi, Kenya. His research interests include optical communications systems, high speed networks, and renewable energy.

STANLEY GITHINJI is currently an Assistant Professor in information security and forensics with United States International University-Africa. He has vast knowledge and experience in applied cryptography, computer forensics enterprise risk management of information systems, information security audit, IT governance, and implementation of ISO 27001 standards. His research interests include information security, distributed systems, and integration of technology for improving business process. He has consulted and trained professionals in information technology and cybersecurity, both in private and public sector.