International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

Empowering Cybersecurity: An Adaptive Approach with Hybrid Machine Learning for Anomaly Detection

Aparna N.[1*], Dr. Chetana Tukkoji[2]

Submitted: 02/05/2024 **Revised**: 15/06/2024 **Accepted**: 22/06/2024

Abstract: The fast development in the use of computer networks raises concerns about network availability, integrity, and confidentiality. This requires network managers to use various types of intrusion detection systems (IDS) to monitor network traffic for unauthorized and malicious activity. In this research, a hybrid machine learning-based framework is introduced for anomaly detection in the system. The suggested hybrid machine learning model, consisting of C4.5, a convolutional neural network (CNN), and a random forest (RF), was applied to the Bot-IoT dataset. The proposed hybrid intrusion detection framework achieved 99.8% accuracy, 96.4% precision, 100% recall, and an F1 score of 98.1% in the classification of malicious activities. This research suggests a more reliable and comprehensive approach to managing the ever-changing landscape of cyber threats by demonstrating the exceptional performance of the proposed framework.

Keywords: *Intrusion detection, Malicious activity, Anomalies, Machine Learning, Cyber security, Deep Learning, Network security.*

1. Introduction

Cyber security is becoming more crucial as the role of networks expands in contemporary society. Cyber security is based on three key things, which are anti-virus programs, firewalls and IDS (Intrusion Detection System). Through these means, hackers are prevented from intruding into networks. One of the detection systems is IDS, which is critical in maintaining network security by watching over every device and program settings linked to it. In 1980, Jim Anderson was the first to introduce the concept of IDS [1]. Since then, several IDS systems have been developed and improved to fulfill the needs of network security [2]. However, due to the tremendous advances in technology over the past decade, both the size of networks and the variety of applications processed by network nodes have dramatically increased.

As a result, a substantial amount of critical information is generated and transmitted across multiple network nodes. Safeguarding the data and network nodes has become a challenging task due to the increasing occurrence of novel attacks that are either generated by modifying current attacks or introducing entirely new ones. Every network node possesses fundamental importance that an attacker may potentially use. For example, the data node might have considerable significance for a business. The potential consequences for that business due to an attack on the node's data may pose a threat in terms of both reputation and cash. Current intrusion detection systems (IDSs) have been ineffective in accurately detecting a diverse range of threats, including zero-day attacks, while also effectively reducing false alarm rates (FAR) [3].

Therefore, there will always be a need for a network-based detection system (NIDS) that is both effective and economical in its protection of the network [4].

The three main categories of ID systems could be broken down into subcategories depending on the methods used for detection. The first group includes systems like the misuse detection method; this group is called Signature-Based Systems (SBS). The second kind of system is the Anomaly-Based System (ABS), sometimes shortened to "anomaly." The final kind of protocol analysis detection is the stateful type [5]. SBS uses a pattern-matching approach, comparing attack signatures from a database with those in the observed data. The moment a match is found, an alert sounds. In addition to the fact that SBS may identify potential threats by referencing previously acquired information, the misuse detection approach is also considered to be a knowledge-based method. The misuse detection method has a high rate of accuracy and a low FAR, but it cannot recognize novel assaults.

The behavior-based ID system, which is commonly referred to as ABS, can identify intrusion by comparing typical behavior with that which is out of the ordinary. Using signature and anomaly-based ID approaches, the stateful protocol ID approach analyzes known harmful actions and determines the eccentricity of protocol activity. The three architectural subtypes in an ID system include a hybrid approach, HIDS (host-based detection systems) and NIDS (network-based detection systems) [6,7].

[1*] *CSE Department, Research Scholar, Gitam School of Technology, Bengaluru, 561203 , India) Email address: aparnaapps51@gmail.com*
[2] *CSE Department, Assistant Professor, Gitam School of Technology, Bengaluru, 561203 , India)*

In times when cyber threats are evolving at an alarming rate, IDS cannot be underestimated for the protection of sensitive digital assets. Unfortunately, current IDS solutions hardly cope with the complexity and diversity of modern attacks. The principal aim of this investigation is to suggest a new hybrid framework that builds on a learning approach in order to simplify IDS's complexities and make it more effective. The major challenge is to integrate ML approaches like deep learning and reinforcement learning, among others, into traditional rule-based methods to develop a more adaptive and intelligent IDS. The framework should have the capability to identify an extensive variety of intrusions, including novel threats, while minimizing false positives and optimizing resource utilization. This research aims to balance detection accuracy, scalability, and simplicity to address the growing need for strong, user-friendly IDS in the age of cyberspace terrorism. Here's a list of objectives given below:

- To improve the accuracy of intrusion detection by integrating hybrid learning techniques.
- To develop an ML that can handle large-scale networks and adapt to evolving attack patterns.
- To Optimize resource usage to ensure efficient intrusion detection without overwhelming system resources.
- The system should be implemented in such a way that it provides a user-friendly interface for security analysts, making it easy for them to interact with the system and ensuring that they are effective.

1.1 Intrusion Detection System

IDSs are hardware or software systems that automate the technique of monitoring and evaluating events occurring within a computer system or network to identify indications of security issues [8]. Most commonly, a traditional firewall cannot detect different types of attacks on network traffic; therefore, intrusion detection systems (IDSs) are required [9]. A sample structure of the IDS systems is presented in Figure 1.

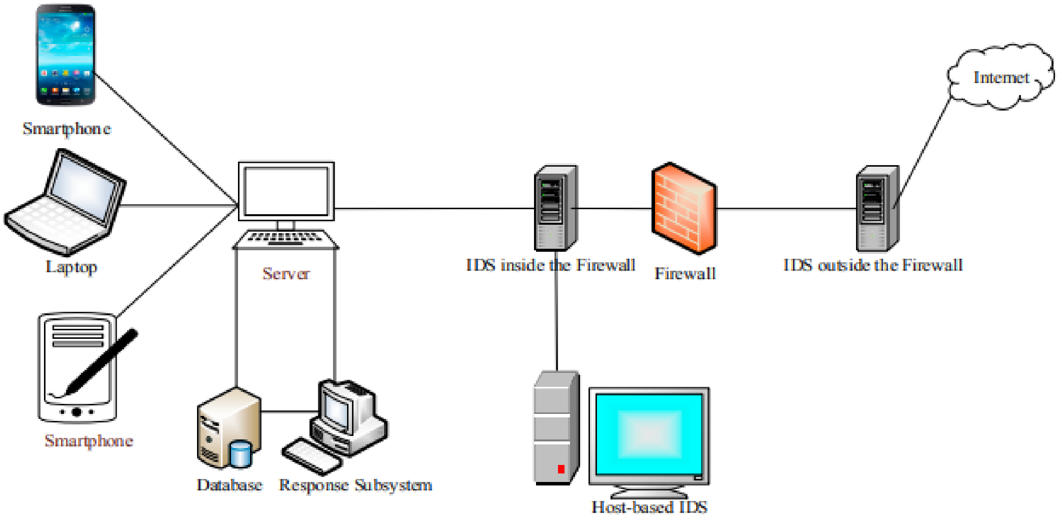


Fig 1: Structure of Intrusion Detection System [10].

1.1.1. Classification of Intrusion Detection System

Several distinct categories of intrusion detection systems are given below:

a.) Anomaly-Based Intrusion Detection System

Anomaly detection is categorized into three main components: supervised anomaly detection, semi-supervised anomaly detection, and managed anomaly detection. The supervised anomaly detection method aims to construct the model by generating anomalous and normal records separately. Conventional data is utilized in the construction of models using semi-supervised anomaly detection techniques [11]. Semi-supervised detection frequently yields a substantial number of false positives and requires a labeled database to be effective. To address this problem and detect any additional irregularities, an unsupervised anomaly detection system is deployed. Figure 2 displays the architectural design of the anomaly-based intrusion detection system [12-14].

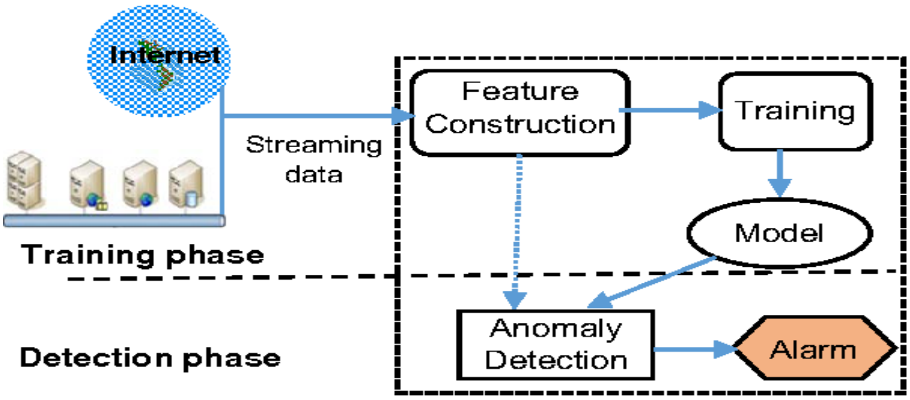


Fig 2. Architecture of Anomaly Based detection [13]

b.) Network-Based Intrusion Detection System

Network-based intrusion detection systems (IDSs) adopt an alternative standpoint, redirecting their attention from the communication infrastructure to the computational infrastructure (hospitable systems and their operating systems). These systems access the network to obtain security-related information [15].

NIDS systems are meant to keep a system safe against network-based attacks by tracking and analyzing data on the network. A NIDS examines every incoming packet for potentially malicious patterns [16]. Figure 3 presents an ordinary NIDS architecture.

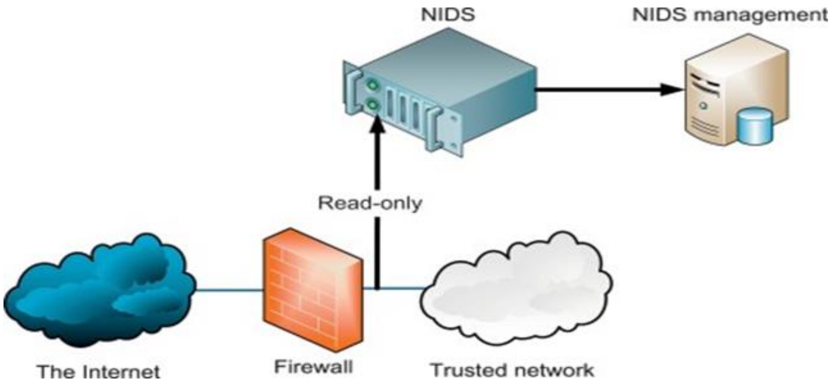


Fig 3. The architecture of a Network-based Intrusion detection system [17].

c.) Host-Based Intrusion Detection System

Host Based Intrusion Detection Systems (HIDS) are software applications that operate on a designated machine, referred to as the host as it can secure the whole system and provide notification in the event of compromise [18]. HIDS ascertains whether a system has been compromised and issues appropriate alerts to administrators [16]. HIDS monitors host activities, including system and shell records, to identify unauthorized actions. To identify intrusions, HIDS may apply a range of data mining techniques, such as artificial neural networks, for monitoring data hosting. With a system-invoked method, HIDS detects anomalous system call sequences by monitoring real-time system call traces. The working of the host intrusion detection system is depicted in Figure 4 [19-21].

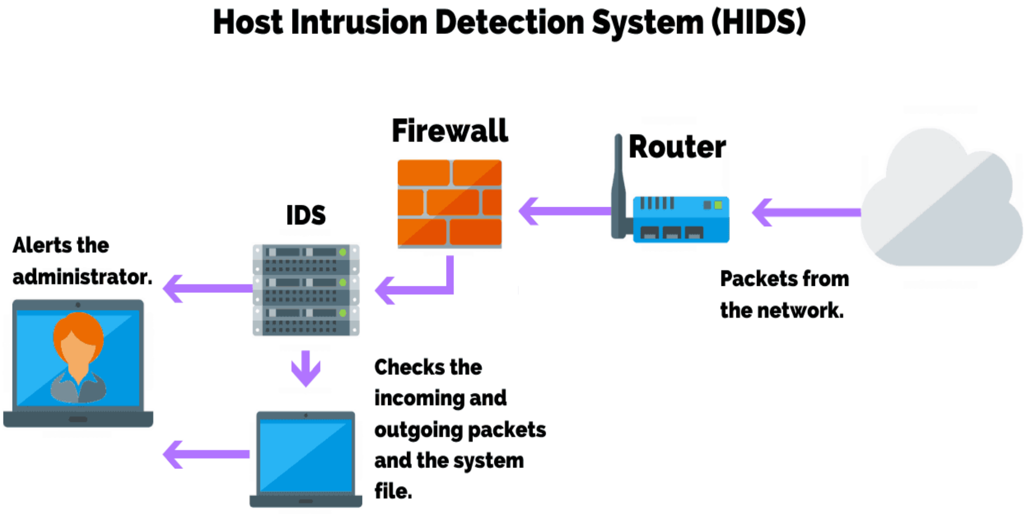

Host Intrusion Detection System (HIDS)
Firewall
Router
IDS
Alerts the administrator.
Packets from the network.
Checks the incoming and outgoing packets and the system file.

Fig 4. Working of HIDS system [20].

d.) Signature-Based Intrusion Detection System

Signature-based techniques are utilized to observe and contrast network connections or packets against pre-established patterns referred to as signatures. The processing of audit data using this method is straightforward and effective. To identify newly identified anomalies that are not explicitly specified in the signatures, signature-based methods are inadequate; consequently, system administrators are required to regularly update the signatures [21]. As shown in Figure 5, the traffic collector collects real-time traffic and restructures it for the signature-based intrusion detection system block, thereby serving as the central component of the system.

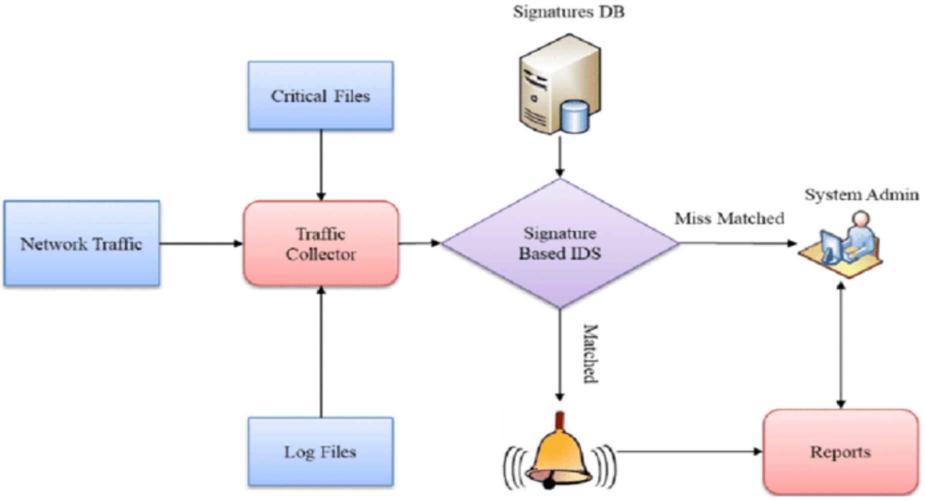


Fig 5. The architecture of signature-based intrusion detection system [22]

Signature-based detection involves the utilization of pre-established attack patterns represented as signatures, which are subsequently employed to identify network attacks. Network traffic is typically analyzed utilizing pre-established signatures, and the database is updated periodically. Source fire non-intrusive detection system (SNORT) serves as an example of a signature-based intrusion detection system [23].

The objective of this paper is to address the limitations of conventional AIDS (anomaly-based intrusion detection systems) and SIDS (signature intrusion detection systems) by combining their strengths. The fundamental goal of this study is to tackle the urgent requirement for efficient intrusion detection in IoT environments, where conventional methods may be inadequate due to the large quantity and variety of IoT devices. The proposed hybrid model, designed by a combination of random forest, C 4.5, and CNN classifiers for HIDS (host-based intrusion detection system), intends to offer a comprehensive solution capable of accurately detecting both known and new assaults while minimizing false positives. The assessment of the Bot-IoT dataset will offer valuable insights into the practical efficacy of the proposed HIDS, hence contributing to the progress of IoT security.

2. Related Work

This section includes the previous studies of several authors built on a hybrid learning-based framework designed to enhance IDS.

2.1 Machine Learning Based Hybrid Intrusion Detection Architecture.

Jadhav K. et al., (2023) [24] employed ML (machine learning) techniques for both network and host intrusion identification, and a heterogeneous extraction of attribute or feature and selection approach for IDS was developed. A large dataset of network logs was utilized to identify the intruder in a potentially dangerous system. In order to construct a reliable module, several different types of feature extraction have been used. Significant testing has been performed on three classifiers, including Recurrent Neural Networks (RNN), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) of various network log datasets for validation purposes. When compared to SVM and ANN, RNN's observational detection and classification accuracy are superior. Also, it helps reduce both the time complexity and error rates for all datasets.

Abbas Q. et al., (2023) [25] proposed a unique hybrid ensemble model based on the RF-RFE (random forest-recursive feature elimination) technique to enhance the accuracy of the prediction of IDS. The proposed ML ensemble approach outperforms DL with low computational cost and shorter training time. For the UNSW-NB15, CSE-CIC-IDS2018 and NSL-KDD datasets, the suggested ensemble ML approach achieves an overall accuracy of 98.53%, 99.9% and 99%, respectively. From these results, the indicated ensemble technique effectively enhances the operation of IDS.

Talukder A. et al., (2023) [26] suggested the latest hybrid model that uses deep learning and machine learning to boost both detection rates and model reliability. In addition, the author has compared the developed method with several deep learning and machine learning algorithms to identify an optimal one that can be integrated into the pipeline. Furthermore, the author selected this model by evaluating its performance through several benchmarks for network intrusion. Authors found this method works very well on two different datasets (CIC-MalMem-2022 and KDDCUP’99), obtaining 99.99% accuracy for one dataset and 100% for another with no Type-1, Type-2, or overfitting problems.

Saraladeve L. et al., (2023) [27] modified an existing hybrid IDS architecture to improve the identification and classification of threats. Specifically, they have incorporated the machine learning library Random Nearest Neighbor (RNN) for classification and used the metaheuristic technique Binary Enhanced-Whale Optimization Approach (BEWOA) for feature selection. When tested using the NSL-KDD dataset, the framework realized an f-measure of 99.22%, a precision of 99.64%, a specificity of 99.63% and an accuracy of 99.22%. On the UNSW-NB15 dataset, the framework obtained an f-measure of 99.07%, a precision of 99.24%, a specificity of 99.22% and a detection rate of 99.90%.

Balyan et al., (2022) [28] created a productive hybrid network-based intrusion detection system (HNIDS) model that utilizes the IRF (improved random forest) and EGA-PSO (enhanced genetic algorithm and particle swarm optimization) methods. In the preliminary stage, the HNIDS system applies hybrid EGA-PSO techniques to improve the quality of the minor data samples, resulting in a more balanced data set that allows more precise learning of small sample attributes. Subsequently, an IRF does the following: eliminates less important characteristics, integrates a decision tree list for each iterative procedure, monitors the classifier's performance and mitigates overfitting concerns. According to experimental results, 88.149% MCC accuracy and 98.979% BCC accuracy were achieved by the proposed HNIDS method for the NSL-KDD dataset. This is significantly better than other machine learning techniques like SVM, CART, NB LDA, RF, and LR.

Saba et al., (2022) [29] introduced a CNN-based technique for anomaly-based intrusion detection systems (IDS) that leverages the capabilities of the Internet of Things (IoT) to efficiently analyze all IoT traffic. The model under consideration exhibits the capability to identify potential intrusions and anomalous traffic patterns. The model underwent training and evaluation utilizing the BoT-IoT dataset and the NID Dataset, attaining respective accuracy rates of 99.51% and 92.85%.

Megantara A. et al., (2021) [30] proposed a hybrid ML approach that integrates feature selection with data reduction techniques, resulting in a better model. To achieve this, the author utilizes a decision tree that is based on feature significance and performs a recursive reduction of features until only the most important and relevant ones are obtained; it further uses the local outlier factor (LOF) to identify anomalies/outliers in the dataset. Experimental findings show that on the NSL-KDD database, our recommended approach outperforms several other previous research in terms of the detection rate of R2L (99.89%) and other attack types. This leads to a more robust performance compared with other alternatives. Binary classes have made the UNSW-NB15 dataset more challenging.

Ozer et al., (2021) [31] developed a novel method for identifying the most optimal and effective feature pairings of datasets to facilitate the construction of lightweight intrusion detection systems. To accomplish this goal, ten machine-learning algorithms were tested with the BoT-IoT (2018) dataset. The creators of the dataset recommend twelve of the ideal traits for this study, and from those twelve created sixty-six unique pairs. After that, all ten algorithms were trained using all twelve full features to create ten intrusion detection systems based on their full features. With this unique process, they found not only the best and lightest features but also lightweight intrusion detection systems with this approach. In fact, the most lightweight one had a detection accuracy of over 90%.

Khonde S. et al., (2020) [32] presented an advanced framework combining anomaly-based detection and signature-based detection. The framework aims to enhance the detection rate and reduce the false alarm rate. The framework employs different supervised learning algorithms and unsupervised learning techniques to analyze live internet traffic. It was validated using the Intrusion Detection Evaluation Dataset (CICIDS-17). It increases the anomaly-based detections by 2% and signature-based detections by 5%. It also reduces the false alarm rate by 0.

NG et al., (2020) [33] suggested a technique of VCDL (vector convolutional deep learning) for anomaly detection in fog. The findings of scientific experiments conducted on the Bot-IoT dataset from UNSW demonstrate that the proposed approach to distributed deep learning is more effective in handling big data than current centralized deep learning methods. Experimental results also indicate that the method is far superior in terms of accuracy, precision and recall when compared with current systems for detecting anomalies.

Ren et al., (2019) [34] designed an IDSOIDS intrusion detection system (IDS), which combined hybrid data optimization and included feature selection and data sampling. This model aimed at finding the best training dataset by feature selection, Isolation Forest (IF) for removing outliers and genetic algorithm (GA) to optimize the proportion of sampling in data sampling. The above issues are covered by solving them again for the best feature subset in this model using RF and GA. For this study, UNSW-NB15 was used as a reference dataset. This model is different from others because it can identify uncommon abnormal behaviors that are novel.

Cavusoglu et al., (2019) [35] proposed new hybrid layered IDS (intrusion detection system) proposed that combined machine learning and feature selection methods to improve attack detection. The developmental system starts by preprocessing the NSL-KDD dataset, which is further reduced using several feature selection algorithms. Two novel approaches have been suggested to accomplish this. Depending on the attack's type, suitable machine learning algorithms can be selected to produce a layered architecture. In performance evaluation, the author found that the suggested system had over 95% accuracy across all types of attacks.

Foroushani et al., (2018) [36] introduced a novel hybrid intrusion detection system that is dependent on K's nearest neighbor and decision trees for anomaly detection. The authors improve the proposed method by making information extraction from the NSL-KDD dataset more efficient using the feature selection technique. It has been experimentally shown that 99.6% accuracy, 0.2% false alarm rate and 99.7% positive detection rate can be achieved through this approach.

2.2 Neural Network Based Intrusion Detection System.

Qazi E. et al., (2023) [37] developed a new DL (deep learning) based IDS (intrusion detection system) and Convolutional Recurrent Neural Network (CRNN). It's designed to find threats in a network. With the system, there's an increase in effectiveness and prediction by using deep RNNs for feature extraction, while CNNs are applied for local feature synthesis using convolutions. The team evaluated the effectiveness of the method using publicly accessible benchmark CICIDS-2018 data. The HDLNIDS discovered malicious assaults with an average accuracy of 98.90%, outperforming all other IDS systems on the market.

Aldallal et al., (2022) [38] proposed a novel IDS system called Cu-LSTMGRU. The author explained that it's a combination of gated recurrent units and a modified long short-term memory computing unit. This system accurately differentiates between benign and malicious network flows. This current system is evaluated using the latest dataset, CICIDS2018. For computational efficiency improvement, this dataset undergoes optimization by the Pearson correlation feature selection algorithm. Various metrics are used in assessing the proposed model. According to the results, this model outperforms benchmarks by far, with up to a 12.045% gap. **Gamal M. et al., (2020) [39]** created a method that combines CNN with ML (SVM, K-Nearest Neighbor (KNN)). In this case, CNN is effectively utilized to extract necessary characteristics from the collected data. Then, the data was classified using ML. Combining the strengths of ML (high accuracy, Low false alarms) with DL (handles plenty of data, cuts down on dataset characteristics) is a smart way to get the most out of both. For this work, they utilized 10% of the KDDcup1999 dataset. The study findings indicated a 99.3% increase in detection accuracy and a 0.03% decrease in losses.

Naseer et al., (2018) [40] proposed a deep convolutional neural network (CNN) based on intrusion detection. The suggested IDS incorporates a deep Convolutional Neural Network (CNN) as its core is optimized through randomized search over the configuration space. GPUs are used for training and evaluation of the proposed approach on NSLKDD training and testing datasets. Comparatively, this study assesses the DCNN model's effectiveness among other classifiers with the help of common metrics used that include accuracy, MAP (Mean Average Precision), and AuC (area under the root-of-curve). The experimental results of the proposed IDS based on DCNN are encouraging enough to be deployed in real-world anomaly detection systems.

Dias et al., (2017) [41] introduced an intrusion detection system (IDS) technology based on artificial neural network (ANN) and KDDCUP'99 dataset that is far better than the rest. The experimental results compared with conventional methods clearly show that the suggested system can classify predefined classes of intrusion attacks with an overall accuracy of 99.9%, which is a highly encouraging result compared to others using ordinary approaches.

3. Research problem

Intrusion detection is an ongoing research field since the problem is constantly changing and influenced by important aspects that evolve over time. The primary obstacle in this domain is to develop a model capable of accurately forecasting malware by utilizing several parameters. Selecting efficient and crucial attributes for intrusion detection is a highly significant subject in the realm of information security. The improvement of intrusion detection systems is challenging due to the complex nature of the problem and the requirement for both exceptional precision and efficiency.

4. Proposed Methodology

The proposed methodology discussed all the steps and techniques used in this research to enhance the IDS system. In the research methodology, the presented method utilizes Principal Component Analysis (PCA) for feature extraction and SelectKBest with F regression techniques for feature selection, and then it utilizes a hybrid model of CNN, random forest, and C4.5 for anomaly detection.

4.1 Dataset Description

The Bot-IoT dataset is utilized for the assessment of the hybrid IDS that has been suggested. This research utilized the Bot-IoT dataset [42] to assess the performance of the suggested framework. This dataset consists of both regular IoT network traffic and a range of attack scenarios. The BoT-IoT collection comprises about 7.2 million records. These data are utilized to identify different types of assaults, such as Data exfiltration, DOS OS, and DDoS attacks, based on the protocol employed in the system. In this research, this dataset is selected because it accurately reflects the real-world IoT ecosystem context. This dataset consists of service scans, DoS, data exfiltration threats, DDoS, OS, and Keylogging. In this case, the data is preprocessed to find network-level patterns associated with different types of traffic emanating from devices. Later, these patterns are used to detect any malicious activities in the IoT infrastructure [43].

4.2 Technique Used

In this section, several methods are explained that are used in research methodology. This includes performing feature extraction using the PCA technique, feature selection via the Select K Best with F regression model and using classification techniques in the hybrid model.

4.2.1. PCA for feature extraction

Principal Component Analysis (PCA) [44-45] is an effective initial approach for identifying botnets in real-time log data. An uncomplicated method for accomplishing this operation might be outlined as follows:

1.) The streaming logs are divided into time windows, which are based on a defined interval.

2.) For each time frame, create a host-request matrix 'x' by converting the log entries within a given period. Hence, the integer value at the ith row and jth column denoted $X(i,j)$ is the rate at which host j makes request i.

3.) Perform Principal Component Analysis (PCA) on each time window. Apply PCA to the variable 'x' to determine if the principal weight surpasses a specific threshold. This approach can identify either individual bots generating a significant amount of traffic or a botnet where each bot may not generate many requests, but they exhibit correlation with one another and collectively generate a substantial volume of traffic [46].

4.2.2. SelectKBest with F Regression for feature selection

The purpose of this method is to increase the efficiency and pertinence of the feature set utilized in IoT networks for intrusion detection. A subset of the most informative features is chosen during this procedure from a BOT-IOT dataset that has been subjected to principal component analysis (PCA). By employing the f regression scoring function, the Select K Best algorithm evaluates the statistical correlation between every feature and the objective variable 'attack'. Following that, the k most important features are selected, with k set to 3 in this instance. Furthermore, index numbers and names have been obtained for these selected components (PCA features). As a result, in order to facilitate interpretation, the author synchronized the names of the chosen PCA features with their literal column names. Subsequently, a refined data frame is generated, comprising solely the most pertinent principal component analysis (PCA) characteristics that support the creation of a more specialized and efficient intrusion detection system tailored to IoT environments.

4.2.3. Classification Techniques

This section describes the various classification techniques used in this research.

- **Random Forest**

Training a multitude of trees as opposed to a single tree using a family of tree-based algorithms. Random subsets of training data are used to construct these trees, which apply to both regression and classification tasks. These systems can scale and handle extensive datasets, effectively reducing the total generalization error with precision [47-48].

- **Convolutional Neural Network**

The objective of this phase is to utilize 1D CNN to produce a resilient detection model capable of precisely identifying IoT attacks; this model should achieve exceptional performance concerning recall, accuracy, and precision. CNNs independently derive relevant characteristics in contrast to conventional approaches, operating without the need for external input or predetermined criteria. As a result, convolutional neural networks (CNNs) possess the capability to reveal associations and trends that might escape traditional feature selection methods. Nevertheless, using CNNs for feature selection and attack detection may be constrained by the difficulty of training deep convolutional neural networks, overfitting, hyperparameter sensitivity, the requirement for vast quantities of labeled data, and sensitivity to hyperparameters [49].

- **C 4.5**

The C4.5 decision tree architecture is, therefore, a strong classification instrument. This means that the C4.5 model can handle both continuous and categorical data that are often seen in IoT systems showing wide characteristics. In this case, the C4.5 model has each node as a decision about a given attribute, which makes it easier to identify malicious behavior. This model also finds complicated relationships between variables in the dataset, which can be difficult to understand and provides understandable rules for humans. As a result, it helps secure IoT networks by accurately classifying normal and unusual activities.

The structure of the proposed methodology is outlined in Figure 6. This describes the proposed hybrid intrusion-based system that uses the BOT-IOT dataset. This collection contains over 7.2 million records pertaining to DDOS, DOS, OS, service scans, and other related activities. The dataset is initially divided into two classes in this research: training data and testing data, which assist in the training and testing stages of the developed model. Then, it goes through a data preprocessing phase. Herein includes data cleaning, noise removal and dimension reduction, making the data more efficient. The preprocessed data then undergoes feature extraction using PCA (principal component analysis). After that feature selection was carried out by using Select K Best with F regression model.

In this research, a hybrid IDS (intrusion detection system) is introduced using convolutional neural networks (CNN), C4.5, and random forests to classify both unknown and known attacks. The BOT-IOT dataset was used in the training of the proposed hybrid IDS system, which was then assessed for its effectiveness and performance in the test phase.

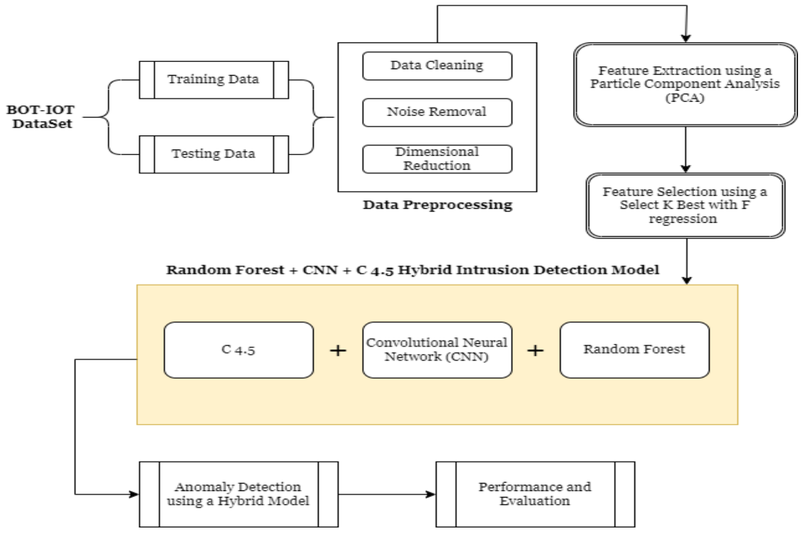

BOT-IOT DataSet
Training Data
Testing Data
Data Cleaning
Noise Removal
Dimensional Reduction
Data Preprocessing
Feature Extraction using a Particle Component Analysis (PCA)
Feature Selection using a Select K Best with F regression
Random Forest + CNN + C 4.5 Hybrid Intrusion Detection Model
C 4.5
+
Convolutional Neural Network (CNN)
+
Random Forest
Anomaly Detection using a Hybrid Model
Performance and Evaluation

Fig 6. The proposed architecture of the suggested hybrid machine learning model is based on CNN, C 4.5, and Random Forest.

5. Performance Evaluation and Analysis

This section pertains to the evaluation and effectiveness of the proposed hybrid machine model.

5.1 Experiment Setup and Selection

In this section, the experiment setup is described, which involves a computer with 8GB of RAM and an ordinary CPU as well as Google Colaboratory TPUs and GPUs for training and testing the system. This study employs different models such as CNN, C4.5, Random Forest etc. The author uses the Bot-IoT dataset in this research, which includes more than 72 lakh records of DDOS, DOS, OS, service scans etc. Python was chosen as the programming language for the development of testing and training systems on account of its extensive collection of versatile scientific frameworks, including Scikit-learn (utilized for ML training and testing, data preprocessing), NumPy (designed for matrix processing), Pandas (utilized for reading data from a file, data handling, and writing the processed data), and Matplotlib (utilized for displaying data and results).

5.2 Feature selection and extraction from the BoT-IoT dataset

The table below 1 represents all the characteristics of the BoT-IoT dataset and the extracted and selected features from this dataset with the help of the PCA technique and Select K Based with F regression techniques.

Table 1. Features of the BOT-IOT dataset

Features	**Features extracted using a PCA technique**	**Features selection using a Select K-Based with f regression**
State, ltime, stime, dur, mean, stddev, sum, minimum, maximum, srate, daddr,pkts,ltime,State, pkts, bytes, state, drate, drate, pkSeqID, state, bytes, dport, saddr, pkSeqID, srate, drate, state_number, stime, dur, mean, stddev, sum, min, max, spkts, dpkts, sbytes, drate, rate, srate, drate, pkSeqID, saddr, sport, daddr, dport, pkts, bytes, state_number, state, srate, drate, rate, srate, drate, ArcIp, Protocol _DestIP,TnBPSrcIp, TnBpDstIP,TnP_PDstIP,TnP_perProto, AR_P_Pro, AR_P_DstIP, AR_P_Proto_P_DstIP, AR_P_Proto_P_DstIP, AR_P_Proto _P_DstIP, AR_P_Proto_P_DstIP, AR _P_Proto_P_DstIP, AR_P_Proto_P_DstIP, AR_P _Proto_P_Sport, AR_P_Proto_P_Dport, Pkts_P_State_P_ Protocol_P_DestIP, and AR_P_Proto_P_DstIP.	STDEV, min, sum, rate, mean, maximum, drate, seq, N_IN_Conn_P _SrcIP, stddev, srate, N_IN_ Conn_P_DstIP, dbyte, AR_P_Protostate _number,TnP_Per_Dport, _P_Sport, and srate.	stddev, srate; N_IN_Conn_ P_DstIP, seq, state_number, drate, minimum, mean, and maximum; N_IN_Conn_P_SrcIP.

Below table 2 depicts the hyperparameters of the different models like random forest, CNN, and C4.5. It describes the various parameters like Random state, kernel size, batch size, epochs etc.

Table 2. Hyperparameters of the hybrid model

Models	**Hyperparameter**	**Values**
Random Forest	N_Estimator	100
	Random State	42
Convolutional Neural Network	Kernel Size	3
	Activation	Relu
	Epochs	50
	Batch Size	32

C 4.5	Criterion	Entropy
	Random State	42

5.3 Performance Measures

They considered metrics such as F1-Score, Recall, Precision, and accuracy to evaluate and analyze the ML models' performances.

- **Accuracy**

The proportion of true positive and negative values to the overall number of values.

$$Accuracy\ = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Positive} \quad (1)$$

True Positive (TP): Detected the altered images without error.

False Positive (FP): Images mistakenly recognized as genuine or manipulated.

True Negative (TN): Validated as authentic on visual inspection.

False Negative (FN): Falsely recognized manipulated images or images mistakenly thought genuine.

- **Precision**

The word precision is used to describe the unavoidable variation in measuring results. Thermal effects probably cause a random fluctuation in the observed value. It can be calculated as:

$$Precision = \frac{TP}{TP+FN} \quad (2)$$

- **Recall**

One of the other most crucial parameters for testing an ML model is recall. The formula for determining the recall is:

$$Recall = TP/(TP + FN) \quad (3)$$

- **F1-Score**

The F1 score is a unified metric utilized for binary classification tasks that balance the evaluation of a model's performance by combining its precision and recall.

$$F1\ score = \frac{2(Precision * Recall)}{Precision + Recall} \quad (4)$$

The confusion matrix of the suggested hybrid model is represented in Figure 7. It shows the efficiency of a classification algorithm. It is divided into four squares, each of which represents a different combination of actual and predicted class labels. The diagonal squares represent the count of accurate predictions, whereas the off-diagonal squares represent the count of inaccurate predictions. The confusion matrix is partitioned into four rectangles in the image. The number of true positives (TP), denoting instances that were accurately classified as assaults, is displayed in the upper left square. In the upper right corner, the count of false positives (FP) is displayed; FP represents instances that were erroneously classified as assaults. The number of false negatives (FN), or instances that were erroneously classified as non-attacks, is displayed in the bottom left square. The figure in the lower right corner represents the true negatives (TN), or the number of occurrences that were accurately classified as non-attacks.

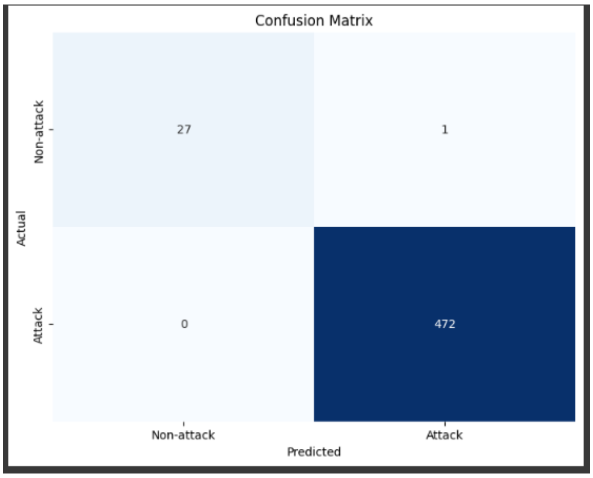


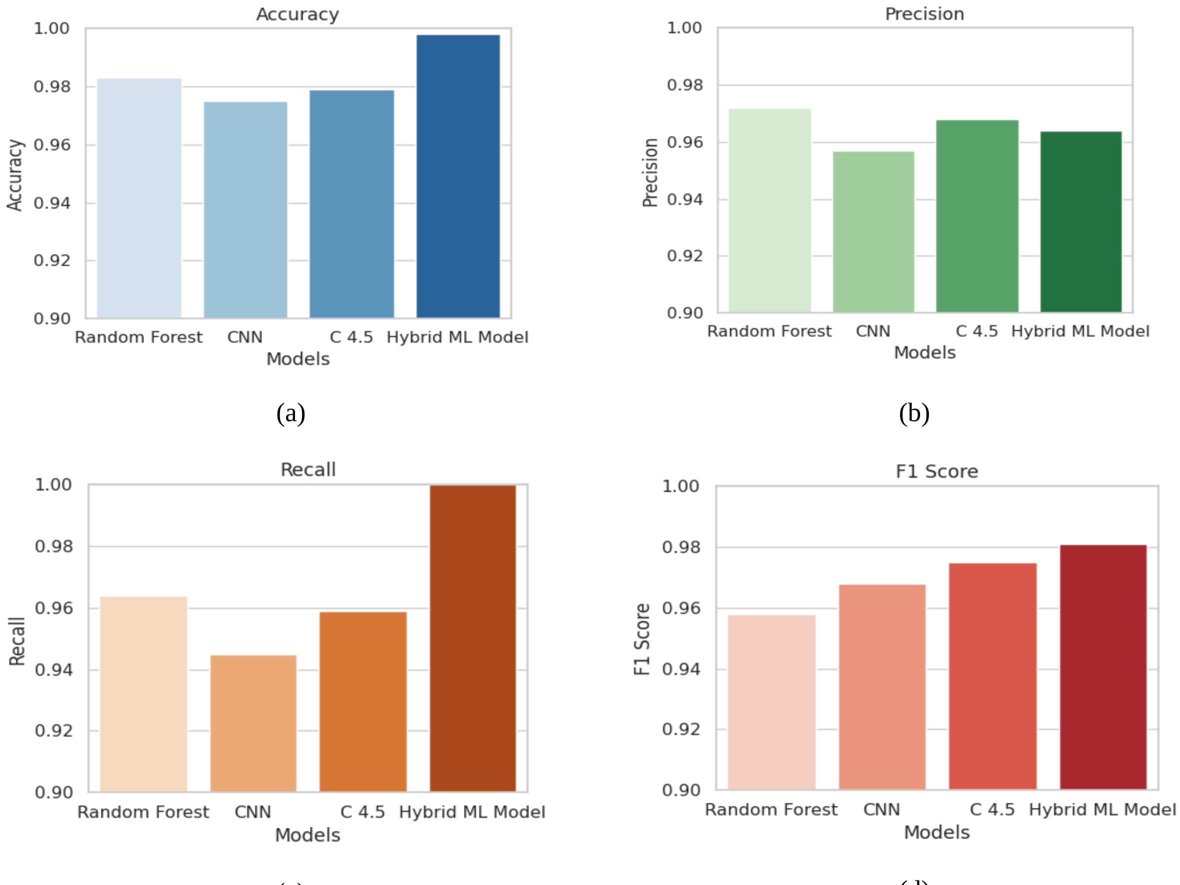
Fig 7. The confusion matrix shows the FN, TP, TN, and FP ratio on the BoT-IoT dataset of the hybrid model.

The comparison analysis of the different models, including random forest, CNN, C4.5, and the hybrid model, based on F1 result, accuracy, precision, and recall is described in Table 3.

Table 3. Model accuracy, precision, recall, and F1 score.

Models	Accuracy	Precision	Recall	F1 Score
Random Forest (RF)	0.983	0.972	0.964	0.958
Convolutional Neural Network (CNN)	0.975	0.957	0.945	0.968
C 4.5	0.979	0.968	0.959	0.975
Proposed Hybrid model (RF+CNN+C 4.5)	0.998	0.964	1.000	0.981

The comparison analysis of the proposed hybrid machine learning model with the different models such as C 4.5, CNN, and Random Forest is depicted in Figure 8. Figure 8 (A) depicts the accuracy of different models. Figure 8 (B) shows the analysis of precision, and figures 8 (C) and (D) represent the recall and the F1 score of different models.



(a)

(b)

(c)

(d)

Figure 8. Comparison of the proposed hybrid model with CNN, Random Forest and C 4.5.

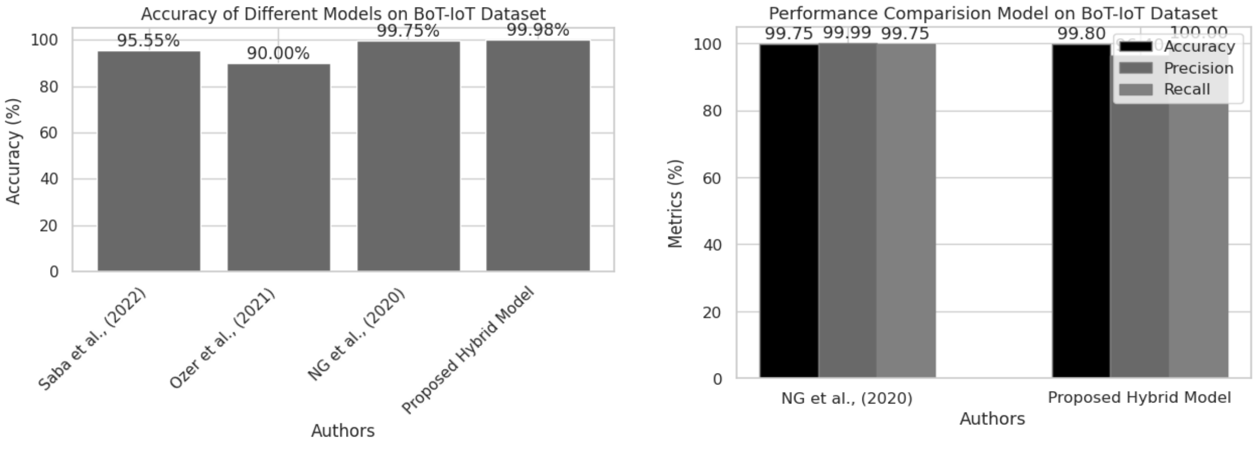
5.4 Comparison Analysis

In this section, a comparison is made between the existing model and the proposed hybrid machine learning model. Table 4 presents a comparison between previous research results and the proposed hybrid model as per factors including precision, recall, accuracy, and F1 score.

Table 4. Comparison Analysis of the proposed hybrid model with existing research.

Author	Dataset	Accuracy	Precision	Recall
Saba et al., (2022) [29]	BoT-IoT	95.55	-	-
Ozer et al., (2021) [31]	BoT-IoT	90.00	-	-
NG et al., (2020) [33]	BoT-IoT	99.755	99.99	99.75
Proposed Hybrid Model	BoT-IoT	0.998	0.964	1.000

The accuracy and performance of the different models based on the BoT-IoT dataset are depicted in Figure 9. Figure 9 (A) depicts the comparison of the accuracy of different models, and Figure 9 (B) represents the comparison of the performance of different models based on BoT-IoT dataset.



(A)

(B)

Fig 9. Comparison of the proposed hybrid machine learning model with previous studies.

6. Conclusion and Future Scope

The rapid expansion of computer networks has caused uncertainty regarding the security of digital domains and systems, according to the cybersecurity community. Availability, confidentiality, and integrity are very important in securing computer networks. Therefore, a novel hybrid machine learning model is introduced for anomaly detection. The main aim of this study is to improve abnormal activity detection in network traffic. For identifying malicious activities in the BoT-IoT dataset, the proposed hybrid model uses C4.5, Convolutional Neural Network (CNN), and Random Forest (RF). In this research, the author observed that the proposed model worked well with 99.8% accuracy in classifying the anomaly using a BoT-IoT dataset. As for malicious activity, by utilizing a proposed hybrid machine learning model, the author achieved 100% recall, 96.4% precision, and a 98.1% F1 score. In a few years, the framework's evolution may explore self-learning mechanisms to develop adaptive responses to emergent threats. The proposed hybrid machine learning-based framework holds the potential to enhance intrusion detection systems' capacity to protect against constantly changing cyber threats. The study proved that the hybrid framework beats conventional models, providing a more dependable and robust strategy for handling the evolving environment of cyber threats.

References

[1] Anderson JP. Computer Security Threat Monitoring and Surveillance. Fort Washington, PA: James P Anderson Co; 1980.

[2] Debar H, Dacier M, Wespi A. Towards a taxonomy of intrusion-detection systems. Computer Network. 1999;31(8):805-822. https://doi.org/10. 1016/S1389-1286(98)00017-6.

[3] Hoque MS, Mukit M, Bikas M, Naser A, An implementation of intrusion detection system using genetic algorithm; 2012. arXiv preprint arXiv:1204.1336.

[4] Ahmad, Zeeshan, Adnan Shahid Khan, Cheah Wai Shiang, Johari Abdullah, and Farhan Ahmad. "Network intrusion detection system: A systematic study of machine learning and deep learning approaches." Transactions on Emerging Telecommunications Technologies 32, no. 1 (2021): e4150.

[5] K. Liu, S. Xu, G. Xu, M. Zhang, D. Sun et al. (2020). , "A review of android malware detection approaches based on machine learning," IEEE Access, vol. 8, pp. 124579–124607.

[6] M. A. Khan and J. Kim. (2020). "Toward developing efficient Conv-AE-based intrusion detection system using the heterogeneous dataset," Electronics, vol. 9, no. 11, pp. 1–17.

[7] Khan, Muhammad Ashfaq, and Yangwoo Kim. "Deep Learning-Based Hybrid Intelligent Intrusion Detection System." Computers, Materials & Continua 68, no. 1 (2021).

[8] Bace, Rebecca Gurley, and Peter Mell. "Intrusion detection systems." (2001).

[9] Khraisat, Ansam, Iqbal Gondal, Peter Vamplew, and Joarder Kamruzzaman. "Survey of intrusion detection systems: techniques, datasets and challenges." Cybersecurity 2, no. 1 (2019): 1-22.

[10] Anwar, Shahid, Jasni Mohamad Zain, Mohamad Fadli Zolkipli, Zakira Inayat, Suleman Khan, Bokolo Anthony, and Victor Chang. "From intrusion detection to an intrusion response system: fundamentals, requirements, and future directions." Algorithms 10, no. 2 (2017): 39.

[11] Samrin, Rafath, and D. Vasumathi. "Review on anomaly based network intrusion detection system." In 2017 international conference on electrical, electronics, communication, computer, and optimization techniques (ICEECCOT), pp. 141-147. IEEE, 2017.

[12] Subba, Basant, Santosh Biswas, and Sushanta Karmakar. "Enhancing performance of anomaly-based intrusion detection systems through dimensionality reduction using principal component analysis." In 2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), pp. 1-6. IEEE, 2016.

[13] https://www.semanticscholar.org/paper/Accelerating-Anomaly-Based-IDS-Using-Neural-Network-Van-Thinh/553bb5fe62092b0b9ec5e601a4e06fd1fd2f95dc/figure/0

[14] Jyothsna, V. V. R. P. V., Rama Prasad, and K. Munivara Prasad. "A review of anomaly-based intrusion detection systems." International Journal of Computer Applications 28, no. 7 (2011): 26-35.

[15] Vigna, Giovanni, and Richard A. Kemmerer. "NetSTAT: A network-based intrusion detection approach." In Proceedings 14th Annual Computer Security Applications Conference (Cat. No. 98EX217), pp. 25-34. IEEE, 1998.

[16] Singh, Amrit Pal, and Manik Deep Singh. "Analysis of host-based and network-based intrusion detection system." International Journal of Computer Network and Information Security 6, no. 8 (2014): 41-47.

[17] https://www.sciencedirect.com/topics/computer-science/network-based-intrusion-detection-system

[18] Gangwar, A., and S. Sahu. "A survey on anomaly and signature based intrusion detection system (IDS)." International Journal of Engineering Research and Applications 4, no. 4 (2014).

[19] Liu, Ming, Zhi Xue, Xianghua Xu, Changmin Zhong, and Jinjun Chen. "Host-based intrusion detection system with system calls: Review and future trends." ACM Computing Surveys (CSUR) 51, no. 5 (2018): 1-36.

https://www.liquidweb.com/blog/host-based-intrusion-detection-system/

[20] Syed Shariyar Murtaza, Wael Khreich, Abdelwahab Hamou-Lhadj, and Stephane Gagnon. 2015. A trace abstraction approach for host-based anomaly detection. In 2015 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA’15). IEEE, 1–8.

[21] Padmaja, B., K. Sai Sravan, E. Krishna Rao Patro, and G. Chandra Sekhar. "A System to automate the development of anomaly-based network intrusion detection model." In Journal of Physics: Conference Series, vol. 2089, no. 1, p. 012006. IOP Publishing, 2021.

[22] Kumar, Vinod, and Om Prakash Sangwan. "Signature based intrusion detection system using SNORT." International Journal of Computer Applications & Information Technology 1, no. 3 (2012): 35-41.

[23] Jadhav, Kishor P., Tripti Arjariya, and Mohit Gangwar. "Hybrid-Ids: An Approach for Intrusion Detection System with Hybrid Feature Extraction Technique Using Supervised Machine Learning." International Journal of Intelligent Systems and Applications in Engineering 11, no. 5s (2023): 591-597

[24] Abbas, Qaiser, Sadaf Hina, Hamza Sajjad, Khurram Shabih Zaidi, and Rehan Akbar. "Optimization of predictive performance of intrusion detection system using hybrid ensemble model for secure systems." PeerJ Computer Science 9 (2023): e1552.

[25] Talukder, Md Alamin, Khondokar Fida Hasan, Md Manowarul Islam, Md Ashraf Uddin, Arnisha Akhter, Mohammand Abu Yousuf, Fares Alharbi, and Mohammad Ali Moni. "A dependable hybrid machine learning model for network intrusion detection." Journal of Information Security and Applications 72 (2023): 103405.

[26] SARALADEVE, L., and A. CHANDRASEKAR. "A NOVEL HYBRID INTRUSION DETECTION MODEL FOR INTERNET OF THINGS USING MACHINE LEARNING." Journal of Theoretical and Applied Information Technology 101, no. 14 (2023).

[27] Balyan, Amit Kumar, Sachin Ahuja, Umesh Kumar Lilhore, Sanjeev Kumar Sharma, Poongodi Manoharan, Abeer D. Algarni, Hela Elmannai, and Kaamran Raahemifar. "A hybrid intrusion detection

model using ega-pso and improved random forest method." Sensors 22, no. 16 (2022): 5986.

[28] Saba, Tanzila, Amjad Rehman, Tariq Sadad, Hoshang Kolivand, and Saeed Ali Bahaj. "Anomaly-based intrusion detection system for IoT networks through deep learning model." Computers and Electrical Engineering 99 (2022): 107810.

[29] Megantara, Achmad Akbar, and Tohari Ahmad. "A hybrid machine learning method for increasing the performance of network intrusion detection systems." Journal of Big Data 8, no. 1 (2021): 1-19

[30] Özer, Erman, Murat İskefiyeli, and Jahongir Azimjonov. "Toward lightweight intrusion detection systems using the optimal and efficient feature pairs of the Bot-IoT 2018 dataset." International Journal of Distributed Sensor Networks 17, no. 10 (2021): 15501477211052202.

[31] Khonde, S. R., and V. Ulagamuthalvi. "Hybrid framework for intrusion detection system using ensemble approach." International Journal of Advanced Trends in Computer Science and Engineering 9, no. 4 (2020).

[32] NG, Bhuvaneswari Amma, and S. Selvakumar. "Anomaly detection framework for Internet of things traffic using vector convolutional deep learning approach in fog environment." Future Generation Computer Systems 113 (2020): 255-265.

[33] Ren, Jiadong, Jiawei Guo, Wang Qian, Huang Yuan, Xiaobing Hao, and Hu Jingjing. "Building an effective intrusion detection system by using hybrid data optimization based on machine learning algorithms." Security and communication networks 2019 (2019).

[34] Çavuşoğlu, Ünal. "A new hybrid approach for intrusion detection using machine learning methods." Applied Intelligence 49 (2019): 2735-2761.

[35] Foroushani, Zohreh Abtahi, and Yue Li. "Intrusion detection system by using hybrid algorithm of data mining technique." In proceedings of the 2018 7th international conference on software and computer applications, pp. 119-123. 2018.

[36] Qazi, Emad Ul Haq, Muhammad Hamza Faheem, and Tanveer Zia. "HDLNIDS: Hybrid Deep-Learning-Based Network Intrusion Detection System." Applied Sciences 13, no. 8 (2023): 4921.

[37] Aldallal, Ammar. "Toward efficient intrusion detection system using hybrid deep learning approach." Symmetry 14, no. 9 (2022): 1916.

[38] Gamal, Merna, Hala Abbas, and Rowayda Sadek. "Hybrid approach for improving intrusion detection based on deep learning and machine learning techniques." In Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), pp. 225-236. Springer International Publishing, 2020

[39] Naseer, Sheraz, and Yasir Saleem. "Enhanced network intrusion detection using deep convolutional neural networks." KSII Transactions on Internet and Information Systems (TIIS) 12, no. 10 (2018): 5159-5178.

[40] Dias, Leonardo P., Jés de Jesus Fiais Cerqueira, Karcius DR Assis, and Raul C. Almeida. "Using artificial neural network in intrusion detection systems to computer networks." In 2017 9th Computer Science and Electronic Engineering (CEEC), pp. 145-150. IEEE, 2017.

[41] Samdekar, Ramanand, S. M. Ghosh, and Konda Srinivas. "Efficiency enhancement of intrusion detection in iot based on machine learning through bioinspire." In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 383-387. IEEE, 2021.

[42] Khraisat, Ansam, Iqbal Gondal, Peter Vamplew, Joarder Kamruzzaman, and Ammar Alazab. "A novel ensemble of hybrid intrusion detection system for detecting internet of things attacks." Electronics 8, no. 11 (2019): 1210.

[43] Pevny, T., M. Rehák, and M. Grill. Detecting anomalous network hosts by means of pca. in Information Forensics and Security (WIFS), 2012 IEEE International Workshop on. 2012. IEEE.

[44] Li, W., et al., Recursive PCA for adaptive process monitoring. Journal of process control, 2000. 10(5): p. 471-486.

[45] Chen, Zheng, Xinli Yu, Chi Zhang, Jin Zhang, Cui Lin, Bo Song, Jianliang Gao, Xiaohua Hu, Wei-Shih Yang, and Erjia Yan. "Fast botnet detection from streaming logs using online lanczos method." In 2017 IEEE International Conference on Big Data (Big Data), pp. 1408-1417. IEEE, 2017.

[46] Hussain, F.; Hussain, R.; Hassan, S.A.; Hossain, E. Machine learning in IoT security: Current solutions and future challenges. IEEE Commun. Surv. Tutor. 2020, 22, 1686–1721.

[47] Alkadi, Sarah, Saad Al-Ahmadi, and Mohamed Maher Ben Ismail. "Toward Improved Machine Learning-Based Intrusion Detection for Internet of Things Traffic." Computers 12, no. 8 (2023): 148.

[48] Alabsi, Basim Ahmad, Mohammed Anbar, and Shaza Dawood Ahmed Rihan. "CNN-CNN: Dual Convolutional Neural Network Approach for Feature Selection and Attack Detection on Internet of Things Networks." Sensors 23, no. 14 (2023): 6507.

SJST MANUSCRIPT TEMPLATE FOR A FIGURE FILE

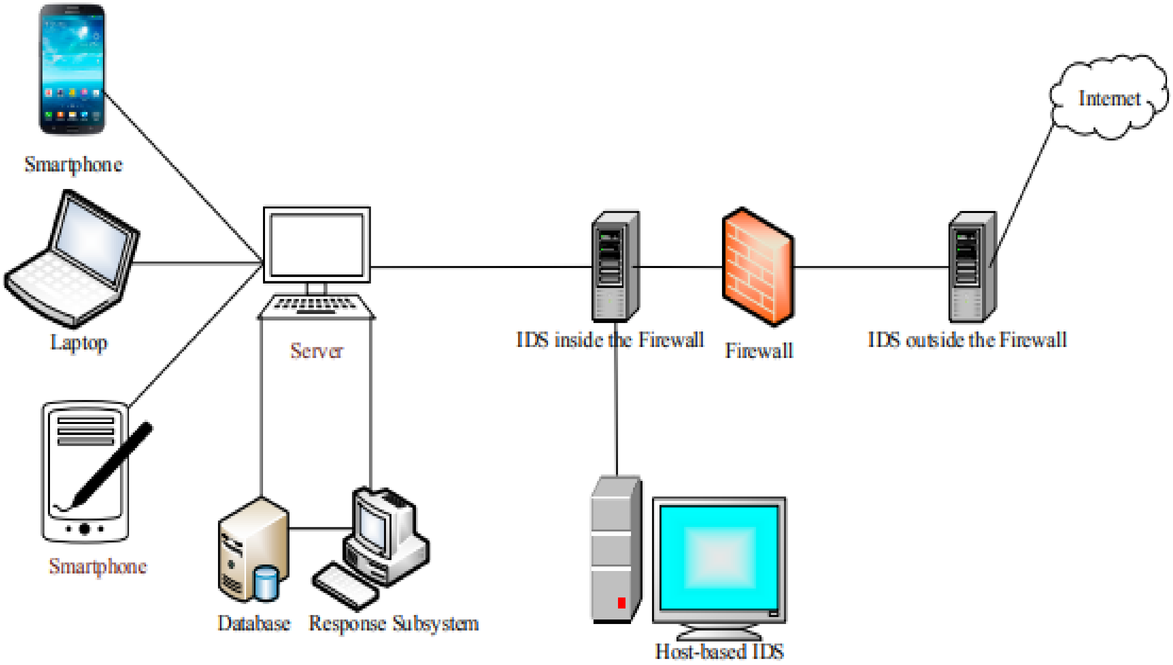


Figure 1 Structure of Intrusion Detection System [10]

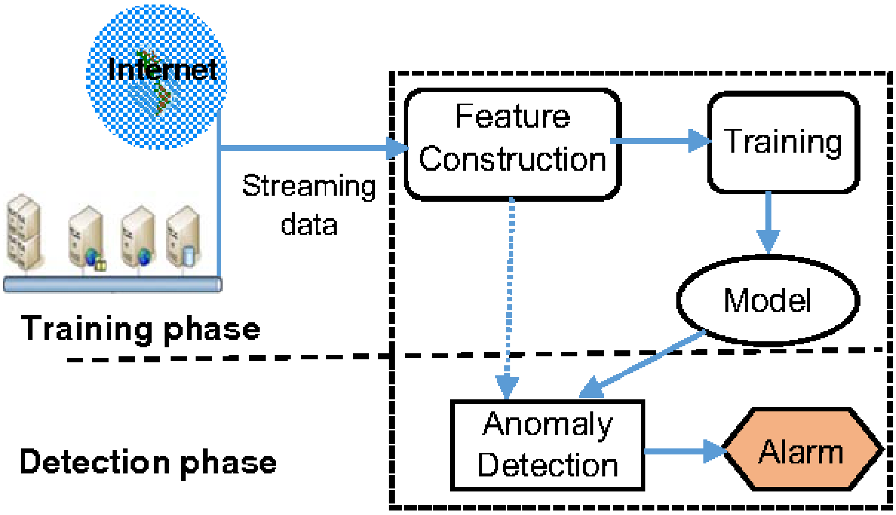


Figure 2 Architecture of Anomaly Based detection [13]

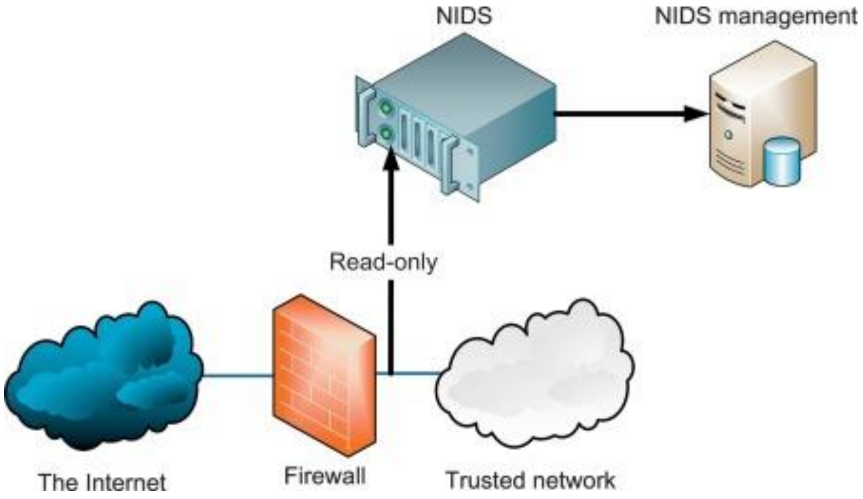


Figure 3 The architecture of a Network-based Intrusion detection system [17]

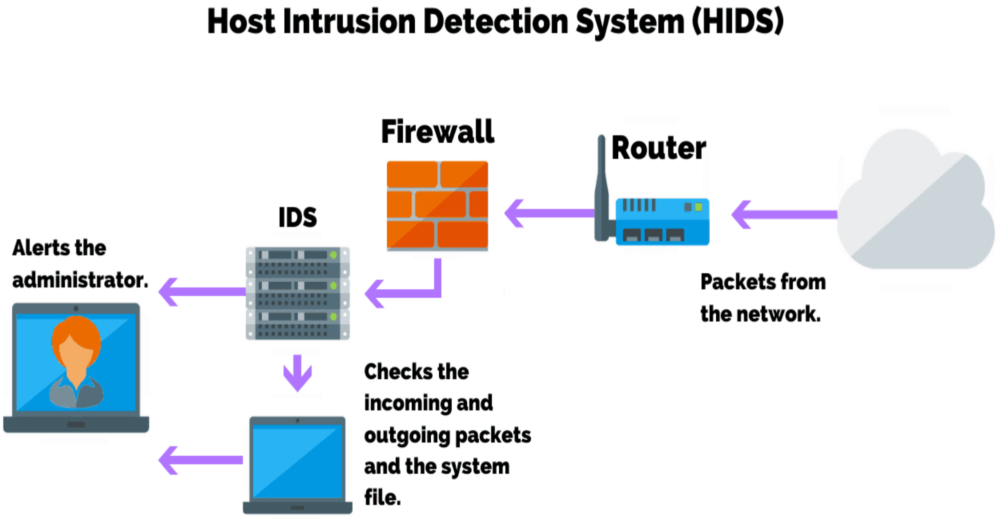

Host Intrusion Detection System (HIDS)
Firewall
Router
IDS
Alerts the administrator.
Packets from the network.
Checks the incoming and outgoing packets and the system file.

Figure 4 Working of HIDS system [20]

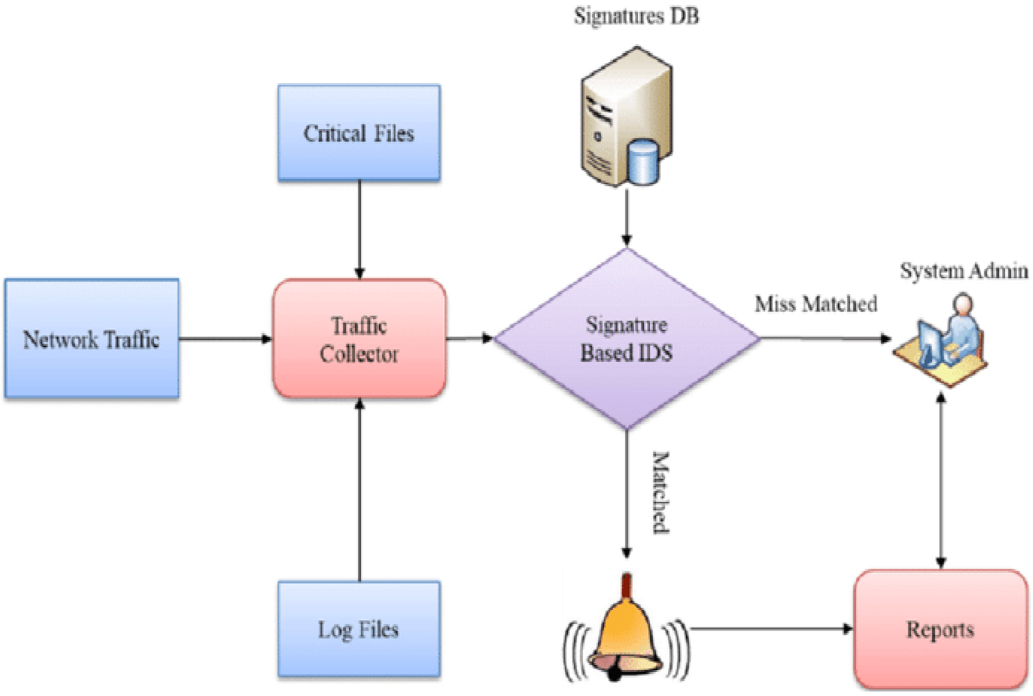


Figure 5 The architecture of signature-based intrusion detection system [22]

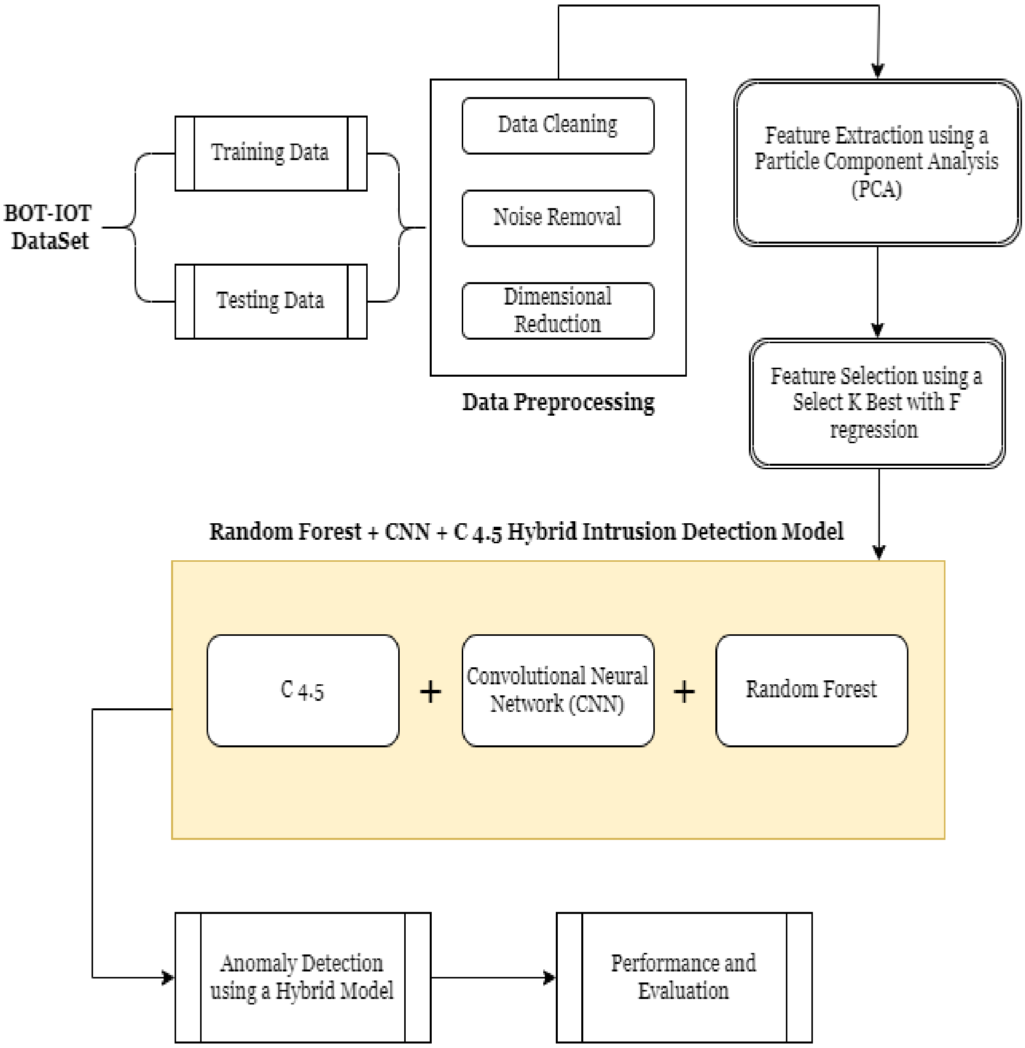

BOT-IOT DataSet
Training Data
Testing Data
Data Cleaning
Noise Removal
Dimensional Reduction
Data Preprocessing
Feature Extraction using a Particle Component Analysis (PCA)
Feature Selection using a Select K Best with F regression
Random Forest + CNN + C 4.5 Hybrid Intrusion Detection Model
C 4.5
+
Convolutional Neural Network (CNN)
+
Random Forest
Anomaly Detection using a Hybrid Model
Performance and Evaluation

Figure 6 The proposed architecture of the suggested hybrid machine learning model is based on CNN, C 4.5, and Random Forest.

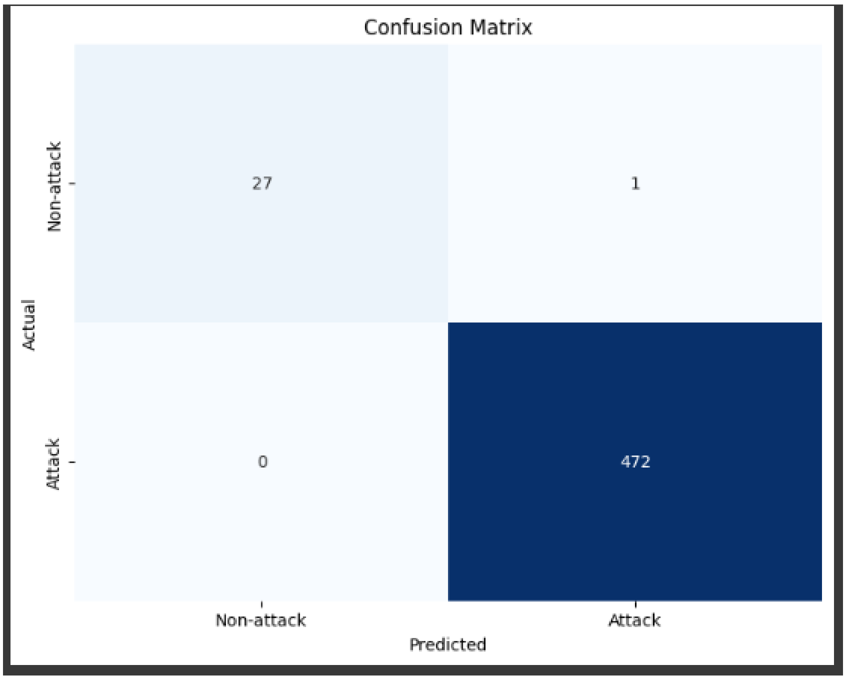


Figure 7 The confusion matrix shows the FN, TP, TN, and FP ratio on the BoT-IoT dataset of the hybrid model.

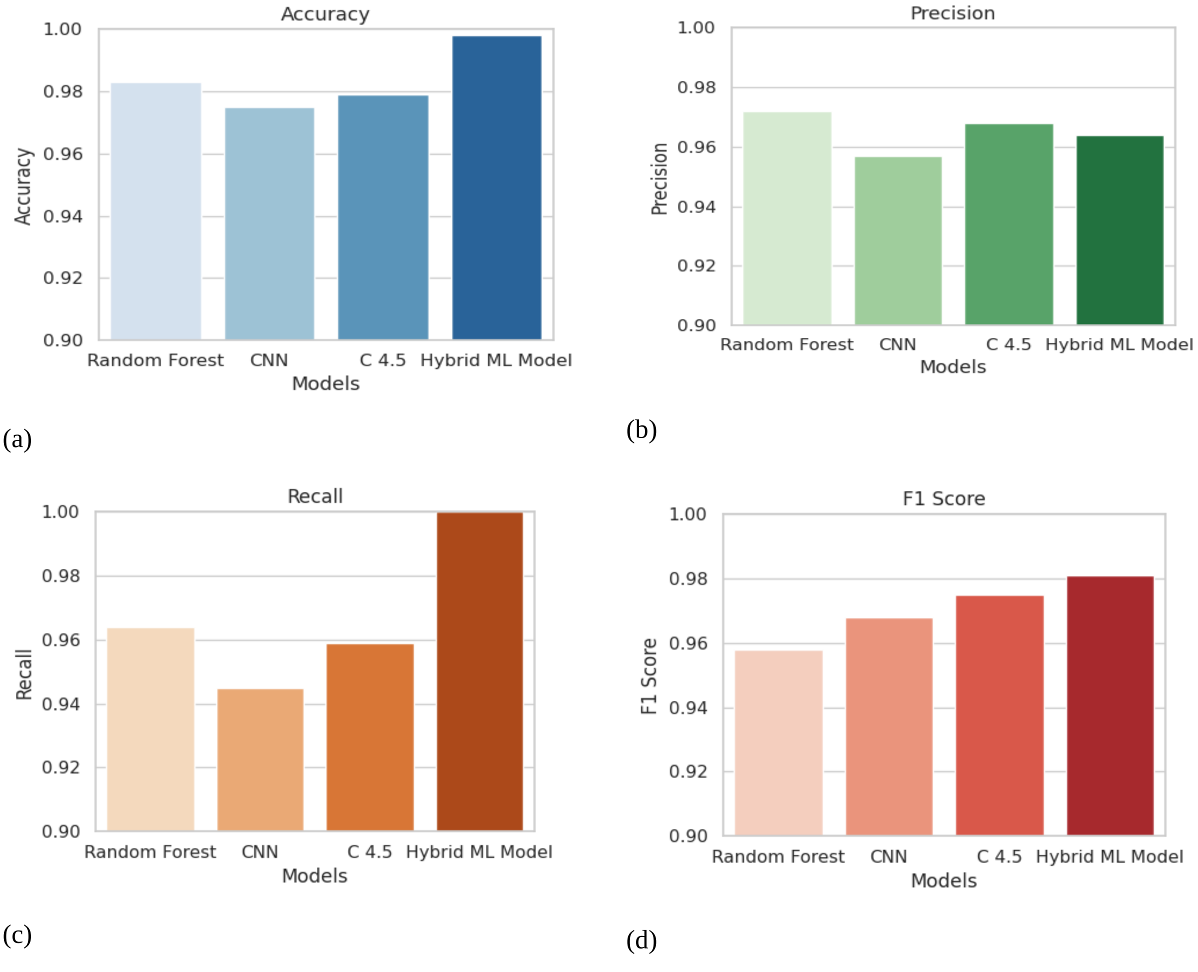
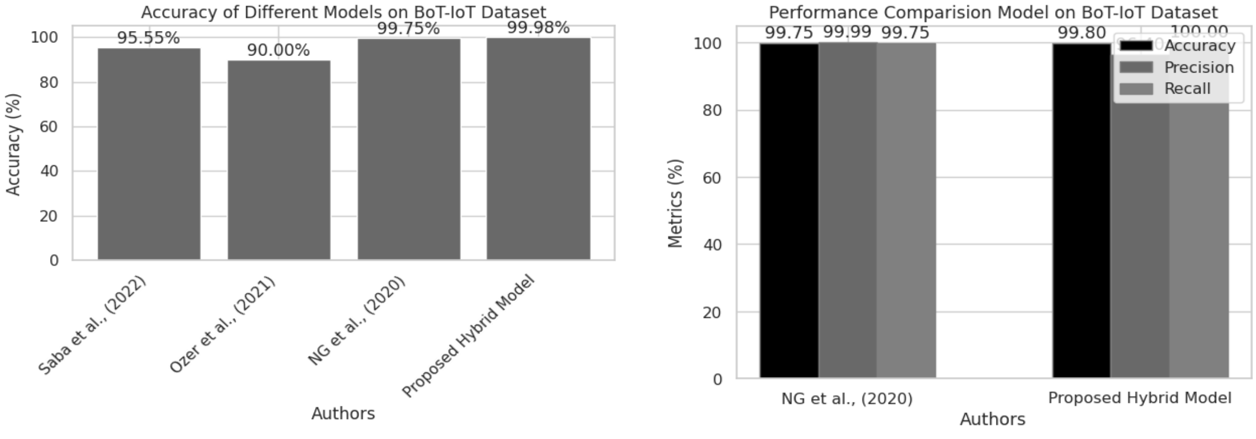


Figure 8 Comparison of the proposed hybrid model with CNN, Random Forest and C 4.5.



(A)

(B)

Fig 9 Comparison of the proposed hybrid machine learning model with previous studies.

SJST MANUSCRIPT TEMPLATE FOR A TABLE FILE

Features	**Features extracted using a PCA technique**	**Features selection using a Select K-Based with f regression**
State, ltime, stime, dur, mean, stddev, sum, minimum, maximum, srate, daddr,pkts,ltime,State, pkts, bytes, state, drate, drate, pkSeqID, state, bytes, dport, saddr, pkSeqID, srate, drate, state_number, stime, dur, mean, stddev, sum, min, max, spkts, dpkts, sbytes, drate, rate, srate, drate, pkSeqID, saddr, sport, daddr, dport, pkts, bytes, state_number, state, srate, drate, rate, srate, drate, ArcIp, Protocol _DestIP,TnBPSrcIp, TnBpDstIP,TnP_PDstIP,TnP_perProto, AR_P_Pro, AR_P_DstIP, AR_P_Proto_P_DstIP, AR_P_Proto_P_DstIP, AR_P_Proto _P_DstIP, AR_P_Proto_P_DstIP, AR _P_Proto_P_DstIP, AR_P_Proto_P_DstIP, AR_P _Proto_P_Sport, AR_P_Proto_P_Dport, Pkts_P_State_P_ Protocol_P_DestIP, and AR_P_Proto_P_DstIP.	STDEV, min, sum, rate, mean, maximum, drate, seq, N_IN_Conn_P _SrcIP, stddev, srate, N_IN_ Conn_P_DstIP, dbyte, AR_P_Protostate _number,TnP_Per_Dport, _P_Sport, and srate.	stddev, srate; N_IN_Conn_ P_DstIP, seq, state_number, drate, minimum, mean, and maximum; N_IN_Conn_P_SrcIP.

Table 1 Features of the BOT-IOT dataset

Models	**Hyperparameter**	**Values**
Random Forest	N_Estimator	100
	Random State	42
Convolutional Neural Network	Kernel Size	3
	Activation	Relu

	Epochs	50
	Batch Size	32
C 4.5	Criterion	Entropy
	Random State	42

Table 2 Hyperparameters of the hybrid model

Models	Accuracy	Precision	Recall	F1 Score
Random Forest (RF)	0.983	0.972	0.964	0.958
Convolutional Neural Network (CNN)	0.975	0.957	0.945	0.968
C 4.5	0.979	0.968	0.959	0.975
Proposed Hybrid model (RF+CNN+C 4.5)	0.998	0.964	1.000	0.981

Table 5 Model accuracy, precision, recall, and F1 score.

Author	Dataset	Accuracy	Precision	Recall
Saba et al., (2022) [29]	BoT-IoT	95.55	-	-
Ozer et al., (2021) [31]	BoT-IoT	90.00	-	-
NG et al., (2020) [33]	BoT-IoT	99.755	99.99	99.75
Proposed Hybrid Model	BoT-IoT	0.998	0.964	1.000

Table 6 Comparison Analysis of the proposed hybrid model with existing research.