Mathematical Modelling of Engineering Problems
Vol. 11, No. 9, September, 2024, pp. 2355-2367
Journal homepage: http://iieta.org/journals/mmep

Ensemble of Tree Classifiers for Improved DDoS Attack Detection in the Internet of Things

Jyoti Mante*, Kishor Kolhe

School of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, Pune 411052, India

Corresponding Author Email: jyoti.khurpade@mitwpu.edu.in

https://doi.org/10.18280/mmep.110909

Received: 4 January 2024
Revised: 4 May 2024
Accepted: 15 May 2024
Available online: 29 September 2024

Keywords:
stacked tree-based classifiers, feature importance, XGBoost, Ensemble Voting Classifier, Step Forward Feature Selection, artificial intelligence, Stacking, Random Forest (RF)

ABSTRACT

IoT networks are made up of devices that have few computer resources, like less battery life, less processing power, less memory, and most importantly, minimum security, defense mechanisms, and infrastructure to protect them. As the number of IoT devices are growing fast, we can't ignore the impact of large-scale DDoS attacks that come from IoT devices. Artificial intelligence, including Deep Learning and machine learning (ML), is critical in the categorization and detection of DDoS attacks in the Internet of Things. We hope to contribute to current research by enhancing the efficiency with which Intrusion Detection Systems (IDS) identify DDoS attacks. This research paper focuses on exploring the effectiveness of tree-based classifiers and ensemble classifiers. The ensemble approaches used are voting and Stacking with a particular focus on Step Forward Feature Selection and average feature importance methods by performing evaluation to improve the classification of DDoS attacks, Decision Tree (DT), Extra Tree (ET), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) were selected as the best tree classifiers. Hyper-parameter tuning was performed to improve the performance of the classifiers. The proposed models are trained on Bot-IoT, CICIoT2023, and DS2OS datasets. The tree-based classifiers with ensemble of Stacking and voting demonstrated their capability to effectively detect and classify DDoS attacks. The result obtained from the proposed approach showcase an impressive accuracy rate of over 99%.

1. INTRODUCTION

The IoT has seen amazing growth in recent years, reshaping our understanding of the physical world by interconnecting a vast network of devices [1], all endowed with digital identities. These devices span the spectrum from micro sensors to substantial entities like mobile phones, televisions, light bulbs, thermostats, clinical equipment, smartwatches, and various software applications. This exponential proliferation involves the integration of an expanding array of physical devices into the digital realm, forging new dimensions of connectivity and convenience in our daily lives. In recent times, the IoT has emerged as a pervasive network, seamlessly integrating an extensive array of devices to streamline various aspects of human life [2]. Remarkably, the research indicates that there are now more than 5 trillion internet-connected devices in existence. The economic impact of this technology has been substantial, with market revenues exceeding $100 billion in 2017, marking a historic milestone. Furthermore, the global market for IoT end-node products surged to a staggering $212 billion USD by the close of 2019. Projections for the future are equally promising, as experts anticipate this figure to skyrocket to approximately 1.6 trillion by the year 2025 [3]. This relentless growth underscores the transformative power and increasing significance of IoT in our interconnected world. The widespread proliferation of IoT applications has made this technology increasingly susceptible to vulnerabilities and potential cyberattacks. Among the various types of attacks targeting IoT networks, one prevalent threat is the DDoS attack. Detecting DDoS attacks proves to be particularly challenging, especially within the IoT context, as these attacks leverage diverse locations and a multitude of devices to execute their malicious activities. In IoT scenarios, assailants exploit IoT devices as bots to orchestrate attacks, rendering detection and prevention even more formidable tasks. Because of the growing frequency of cyberattacks and threats, businesses are confronted with the formidable challenge of securing their systems and data against unauthorized access and malicious actions. IDS have become an essential part of network security because of their ability to detect and react to potential intrusions in a timely and efficient manner [4]. These systems employ a wide variety of artificial intelligence methodologies and algorithms to examine the IoT traffic and identify any abnormal patterns of behavior that may point to a breach in security. Exploration of various classification algorithms and approaches is one area that is still undergoing research and development in the field of intrusion detection. Traditional methods of network traffic classification have relied on singular classifiers, such as DT and RF to determine whether or not network traffic is benign or malicious. Despite the fact that these classifiers have shown reasonable performance, there is a growing interest in investigating more advanced techniques that have the potential to enhance the efficiency of IDSs [4].

In recent years, ensemble learning and Stacking techniques have gained prominence in the field of ML and data mining [5]. Ensemble learning involves combining multiple classifiers to make more accurate predictions than any individual classifier alone. Stacking, a popular ensemble technique, takes the concept of ensemble learning a step further by training a meta-classifier to combine the predictions of the individual classifiers. By leveraging the diversity of the classifiers, Stacking aims to achieve higher predictive performance. In the context of IDSs, stacked tree-based classifiers have shown promise in enhancing the performance of the detection process. Tree-based classifiers, such as DT and RF, are particularly well-suited for this task due to their ability to handle "capture complex relationships", "high-dimensional data" and provide "interpretable decision rules". By combining these classifiers using Stacking techniques, the strengths of each individual classifier can be leveraged to increase the effectiveness of the whole IDS.

The main purpose of this research work is to investigate the efficacy of stacked tree-based classifiers in IDSs, with a particular emphasis on feature importance and performance evaluation as the two primary areas of investigation. Within the context of the stacked ensemble, the purpose of this work is to investigate the performance of various ML algorithms such as RF, DT, ETC and XGBoost as individual base classifiers. In order to further improve the system's capacity for accurate prediction, this research makes use of an Ensemble Voting Classifier. Experiments were carried out with the datasets Bot-IoT CICIoT2023 and DS2OS, which are benchmark datasets in the field of intrusion detection research. The purpose of these experiments was to evaluate how well the proposed method performed. The Bot-IoT dataset depicts a network environment with a total of 18 features, each of which captures a unique characteristic of network traffic as well as a system parameter. The Step Forward Feature Selection approach (Wrapper), which picks the most relevant features based on their influence on classification accuracy, was used to increase the efficacy and efficiency of the feature selection process [6].

For the Bot-IoT dataset, here employed a Tree-Based Classifiers with Stacking and Feature Importance technique to reduce the feature dimensionality to 9. Similarly, for the DS2OS dataset, applied the same approach to obtain a reduced feature set of 10. By utilizing these feature selection methods, the objective should be to improve the IDS's efficiency and effectiveness by concentrating on the characteristics that provide the most relevant information. The evaluation of the approach that was suggested involved measuring a number of different performance metrics as accurately as possible. Furthermore, an in-depth analysis of feature importance was carried out to determine the primary characteristics that significantly contribute to the accurate detection of intrusions. This was done in order to identify the key attributes. This analysis reveals some very useful insights into the fundamental patterns and characteristics of network traffic that are associated with intrusions.

The results obtained from experiments demonstrated exceptional performance, with an overall rate exceeding 99%. This indicates the effectiveness of the stacked tree-based classifiers in accurately detecting and classifying intrusions in the network environment. In addition, the feature importance analysis uncovered the crucial factors that contribute to the accurate detection of intrusions, which enabled a better understanding of the fundamental characteristics of network traffic that are associated with security breaches. This was made possible by the fact that the analysis uncovered the crucial factors.

The implications that these findings have for the investigation of IDSs as a field as a whole are substantial and significant. The high performance that was achieved by the proposed method using "stacked tree-based classifiers" suggests that it may have the potential to be implemented in the real world in a variety of network security settings. The improved and robustness of the IDS can help organizations mitigate the risks associated with cyber threats and protect their critical assets.

This study demonstrates how the "ensemble learning" and "Stacking techniques" which uses DT, RF, ET, Extreme GB and AdaBoost in the process of DDoS attack detection is used, which is a significant contribution to the existing knowledge in the fields of ML The Stacking ensemble learning model is used to enhance the accuracy of predictions, while feature selection approaches are utilised to mitigate computing time requirements. The performance of the proposed research is executed on latest publicly available IoT datasets. Our contributions in this research paper can be summarized as follows:

- This study proposes a binary classification approach for categorizing network traffic of IoT devices into two distinct categories: normal and abnormal.
- Novel Approach: This paper proposes a unique approach for DDoS attack detection in IoT environment using stacked tree-based classifiers with average feature importance analysis and Ensemble learning methods with Step Forward Feature Selection. This approach combines Decision Tree, Random Forest, Extra Tree and XGBoost through Stacking, leveraging their individual strengths for improved predictive performance. The analysis of feature importance helps identify key factors contributing to accurate DDoS attack detection in IoT network.
- Evaluation on Real-world Datasets: Here conducted comprehensive evaluations on real-world datasets, Bot-IoT, CICIoT2023 and DS2OS. This is the first paper giving analysis on CICIoT2023 dataset with different classifiers by employ feature selection methods, such as Step Forward Feature Selection and Tree-Based Classifiers with Stacking and Feature Importance, to reduce the feature sets.
- The results show a rate exceeding 99%, demonstrating the effectiveness of proposed approach for accurately detecting and classifying DDoS attacks in IoT.

This research paper aims to explore the effectiveness of stacked tree-based classifiers in IDSs, with a focus on feature importance and performance evaluation. Through extensive experiments on benchmark datasets, proposed approach demonstrates the superior performance of the proposed approach and provide understanding of the underlying characteristics of network traffic associated with intrusions. The results highlight the potential of ensemble learning techniques in enhancing the effectiveness of IDSs, thereby contributing to the advancement of network security.

2. RELATED WORK

The IDS are necessary for the proper defence of IOT networks against DDOS attack detection. When it comes to protecting their data and systems, businesses are confronted with a myriad of challenges as a direct result of the proliferation of cybercrimes and cyber threats. Researchers have been hard at work developing novel strategies and methods to enhance the efficiency of AI based IDS to find solutions to the problems that have been identified. The purpose of this literature review is to offer a summary of the most recent research that has been conducted. In addition to this, it delves into a wide variety of subjects and approaches. Among the highlighted papers are those that focus on the investigation of IoT network traffic, as well as the application of DL models and ML algorithms. Our research included an examination of many scholarly articles that explored the use of ML and DL methodologies in constructing models for the identification and categorization of DDoS attacks. These models were developed utilising both publicly available and privately sourced datasets.

The reviews offer insightful information regarding the most recent developments, methodologies, and potential future trajectories for research in this area of study. The reviews draw attention to the inventiveness and variety that exist within the field of DDoS attack detection. The papers examine the application of a variety of methods, including optimization strategies, ML algorithms and DL frameworks, among others. These methodologies hold promise for enhancing the efficiency, robustness, and performance of systems. Furthermore, they are adaptable across various networks, encompassing healthcare systems and those within the IoT infrastructure.

Almaraz-Rivera et al. [5] worked on Bot-IoT dataset to address class imbalance issues and develop an innovative IDS employing both ML and DL models. To assess the impact of timestamp data on predictions, three distinct feature sets were employed for binary and multiclass classifications, mitigating potential feature dependencies inherited from the Argus flow data generator. The meticulous feature engineering approach resulted in an impressive average accuracy exceeding 99%. Moreover, extensive experimentation, encompassing evaluation of time performance, showcased the superiority of "DT and MLP models" in detecting DDoS and DoS attacks within IoT networks, outperforming established state-of-the-art benchmarks.

Moustafa et al. [7] proposed an ensemble intrusion detection approach for limiting harmful activities, namely botnet attacks on the HTTP, DNS and MQTT protocols often used in IoT networks. The technique develops unique statistical flow characteristics based on protocol attributes and evaluates their efficiency on the NIMS botnet and UNSW NB15 datasets. This approach employs the AdaBoost ensemble learning framework, which combines ML techniques.

Kumar et al. [8] introduces a novel IDS that leverages “fog computing” to enhance security in IoT networks. The distributed ensemble design allows the system to process data at the edge, reducing latency and conserving network bandwidth. However, the specific ML algorithms used in the ensemble are not specified. The paper presents an experimental evaluation of the system's performance in detecting anomalies and cyber threats in IoT environments. The results demonstrate the system's capability to efficiently handle large-scale IoT networks and effectively detect suspicious activities.

Thakkar and Lohiya [9] focused on the critical task of feature selection for attack classification in IoT networks. The authors conduct a comparative study of various feature selection techniques to identify informative attributes that aid in accurate attack classification. Although the specific algorithms are not explicitly mentioned, the study evaluates the performance of multiple feature selection methods and their impact on classification accuracy. The findings provide insights into the most effective feature selection techniques for improving the performance of Intrusion Detection in IoT.

Leevy et al. [10] focused on the “Bot-IoT dataset”, which is widely used for evaluating IDSs in IoT environments. The authors provide review and analysis of various datasets, including its characteristics, features, and potential applications. Although not a primary research paper, this work acts as a important resource for researchers, offering insights into the dataset's strengths and limitations for IoT security research.

Arthi and Krishnaveni [11] presented the design and development of an IoT testbed equipped with capabilities to simulate and analyze DDoS attacks for cybersecurity research. The testbed allows researchers to experiment with different attack scenarios in controlled environments. Although the paper does not explicitly discuss the algorithms or methodologies used for intrusion detection, it serves as a foundational piece in the context of developing experimental frameworks for assessing IoT network security.

Churcher et al. [12] focused on the practical application of ML algorithms for attack classification in IoT networks. The authors conduct experimental analyses, evaluating the performance of various ML techniques in accurately classifying different types of attacks. While specific algorithms are not mentioned, the research aims to determine the most effective ML methods for building robust IDSs in IoT environments.

Kumar et al. [13] proposed an “anomaly-based IDS leveraging fog computing” to enhance IoT network security. The system aims to detect anomalous activities and potential threats effectively in IoT environments. While the specific algorithms are not detailed, the study showcases the potential of combining fog computing and “anomaly-based intrusion detection” to protect IoT networks from emerging cyber threats.

Prasad and Chandra [14] introduced VMFCVD, an optimized framework that utilizes ML techniques to counter volumetric “DDoS attacks” in IoT networks. Although the specific ML algorithms are not explicitly mentioned, the paper highlights the framework's capabilities in mitigating and preventing volumetric DDoS attacks, thereby enhancing the security of IoT networks.

Kumar et al. [15] introduced a novel approach to mitigating DDoS attacks within blockchain-enabled IoT networks. Their solution involves leveraging fog computing to develop an IDS tailored to effectively identify and thwart DDoS attacks, with a focus on those targeting mining pools. The authors utilize XGBoost and Random Forests algorithms to assess the efficacy of their proposed model, employing the Bot-IoT dataset for evaluation purposes. While XGBoost demonstrates superior performance in binary attack detection, RF exhibits greater efficacy in multi-attack detection, with the added advantage of shorter training and testing durations.

Douiba et al. [16] came up with a better IDS that uses methods called GB and DT. The idea is to make the IDS more accurate and reliable by using these advanced ML techniques. These techniques are made to handle the difficulties and problems of IoT security better.

Kareem et al. [17] and Kim et al. [18] centered their study on refining feature selection for IoT intrusion detection through hybrid metaheuristic algorithms. Their model aims to streamline the selection of relevant features to bolster the accuracy of IDSs within IoT settings. The research showcases experimental findings that underscore the efficacy of employing hybrid metaheuristic strategies in elevating detection performance.

A suggested IDS by Roopak et al. [19] is intended to defend against DDoS attacks on IoT networks. The system makes use of a number of strategies; however, the precise methods are not disclosed. The study emphasises the necessity for specialised defence methods as well as the growing threat posed by DDoS attacks on IoT devices. In order to maintain the stability and availability of IoT services even in the face of attacks, the suggested IDS seeks to identify and counteract DDoS attacks.

Shafiq et al.'s [20] method addressed the issue of suitable feature selection. The author proposed CorrAUC feature selection techniques based on the "wrapper technique" and employs the AUC measure to filter and pick the important features. This approach was tested using four different ML algorithms and the "Bot-IoT dataset". On average, the proposed strategy can deliver results of more than 96%.

An improved feature selection approach for intrusion detection was put out by Alghanam et al. [21], with a particular emphasis on the use of ensemble learning to detection. These models have quite simple architectures and datasets.

The reviewed literature encompasses various aspects of intrusion detection in IoT networks, including algorithmic approaches, feature selection techniques, and experimental evaluations. The studied literature shows how important and useful it could be to use Artificial intelligence methods, ML algorithms, and DL algorithms to improve the accuracy and of IDS and find possible risks and weaknesses in IoT settings.

The literature review presents an overview of the latest studies on the IDS industry. It covers various topics, such as ML algorithms, network traffic analysis, DL models for IoT networks, and feature selection. The papers highlight the use of different methodologies and techniques to improve the systems' efficiency, resilience, and accuracy. The reviewed papers discuss the utilization of various techniques, such as optimization approaches and feature grouping, to improve the efficiency of detecting intrusions and reduce the overall feature space's dimensionality. ML models and algorithms have been proven to accurately identify complex anomalies and patterns in network traffic data.

These studies highlight the need to consider the diverse requirements of different networks, such as those involving wireless sensors and privacy preservation. The reviews provide valuable insight into the current state of the art and the future directions of intrusion detection. The papers exhibited how different approaches and techniques can address the security challenges of networks. The findings of these studies have the potential to contribute to the development of new and improved IDS that can effectively address the evolving threats and attacks in IoT.

3. METHODOLOGY

3.1 Dataset

The "Bot-IoT" [22, 23] dataset is a real-world simulated dataset created in the "Cyber Range Lab at the UNSW Canberra Cyber Centre". It is composed of network traffic within an interconnected IoT environment. The Bot-IoT dataset exhibits various kinds of intrusion and normal traffic patterns and is large enough to cover both attacks and normal network traffic. The dataset has following attacks:

- DDoS - TCP, UDP, or HTTP as the protocol.
- “Denial of Service” (DoS): Attack relies on TCP, UDP, or HTTP.
- Collecting information.
- Theft of personal information.

All 46 features were selected in the dataset by combining 4 files of 5% dataset consisting of 3,668,522 records. Samples related to DDoS and Normal were filtered from the combined dataset for training and testing purpose.

The DS2OS [24] dataset is designed to investigate the anomalies and dynamics observed in a distributed system. The collection of data includes event sequences, logs, and other attributes that can be used to identify malicious activities or unusual operations in the system [25]. It is comparable in size to the Bot-IoT dataset having 355902 records, 11 features and has classes that correspond to both normal and anomalous activities.

CICIoT2023 [26] is a realistic IoT attack dataset made with a large layout made up of many real IoT devices and using IoT devices as both offenders and victims. The information comes from 33 attacks that fall into 7 groups. We have filtered samples related to Normal and DDoS with 477,374 records and 47 columns.

The Bot-IoT, CICIoT2023 and DS2OS datasets serve a vital role in the design, development, and testing of intrusion detection methods. They give researchers the necessary information to create and benchmark approaches and algorithms. The former mainly pertains to system-level behaviors. The availability of such datasets enables researchers to develop better and more accurate methods for detecting attacks on the IoT and complex systems.

3.2 Data pre-processing

The process of pre-processing is essential for creating data suitable for use by an IDS. It involves converting raw information into a format that can be utilized by the algorithms. The pre-processing process aims to address missing or distorted data, fix anomalies, and normalize features.

3.2.1 Handling missing values

One of the fundamental and basic phases in any data pre-processing techniques is dealing with the missing values. Many statistical variables are employed in a variety of data pre-processing procedures to substitute for the missing value or instances, such as the “mean and standard deviation”. All datasets are available in clean versions, which are free of any missing values and have minimal missing values. The minimal missing values, NaN values are deleted from dataset.

3.2.2 Label encoding

Many of the used datasets' features are string values or special letters like saddr, daddr, or sport, which are not quite right for many ML models. Python has encoding methods like one-hot encoding, which is best for categorical columns, and label encoding, which is best for binary columns. Label encoder from sklearn, preparation for binary and one hot encoder for categorical data to replace string values with numeric replacement values based on the classes of the data in that feature. In IDS, it is common to have categorical variables like “saddr”, “daddr”, “sport”, “dport”, “state”, “flgs”, “category”, or “protocol.” For example, label encoding can convert "Normal" and "DDoS" into 0 and 1, respectively. The conversion can be represented mathematically as follows:

Let X_{cat} = “Categorial feature”, X_{enc} = “corresponding numerical representation”. The label encoding expressed as Eq. (1).

$$X_{enc} = LabelEncoding(X_{cat}) \tag{1}$$

3.2.3 Normalization - Min-Max normalization

The Min-Max method normalizes certain numerical features to a certain range, which helps prevent them from dominating others in training. It can be used to calculate the Min-Max for feature *X* as Eq. (2).

$$X_{normalized} = \frac{X - min(X)}{max(X) - min(X)} \tag{2}$$

where, *min*(*X*) and *max*(*X*)=“minimum and maximum value of feature *X*”.

3.2.4 Standard scaling (Standardization)

Standard scaling or standardisation converts numerical characteristics to have a mean of zero and a standard deviation of one. It works well when the features vary in scale. The formula for standard scaling is given by Eq. (3):

$$X_{standardized} = \frac{X - mean(X)}{std(X)} \tag{3}$$

Bot-IoT datasets have different features that are not set up in a way that makes them easy to use with models. IP numbers are used to describe things like the source and target IP addresses. After label encoding, the numbers need to be normalised in a good way. Standard Scaler is used in the suggested model to standardise the data that comes in.

3.3 Data balancing

Peterson et al. [27] delved into the "Bot-IoT dataset," a widely utilized resource for assessing IDSs within IoT settings. Their work entails an exhaustive examination and appraisal of the dataset, encompassing its attributes, functionalities, and prospective uses. While not constituting a primary research contribution, this study serves as a pivotal asset for scholars, furnishing discernments into the dataset's advantages and constraints for IoT security investigations. The combination of both methods helps balance the class distribution. Mathematically, SMOTE generates synthetic samples by interpolating between neighboring instances of the minority class. Suppose minority $X_{minority}$ represents the feature matrix of the minority class, and synthetic $X_{synthetic}$ represents the synthetic samples generated. Then, the SMOTE operation can be defined as Eq. (4).

$$X_{synthetic} = X_{minority} + \alpha \times \left(Neighbor - X_{minority}\right) \tag{4}$$

where, α= “random value between 0 and 1”, Neighbor = “randomly chosen neighboring instance of the minority class”.

The Bot-IoT dataset is imbalanced [19] having less Normal; sample and more attack samples. SMOTE is used to balance the dataset.

Table 1. Feature selection details

Method	Dataset	Original Features	Manual Feature Select	Features Selected Method	Features Selected	Feature Names
Ensemble Voting	Bot-IoT	46	41	SFS	18	[“stime, bytes, pkts, state_number, seq, ltime, dur, stddev, mean, sum, max, min, spkts, sbytes, dpkts, dbytes, srate, rate”]
Ensemble Voting	CICIoT2023	47	38	SFS	18	[“Header_Length, Srate, Drate, fin_flag_number, syn_flag_number, ack_count, fin_count, urg_count, HTTPS, DNS, UDP, ARP, LLC, Std, Tot size, IAT, Number, Weight”]
Stacking	Bot-IoT	46	41	Feature Importance	9	[“category_enc, ltime, stime, N_IN_Conn_P_DstIP, dport_enc, N_IN_Conn_P_SrcIP, seq, max, TnP_PerProto”]
Stacking	DS2OS	13	13	Feature Importance	10	[“dloc, sid, sloc, ana, dtype, stype, saddr, daddr”]

3.4 Feature selection (FS)

Various methods are employed to identify the most relevant features from datasets. In the context of Ensemble Voting, Sequential Forward Selection (SFS) was utilized for both Bot-IoT and CICIoT2023 datasets, resulting in 18 selected features for each. These features include parameters like packet count, byte count, and state information. In contrast, Stacking employed Feature Importance to distill features, yielding 9 and 10 significant features for Bot-IoT and DS2OS datasets, respectively. Features with most importance selected encompass temporal aspects, network protocols, and connection attributes, offering valuable insights into the underlying data dynamics shown in Table 1.

4. PROPOSED APPROACH

4.1 DDoS attack detection using Stacking approach and average feature importance

The proposed approach begins with data pre-processing to ensure the quality and suitability of the input data for subsequent analysis. This includes several steps such as checking for missing values in the dataset and applying appropriate techniques to handle them, such as imputation or removal. Furthermore, label encoding is used to encode categorical information and translate them into numerical representations. To normalize the numerical features and bring them to a comparable scale, the data is standardized using a standard scaler. Additionally, as seen in Figure 1, SMOTE is used with the average feature significance to create artificial instances of the minority class and balance the distribution of classes in order to address the problem of class imbalance.

4.1.1 Dimensionality reduction

Dimensionality reduction is a critical step to overcome the curse of dimensionality and improve the IDSs effectiveness and efficiency. The proposed approach employs feature mapping techniques to transform the original feature space into a lower-dimensional space, aiming to capture the most informative features for intrusion detection. The approach considers the average feature importance derived from the ensemble of ML classifiers to identify and retain the most relevant features.

4.1.2 Data partitioning

By dividing the dataset into train and test subsets, the accuracy of the IDS performance evaluation is ensured. 70% of the data are utilised to train the ML classifiers and the remaining 30% are used to assess the classifiers' performance in the suggested approach. When working with the "Bot-IoT dataset," we applied the stratified k-fold cross-validation strategy to improve our model's performance and avoid overfitting.

4.1.3 ML classifiers

The proposed approach utilizes a set of powerful tree [28] based classifiers, including “DT, ETC, RF, and XGBoost” [29-31]. These classifiers are trained on the training subset of the data, leveraging their ability to capture complex patterns and anomalies in the network traffic features. The classifiers are individually optimized and tuned using RandomizedSearchCV to achieve their best performance in DDoS attack detection.

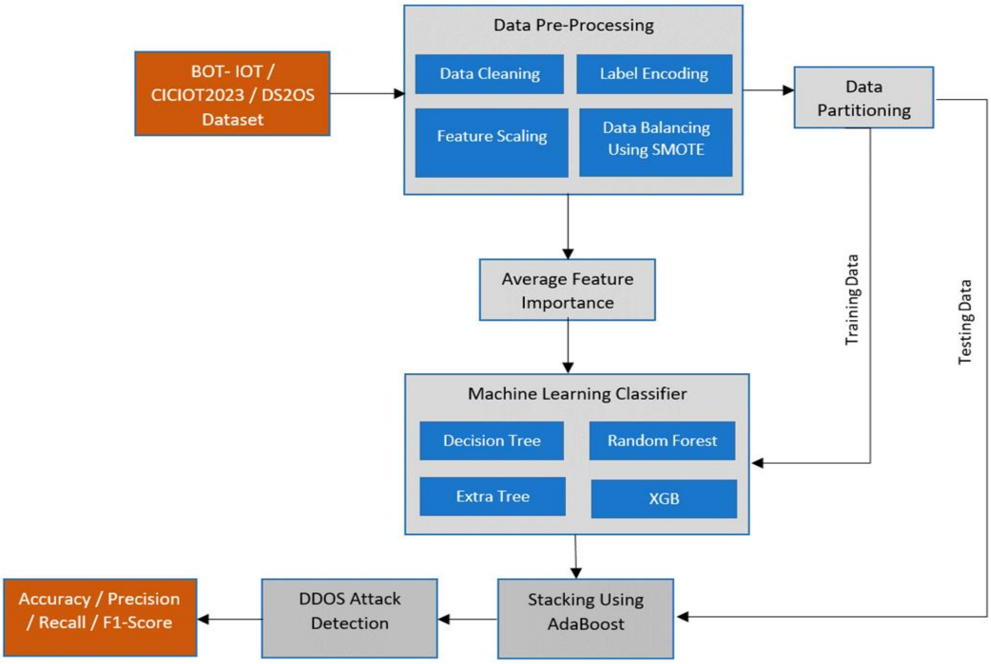

Figure 1. Proposed framework for “DDoS attack” detection using Stacking with average feature importance

4.1.4 Stacking using AdaBoost

Stacking is an ensemble method used to merge predictions from various ML classifiers in order to enhance the overall predictive accuracy. The proposed method utilizes AdaBoost as the Stacking algorithm, which assigns weights to the predictions of each base classifier according to their individual performance. The weighted predictions are combined to create the final ensemble prediction, which is then evaluated to determine its effectiveness in classifying intrusions.

4.2 DDoS attack detection using ensemble approach and step forward feature selection

This proposed approach aims to enhance the performance of IDSs through a comprehensive pipeline that includes data pre-processing, dimensionality reduction, data partitioning, and the utilization of ML classifiers with ensemble voting as shown in Figure 2. The following sections outline each step in detail:

4.2.1 Feature selection

FS is crucial for improving the efficiency of ML models by reducing overfitting, decreasing computational learning time, and ultimately enhancing predictive accuracy. It is a vital technique for optimizing model performance in different applications.

The proposed method uses the “Step Forward Feature Selection” (SFS) technique to determine the most important features for intrusion detection based on their individual effectiveness. Hyperparameter tuning is conducted to enhance the performance of the FS technique.

4.2.2 Data partitioning

Stratified k-fold cross-validation represents an advancement over standard k-fold cross-validation, tailored specifically for classification tasks. Unlike traditional k-fold cross-validation where splits are entirely random, this technique maintains the proportion of target classes within each fold consistent with the original dataset. To tackle the class imbalance inherent in the Bot-IoT dataset, we've adopted stratified k-fold cross-validation. By employing this method instead of a simple train-test split, we ensure a more representative sampling and mitigate the risk of overfitting. This approach is crucial for accurately evaluating the performance of IDS.

4.2.3 ML classifiers

The proposed approach utilizes a set of ML classifiers, including “DT, RF, ETC and XGBoost” [29-32].

4.2.4 Ensemble Voting Classifier

To leverage the strengths of multiple classifiers, the proposed approach utilizes an Ensemble Voting Classifier. The predictions of the individual ML classifiers are combined using ensemble techniques, such as majority voting or weighted voting, to generate the final ensemble prediction. The Ensemble Voting Classifier is evaluated using various performance evaluation metrics to assess its effectiveness in intrusion detection.

4.2.5 Binary classification

Intrusion detection involves binary classification, where normal and DDoS instances are identified and categorized. The proposed approaches evaluate the performance of the system using various performance evaluation metrics, including “recall, accuracy, precision, F1-score”. These metrics provide a comprehensive assessment of the system’s ability to correctly classify intrusions and distinguish them from normal network traffic.

By employing this proposed approach, the IDS aims to achieve high accuracy and effectiveness in identifying and classifying intrusions, thereby enhancing the security and resilience of computer networks.

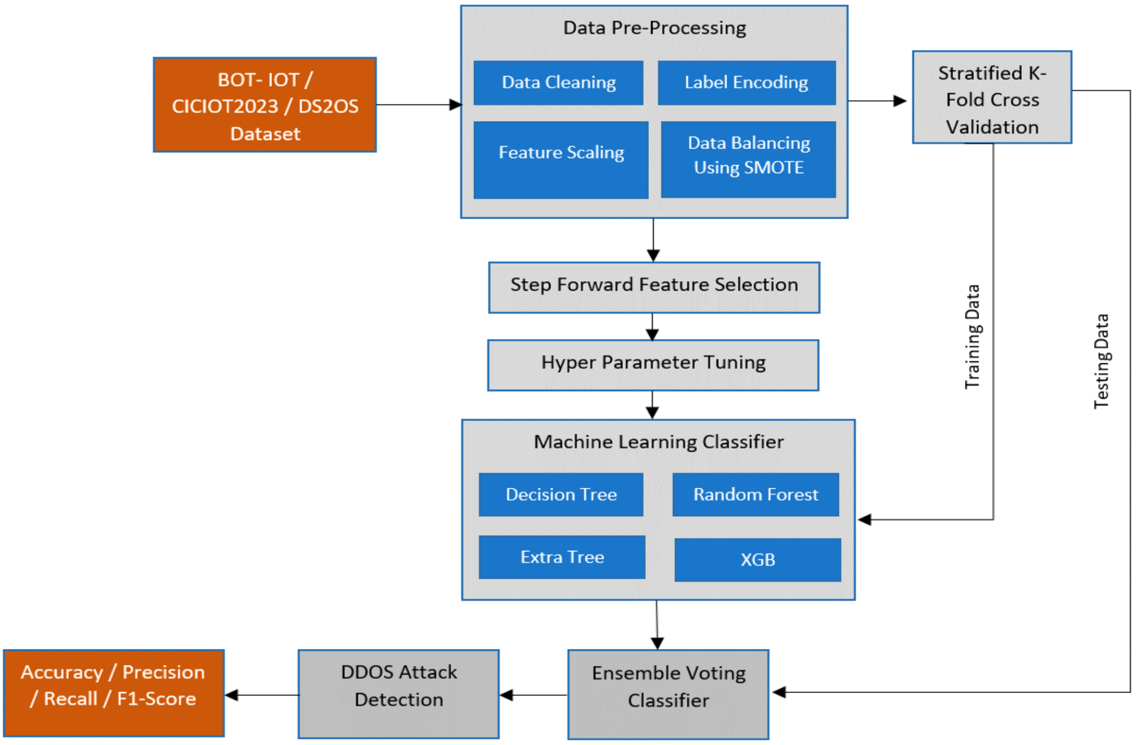
BOT- IOT / CICIOT2023 / DS2OS Dataset
Data Pre-Processing
Data Cleaning
Label Encoding
Feature Scaling
Data Balancing Using SMOTE
Stratified K-Fold Cross Validation
Step Forward Feature Selection
Hyper Parameter Tuning
Training Data
Testing Data
Machine Learning Classifier
Decision Tree
Random Forest
Extra Tree
XGB
Accuracy / Precision / Recall / F1-Score
DDOS Attack Detection
Ensemble Voting Classifier

Figure 2. Proposed framework for DDoS attack detection using Ensemble Voting Classifier with SFFS

Table 2. Hyper parameter tuning with RandomSearchCV

Model	Name of Parameters	Params Tuned Using RandomizedSearchCV

RF	'max_depth': [1,3,5,7,9, None],
	'max_features': ['auto', 'sqrt','log20', None],
	'min_samples_leaf': [1,3,5]
	'min_samples_split': range (2,10),
	'n_estimators': [100,120,150]
		n_estimators=150,
		min_samples_split=3,
		min_samples_leaf=5,
		max_features='auto',
		max_depth=5

DT	'max_depth': [2, 3, 5, 10, 20],
	'min_samples_leaf': [5, 10, 20, 50, 100],
	'criterion': ["gini", "entropy"]
		min_samples_leaf=100,
		max_depth=10,
		criterion='entropy'

XGB	'n_estimators': range (100,1200),
	'max_depth': range (1,11),
	'learning_rate': [1 * 10**(-3), 1 * 10**(-2), 0.1, 0.5, 1],
	'subsample': [0.05,1.01],
	'min_child_weight': range (1,21)
		subsample=0.05,
		n_estimators=419,
		min_child_weight=9,
		max_depth=9,
		learning_rate=0.1

ET	‘n_estimator’: [100,200,500],
	'min_samples_leaf': [5, 10, 20],
	‘max_features’: [2,3,4]
		n_estimators=200
		max_features=3
		min_sample+leaf=5

4.2.6 Hyper parameter tuning

Tuning hyper parameters is a process that can improve the performance of learning models and help prevent IDSs from being inaccurate. It involves identifying the optimal combinations of these parameters that are not derived from the data to reduce overfitting, leading to a better generalization of the model. The proposed approach involves tuning the parameters of various ML models, such as “XGBoost, decision tree and Random Forest” using RandomizedSearchCV. While tuning hyperparameters “max_depth, min_samples_leaf, min_sample_split, n_estimators, criterion, max_features, learning rate and min_child_weight” have been optimized to increase performance parameters as shown in Table 2.

Description of Parameters

max_depth= “longest Path from the root node to leaf node”.

max_features = “maximum number of features in every tree”.

min_samples_leaf = "The leaf node should contain the minimum number of samples following a node split."

min_samples_split = “To separate it, at least several observations were needed in each node”.

n_estimators = “several trees required for the Random Forest”.

criterion = “function to measure the quality of a split”.

5. RESULTS AND DISCUSSION

5.1 Experimental environment

The work is analysed in a cloud-based setting called Google Colab. Through Google Collaboratory. The models are trained using the CPUs, GPUs in this context. Python 3.11.4 is used to write our work. ML methods have been put into place and trained with the help of Scikit learn 1.3.0.

5.2 Performance parameters

Accuracy, Recall, Precision and F1 Score are used to measure how well a classifier works. True-Positive (TP) means that a true positive is expected to be true. “False Positive” (FP) is when a bad event is mistakenly thought to be positive. “True-Negative” (TN) is an instance of negative that is expected to be negative. “False Negative” (FN) is when a positive case is thought to be negative [33].

Accuracy: Accuracy can be measured by how many guesses were right out of all the predictions made.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

Precision: Precision measures the positive prediction made by the model.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Recall: Recall measures the relevant data points that were correctly identified by the model.

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

F1 Scores: F1 - Measure computes the harmonic mean of “precision and recall”.

$$F1\ Score = 2*\frac{(Recall*Precision)}{(Recall+Precision)} \tag{8}$$

The proposed approach has been evaluated using the latest IoT datasets, environmental setup, and performance metrics.

Table 3. Evaluation parameter of various proposed approach

	Algorithm	Dataset (s)	Features Selected	ACC
Almaraz-Rivera et al. [5]	RF DT LSVM	Bot-IoT	15	99.97 99.94 98.40
Shafiq et al. [20]	DT NB RF SVM	Bot-IoT	5	99.99 97.5 99.98 97.8
Kumar et al. [15]	KNN RF XGBoost	Bot-IoT	10	85.92 99.99 99.99
Proposed Work	RF DT ET XGBoost Ensemble-Voting	Bot-IoT	18	100 100 100 100 99.95
Proposed Work	RF DT ET XGoostT Ensemble - Voting	CICIoT 2023	18	99.09 99.12 98.84 99.06 99.30
Proposed Work	DT RF ET XGBoost Stacking using XGBoost	Bot-IoT	9	99.8 99.89 99.90 99.41 99.92
Proposed Work	DT RF ET XGBoost Stacking using XGBoost	DS2OS	10	97.58 97.58 97.58 98.7 98.88

Table 4. Total train and test time for CICIoT2023 dataset

	Total Train and Test Time for CICIoT2023	
Classifiers	**Before Feature Importance-CICIoT2023 (sec)**	**After Feature Importance-CICIoT2023 (sec)**
DT	3.59	2.281
RF	17.57	13.718
ET	8.517	7.77
XGBoost	9.05	6.43
Stacking using XGBoost	8.19	7.06

5.3 Result analysis

In the result discussion, the performance of the proposed approach using different techniques and algorithms for intrusion detection on the “Bot-IoT, CICIoT2023 and DS2OS” datasets are analyzed (Table 3). Table 4 shows the training and testing time in total required for CICIoT2023 dataset before and after average feature importance. Figures 3-5 show the time comparison graphs of various algorithms with and without feature selection techniques on multiple datasets.

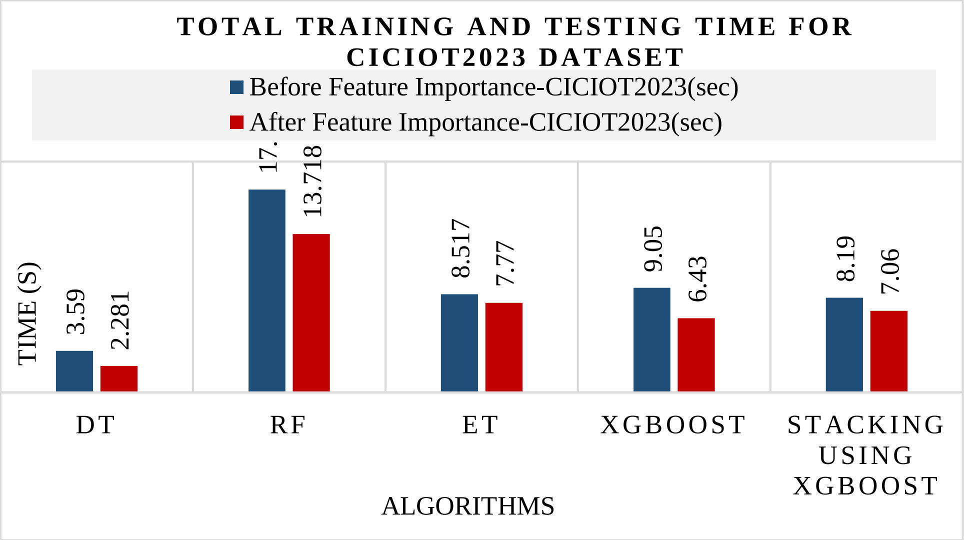

Figure 3. Total training and testing time for CIC-IOT2023 dataset with and without feature importance

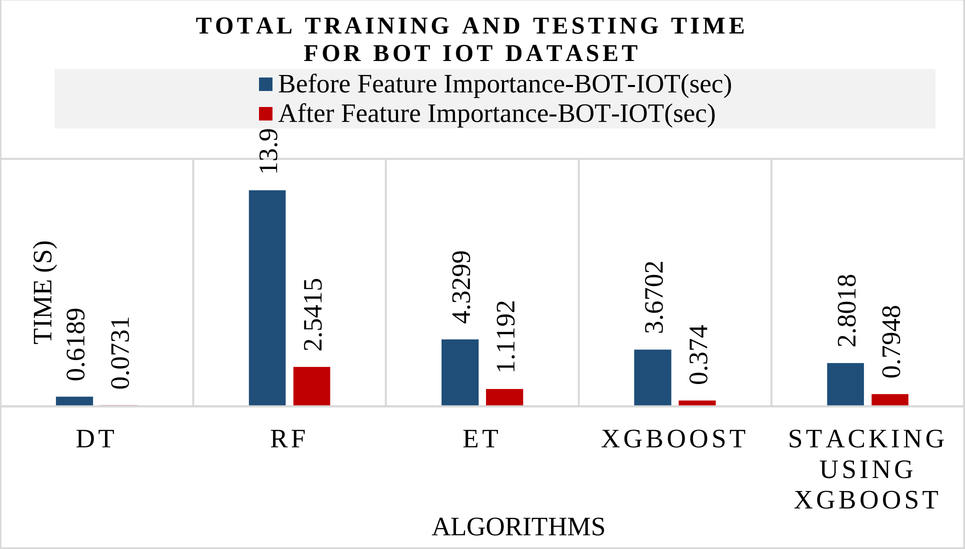

Figure 4. Total training and testing time for Bot IoT dataset with and without feature importance

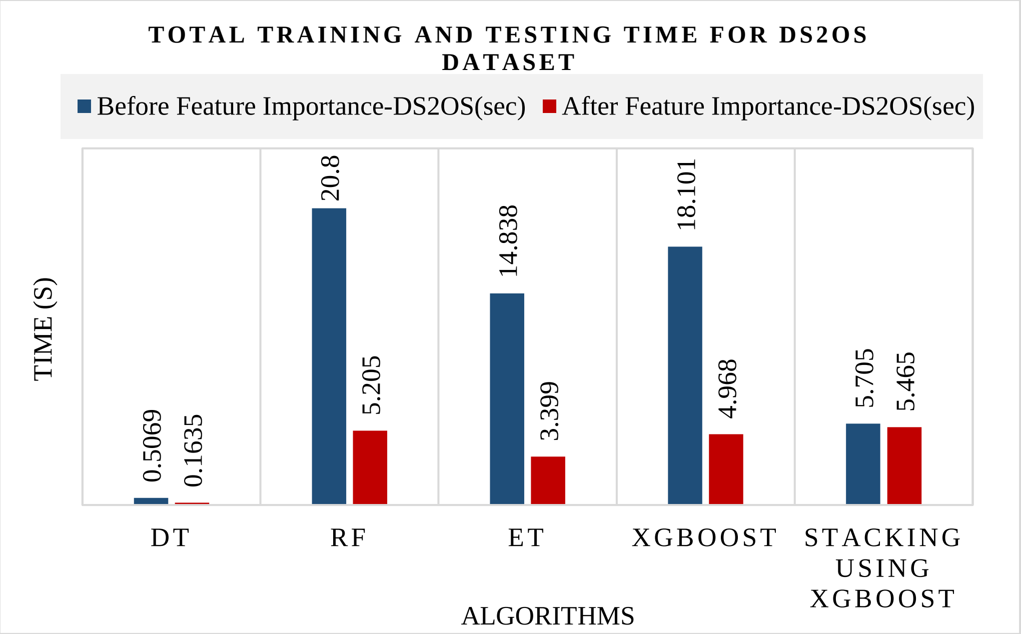

Figure 5. Total training and testing time for DS2OS dataset with and without feature importance

5.3.1 Proposed-step forward feature selection method (Wrapper) with Ensemble Voting Classifier

When applied to the Bot-IoT dataset, all three algorithms (RF, DT, XGBOOST) achieved a remarkable accuracy of 100% using the Step Forward Feature Selection method. The Ensemble Voting Classifier, which combines the predictions of these algorithms, achieved a slightly lower accuracy of 99.95% but still demonstrated excellent performance in classifying intrusions. The performance metrics are shown in Figure 6.

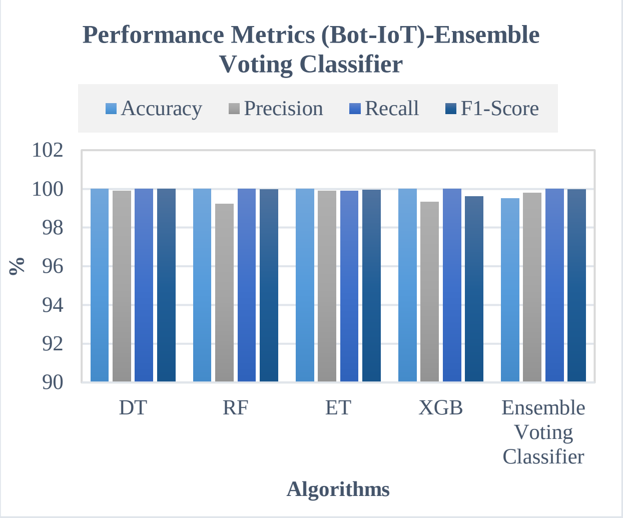

Figure 6. Performance metrics (Bot-IoT)-Step Forward Feature Selection method (Wrapper)

5.3.2 Proposed-tree-based classifiers with Stacking and feature importance (CICIoT2023 dataset)

When applied to the new real time CICIoT2023 dataset, the DT, RF, ET, and "XGBoost" achieved high accuracies ranging from 99.09% to 99.12%. The Stacking approach using XGBoost as the base classifier also yielded a high accuracy of 99.30%. These results demonstrate the effectiveness of the proposed approach in accurately classifying DDoS attacks in the Bot-IoT dataset. The performance metrics are shown in Figure 7. DT requires less time as compared to other classifiers.

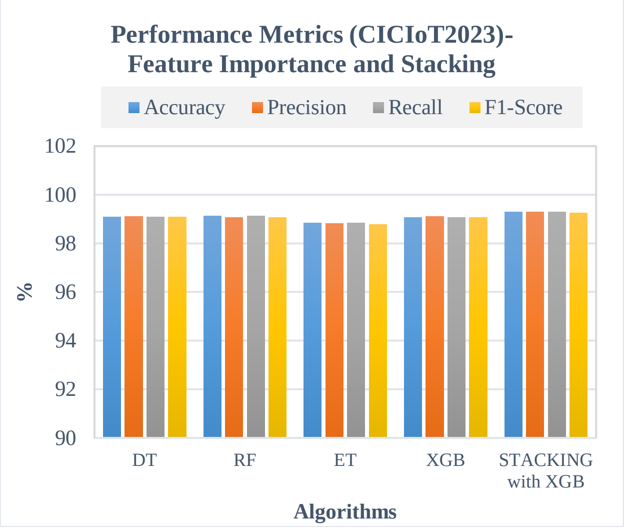

Figure 7. Performance Metrics (CICIoT2023)-tree based classifiers with stacking and feature importance

5.3.3 Proposed-tree-based classifiers with stacking and feature importance (Bot-IoT dataset)

When applied to the "Bot-IoT dataset", the "DT, RF, ETC and XGBoost" achieved high accuracies ranging from 99.41% to 99.90%. The Stacking approach using XGBoost as the base classifier also yielded a high accuracy of 99.92%. The results show that the proposed method is effective in accurately categorizing intrusions in the "Bot-IoT dataset". Table 5 shows the training and testing time in total required for Bot-IoT dataset before and after average feature importance. The performance metrics are shown in Figure 8. The performance metrics (DS2OS)-tree based classifiers with stacking and feature importance is shown in Figure 9. The confusion matrix for multiclass classification for "DS2OS" dataset is shown in Figure 10. DT requires less time as compared to other classifiers.

Table 5. Total training and testing time for Bot IoT dataset

	Total Training and Testing Time for Bot IoT	
Classifiers	Before Feature Importance-Bot-IoT (sec)	After Feature Importance-Bot-IoT (sec)
DT	0.6189s	0.0731s
RF	13.964s	2.5415s
ET	4.3299s	1.1192s
XGBoost	3.6702s	0.3740s
Stacking using XGBoost	2.8018s	0.7948s

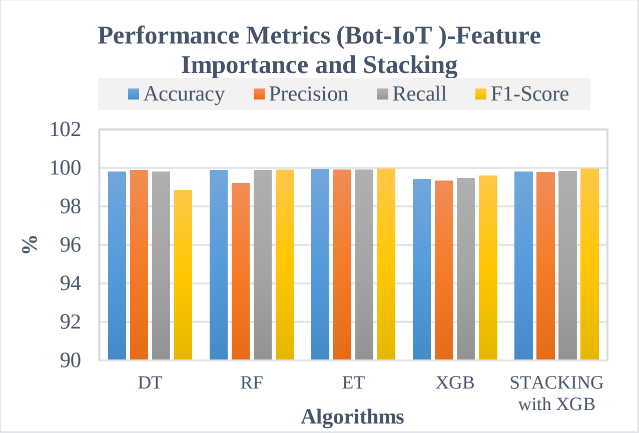

Figure 8. Performance Metrics (Bot-IoT)-tree based classifiers with stacking and feature importance

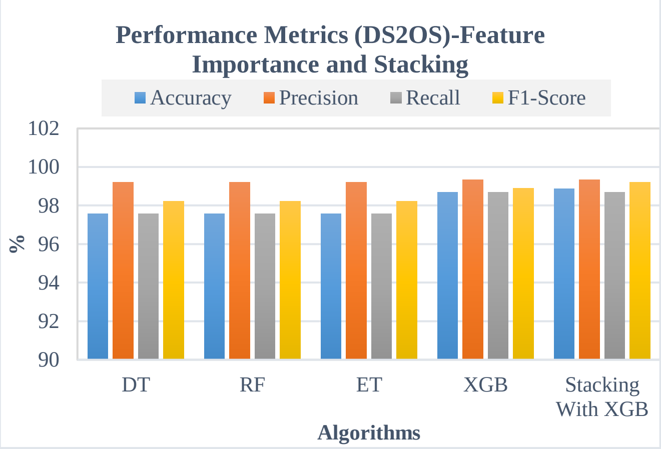

Figure 9. Performance metrics (DS2OS)-tree based classifiers with stacking and feature importance

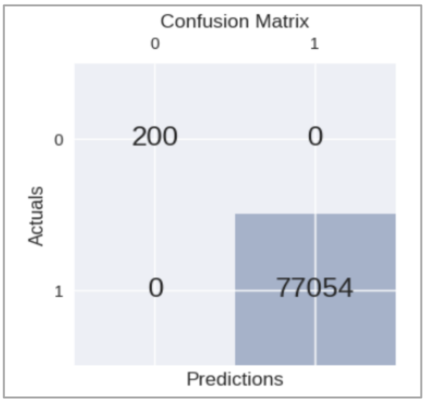

Figure 10. Confusion matrix for (Bot-IoT)-tree based classifiers with Stacking and feature importance

5.3.4 Proposed-tree-based classifiers with Stacking and feature importance (DS2OS dataset)

When applied to the DS2OS dataset, the same set of classifiers (DT, RF, ET, XGBoost) achieved accuracies ranging from 97.58% to 98.7%. The Stacking approach using XGBoost as the base classifier achieved an accuracy of 98.88%. These results indicate that the proposed approach is capable of achieving high accuracies in classifying intrusions in the DS2OS dataset as well. Table 6 shows the training and testing time in total required for DS2OS dataset before and after average feature importance. DT requires less time as compared to other classifiers. The confusion matrix for multiclass classification for DS2OS dataset is shown in Figure 11.

Table 6. Total training and testing time for DS2OS dataset

	Total Training and Testing Time for DS2OS	
Classifiers	Before Feature Importance-DS2OS (sec)	After Feature Importance-DS2OS (sec)
DT	0.5069s	0.1635s
RF	20.845s	5.205s
ET	14.838s	3.399s
XGBoost	18.101s	4.968s
Stacking using XGBoost	5.705s	5.465s

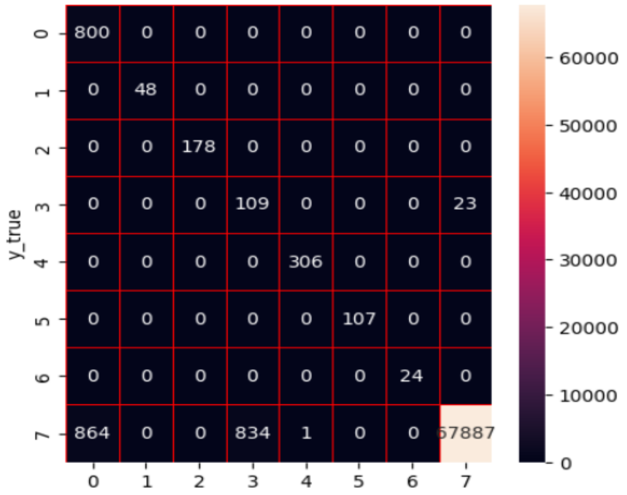

Figure 11. Confusion matrix for (DS2OS)-tree based classifiers with stacking and feature importance

The proposed method was able to classify and detect DDoS attack intrusions in the Bot-IoT, CICIoT2023 datasets. We have used total 46 features of the dataset by considering all the traffic. It was able to achieve high accuracies and shows the potential of this approach in the real world. The Step Forward Feature Selection method and the tree-based classifiers exhibited their ability to extract important features and improve their accuracy.

6. CONCLUSION AND FUTURE SCOPE

In this research study, we proposed a comprehensive approach for DDoS attack detection in IoT using feature selection, ML classifiers, and ensemble techniques. The results demonstrate the effectiveness of the proposed approach in accurately detecting and classifying intrusions in the Bot-IoT and DS2OS datasets. The Step Forward Feature Selection method (Wrapper) showcased its ability to select relevant features and achieved outstanding accuracies of 100% when applied to the Bot-IoT dataset. The tree-based classifiers (DT, RF, ET, XGBoost) with Stacking and feature importance demonstrated high accuracies of ~ 97.0% to ~ 99.0% on multiple datasets. The Ensemble Voting Classifier and Stacking using XGBoost as the base classifier further improved the overall performance. The achieved accuracies demonstrate its effectiveness in accurately identifying and classifying intrusions.

Future research can build upon this work to explore new hybrid feature selection and feature extraction methods with hyper parameter optimization and DL techniques, evaluate on diverse datasets, intend to experiment with advanced ensemble techniques, focus on new real-time datasets for analysis. CICIoT2023 is a new publicly available dataset and need to be explored in future using various hybrid feature selection and metaheuristic optimization algorithms for feature extraction.

REFERENCES

[1] Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M. (2015). Internet of Things: A survey on enabling technologies, protocols, and applications. IEEE Communications Surveys & Tutorials, 17(4): 2347-2376. https://doi.org/10.1109/COMST.2015.2444095

[2] Gantz, J., Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the Future, 2007(2012): 1-16.

[3] Vailshery, L.S. (2023). Number of IoT connected devices worldwide 2019-2023, with forecasts to 2030. Statista, accessed on 10 Nov. 2023.

[4] Rupa Devi, T., Badugu, S. (2019). A review on network intrusion detection system using machine learning. In International Conference on E-Business and Telecommunications, Jordan, pp. 598-607. https://doi.org/10.1007/978-3-030-24318-0_69

[5] Almaraz-Rivera, J.G., Perez-Diaz, J.A., Cantoral-Ceballos, J.A. (2022). Transport and application layer DDoS attacks detection to IoT devices by using machine learning and deep learning models. Sensors, 22(9): 3367. https://doi.org/10.3390/s22093367

[6] Soe, Y.N., Feng, Y., Santosa, P.I., Hartanto, R., Sakurai, K. (2020). Towards a lightweight detection system for cyber-attacks in the IoT environment using

corresponding features. Electronics, 9(1): 144. https://doi.org/10.3390/electronics9010144

[7] Moustafa, N., Turnbull, B., Choo, K.K.R. (2018). An ensemble intrusion detection technique based on proposed statistical flow features for protecting network traffic of Internet of Things. IEEE Internet of Things Journal, 6(3): 4815-4830. https://doi.org/10.1109/JIOT.2018.2871719

[8] Kumar, P., Gupta, G.P., Tripathi, R. (2021). An ensemble learning and fog-cloud architecture-driven cyber-attack detection framework for IoMT networks. Computer Communications, 166: 110-124. https://doi.org/10.1016/j.comcom.2020.12.003

[9] Thakkar, A., Lohiya, R. (2023). Attack classification of imbalanced intrusion data for IoT network using ensemble-learning-based deep neural network. IEEE Internet of Things Journal, 10(13): 11888-11895. https://doi.org/10.1109/JIOT.2023.3244810

[10] Leevy, J.L., Hancock, J., Khoshgoftaar, T.M., Peterson, J.M. (2022). IoT information theft prediction using ensemble feature selection. Journal of Big Data, 9(1): 6. https://doi.org/10.1186/s40537-021-00558-z

[11] Arthi, R., Krishnaveni, S. (2021). Design and development of IoT testbed with DDoS attack for cyber security research. In 2021 3rd International Conference on Signal Processing and Communication (ICPSC), Coimbatore, India, pp. 586-590. https://doi.org/10.1109/ICSPC51351.2021.9451786

[12] Churcher, A., Ullah, R., Ahmad, J., Ur Rehman, S., Masood, F., Gogate, M., Buchanan, W.J. (2021). An experimental analysis of attack classification using machine learning in IoT networks. Sensors, 21(2): 446. https://doi.org/10.3390/s21020446

[13] Kumar, P., Gupta, G.P., Tripathi, R. (2021). Toward design of an intelligent cyber attack detection system using hybrid feature reduced approach for IoT networks. Arabian Journal for Science and Engineering, 46(4): 3749-3778. https://doi.org/10.1007/s13369-020-05181-3

[14] Prasad, A., Chandra, S. (2022). VMFCVD: An optimized framework to combat volumetric DDoS attacks using machine learning. Arabian Journal for Science and Engineering, 47(8): 9965-9983. https://doi.org/10.1007/s13369-021-06484-9

[15] Kumar, R., Kumar, P., Tripathi, R., Gupta, G.P., Garg, S., Hassan, M.M. (2022). A distributed intrusion detection system to detect DDoS attacks in blockchain-enabled IoT network. Journal of Parallel and Distributed Computing, 164: 55-68. https://doi.org/10.1016/j.jpdc.2022.01.030

[16] Douiba, M., Benkirane, S., Guezzaz, A., Azrour, M. (2023). An improved anomaly detection model for IoT security using decision tree and gradient boosting. The Journal of Supercomputing, 79(3): 3392-3411. https://doi.org/10.1007/s11227-022-04783-y

[17] Kareem, S.S., Mostafa, R.R., Hashim, F.A., El-Bakry, H.M. (2022). An effective feature selection model using hybrid metaheuristic algorithms for IoT intrusion detection. Sensors, 22(4): 1396. https://doi.org/10.3390/s22041396

[18] Kim, Y.E., Kim, Y.S., Kim, H. (2022). Effective feature selection methods to detect IoT DDoS attack in 5G core network. Sensors, 22(10): 3819. https://doi.org/10.3390/s22103819

[19] Roopak, M., Tian, G.Y., Chambers, J. (2020). An intrusion detection system against DDoS attacks in IoT networks. In 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, pp. 0562-0567. https://doi.org/10.1109/CCWC47524.2020.9031206

[20] Shafiq, M., Tian, Z., Bashir, A.K., Du, X., Guizani, M. (2020). CorrAUC: A malicious Bot-IoT traffic detection method in IoT network using machine-learning techniques. IEEE Internet of Things Journal, 8(5): 3242-3254. https://doi.org/10.1109/JIOT.2020.3002255

[21] Alghanam, O.A., Almobaideen, W., Saadeh, M., Adwan, O. (2023). An improved PIO feature selection algorithm for IoT network intrusion detection system based on ensemble learning. Expert Systems with Applications, 213: 118745. https://doi.org/10.1016/j.eswa.2022.118745

[22] Zeeshan, M., Riaz, Q., Bilal, M.A., Shahzad, M.K., Jabeen, H., Haider, S.A., Rahim, A. (2021). Protocol-based deep intrusion detection for DOS and DDoS attacks using unsw-nb15 and bot-IoT data-sets. IEEE Access, 10: 2269-2283. https://doi.org/10.1109/ACCESS.2021.3137201

[23] Koroniotis, N., Moustafa, N. (2021). The Bot-IoT Dataset | UNSW Research. Retrieved from https://research.unsw.edu.au/projects/bot-iot-dataset.

[24] Aubet, F.X., Pahl, M. (2018). DS2OS traffic traces. https://www.kaggle.com/datasets/francoisxa/ds2ostraffictraces/data, accessed on Apr. 23, 2022.

[25] Pahl, M.O., Aubet, F.X. (2018). All eyes on you: Distributed Multi-Dimensional IoT microservice anomaly detection. In 2018 14th International Conference on Network and Service Management (CNSM), Rome, Italy, pp. 72-80.

[26] Neto, E.C.P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., Ghorbani, A.A. (2023). CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. Sensors, 23(13): 5941. https://doi.org/10.3390/s23135941

[27] Peterson, J.M., Leevy, J.L., Khoshgoftaar, T.M. (2021). A review and analysis of the Bot-IoT dataset. In 2021 IEEE International Conference on Service-Oriented System Engineering (SOSE), Oxford, United Kingdom, pp. 20-27. https://doi.org/10.1109/SOSE52839.2021.00007

[28] Al Hamad, M., Zeki, A.M. (2018). Accuracy vs. cost in decision trees: A survey. In 2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Sakhier, Bahrain, pp. 1-4. https://doi.org/10.1109/3ICT.2018.8855780

[29] Khetani, V., Gandhi, Y., Bhattacharya, S., Ajani, S.N., Limkar, S. (2023). Cross-domain analysis of ML and DL: Evaluating their impact in diverse domains. International Journal of Intelligent Systems and Applications in Engineering, 11(7s): 253-262.

[30] Asgharzadeh, H., Ghaffari, A., Masdari, M., Gharehchopogh, F.S. (2023). Anomaly-based intrusion detection system in the Internet of Things using a convolutional neural network and multi-objective enhanced Capuchin Search Algorithm. Journal of Parallel and Distributed Computing, 175: 1-21. https://doi.org/10.1016/j.jpdc.2022.12.009

[31] Shukla, P. (2017). ML-IDS: A machine learning approach to detect wormhole attacks in Internet of

Things. In 2017 Intelligent Systems Conference (IntelliSys), London, UK, pp. 234-240. https://doi.org/10.1109/IntelliSys.2017.8324298

[32] Gaur, V., Kumar, R. (2022). Analysis of machine learning classifiers for early detection of DDoS attacks on IoT devices. Arabian Journal for Science and Engineering, 47(2): 1353-1374. https://doi.org/10.1007/s13369-021-05947-3

[33] Khanday, S.A., Fatima, H., Rakesh, N. (2023). Implementation of intrusion detection model for DDoS attacks in lightweight IoT networks. Expert Systems with Applications, 215: 119330. https://doi.org/10.1016/j.eswa.2022.119330