Cyber Attacks Detection using Machine Learning

1. R.Deepthi Reddy,
Research Scholar, GITAM University, Hyderabad, Telangana, India,
draavi@gitam.in.
2. Dr Srinivas Katkam,
Professor, Department of CSE, Geethanjali College of Engineering and Technology,
Medchal District, Telangana, India,
katkamsrinu@gmail. com.
3. Channapragada Rama Seshagiri Rao
Professor & Principal Vignana Bharathi Engineering College, R. R. District, Telangana, India
crsgrao@gmail.com

Abstract—
The Internet of Things, sometimes known as IoT, is a network that makes it possible for inanimate objects to communicate with one another virtually entirely independent of the intervention of humans. The number of things that can be connected by this network is in the hundreds of millions. The Internet of Things (IoT) is one of the computing disciplines that is increasing at one of the fastest rates; nevertheless, the harsh reality is that the internet is a very hostile environment, which makes the IoT vulnerable to a wide variety of different kinds of attacks. One of the realistic defences that must be created to protect IoT networks and address this problem is network anomaly detection. While preventing all attacks is challenging in the long run, successful defence in the present depends on the early discovery of an attack Because IoT devices have limited storage and processing, traditional high-end security solutions cannot secure them. Additionally, devices connected to the Internet of Things may now maintain a connection for longer periods of time without human input. Therefore, intelligent network-based security solutions, such as those based on machine learning, are necessary. The application of machine learning strategies to problems relating to attack detection has been the primary focus of a significant amount of research that has been published over the course of the past several years. This study has been carried out over a period of several years. Nevertheless, this is contingent upon the identifying of the assaults, especially those targeted at Internet of Things (IoT) networks, has received very little attention. By examining different machine learning algorithms that are capable of promptly and effectively identifying assaults on Internet of Things networks, this study seeks to further knowledge. Using a brand-new dataset named Bot-IoT, the effectiveness of several detection techniques is evaluated. Seven distinct machine learning algorithms were tried throughout the implementation process, and most of them were successful in achieving high performance. The Bot-IoT dataset's most recent properties were taken out and used during installation. The results of the new features were superior when they were compared to studies from previously published research.

Keywords—Network anomaly detection; machine learning; Internet of Things (IoT); cyber-attacks; bot-IoT dataset
DOINUMBER:10.48047/NQ.2022.20.19.NQ99404 **NEUROQUANTOLOGY2022;20(19):4388-4398**

1. INTRODUCTION

Is certainly a substantial addition to the mass of literature already in existence. As information technology grows increasingly widespread in daily life, the security of computer networks and privacy is becoming a more significant global issue. Furthermore, the need for computer security has grown in significance. A surge in the amount of attempts to break into computer networks and systems has occurred concurrently

with the spread of Internet-based apps and other cutting-edge technology, such as the Internet of Things. As a result, the overall number of successful hacking attempts has increased significantly (IoT). On a global scale, one of the most serious challenges right now is how to preserve internet users' privacy without jeopardising the integrity of computer networks. The devices may converse with one another while a person is around.. These devices can be remotely or manually controlled by a person. Furthermore, the need for computer security has grown in significance. There has been a significant uptick in the number of attempts made to break into computer networks and systems, which has been matched by the proliferation of Internet-based applications and other cutting-edge technology, such as the Internet of Things (IoT). The Internet of Things is a network of connected electronic items that can communicate with one another either accidently or voluntarily at the command of a person (IoT).The Internet of Things (IoT) allows a wide range of sensor-equipped items (such as bicycles, coffee makers, lights, and many other things) to connect to the Internet. These devices are used in a wide range of industries, including healthcare, agriculture, transportation, and others. [1]. Internet of Things applications are revolutionising both our personal and professional lives by enabling us to save time and money. It also provides an unending array of advantages and a myriad of chances for knowledge sharing, stimulating innovation, and improving society. The Internet, which acts as the foundation and central hub of the Internet of Things, is affected by every security issue that affects the Internet. Internet of Things nodes have limited resources and no manual controls. This network is unlike others.. Because of the fast expansion of Internet of Things devices and their pervasive use in everyday life, as well as the fact that IoT security challenges are extremely challenging to handle, security solutions based on network architecture have been developed. While certain attacks may be detected relatively well by modern security systems, others remain difficult to identify.

Because of the tremendous expansion in the amount of information that is carried by networks as well as the increase in the number of threats aimed at those networks, it is without a doubt that there is a rising demand for improved and more creative network security approaches. Because of this, more effective and quick methods of attack detection are required [2], and there is no doubt that such mechanisms already exist. One of the best computational techniques for supplying embedded intelligence in the Internet of Things environment in this scenario is machine learning (ML). If we proceed in this fashion, we will be able to include embedded intelligence into the Internet of Things environment. If we continue in this direction, we will be able to complete our objectives. Many diverse network security jobs have been done effectively with the help of machine learning algorithms. These responsibilities include a wide range of responsibilities. Monitoring network traffic [3, 4], identifying botnets [6, 7], and detecting intrusions [6, 7] are just a few of the duties that fall into this category. The creation of an Internet of Things (IoT) solution depends on an intelligent object's capacity to modify or automate a knowledge-based condition or behaviour.This ability is also known as "knowledge-based automation. “This capacity is known as "machine learning." Due to its ability to extract crucial information from data generated by either machines or people, machine learning is employed in procedures like regression and classification. Machine learning could be used to secure Internet of Things networks. The application of machine learning in the field of cyber security is experiencing explosive growth, and one of the hottest research topics currently is the use of ML to address problems with attack detection. Despite the fact that a large number of studies have been published that employ ML algorithms to identify the most effective ways to recognise assaults, only a tiny amount of research has been done on effective detection strategies that are ideal for IoT contexts. Signature-based cyber analysis, also known as misuse-based cyber analysis, and anomaly-based cyber analysis are the two basic types of cyber analysis that can be used to incorporate machine learning into the process of attack detection. Using one of these cyber analysis approaches, it is feasible to detect suspicious activity. These two methods of cyber analysis are also referred to as misuse-based cyber analysis. Both of these techniques to cyber security analysis are referred to as misuse-based

cyber analysis. Signature-based systems attempt to identify known assaults by utilising specific traffic features known as "signatures" in such assaults. These strategies are known as "signature-based." These methods are designed to detect previously reported assaults. These solutions were developed to counteract the consequences of the aforementioned hazards. The ability of this kind of detection system to promptly identify all known dangers while decreasing the likelihood of too many false alarms is one of its primary advantages.

The work in [3] used four unique machine learning algorithms in the earliest phases of the network traffic analysis in order to grasp the characteristics of a wide variety of well-known assaults. This was done in order to protect the network from being compromised. Other works [3] and [7] use signature-based techniques to identify assaults. Additionally, [7] used signature-based techniques to identify compromised workstations by seeing trends in botnet-generated network traffic. To do this, traffic patterns were examined. The main shortcomings of signature-based systems are their inability to identify assaults that were previously undetected and their frequent dependence on human updates of attack traffic signatures to operate. The second type of methodology used in the detection process is anomaly-based detection. In this class, which mimics regular network behaviour, any anomalous behaviour is viewed as an attack. What makes this class intriguing is its ability to discover previously undetected attacks. Anomaly-based techniques have a variety of drawbacks, the most notable of which being the potential for significant false alarm rates (FARs). FARs occur when previously unknown behaviours, even if lawful, are labelled as system abnormalities. When there is a shortage of information, this can occur. The use of signature and anomaly detection techniques can enable the construction of a hybrid strategy. [8] Uses a hybrid strategy as an example. The hybrid technique is employed in this instance to lower false positives (FP) for unidentified assaults while raising detection rates for identified attacks. This research was carried out in order to add to the current body of knowledge.

[3] In order to evaluate the effectiveness of the detection algorithms, a recent dataset called Bot-IoT is utilised. This data collection includes both actual and simulated network traffic from Internet of Things (IoT) devices, in addition to various different kinds of assaults. [9]. In order to identify characteristics in this dataset, the Random Forest Regressor technique was utilized. During the implementation phase, seven distinct machine learning approaches were applied to obtain remarkable performance. [4]: KNN, ID3, QDA, RF, AdaBoost, MLP, and Naive Bayes are a few machine learning methods (NB).

In a nutshell, we contributed the following to this study:

• The ability to detect assaults on Internet of Things networks will be enhanced by a deeper knowledge of how machine learning algorithms function on recently gathered IoT datasets.

• Two steps that can be taken to improve the efficiency of the machine learning algorithm are in order to derive new characteristics from the data and to choose those characteristics that are most relevant to the problem at hand.

• Increase our understanding of the Internet of Things. Due to the limited number of previous research projects that have used the Bot-IoT dataset, working with it presents a number of difficulties.

2. RELATED WORKS

Machine learning research has increased recently.. This expansion can be attributed to several factors. This is because the field is steadily gaining significance as time goes on. This rise has been accompanied by: [6]. In addition, a number of scholarly papers on the application of artificial intelligence and data mining to intrusion detection have been published [10]. On the other hand, the majority of these older studies relied primarily on machine learning algorithms to detect intrusions in traditional networks. These investigations were designed to discover potential security flaws. As a result, the primary focus of our work to advance research in this area is on the application of machine learning to the challenge of determining whether or not an attack has been carried out within the framework of the Internet of Things (IoT).Machine learning research in the Internet of Things (IoT) industry, especially in IoT security, has a lot of potential. IoT security needs improvement. This subject offers a lot of potential for advancement. There is a lot of room for progress in the subject of IoT security, in

particular. There is a significant potential for it to reveal insights from IoT data [11]. This is especially true in terms of IoT security. To identify and stop potentially risky behaviours, IoT networks can make use of tools like behaviour analysis, pattern recognition, and anomaly detection. Pattern detection is another application for these techniques. These techniques can also be used to find patterns in data.

We read a number of articles in order to conduct a review of recent research on the topic of detecting dangers in IoT networks using machine learning. This enabled us to conduct the review. Table I summarises the findings in their entirety. Each study describes the datasets, machine learning techniques, and detection methods that were used. When selecting from among the available research projects, we gave precedence to those who made use of a wide range of different datasets and machine learning approaches. The findings suggest that machine-learning-based detection systems have the potential to be useful in the future. Unsupervised approaches are classified as methods [10], [12], [13], and [14], whereas supervised techniques are classified as detection methodologies [15], [16], [17], and [9]. For the purpose of deriving these categories, research on the topic of utilising machine learning for the security of Internet of Things (IoT) devices can be used.

Several authors have turned to unsupervised machine learning strategies in an effort to find solutions to detection-related issues. In a number of studies, K-means, ANNs, RFs, and auto-encoders can help identify assaults. One of the most widely used unsupervised methods is auto-encoders to extract features from datasets. Mirsky et al. [10] proposed this method as one of the most prominent unsupervised strategies that has been employed. This move was done to improve the detection of potential cyber risks. This was done to increase the precision with which cyber hazards could be identified. This phrase requires elaboration. UN - supervised detection of network intrusions system Kitsune was shown. Kitsune autonomously identifies network intrusions. Kitsune was created to identify network intrusions without the need for human intervention. Kitsune can detect network intrusions even in the absence of human supervision. Kitsune was created to identify network intrusions without the need for human intervention. Kitsune gets his name from the Japanese word for "tiger." By merging a variety of neural networks that are referred to as "autoencoders," KitNET, which is the foundation of Kitsune's method, is able to differentiate between normal and aberrant patterns of traffic. Meidan et al. developed and evaluated a unique detection approach in their study [12], which makes use of auto-encoders to identify abnormal network traffic that is caused by infected devices and collects network behavioural snapshots. This detection method also takes network snapshots. This method was named after being described as the "extraction of behavioural snapshots from the network." The use of unsupervised machine learning techniques for issue identification has a number of drawbacks The fact that most network traffic flows are normal and attacks and abnormalities are rare is crucial. This is one of the reasons why these elements are so powerful. As a result, neither success rates nor the ability to detect anomalies in data are improved, which is a big setback. As a result, employing tactics that necessitate supervision may produce better results.

In order to detect assaults, however, a variety of supervised learning methods are used. The datasets used to train these algorithms include labels that indicate whether or not the occurrences have previously been labelled as assaults. The algorithms were trained using these labels. Elike Hodo utilised ANN or support vector machine approaches, as described in [19], in order to identify assaults that did not involve Tor traffic. ML approaches were used to apply these strategies to datasets gathered from UNBCIC. Identifying the Internet of Things device subsets that are permitted on the white list the random forest method was employed in the study [15] to process the network traffic features derived from the data. Moustafa et al. [9] presented a recent study that used a methodology similar to the one we used in our analysis in the first publication and outlined how the Bot-IoT dataset was obtained. The context of the original paper was utilised to present this investigation. They analysed the IoT dataset using LSTM, SVM, and RNN machine learning models, but they did not test their models' susceptibility to adversarial attacks.

Machine learning models come in a variety of shapes and sizes, including the LSTM, SVM, and RNN.

In spite of the fact that we make use of the dataset that was presented in [9], the primary objective of our research is not to perform an analysis of the dataset but rather to assess how well various machine learning techniques perform on this dataset. This dataset was published in [9]. If you look at [22], you may come across a study that used the BoT-IoT dataset. If you come across such a study, make sure to properly reference it. They compared the classification of potential intrusion hazards in a network connected to the internet of things to the performance of a self-normalizing neural network (SNN) and a fully connected neural network (FNN). The many distinct performance measurements used in this experiment provided the foundation for the comparison. Ferrag examined how well the Deep Coin framework performed in Internet of Things (IoT) traffic using the Bot-IoT dataset from [14]. Deep Coin is a ground-breaking new energy framework built on technologies such as deep learning and block chain. Through the execution of performance experiments on the Bot-IoT dataset, they were able to prove that the suggested Deep Coin system was technically feasible. The BoT-IoT dataset was used by researchers who were working on a separate project [17] in order to construct the rules that are a part of IoT-IDS. They developed effective rules for the purpose of assisting in the creation of compact intrusion detection systems that are suitable for Internet of Things devices by using the J48 machine learning technique.

3. PROPOSED WORK

This section provides a concise overview of the dataset that was utilized and suggested method for identifying threats in IoT networks. The method we have provided uses a number of pre-processing steps and real-world applications to discover anomalies using machine learning techniques. The method's initial step was to do data pre-processing in order to get the dataset ready for further splitting into two parts: training and testing. The CIC Flow Meter [25] was used to start the process of extracting flow-based characteristics from the raw dataset. The "feature selection" step comes after these processes, during which the algorithms choose which properties to use. The apex of our process is the deployment of various machine learning algorithms. A high-level overview of the suggested technique is shown in Figure 1. Because of its potential to derive novel characteristics from raw data, the Bot-IoT dataset was chosen for the studies, the frequency with which it is updated, the high attack variety it possesses, and the fact that it includes traffic generated by IoT devices. In addition, the dataset has the capability of frequently obtaining novel attributes from raw data. The Cyber Range Lab at the Australian Centre for Cyber Security is responsible for the creation of the dataset that is now known as Bot-IoT [10]. This dataset accounts for three different kinds of cyber attacks: information theft, denial-of-service attacks, and probing. All three of these forms are known as "probing." These three assaults are all rooted in the functioning of botnets in one way or another. We were able to extract flow-based information from unprocessed traffic traces by utilising the CIC Flow Meter software. A CIC-provided network traffic flow generator is called CIC FlowMeter [26]. It is capable of producing 84 different kinds of network traffic.

The study's main objective is to assess how effectively machine learning algorithms can identify intrusions into Internet of Things (IoT) networks, as was described in the sections above. This section will outline the machine learning techniques we utilised, describe the dataset we used, and show how we implemented our plan.

A. Datasets

For network flow analysis to discern between regular and abnormal traffic, large datasets are necessary. This is due to the inclusion of machine learning approaches in the apps used for various network security jobs. To create network datasets, a variety of experiments have been carried out over time. The majority of machine learning research has used either simulated or actual network data to confirm its findings, as indicated in Table I. The public has access to the datasets from DARPA 98, KDD 99, UNSW-NB 15, ISCX, CICIDS 2017, and N-BaIoT, among others. Even if the vast majority of these datasets are still secret, security issues are the main cause of this. There has been very little progress made in the production of realistic IoT & network traffic statistics that include unique botnet

circumstances, despite the development of a number of datasets. What's more, some databases don't include IoT-related traffic, while others don't produce any new features at all. Both the test bed used in some instances and the assault scenarios that were researched in others were not sufficiently varied. As an illustration, Meidan et al. [12] produced and made public the N-BaIoT Internet of Things dataset. This dataset was used in many future studies to develop and validate their classifier models. This dataset is uneven, with a far lower ratio of normal data to attack data than there should be, despite being rather large and having been cleaned up. The Bot-IoT dataset was developed by Moustafa et al. [9] in an effort to address the problem, and it was later employed in the tests that we carried out. The Bot-IoT collection includes a variety of cyber attacks as well as real and simulated Internet of Things (IoT) network traffic [14]. The following categories apply to different types of attacks on the BotIoT database: Attacks that involve probing, data stealing, and denial of service

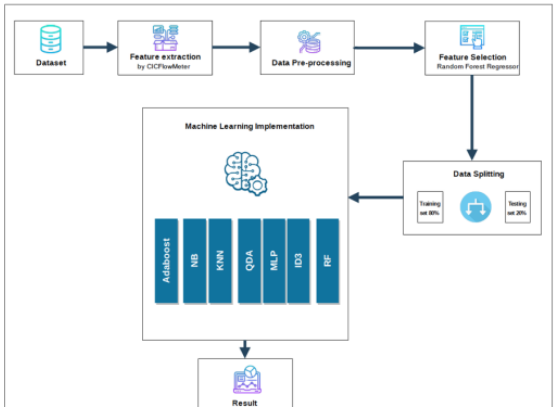

Figure 1: An Overview of the Proposed Approach

A. Machine Learning Algorithms

On the Bot-IoT dataset, we evaluated seven of the most popular machine learning classifiers, including K-Nearest Neighbors (KNN), ID3 (Iterative Dichotomiser 3), Random Forest, AdaBoost, Quadratic Discriminant Analysis (QDA), Multilayer Perceptron (MLP), and Nave Bayes (NB). The selection of these classifiers places an emphasis on combining well-known techniques with a wide range of different characteristics. The following is a condensed discussion of the algorithms used in this setting.

• K-Nearest Neighbors (KNN): KNN is one of the supervised learning algorithms that is both easy to understand and successful. It is put to use to search through the dataset that has been provided and correlate new data points with similar ones that already exist [24]. KNN is an efficient method that works well with multidimensional data and moves quickly during the training phase. However, when it comes to the estimation step, KNN is a pretty slow algorithm.

• Quadratic discriminant analysis (QDA): QDA is a fantastic approach for solving problems involving supervised classification. The statistical method known as discriminant analysis is used to assign data that has been measured to one of several different categories. In circumstances in which a category lacks data, qualitative data analysis (QDA) is acceptable. Quadratic discriminant analysis requires more samples than groups.

•The Iterative Dichotomizer 3, more commonly referred to as the ID3 algorithm, is a sort of algorithm that, when applied to a dataset, produces a decision tree. Ross Quinlan [27] was the one who came up with it. A classification algorithm that has a decision structure that is similar to a tree is called a decision tree. It is a way of displaying an algorithm that simply uses conditional control statements and is one of the methods available. The attributes are tree nodes, and the class values assigned to a record are the "leaves."[16]. ID3 is a well-known method that is utilised in areas such as natural language processing and machine learning. It also served as foundation for C4.5 algorithm. The (RF), often known as "random forest," is a method of machine learning that makes use of decision trees. In this method, a "forest" is created by constructing a huge number of different decision tree structures that are built in a variety of different ways [28]. This method has a number of benefits, the most notable of which are its speed when operating on large datasets, its light weight in comparison to other approaches, and AdaBoost is a machine learning algorithm that focuses on classification issues and seeks to transform ineffective classifiers into effective ones. Introduced in 1996 by Schapire and Freund, it can be used in conjunction with many different learning algorithms to achieve better results. The capacity of the AdaBoost algorithm to handle missing values in a dataset is the most important quality that it possesses.

• Multilayer Perceptron (MLP): sometimes known as an MLP, this is an example of a feedforward

artificial neural network (ANN). A technique for machine learning known as "artificial neural networks" (ANNs),As a point of departure, this strategy looks to how the human brain tackles problems like learning and coming up with original content. At a minimum, an MLP will have three layers, labelled input, output, and hidden, respectively. A is utilised by the MLP.

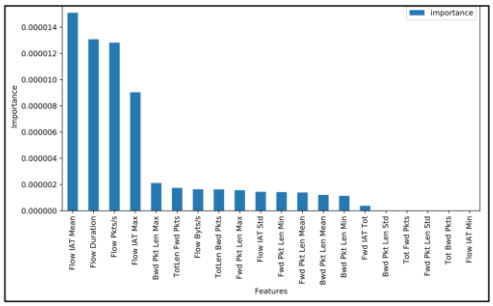

Fig. 2. Characteristics Graph The entire dataset is important in the back-propagation supervised learning method used for training.

• **NaiveBayes (NB):** The supervised algorithm known as the NB is well renowned for its straightforward guiding principles. The works of Thomas Bayes are used as the foundation for the nave Bayes approach [29]. For instance, using NB to classify traffic as normal or abnormal might be used for intrusion detection. Despite the fact that the NB classifier's traffic classification features may be interdependent, it maintains these features independently. The simplicity of NB, the minimal sample demand, and the ease with which it may be implemented are only a few of the many qualities that contribute to its user-friendliness [27]. On the other hand, because NB analyses each feature on its own, it is unable to derive actionable insights from the interactions and correlations that exist between traits.

Implementation Steps

Our methodology is comprised of five primary components: the application of machine learning strategies, the selection of features, the pre-processing of data, and the extraction of features. The first step of the method is the extraction of features.

• Extraction of Features: The raw network traffic data was processed with CIC Flow Meter [25] so that flow-based features could be extracted from it and saved in pcap format. A CIC-distributed network traffic flow generator, the CIC Flow Meter. It generates 84 unique types of network traffic. Along with reading the PCA file, it creates a visual showing the features gathered and outputs a CSV file containing the dataset. By extracting more dataset features, this technique was developed primarily for the goal of increasing classifier prediction abilities.

• Data pre-processing: Before the dataset is turned into a structure that is conducive to machine learning, pre-processing data transformation activities must be carried out. In addition, the dataset needs to be cleaned up at this point by deleting any inaccurate or irrelevant data that can have an impact on the accuracy of the dataset. Furthermore, transformation operations need to be carried out. In addition, the dataset needs to be cleaned up at this step by getting rid of any inaccurate or irrelevant data that might have an impact on how accurate the dataset is. By eliminating this kind of data, the dataset is made more effective.

• Data Segmentation: The machine learning process is dependent on having access to data in order for there to be any learning at all. Test data are necessary in addition to the training data needed to assess the algorithm's performance and determine how well it operates. During the course of our investigation, we came to the conclusion that only twenty percent of the Bot-IoT dataset would be used for testing, while the remaining eighty percent would be utilised for training purposes.

• Feature Selection: In order to find a lightweight security solution that is appropriate for Internet of Things (IoT) systems, it is essential to limit the total amount of features and use just those features that are required to train and test the algorithms. This is known as the feature selection.. This will assist you in locating a solution suitable for IoT systems. The feature selection process is the method for doing so.[13].The random forest regression technique was employed for feature selection. It has been demonstrated that the random forest regressor is an effective way to reduce the parameters that describe a dataset. When the amount of input data features. The model trains and reacts faster with seven network traffic features instead of 80. Figure 2 shows attribute relevance weights as percentages across the dataset. Python, as well as a number of its own machine learning tools, were used at various phases of the process. The use of algorithmic approaches to machine learning was employed (Scikit-Learn, Matplotlib, Pandas, and

NumPy). Table II has a comprehensive list of all of these characteristics and qualities. Third, we applied the algorithms to the entire dataset by aggregating the best qualities for each type of assault. Finally, we ran the algorithms on the entire dataset, focusing on the top seven attributes

4. RESULTS AND DISCUSSION

A. Evaluation Metrics

Performance criteria must be tailored to the project when assessing machine-learning models. We used the most significant performance metrics available to us while analysing the conclusions of our investigation. These measures were accuracy, precision, and recall, and their equations are listed below. The following measurements were used:

Precision=TP/ TP+FP

Recall=TP/TP+FN

Accuracy=TP+TN/TP+FP+TN+FN

We divided the assessment of the performance of the machine learning algorithms on the dataset into three distinct stages, as was covered in the section that came before this one. Phase 1: Machine learning will be used to analyse each assault in the dataset. Phase 2: Utilizing machine learning, discover the most useful aspects of each assault from the entire data set. Phase 3: Using the top seven features from the feature selection step, apply machine learning techniques to the complete dataset. Phase 1: Apply machine learning on the data using the top 7 features from feature selection. The tables below provide a summary of the findings from all of the experiments. The numbers that are presented in the tables represent the arithmetic mean of the outcomes of the 10 iterations of the performance evaluation approaches for each machine learning algorithm. Phase 1: Each assault in the dataset is being subjected to a single application of machine learning methods. Table III presents the findings of an experiment in which each of ten different types of attacks was subjected to seven different machine learning approaches. The algorithms look at the values of precision, accuracy, recall, and time in order to remove the equality if the F-measure is equal. With the exception of the Naive Bayes (NB) and Quadratic algorithms (QDA), all of the algorithms had a success rate of more than 90% in identifying practically all forms of attacks, according to Table III's data. The ID3 algorithm had the highest score, completing six out of ten objectives (including DOS-TCP, data exfiltration, and service scanning). Indeed, despite all the

Table III: Results are distributed by attack type.

Attack name	F-Measures					
	NB	QDA	RF	ID3	AB	MLP
DDOS_HTTP	0.62	0.95	0.86	0.86	0.85	0.86
DDOS_UDP	0.63	0.82	0.88	0.88	0.88	0.88
DDOS_TCP	0.61	0.75	0.89	0.5	0.64	0.89
DOS_HTTP	0.62	0.62	0.84	0.84	0.84	0.85
DOS_UDP	0.62	0.73	0.76	0.87	0.88	0.88
DOS_TCP	0.54	0.64	0.5	0.5	0.5	0.87
Data extraction	0.62	0.65	0.83	0.87	0.87	0.88
Keylogging	0.62	0.72	0.85	0.85	0.83	0.87
Service_scan	0.63	0.72	0.84	0.83	0.83	0.83
OS_Scan	062	0.65	0.84	0.84	0.82	0.81

At least one other method has a score that is lower than or equal to ID3. But because it processes more quickly than the other algorithms, it is preferable to them. Since the Naive Bayes method had the lowest F-measure of all the algorithms, it was chosen for the final challenge. It received a generally poor rating, primarily because of the DOS TCP attack. The Naive Bayes algorithm outperformed the alternatives in terms of performance, but it did so at a much faster rate. However, since the QDA performed the second-worst of all the algorithms at this point, it is also necessary to bring it up.In the second phase, the entire dataset and a set of features that are a composite of the best characteristics for each assault are applied to machine learning techniques. The entire dataset is being used at this time. We used feature sets that were isolated for each type of assault and seven distinct machine learning algorithms were applied to the complete dataset. The findings from using the 13 features retrieved from the attacks are displayed in Table 4.

Table 4. Implementation of features obtained from phase

ML Algorithm	Accuracy	Precision	Recall	F-Measure	Time
NB	0.68	0.74	0.68	0.65	4.045
QDA	0.78	0,79	0.77	0.77	5.2867
RF	0.88	0.88	0.87	0.87	18.2336
ID3	0.78	0.88	0.73	0.76	1290.3001
Adaboost	05	0.5	0.5	0.5	206.987
MLP	0.76	0.76	0,74	0.72	2081.9821
KNN	0.88	0.88	0.88	0.88	1093.1098

Table 4 makes it clear that Adaboost, KNN, and ID3 were the top-performing algorithms. In this feature, ID3 is given preference over KNN because it is much faster. The algorithm with the lowest total score, the Naive Bayes method, had a score of 0.75. In the third step, machine learning algorithms that use the seven top features from the first round of feature selection are applied to the whole data set. The algorithms' performance did not significantly change when viewed through the F-measure lens, but when viewed through the lens of speed, the time it took for each algorithm to complete its cycle significantly decreased. The approach used 13 properties, but the research that was compared with only used 7 features in a study that was published in the literature, which resulted in a reduction in execution time. The experiment conducted by Ferrag and his colleagues [14] in 2019 provided as a point of reference for the research in this study. The previously cited study employed the same dataset and two machine learning techniques similar to ours. This is due to the fact that this was the case. Random Forest and Naive Bayes are two machine learning algorithms that are very similar to one another. The key distinction between the work they completed and the work we completed is the feature set that was chosen to be implemented. They used the set of features that had been built from the beginning, whereas we used one that had just been retrieved from CiFlowMeter. The detection rate, sometimes referred to as the recall rate, was chosen as the main criterion for evaluation. In Table VI, which is included below, the findings from the two investigations are compared. The conclusion that can be reached from comparing the two methods is that the Random Forest method is superior to the one described in [14]. This is evident from the analysis of the data. For the great majority of the various attack types that can be mounted, the NB algorithm shows a consistent pattern. This makes it quite evident that the supplemental features created as a result of our research have increased the overall effectiveness of both the algorithms.

5. CONCLUSION

This study uses machine learning techniques to detect assaults on IoT networks. Due to its frequent updates, wide variety of attacks, and numerous unique network protocols [9], this experiment will make use of the Bot IoT dataset, which was specifically chosen for the purpose of this inquiry. Using the CICFlowMeter [25] programme, unprocessed traffic traces allowed us to retrieve flow-based features, which we were able to do. The 84 network traffic components included in the dataset were all created by CIC Flow Meter, which was in charge of their development. Each of these elements helps define the network flow in some way. The Random Forest Regressor technique was used throughout implementation to demonstrate the importance of weight computations. As part of the implementation phase, this was done. To choose which features to include in the machine learning algorithms, this was done. These calculations were carried out using two different approaches. The importance weights for each type of attack were computed separately for the first technique. The common characteristics that were relevant for all attacks were discovered using the second technique, which grouped all attacks into one and used that group's data to determine its important weights. When calculating the important weights for the first technique, each type of attack was analysed and evaluated independently. Following that, the data was run through a total of seven different machine learning algorithms, each of which is well-known in its own right and provides a unique set of benefits. The following is a list of many algorithms, together with the F-measure performance ratios that accompany them: While the F-measure ranged from 0 to 1, The Naive Bayes technique received a score of 0.77, the QDA

approach received a score of 0.86, Random Forest received a score of 0.97, ID3 received a score of 0.97, AdaBoost received a score of 0.97, MLP received a score of 0.83, & K Nearest Neighbors received a score of 0.99. Within the scope of this study, seven distinct supervised algorithms were evaluated. Analyzing the performance of a number of unsupervised algorithms would be interesting and would fall under the category of "future work." In addition, we applied a number of unique machine learning techniques independently of one another. Future detection performance should be enhanced by combining various machine learning techniques into a multi-layered model. More freedom would be offered by this.

REFERENCES

[1] J. Deogirikar and A. Vidhate, "Security attacks in IoT: A survey,"*International Conference on I-SMAC (I-SMAC)*, pp. 32–37, 2017.

[2] T. Bodstrom and T. H ¨ am ¨ al ¨ ainen, "State of the art literature review ¨on network anomaly detection with deep learning," *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*, pp. 64–76,2018.\

[3] I. Arnaldo, A. Cuesta-Infante, A. Arun, M. Lam, C. Bassias, andK. Veeramachaneni,"Learning representations for log data in cybersecurity," *International Conference on Cyber Security Cryptography andMachine Learning*, pp. 250–268, 2017.

[4] M. Du, F. Li, G. Zheng, and V. Srikumar, "Deeplog: Anomaly detectionand diagnosis from system logs through deep learning," *Proceedings of*
the 2017 ACM SIGSAC Conference on Computer and CommunicationsSecurity, pp. 1285–1298, 2017.

[5] B. J. Radford, B. D. Richardson, and S. E. Davis, "Sequence aggregationrules for anomaly detection in computer network traffic," *arXiv preprintarXiv:1805.03735*, 2018.

[6] I. Lambert and M. Glenn, "Security analytics: Using deep learning todetect cyber attacks," 2017.

[7] M. Stevanovic and J. M. Pedersen, "Detecting bots using multi-leveltraffic analysis." *IJCSA*, vol. 1, no. 1, pp. 182–209, 2016.

[8] H. Sedjelmaci, S. M. Senouci, and M. Al-Bahri, "A lightweight anomalydetection technique for low-resource iot devices: A game-theoreticmethodology," *IEEE International Conference on Communications (ICC)*, pp. 1–6, 2016.

[9] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towardsthe development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *Future GenerationComputer Systems*, vol. 100, pp. 779–796, 2019.

[10] Y. Mirsky, T. Doitshman, Y. Elovici, and A.Shabtai, "Kitsune: anensemble of autoencoders for online network intrusion detection," *arXiv preprint arXiv:1802.09089*, 2018.

[11] X. Yuan, C. Li, and X. Li, "Deepdefense: identifying ddos attack viadeep learning," *IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 1–8, 2017.

[12] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y.Elovici, "N-baiot—network-based detection of iot botnet attacks using deep autoencoders," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12–22, 2018.

[13] M. K. Putchala, "Deep learning approach for intrusion detection system (ids) in the internet of things (iot) network using gated recurrent neural networks (gru)," 2017.

[14] M. A. Ferrag and L. Maglaras, "Deepcoin: A novel deep learning andblockchain-based energy exchange framework for smart grids," *IEEE Transactions on Engineering Management*, 2019.

[15] Y. Meidan, M. Bohadana, A. Shabtai, M. Ochoa, N. O. Tippenhauer, J. D. Guarnizo, and Y. Elovici, "Detection of unauthorized iot devices using machine learning techniques," *arXiv preprint arXiv:1709.04647*, 2017.

[16] N. Koroniotis, N. Moustafa, E. Sitnikova, and J. Slay, "Towards developing network forensic mechanism for botnet activities in the iot based on machine learning techniques,"*International Conference onMobile Networks and Management*, pp. 30–44, 2017.

[17] Y. N. Soe, Y. Feng, P. I. Santosa, R. Hartanto, and K. Sakurai, "Rulegeneration for signature based detection systems of cyber attacks in iot environments," *Bulletin of Networking, Computing, Systems, andSoftware*, vol. 8, no. 2, pp. 93–97, 2019.

[18] V. H. Bezerra, V. G. T. da Costa, S. B. Junior, R. S. Miani, and B. B.Zarpelao, "One-class classification to detect botnets in iot devices," *Anais do XVIII Simposio Brasileiro em Seguranc¸a*

da Informac¸ ´ao e ˜de Sistemas Computacionais, pp. 43–56, 2018.
[19] E. Hodo, X. Bellekens, A. Hamilton, P.-L. Dubouilh, E. Iorkyase,C. Tachtatzis, and R. Atkinson, "Threat analysis of iot networks using artificial neural network intrusion detection system," *InternationalSymposium on Networks, Computers and Communications (ISNCC)*, pp. 1–6, 2016.
[20] D. H. Summerville, K. M. Zach, and Y. Chen, "Ultra-lightweight deeppacket anomaly detection for internet of things devices," *IEEE 34th international performance computing and communications conference(IPCCC)*, pp. 1–8, 2015.
[21] F. Y. Yavuz, "Deep learning in cyber security for internet of things,"Ph.D. dissertation, 2018.
[22] O. Ibitoye, O. Shafiq, and A. Matrawy, "Analyzing adversarial attacksagainst deep learning for intrusion detection in iot networks," *arXivpreprint arXiv:1905.05137*, 2019.
[23] I. Cvitic, D. Perakovi ´ c, M. Peri ´ sa, and M. Botica, "Novel approach fordetection of iot generated ddos traffic," *Wireless Networks*, pp. 114,2019.
[24] Z. A. Baig, S. Sanguanpong, S. N. Firdous, T. G. Nguyen, C. So-In*et al.*, "Averaged dependence estimators for dos attack detection in iot networks," *Future Generation Computer Systems*,vol. 102, pp. 198–209, 2019.
[25] A. H. Lashkari, G. Draper-Gil, M. S. I. Mamun, and A. A. Ghorbani,"Characterization of tor traffic using time based features." *ICISSP*, pp.53–262, 2017.
[26] S. Yu, "Study on the internet of things fromapplications to securityissues," *International Conference on Cloud Computing and Security*, pp. 80–89, 2018.
[27] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machinelearning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*,vol. 160, pp. 3–24,2007.
[28] K. Kostas, "Anomaly detection in networks using machine learning,"Ph.D. dissertation, 08 2018.
[29] M. Panda and M. R. Patra, "Network intrusion detection using naïve bayes," *International journal of computer science and network security*, vol. 7, no. 12, pp. 258–263, 2007.