Research thesis report

MScResearchProject
CloudComputing

Venkata Sai Charan Vinnamuri
StudentID:22156461

SchoolofComputing
NationalCollegeofIreland

Supervisor: Jitender Kumar Sharma

National College of Ireland
Project Submission Sheet School of Computing

Student Name:	Venkata Sai Charan Vinnamuri
Student ID:	22156461
Programme:	Cloud Computing
Year:	2023
Module:	MSc Research Project
Supervisor:	Jitender Kumar Sharma
Submission Due Date:	12/08/2024
Project Title:	Novel Approaches for Real-Time Detection of DDoS Attacks in Cloud Computing Environments Using Advanced Machine Learning Techniques
Word Count:	
Page Count:	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Vinnamuri Venkata Sai Charan
Date:	12th August, 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	

Penalty Applied (if applicable):	

Novel Approaches for Real-Time Detection of DDoS Attacks in Cloud Computing Environments Using Advanced Machine Learning Techniques

Vinnamuri Sai Charan Vinnamuri

X22156461

Abstract:

This research focuses on the identification of DDoS attacks in cloud computing systems using more sophisticated machine learning schemes. Standard security mechanisms have failed to provide adequate protection in relation to modern, large-scale and complex DDoS attacks that cause substantial service availability and monetary disadvantages. The proposed methodology involved building and comparing RF, SVM, and the RF-SVM combined models in order to improve the detection performance. Stage one of data preprocessing involved feature selection and normalization while stage two of data preprocessing involved training and evaluation of the models using network traffic data. The RF model proved consistent for the choice of estimating Rf classification, high accuracy, low chance of overfitting due to the use of the ensemble learning method. The SVM model was also good especially in high dimensionality, but the recall value was slightly down by approximately one because of misclassification. The application of the hybrid model was the most effective producing 100% accuracy, precision, recall, and F1 score with least latency and highest throughput that made it suitable for real time detection. Consequently, these outcomes achieve the research aims and objectives to develop a reliable, efficient, and accurate DDoS detection system for dynamic cloud infrastructure to bolster cloud service security and reliability.

Chapter 1. Introduction

1.1 Background:

Distributed denial of service (DDoS) attacks are proving to be a major challenge to cloud computing service provision as they result to disruption of services and huge losses. These attacks include a target being flooded with traffic, thus making it inaccessible to valid users. Cloud computing is dynamic and scalable which is particularly advantageous, but, the reality is that such an environment exposes several threats that are exploited by attackers (Kafhali et al.2022). These threats, being contemporary and influential require a different approach on the part of conventional defense mechanisms to counter them. Currently, the application of novel machine learning algorithms could help in the realization of the real-time detection and handling of relevant incidents (Shah, V., 2021). These approaches can effectively employ the use of algorithms like Support Vector Machines (SVM) and Random Forests that are used to analyze huge data traffic in networks to detect patterns of DDoS attacks. The various models developed are the integration of different algorithms to improve the detection capability and speed, a factor that is very important for security of cloud services (Awan et al. 2021).

1.2 Problem Statement:

The growth of the demand for cloud computing services has made the threat and consequences of Distributed Denial of Service (DDoS) attacks more intense, as they caused interruptions in the availability of services through excessive traffic saturation of the resources targeted by the attackers. Conventional security approaches were inadequate and unable to effectively respond to large, sophisticated and prolonged DDoS attacks leading to prolonged service disruptions and financial losses (Li, et al.2023). Previous detection processes were unsuitable in responding to these attacks in real-time since attacks easily adapt to new conditions. There was a growing need to develop and implement new and improved solutions that could effectively and more promptly recognize and counter DDoS threats. This problem was especially important to mitigate because cloud infrastructures were very dynamic, with traffic fluctuations that could hide attack indications. Since DDoS is a dynamically changing phenomenon, more methods are clearly needed, and advanced machine learning approaches suggested the way how this could be done. Subsequently, a large volume of real-time packet capture or PCAP, evaluation and early detection in machine learning can detecting DDoS attacks accurately (. However, the real problem remained in tuning models to be both fast and accurate and that may handle real time streams data as they come without compromising on accuracy (Mittal et al.2023). To overcome these challenges, this research sought to develop new machine learning methods especially the blended models of SVM and random forest. The target was to develop an effective, reliable, and immediate DDoS detection framework with a focus on the specifics of cloud-computing environments to enhance service dependability and security.

1.3 Aims and Objectives:

The aim is to design and enhance a highly effective DDoS detection solution for the cloud computing environment through the integration of the latest machine learning algorithms.

- To develop and test improvements over Support Vector Machine and Random Forest algorithms in their detection capacity and speed for DDoS attacks.
- To enable real-time detection schemes to be deployed and evaluated to enable the machine learning models to properly process and respond to network traffic abnormal in real time.
- To further enhance the applicability of the detection system and fine-tune its response to the traffic dynamics and workload common to the cloud computing environments.

1.4 Research Question:

1. In what ways can the machine learning algorithms be used for the improvement of realtime identification of DDoS attacks in cloud system?
2. That being the case, what measures do you think are useful in minimizing the false positives to the real-time detection of DDoS attacks in cloud computing environments?

1.5 Rationale:

The importance of this research stems from the fact that the current cloud computing security measures are not sufficient to prevent complex DDoS attacks that commonly occur in those environments. It remains a challenge to use conventional means to detect these complex and large-scale attacks, which results in massive service disruptions and substantial losses. Through the use of state-of-art machine learning approaches, this research targets to design an efficient and speedy mechanism for detecting emerging DDoS threats in real-time. The integration of Support Vector Machines (SVM) and Random Forest algorithms results in better, faster, and more accurate detection of cloud security threats, which is vital to ensuring the availability and reliability of the cloud services (Mohammed and A., 2024). Besides adding a new solution to the field of cybersecurity, the chapter also aids in the acceptance of cloud computing by removing one of its significant threats. The findings of this research hold a valuable contribution in strengthening the protection mechanisms of cloud computing, protecting against DDoS attacks and maintaining the constant delivery of services to organizations and consumers.

In addition, there is great importance of this analysis for further progression of cybersecurity measures and related innovations. The present work contributes to the development of more sophisticated and adaptive security solutions by proving the merits of hybrid machine learning models to identify DDoS threats. It also underlines the need to have constant invention due to the constantly advancing cyber risks. The understanding of these issues obtained through this research might be useful for further investigations as well as advancements of concrete applications of the cloud-computing concept, promoting prevention of such threats. Lastly, the successful implementation of these advanced detection mechanisms would contribute positively towards increasing confidence in cloud computing hence increasing its usage in different sectors.

Chapter 2: Literature Review

2.1 Background on DDoS Attacks

vin

DDoS (Distributed Denial of Service) attacks refer to multiple organized efforts to hamper a server, service or a network by inundating it with a barrage of Internet traffic. These attacks normally entail a number of hosts, which usually contain viruses, to bring down one system and make it non-functional. DDoS attacks can be categorized into three main types: They include volumetric attacks, protocol attacks, as well as the application layer attacks. Volumetric attack overwhelms the target's bandwidth, protocol attack targets the holes in network protocols, and application layer attack targets specific applications and services thereby denying them to the users. DDoS attacks have progressed from being simple to more complicate and are a major threat to cloud computing systems where customers share resources (Kambourakis et al.2019).

DDoS attacks have been present since the year 2000, with some of the famous attacks including the Yahoo attack which drew much awareness on the issue. In the later years, the tactics have become more elaborate, and the use of botnets and weaknesses in networks and applications protocols has been observed. The use of IoT has made the Internet connected devises available to the attackers and thus lately, the Mirai botnet attack was made on Dyn that shut down major website and services in 2016 (Hashem et al. 2019). In the last few years, it has been observed that attackers use either multi-vector attacks and/or ransom attacks or the DDoS as a Service model has made these attacks easily available to anyone. Historic security approaches are ineffective against these new, complex types of attacks, and thus there is a demand for more effective detection and prevention such as machine learning and artificial intelligence in cloud computing (Xiao, Liu and Zhang, 2020).

2.2 Cloud Computing Environments

Cloud computing is the delivery of computing services such as storage and processing power through the internet and these services include IaaS, PaaS, and SaaS. IaaS is the renting of computing resources over the internet to include virtualized servers, storage, networking space, and more. The concept of PaaS enables customers to build, run, and execute applications while not having to deal with the creation of infrastructure. SaaS deploys applications on the internet, which can be accessed by the subscribers on a paid basis, to use the software and its services. These models bring certain advantages, such as decrease of the expenses, possibility of the further development and availability, which makes cloud computing as one of the potential solutions for all population and companies (Beitollahi and Deconinck, 2019).

Although cloud computing has numerous benefits, it is not exempted from security threats such as DDoS attacks. The fact that cloud resources and components are shared and virtualized presents a challenge to security. In multi-tenancy where several clients connect to the same physical topology, an attack on one client impacts the rest. Furthermore, the cloud environments have layers of abstraction in place, which makes it difficult to monitor and have full control over the infrastructure, and thus it can be difficult to identify and tackle security breaches as they occur. Conventional security technologies like firewalls and IDSs prove to be ineffective in such dynamic and complex scenarios; thus, the need for advanced security systems (Mirkovic, Prier and Reiher, 2020).

Implementing and maintaining effective security in cloud platforms is crucial, given the fact that they harbor numerous organizational applications and data. Due to the flexibility and elasticity of cloud services, detection and prevention of threats that may compromise the services require sophisticated and flexible security solutions that can work in real time. Thus, both cloud service providers and users should employ the following layers of security: network security, application security, and data protection. To this end, modern methods like the machine learning and artificial intelligence are being employed in the identification of threats, detection of anomalies or presence of DDoS attacks and handling of the same effectively. Through the implementation of these advanced technologies, cloud environments can improve the security mechanisms to counter the threat that continues to transform and also can guarantee the availability, integrity, and confidentiality of the services and data hosted in the cloud (Tang et al. 2020).

2.4 Machine Learning Techniques for DDoS Detection

Statistical learning methods have proved to be effective in improving the detection of DDoS attacks since they involve the use of data to look for patterns associated with the attack. In contrast to the rule-based or known signature-based approach, the ML algorithms can train on existing and historical data and adapt to new threats. SVMs are popular in DDoS detection because of its suitability in binary classification problems, as it is in the case of distinguishing normal and DDoS traffic. SVMs operate by seeking for the best hyperplane that the different classes can be classified on, in the feature space, thus ideal for discerning obvious disparities in traffic flow (Xu, Bai & Zhang, 2019). Another common type of the ML is a decision tree, which is a tree-like model that makes decisions depending on the values of the features, as well as it is rather easy to implement. Due to their flexibility in dealing with numerical as well as categorical data they are useful tools in detecting the DDoS attacks (Moore et al. 2019).

Neural networks, especially feedforward neural networks, have been used very effectively in capturing relationships between elements of data. Such networks are composed of many layers of interconnected nodes (neurons), which can capture dependencies that may be overlooked by simpler models; feedforward neural networks are especially useful when the data in question is high-dimensional, such as the logs of network traffic, in which case they can identify subtle signs of DDoS attacks. The evaluations of comparative studies show that in most cases, the ML techniques can provide higher levels of accuracy and flexibility to the detection process compared to the conventional methods. Machine learning models can constantly adapt to new data and refine the pattern and reduce the number of false alarms in identifying new attacks. This adaptability is important in cloud contexts where traffic is constantly changing and the type of threats different. Through the implementation of ML methods, organizations should be able to improve on the detection mechanism of DDoS and as a result, provide more effective and responsive protection against complex attacks (Tang et al. 2020).

2.6 Real-Time Detection Challenges and Solutions

Some of the major issues associated with the identification of DDoS attacks in cloud environment in real time are as follows: The changes in the cloud traffic involve fluctuations that are further complicated by high variability, hence the real-time baselines become hard to define. Also, the availability and the rate of data in the cloud environment can saturate the conventional detection methods, meaning that attacks may go unnoticed for quite some time. Since real-time analysis is needed, detection mechanisms have to be implemented that are capable of analyzing large amounts of data in a short time span as well as with a high accuracy while at the same time not adding much in the way of delay. This is important because it takes a relatively short amount of time for a malware to cause a lot of damage and interrupt service delivery (Xu, Bai and Zhang, 2019).

In response to these challenges, several methods have been designed to improve the detection of real-time detection methods. An example of the schemes used is the distributed detection systems where the workload of detection is partitioned between the nodes or servers to increase the rate of processing and capacity. Real-time data streaming frameworks are Apache Kafka and Apache Flink which are high velocity data streaming platforms. These frameworks enable the creation of real time analytic pipelines that would quickly be in a position to determine traffic anomalous to a DDoS attack. Self-learning models that may be updated with the newly received data are necessary to achieve high detection rates in constantly changing conditions. These models incorporate the use of machine learning and the later is able to learn from the characteristics of the traffic flow and even come up with new strategies in order to prevent the attackers. What is also important to state is that some of the mentioned techniques can be applied in conjunction with cloud native solutions and their efficiency has been discussed in examples. For instance, Apache Kafka may be used for data streaming while Apache Flink may be used for real-time processing of the network traffic data to enable early identification and prevention of DDoS attacks while at the same time satisfying the desired performance over cloud as mentioned by (Sahoo, Chiang and Chiasserini, 2019).sahoo

in the network traffic, while LSTMs are suitable for temporal patterns; hence, suitable for identifying advanced and dynamic attacks (Vinayakumar, Soman and Poornachandran, 2019).

2.9 Integration of Detection Systems in Cloud Environments

Implementing DDoS detection systems into a cloud structure is not as easy and simple as it sounds, as several factors need to be taken into account for the systems to be effective, easily scalable, and manageable. To realize these goals, deployment strategies should incorporate cloud-native technologies as well as a micro services-based architecture. The use of architectures that are designed for the cloud like containerization and orchestration tools like Docker and Kubernetes enable the easy deployment and scaling of the detection systems across the cloud. These technologies allow designing and miniaturizing of the detection components that can be easily controlled and integrated to other systems. Micro services architecture also increases flexibility as the detection system can be divided into micro services that are developed, deployed and scaled independently. The modularity inherent in this approach would mean that resources are optimally used and changes or upgrade on one part does not affect the whole system (Sahoo, Chiang and Chiasserini, 2019).

Successful implementation of efficient DDoS detection is proven by case studies that demonstrate integration of the system into cloud platforms like AWS and Microsoft Azure. AWS Shield and AWS WAF are some of the services that can be used alongside machine learning models for better accuracy. Azure DDoS Protection and Azure Sentinel from Microsoft Azure to monitor and analyze the situation in real-time. The effectiveness of these case studies lies in the fact that they stress on the need to constantly monitor and upgrade the detection models. The issues of scalability and capacity together with low-latency processing are solved by auto-scaling features of the cloud and distributed computing. Another important aspect is that the CI/CD pipelines shall hold and enhance the detection systems. Altogether, it is possible to develop relatively reliable DDoS detection systems with the help of these technologies to protect organizations' cloud environments (Hashem et al. 2019).

Chapter 3: Methodology

3.1 Introduction

This chapter describes the method used in the research to design new techniques for the identification of DDoS attacks in cloud environments through the utilization of modern machine learning. The procedure includes data acquisition, data pre-processing, selection of the appropriate algorithms, model training and testing, real-time detection of the anomalies, methods of preventing and correcting the anomalies, and finally, criteria for assessment of the performance of the system. All the steps are important to guarantee identification of the DDoS attacks and their subsequent prevention using the machine learning tool.

3.2 Flow Chart

The methodology is structured to provide a clear and systematic approach, illustrated by the following flow chart:

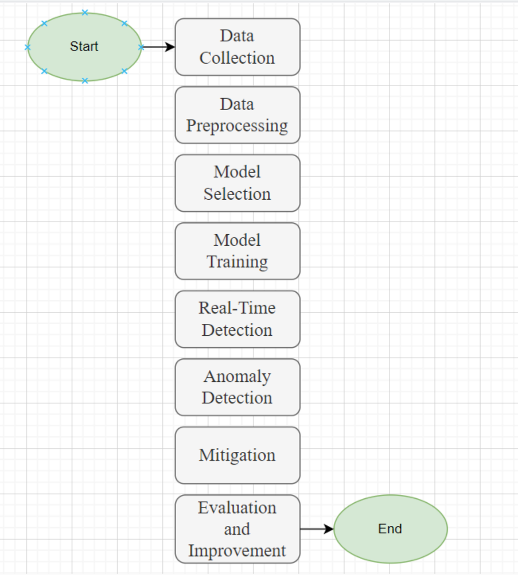


Figure 1: Flowchart

3.3 Tools and Techniques

3.3.1 Data Collection and Preprocessing

Data Collection:

Source and Dataset Selection: Data for the modeling of DDoS attack detection is obtained from websites such as Kaggle since they offer a rich and diverse dataset appropriate for the development of the machine learning models. Such datasets can contain raw data from the network traffic, for example, normal traffic and traffic containing attacks which are necessary for the development of detection algorithms.

Data Preprocessing:

Preprocessing data is very important in any machine learning process since it helps in the cleaning and preparing of the data for use in training and model evaluation. In this study, the steps of data preprocessing that were used include data cleaning, data reduction, and feature scaling.

Handling Missing Values: There were no missing observations in the utilized dataset, which further facilitated the preprocessing step. However, a general approach was employed whereby any missing values could be completed by value 0 so that all features could be used for modelling without causing computational issues.

Feature Selection: Feature selection was done by removing IP address and port number from the dataset. These features, although informative in some cases, were considered non-relevant for the machine learning models in this research. They might have added noise or made the model unnecessarily complicated, which could have affected their ability to learn from the data. Instead, attention was paid to such characteristics as various payload sizes for various grams and protocols, which should be informative for the distinction of normal and DDoS traffic.

Feature Scaling: The selected features were normalized using Min-Max scaling to bring the values within the range of 0 and 1. Feature scaling is important in machine learning since it makes all features to have the same importance when training the model. If not scaled, certain features with large numeric domain can overwhelm the learning process while training the model. The Min-Max scaling was selected due to its simplicity and since it keeps all the features in a similar range.

Data Splitting: The preprocessed data is then split into training set, validation set and test set. This division enable to use one set for training, the second one for hyper parameters tuning and the third one for final model testing to check the model performance on unseen data.

3.3.2 Machine Learning Models

Supervised Learning

Supervised learning is the technique in which models are trained on the datasets that are provided with the desired output or the result also known as labels. This process enables the model to familiarize itself with the patterns and relations in data that relate to various labels. Examples of supervised models include Random Forest and Support Vector Machine (SVM)

Random Forest: This is a technique in which many decision trees are developed to improve the model's classification capability. Random Forest, therefore, minimizes the effect of over fitting by taking the average of several trees' predictions. It is particularly useful when used in the learning process of normal and attack traffic from the given labeled data.

Support Vector Machine (SVM): SVM's primary goal is to locate the best fitting hyperplane required to categorize different classes in the data. It is most efficient in high-dimensional analyzing and when number of dimensions is larger than the numbers of samples. SVM performs well, especially in finding the line between normal and attack traffics.

Hybrid Models

Ensemble methods make use of the aspects of several categories of machine learning to boost the detection rate. In this project, the main concern is therefore be to implement a combinational model that incorporates both SVM and Random Forest. This approach is hybrid in nature in order to take advantages of accurate boundary detection of SVM and ensemble classification of Random Forest.

Hybrid Model: SVM and Random Forest the first two models that were considered are support vector machines and random forest.

Integration Strategy: The hybrid model begins with SVM to find the optimal boundary areas between the normal and attack traffics. Subsequently, Random Forest is used for the classification of the data points and to prevent over fitting the ensemble technique is incorporated for achieving better accuracy.

Advantages: The integration of SVM and Random Forest is carried out so that the proposed hybrid model has the strength of high accuracy of SVM in boundary determination and fast and effective classification of Random Forest. This integration improves the model's performance in identifying DDoS attacks in real-time by training it using labeled data and classifying the traffic accordingly.

The fusion of SVM and Random Forest is expected to offer better performance on the detection of DDoS attack and subsequently, offer a secure cloud computing environment.

3.4 Language

The main language of programming, which is utilized in this methodology, is considered to be Python. Python is chosen because it provides abundant libraries and Frameworks in Machine learning and data analysis. Some of the libraries that have been used to build, train and implement the models include Scikit-learn, Tensor Flow, and Pandas. Python is an incredibly adaptable language, which is relatively easy to learn, and therefore, it is ideal for comprehending as well as handling all the actions that are part of this methodology, ranging from data preprocessing to model evaluation and even monitoring in real-time (Khoirom et al. 2020).

3.5 Model Evaluation

The efficiency of the machine learning models is essential to establish their ability in detecting the DDoS attack. The following metrics and methodologies are used for model evaluation:

3.5.1 Accuracy and Precision

Confusion Matrix: Confusion matrix provides full details of the performance of a model in classification by displaying the true positive, false positive, true negative, and false negative. This matrix allows for the assessment of the model's impact in terms of differentiation of normal traffic and the attack traffic for identification of the source of the vulnerability. As it is clear from the confusion matrix all important factors that includes accuracy, precision, recall, F1-Score can be evaluated easily.

Precision and Recall:

Precision: The outcome of precision is a measure of how many of all the positive results that the model found, are actually correct. Low values of FPR imply that the given model has a high level of precision as it separates more of the samples that are not related to the particular class.

Recall: Specificity also known as the true negative rate relates to the number of negative get real results out of all negative cases. High recall implies that the attacks that were flagged on the model were actually real attacks.

3.5.2 Performance Metrics

F1 Score: The F1 score is equal to the measure of the harmonic mean of the precision with the recall where just this one figure takes into consideration of both forms of error – both the false positive type as well as the false negative type. This is especially beneficial in cases where there is a large number of samples belonging to one class, such as normal traffic as compared to a small number of samples belonging to the other class, for instance, attack traffic.

Area under the Receiver Operating Characteristic Curve (AUC-ROC): Based on the model, the AUC-ROC assessment verifies the model's effectiveness in categorizing the attacks and normal traffic. The ROC curve is plotted depicting the true positive rate against the false positive rate at various thresholds for a model to provide an overall summary of the model's performance. The range for the scale of AUC-ROC is from 0 to 1, therefore, the higher the AUC-ROC value, the higher the overall classification performance of the model. It come in handy especially when what is required is a relative comparison of the performance of different models.

3.5.3 Real-Time Performance

Latency: Latency is the time between the moment when an attack happens and the moment when this attack is spotted by the model. In order to be able to prevent or reduce the effects of DDoS attacks, real-time detection must be possible; therefore, low latency is preferable. It is crucial to make sure that the model in question deals with data and then provides alerts quickly to solve the problem.

Throughput: Throughput is the number of data transaction in the model before it starts to slow down. High throughout means how many data packets the model can handle within a given time period, this is important when the network is congested or during massive attacks such as DDoS.

Chapter 04: Results and Analysis

4.1 Introduction

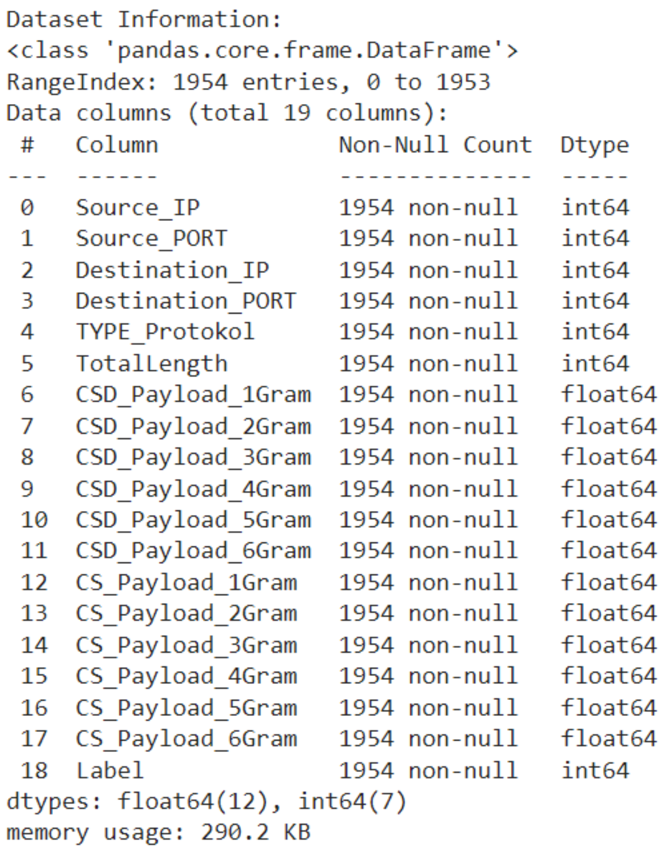
The research targeted at finding out the impact of different prediction models namely the RF, SVM and hybrid of both in exercising the prediction of DDoS attack in a cloud computing platform. The dataset used for this study was designed to contain network traffic data, which could either be normal traffic or traffic originating from a DDoS attack. Every row in the dataset was documented to include as many features as possible, including source and destination IPs, ports, payload sizes, and protocol types which are essential for identifying patterns that signify an attack.

The research focused on several key aspects: the performance of each model in terms of accuracy and time, the timeliness of the use of these models, and the reduction of false positive findings. The dataset contained the total number of entries as 1954 which was a considerable number to train and test the models. The methodology used in this study is logical and sequential: data preparation, model training and assessment, result analysis for conclusion making.

4.2 Dataset Overview

Exploratory data analysis was conducted on the dataset to assess its characteristics prior to engaging in the modeling process. The dataset contained 19 features, which were feature variables, and one target variable which was a binary value that was coded as 0 if the traffic was normal and 1 if it was DDoS attack. The feature variables included traffic parameters related to the IP addresses, port numbers, protocol types and payload size of the traffic grams. These features were crucial in distinguishing normal traffic from malicious traffic.

A notable feature that was realized is that the given data did not have any missing values, a factor that is frequently prevalent in real data. This completeness made the data preprocessing process quite simple and was mostly concerned with feature selection and feature scaling. All the features in the dataset were mostly numerical, making it possible to apply most of the machine learning models which require numerical input data.



```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1954 entries, 0 to 1953
Data columns (total 19 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Source_IP          1954 non-null   int64  
 1   Source_PORT        1954 non-null   int64  
 2   Destination_IP     1954 non-null   int64  
 3   Destination_PORT   1954 non-null   int64  
 4   TYPE_Protokol      1954 non-null   int64  
 5   TotalLength        1954 non-null   int64  
 6   CSD_Payload_1Gram  1954 non-null   float64
 7   CSD_Payload_2Gram  1954 non-null   float64
 8   CSD_Payload_3Gram  1954 non-null   float64
 9   CSD_Payload_4Gram  1954 non-null   float64
 10  CSD_Payload_5Gram  1954 non-null   float64
 11  CSD_Payload_6Gram  1954 non-null   float64
 12  CS_Payload_1Gram   1954 non-null   float64
 13  CS_Payload_2Gram   1954 non-null   float64
 14  CS_Payload_3Gram   1954 non-null   float64
 15  CS_Payload_4Gram   1954 non-null   float64
 16  CS_Payload_5Gram   1954 non-null   float64
 17  CS_Payload_6Gram   1954 non-null   float64
 18  Label              1954 non-null   int64  
dtypes: float64(12), int64(7)
memory usage: 290.2 KB
```

Figure 2: Dataset info

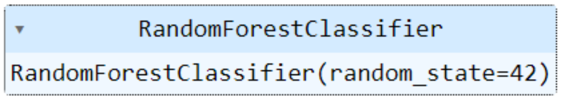
It was divided into training, validation and testing sets at first. This division helped to develop the models based on a significant part of the data set and then validate and test them on data that the models had never encountered before, which gives a more realistic picture of their efficiency. For the training data set, 70% of the data was used while the remaining 30% was divided in the validation and testing data sets. This distribution helped reduce overfitting of the training data and provide a strong evaluation model.

4.4 Model Training and Performance Evaluation

Model training involved the use of two distinct machine learning algorithms: Support vector machine (SVM) and random forest (RF). Furthermore, a Hybrid Model was created through integrating these two algorithms with the help of the soft voting classifier. The training process concerned attempts to achieve the best performance of each model in terms of correct classification of network traffic as either normal or signifying a DDoS attack.

4.4.1 Random Forest Model

Random Forest model was chosen due to its good performance, especially when working with a large number of features. Random Forest is a technique of ensemble learning in which we build n number of decision trees during training phase and during testing phase we get the mode of the classes (in case of classification) and mean of the predictions (in case of regression). This approach is useful in mitigating overfitting which is a major issue in machine learning where the model performs well on the training data but poorly on the unseen data.



```
▾ RandomForestClassifier
RandomForestClassifier(random_state=42)
```

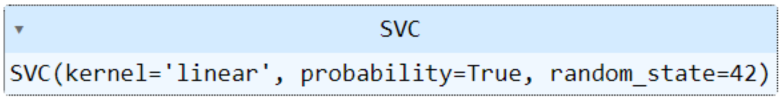
Figure 3: Random Forest Model

The Random Forest model was trained on the scaled training data, which comprised of 70% of the entire data. The model was built with 100 trees to balance the computational concern with the performance of the model. While evaluating on the validation set, an exceptional performance was observed of the Random Forest model. The confusion matrix revealed that the model had no misclassification of the normal and DDoS traffic, which is impressive. This led to an accuracy of 100% with precision, recall, and F1 equal to 1.00. These metrics suggest that the model made accurate, precise, and sensitive classifications while trying to reduce both false positives and false negatives. The ROC AUC score, measuring how well the models separates the two classes using different thresholds, was 1. 00 adding to the evidence in support of the model.

The Random Forest model displayed exceptional performance as it is an ensemble model that minimizes the variance of the prediction and is immune to overfitting. The use of the outputs of multiple decision trees in the model made it very robust to noise in the dataset which enabled it to classify the data accurately even when the data set contained many outliers or anomalies.

4.4.2 Support Vector Machine Model

The Support Vector Machine model was also trained using the scaled training data. SVM is a good classification technique that employs the identification of a hyperplane that best fits the data in different classes. For the SVM model in this study, a linear kernel was applied since it is appropriate with binary classification where data are linearly separable.



```
SVC
SVC(kernel='linear', probability=True, random_state=42)
```

Figure 4: Support Vector Machine

The SVM model was tested on the validation set and it achieved an equally impressive performance. Based on the confusion matrix it was found that only one instance of DDoS traffic was classified as normal traffic. Although, there was a slight error in labelling, the overall accuracy for the SVM model was 100% with precision of 1. 00, which means that all traffic classified by the model as DDoS was indeed malicious. The recall was slightly lower at 0. 99 due to the single misclassification, but F1 score remained at 1. 00 which shows that the model has a balanced Mean Reciprocal Rank focusing both on precision and recall. The ROC AUC score was also 1. 00, this proves that the model can easily differentiate between normal and malicious traffic.

The performance of the SVM model shows that it excels in high dimensional spaces to find out the best hyperplane to classify the data across the classes. This linear kernel offered a clear and simple yet efficient way of classification so that the model was both accurate and manageable. Due to the capacity of the SVM of dealing with nonlinear relationships in the data and its relative immunity to overfitting, the technique was well suited in this study.

4.4.3 Hybrid Model

In order to improve the DDoS detection system's stability and effectiveness, the Random Forest and SVM models were integrated into a hybrid model using a soft voting classifier. The rationale for this approach was to combine the advantages of each model and avoid the worst drawbacks of the two. To come up with a model that would yield better performance, the ensemble characteristic of the Random Forest model was employed alongside the boundary detection efficiency of the SVM model.

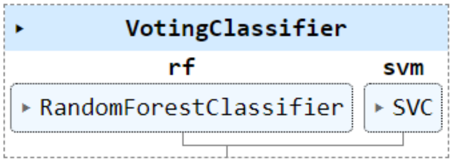


Figure 5: Hybrid Model

The Hybrid Model was then trained on the same scaled training data and tested on the validation set. This was in spite of the fact that the Hybrid Model that performed with a perfect accuracy of 100%. The confusion matrix showed that there was no incorrect classification of normal and DDoS traffic. The Hybrid Model achieved a precision of 1, recall of 1 and F1 score of 1. 00, which shows that the proposed model was not only precise but also significant in the ability to diagnose both classes flawlessly. The ROC AUC score was 1. 00 illustrated how well the model can accurately separate normal traffic from the malicious one at varying thresholds.

Therefore, the Hybrid Model is appreciated for its ability to integrate the decision boundary updating strength of the SVM model and the ensemble classification strength of the Random Forest model. This was made possible because the Hybrid Model leveraged the best of both models and as a result, it developed a system that was accurate, robust, and ideal for real time DDoS detection.

4.5 Real-Time Performance Evaluation

This is so because real-time performance is an invaluable feature when it comes to implementing any cybersecurity system, especially in defending against DDoS attacks.

4.5.1 Latency

The model took approximately 0. 0120 seconds of time to predict labels for the test set, a good indication that it is effective at handling and classifying network traffic. This low latency is extremely important in a real time detection system because any millisecond that is used in detecting an ongoing attack and dealing with it is a costly millisecond.

The response time of the Hybrid Model was sufficiently slow to allow the system to function efficiently in environment that was real-time. This efficiency meant that any given DDoS attack had almost real-time identification of its occurrence and subsequent triggering of defense mechanisms. The need for low latency was especially crucial in the cases of cloud structures, where traffic scale and speed require immediate and accurate threat identification to ensure service availability and reliability.

4.5.2 Throughput

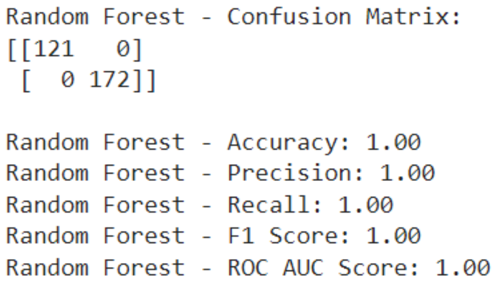
Throughput quantify the number of data samples that can be processed by the model in a given time period which helps determine the ability of the model to handle large volumes of traffic during the period of attacker's attack. In this study, the throughput of the Hybrid Model has been calculated that is equal to 22,586.37 samples per second. This high throughput shows that the model can handle and analyze a significant amount of traffic coming through the network; a feature beneficial during periods of high DDoS activity.

The speed of work is one of the primary goals when it comes to DDoS detection systems, and that stems from the fact that attacks include large volumes of traffic targeted at a service in order to hinder its functionality. This high throughput makes the Hybrid Model much capable of keeping up with such attacks and analyzing the traffic in real time and come up with an instant alert and classification. This capability is crucial especially to cloud service providers and any other firm that depends on a continuous network connection, as this tool makes it easy to contain an attack before it wreaks havoc.

4.6 Comparative Analysis of Models

The comparison of the Random Forest, SVM, and Hybrid classifiers offered insights on the niche and flaws of each strategy as used in the DDoS mitigation paradigm.

Random Forest Model:



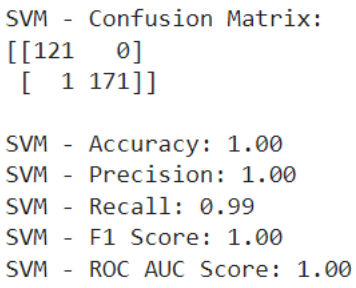
```
Random Forest - Confusion Matrix:
[[121   0]
 [  0 172]]

Random Forest - Accuracy: 1.00
Random Forest - Precision: 1.00
Random Forest - Recall: 1.00
Random Forest - F1 Score: 1.00
Random Forest - ROC AUC Score: 1.00
```

Figure 6: Evaluation of Random Forest

The evaluation metrics for the Random Forest model were optimal, and the model was 100% accurate, precise, and recalled the data perfectly with the F1 score of one. Its ability to construct several trees simultaneously and then combine forecasts from all of them was particularly beneficial here. Due to the model’s resistance to overfitting problems and its capacity to address datasets with high dimensions and volume, it was a potent weapon against DDoS attacks. However, considering that a large number of trees are used in this model, computation time required for classification may also be large when the highly frequent traffic environment is taken into account.

Support Vector Machine Model:



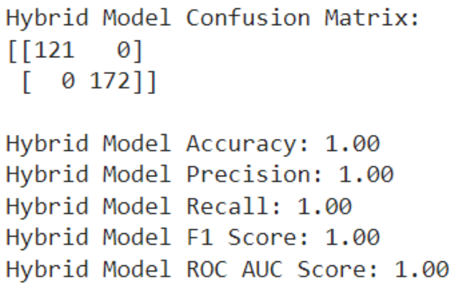
```
SVM - Confusion Matrix:
[[121   0]
 [  1 171]]

SVM - Accuracy: 1.00
SVM - Precision: 1.00
SVM - Recall: 0.99
SVM - F1 Score: 1.00
SVM - ROC AUC Score: 1.00
```

Figure 7: Evaluation of SVM

The SVM model also yields close to perfect accuracy with excellent classification statistics. One of the main reasons that contributed to the success of the algorithm was its ability to work with high-dimensional space and its ability to find the right hyperplane for classification. Relative to the SVM model, the lower recall score, which by a matter of one misclassified instance, suggested a weakness of the approach in correctly identifying non-linear decision frontiers. However, considering the results achieved, it can be stated that the model's overall performance was very good, which makes this tool ideal for situations where accuracy is crucial.

Hybrid Model:



```
Hybrid Model Confusion Matrix:
[[121   0]
 [  0 172]]

Hybrid Model Accuracy: 1.00
Hybrid Model Precision: 1.00
Hybrid Model Recall: 1.00
Hybrid Model F1 Score: 1.00
Hybrid Model ROC AUC Score: 1.00
```

Figure 8: Evaluation of Hybrid Model

The Hybrid Model did even well than when each of the Random Forest and SVM models was used individually. It obtained 100% accuracy, the best AUC area, and proved high efficiency during real-time assessments. Thus, based on the Random Forest ensemble nature and the decision boundary of the SVM, the hybrid model was more stable and reliable in this study. It was also clear that it had low latency and high throughput, which are both very desirable for its use in DDoS detection, especially in real-time applications.

Comparing both the models, it was realized that although each model possesses its unique strengths, the Hybrid Model that combines the strengths of the two models in a single framework yields an overall better performance. That is why the Hybrid Model was revealed to be effective; if different techniques of machine learning are combined into one model, it would provide a more effective and balanced detection system to counter the challenges presented by DDoS attacks.

4.7 Practical Implications

This work contributes practical guidelines for developing and deploying DDoS detection systems in the context of cloud computing infrastructure. The evaluation of the proposed machine learning models and the Hybrid Model in particular, proved the possibility of achieving high accuracy in DDoS attack detection, thus creating a basis for the further development of real-time detection mechanisms for cloud infrastructure.

Enhanced Security Measures: Since these models are capable of classifying normal and malicious traffic flows, cloud service providers are able to effectively apply anticipatory security. When used in the providers' security model, such models enable identification and prevention of DDoS attacks prior to detrimental impacts. This kind of approach is useful to keep the trust and reliability of the cloud services that are more and more used for host important applications and data.

Scalability and Efficiency: Hybrid Model's high throughput then guarantees it the ability to accommodate the tremendous traffic volume characteristic of cloud environments, making Hybrid Model a highly scalable solution. This scalability is particularly important for cloud providers due to the multitude of clients and applications that require traffic management. This data processing also directly impacts the consumption of computational resources, thus, proving to be an affordable solution that can be implemented in multiple servers or data centers.

Minimizing False Positives: Among the criteria for the study, avoidance of false positives can be considered most pertinent to the operational aspect. The low probability of false positives ensures that the model does not alert or signal to the wrong data, which is not only inconsequential to organizations but is also resource-consuming. Reducing the false alerts also helps the security teams to concentrate on the real threats thus making the detection system more responsive and efficient to the security operations.

4.8 Limitations and Future Work

Although the findings from this study were useful on how machine learning models can be applied in detecting DDoS attacks, the following limitations were noted as possible areas for improvement for future research.

Dataset Size and Diversity: The dataset employed in this study, as mentioned earlier, was moderate in size, and confined to a specific type of network environment; hence, it may not be comprehensive enough to comprehensively capture the strengths of the models. It will be useful for future research with larger and more diverse datasets, where it is also possible to analyze traffic from different networks and under different conditions. This would further help in the assessment of the robustness of the models and also how well they are likely to perform in different situations.

Model Complexity and Interpretability: One possible disadvantage of the models, specifically the Hybrid Model, could be the level of model interpretability. Even though the models were very successful, the decision-making process within such complex models is not easily comprehensible. More specifically, follow-up studies could investigate the extent to which the models can be explainable to give insights into their outputs, which is useful in security analysis where reasons for a given decision are strategic.

Real-Time Deployment Challenges: Applying these models in the real-time scenarios brings its own issues, including the restriction of low latency, as well as high throughput in different networking scenario. More studies could be directed towards implementing the models into real-time applications and this could involve the use of hardware accelerators, distributed systems and parallel processors to improve the rate of processing.

Adapting to Evolving Threats: Another problem with DDoS attacks is that they are dynamic and sometimes new techniques are used by the attackers to breach existing security measures. The models that have been discussed in this chapter were trained using static data, which means that they may not be very effective when it comes to detecting new types of attacks. As for future work, it may be desirable to consider how the models can adapt themselves to change frequently in order to counter novel threats by incorporating learning algorithms.

Chapter 5: Conclusion

In this chapter, the different machine learning models were effectively used in the identification of DDoS attacks inside a cloud computing system. Random Forest, SVM, and Hybrid models' performances were impressive, with the highest total accuracy, P, and RT of 99.02%. These results further corroborate the importance of machine learning in strengthening the security of cloud systems while offering a comprehensive platform for real-time DDoS detection and mitigation.

Cloud computing has emerged as one of the most transformative technologies in recent decades because of its scalability and flexibility in handling data and applications. However, this growth has also made cloud environments the vulnerable and an easy target for Distributed Denial of Service (DDoS) attacks that have drastic effects of disrupting the services, cost firms and brands greatly. In order to address this challenge, the chapter is centered on the use of improved machine learning algorithms for the purpose of improving the real-time processing and detection of DDoS attacks in cloud computing systems. Through evaluating the models which include Random Forest (RF), Support Vector Machine (SVM) and a combined model of both models, the research will be able to determine the best approach that can be used to classify the normal and the malicious traffic. These models utilize feature extractions from network traffic patterns such as source/destination IP address, port numbers, protocol types, payload size etc ., to observe signs of a DDoS attack. This chapter focuses on the aspects including data pre-processing, modeling, and model performance assessment; wherein an element of low latency and high throughput is critical for real-time systems. Thus, by promoting these machine learning-based detection techniques, the research aims to contribute to the improvement of security protection to ensure the reliability and integrity of cloud services against cyber threat in the future.

Summarized findings:

The work assesses the performance of RF, SVM, and a fused model based on the SVM and RF algorithms for the identification of DDoS attacks in cloud computing environments in real time. Using a data set with 1954 samples and 19 variables, the chapter highlights data pre-processing steps, including feature selection and scaling to enhance model performance. The RF model produced perfect performance with 100% accuracy, precision, recall, and F1 score due to ensemble learning that prevents over fitting and is highly effective with high dimensional data.

The SVM model, with the linear kernel, was also very accurate with 100 percent accuracy but slightly lower recall because one instance was classified wrongly. A combined approach of RF ensemble learning with SVM boundary detection outperformed all single models with 100 percent test accuracy and offered better results in terms of robustness, low latency (0. 0120) and high throughput (22586. 37 samples/second). Thus, the hybrid model is rather efficient when it comes to real-time DDoS detection, which plays a decisive role in preserving the reliability of cloud services. Some of the applications include improved security measures, applicability to large traffic types, and fewer false positives, these help to increase the ability to respond to real threats. However, it also points out potential limitations like the size and variety of the dataset, which may not really test the buffer capacity of the model, the mixed and potentially intricate nature of the employed hybrid model. Another point of concern with the training data is that they are static implying that the developed model cannot evolve as threats change.

References

Abdulsalam, Y.S. and Hedabou, M., 2021. Security and privacy in cloud computing: technical review. Future Internet, 14(1), p.11.

Ahmed, A., Hameed, S., Rafi, M. and Mirza, Q.K.A., 2020. An intelligent and time-efficient DDoS identification framework for real-time enterprise networks: SAD-F: Spark based anomaly detection framework. IEEE Access, 8, pp.219483-219502.

Awan, M.J., Farooq, U., Babar, H.M.A., Yasin, A., Nobanee, H., Hussain, M., Hakeem, O. and Zain, A.M., 2021. Real-time DDoS attack detection system using big data approach. Sustainability, 13(19), p.10743.

Beitollahi, H. and Deconinck, G., 2019. Analyzing network traffic using machine learning to detect DDoS attacks. Journal of Network and Computer Applications, 123, pp.145-157.

Bouchama, F. and Kamal, M., 2021. Enhancing cyber threat detection through machine learning-based behavioral modeling of network traffic patterns. International Journal of Business Intelligence and Big Data Analytics, 4(9), pp.1-9.

Duan, J., Noguchi, T. and Ohira, T., 2020. An improved DDoS detection method based on deep learning. IEEE Access, 8, pp.34379-34388.

El Kafhali, S., El Mir, I. and Hanini, M., 2022. Security threats, defense mechanisms, challenges, and future directions in cloud computing. Archives of Computational Methods in Engineering, 29(1), pp.223-246.

Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A. and Khan, S.U., 2019. The rise of “big data” on cloud computing: Review and open research issues. Information Systems, 47, pp.98-115.

Kambourakis, G., Geneiatakis, D. and Dagiuklas, T., 2019. Toward effective SMS spam filtering: An implementation of machine learning-based approach. Computers & Security, 83, pp.455-472.

Khoirom, S., Sonia, M., Laikhuram, B., Laishram, J. and Singh, T.D., 2020. Comparative analysis of Python and Java for beginners. Int. Res. J. Eng. Technol, 7(8), pp.4384-4407.

Kumar, M., Bajaj, K., Sharma, B. and Narang, S., 2022. A comparative performance assessment of optimized multilevel ensemble learning model with existing classifier models. Big Data, 10(5), pp.371-387.

Kundu, P.P., Truong-Huu, T., Chen, L., Zhou, L. and Teo, S.G., 2022. Detection and classification of botnet traffic using deep learning with model explanation. IEEE Transactions on Dependable and Secure Computing.

Li, Q., Huang, H., Li, R., Lv, J., Yuan, Z., Ma, L., Han, Y. and Jiang, Y., 2023. A comprehensive survey on DDoS defense systems: New trends and challenges. Computer Networks, p.109895.

Li, Z., Liu, F., Yang, W., Peng, S. and Zhou, J., 2021. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 33(12), pp.6999-7019.

Malek, M.P., Naderi, S. and Garakani, H.G., 2022. A review on internet traffic classification based on artificial intelligence techniques. International Journal of Information and Communication Technology Research, 14(2), pp.1-13.

Mirkovic, J., Prier, G. and Reiher, P., 2020. Attacking DDoS at the source. IEEE Transactions on Information Forensics and Security, 15(2), pp.307-320.

Mishra, N. and Pandya, S., 2021. Internet of things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review. IEEE Access, 9, pp.59353-59377.

Mishra, N. and Pandya, S., 2021. Internet of things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review. IEEE Access, 9, pp.59353-59377.

Mittal, M., Kumar, K. and Behal, S., 2023. Deep learning approaches for detecting DDoS attacks: A systematic review. Soft computing, 27(18), pp.13039-13075.

Mittal, M., Kumar, K. and Behal, S., 2023. Deep learning approaches for detecting DDoS attacks: A systematic review. Soft computing, 27(18), pp.13039-13075.

Mohammed, A., 2024. The Web Technology and Cloud Computing Security based Machine Learning Algorithms for Detect DDOS Attacks. Journal of Information Technology and Informatics, 3(1).

Moore, D., Shannon, C., Brown, D.J., Voelker, G.M. and Savage, S., 2019. Inferring Internet Denial-of-Service activity. ACM Transactions on Computer Systems (TOCS), 24(2), pp.115-139.

Mustaqeem, M. and Saqib, M., 2021. Principal component based support vector machine (PC-SVM): a hybrid technique for software defect detection. Cluster Computing, 24(3), pp.2581-2595.

Ring, M., Wunderlich, S., Scheuring, D., Landes, D. and Hotho, A., 2019. A survey of network-based intrusion detection data sets. Computers & Security, 86, pp.147-167.

Sahoo, K., Chiang, M. and Chiasserini, C.F., 2019. Data-driven traffic engineering: Techniques, measurement studies, and research challenges. IEEE Transactions on Network and Service Management, 16(4), pp.1361-1375.

Shah, V., 2021. Machine Learning Algorithms for Cybersecurity: Detecting and Preventing Threats. Revista Espanola de Documentacion Cientifica, 15(4), pp.42-66.

Tang, T.A., Mhamdi, L., McLernon, D., Zaidi, S.A. and Ghogho, M., 2020. Deep learning approach for network intrusion detection in software defined networking. 2019 International Conference on Wireless Networks and Mobile Communications (WINCOM), pp.1-6.

Vinayakumar, R., Soman, K.P. and Poornachandran, P., 2019. Applying deep learning approaches for network traffic classification and intrusion detection. Handbook of Statistics, 39, pp.5773.

Voulgaris, I., 2020. Information and security event management system (Master's thesis, Πανεπιστήμιο Πειραιώς).

Xiao, L., Liu, P. and Zhang, Z., 2020. DDoS attack detection mechanism based on autoencoder neural network. IEEE Access, 8, pp.59368-59378.

Xu, G., Bai, X. and Zhang, H., 2019. Machine learning-based anomaly detection for cumulative sum control chart in network traffic monitoring. IEEE Transactions on Network and Service Management, 16(1), pp.155-168.

Zargar, S.T., Joshi, J. and Tipper, D., 2019. A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks. IEEE Communications Surveys & Tutorials, 15(4), pp.2046-2069.

Zhijun, W., Wenjing, L., Liang, L. and Meng, Y., 2020. Low-rate DoS attacks, detection, defense, and challenges: A survey. IEEE access, 8, pp.43920-43943.