Global Scientific JOURNALS

GSJ: Volume 11, Issue 12, December 2023, Online: ISSN 2320-9186
www.globalscientificjournal.com

A NAÏVE BAYES-BASED MODEL FOR DETECTING ATTACKS IN CLOUD COMPUTING ENVIRONMENT

Yusuf S. O,
Department of Engineering and ICT, Federal Radio Cooperation of Nigeria, Idofian, Kwara State
solihuyusuf@yahoo.co.uk

KeyWords
Machine learning, Algorithm, Cloud computing, Cloud, CIDDS-001, DDoS Attack, IDS, Security.

ABSTRACT

Cloud computing platforms are getting popular among individuals and corporate organisations. The popularity of this platform has attracted some people with malicious intent to launch attacks on the environment in recent times. Signature-based and the Machine learning (ML) approaches have been widely used for detecting some intrusions in the cloud computing platforms. When there is no sophisticated security the advantages the cloud computing may have to provide as services will have less integrity. However, ML approaches have been reported to be more promising to detect attacks in the cloud environment as they can intelligently identify attacks and report low false positive rate. This study proposes to use innovative approaches to achieve improved Machine Learning methods for the classification of attacks in the chosen dataset. The dataset has four captures with each of them containing network captures. Random sub-sampling technique was used for handling the large size of the data without changing the patterns in the dataset and the selection was unbiased. The dataset was pre-processed and Mutual-info filter technique was used for the selection of promising features from the dataset features, with split train test (size 0.25 at random 42) used to train the model. Results were determined by the hyper parameter settings, so the feature scores were determined using feature scores techniques and it shows that 'Bytes' was promising in case1, while 'Packets' was promising in cases2, 3 and 4,'. The classified attack detection was made possible through a multiclass algorithm, Naive-Bayes Algorithm which was deployed in building four intrusion classification models from the set of four captures. The results of the performance metrics are; accuracy score for cases 1, 2, 3, and 4 were 95.06%, 94.69%, 97.14% , 97.38% respectively. The results obtained are satisfactory and the system achieved an average accuracy of 95.75% for all captures. The final settings were fixed to size = 0.25 and random = 42 by stratifying y. The typical 'weighted' was picked because of the characteristics of the data. Thus, it was concluded that the average setting depend on the kind of procedure used to develop a model. Mutual-info fared better on CIDDS-001 in terms of metrics. The evaluation's findings lead to the conclusion that, in terms of feature scores, 'packets' in the CIDDS-001's features is the most promising.

INTRODUCTION

Cloud computing is a computing model that makes use of Internet to deliver information technology devices such as infrastructure applications, and platforms as a service (Arora & Bal, 2021). Arora et al., 2021 further explained that current internet-based development has improved processing capacity, power, and flexibility. Usual financial constraints and the hike in computational costs necessitate data storage, analysis, and presentation that have brought crucial changes to today's cloud model (Borylo, Tornatore, Jaglarz, Shahriar, Cholda, & Boutaba, 2020).

Cloud computing is a processing and computing infrastructure for all kinds of data resources used to handle large amounts of data. To better understand the relationship between computing power in cloud computing, Butt, Mehmood, Shah, Amin, Shaukat, Raza, and Piran (2020) defined cloud computing as the on-demand accessibility of end-user resources, specifically information storage and processing power, without direct special organization by the customer. The cloud activates new means of offering services with innovative, technical and price options.

The composition of cloud computing consists of infrastructure, platform and software as services in a distributed computing. "Distributed computing" are popular word combinations in cloud computing that have different meanings on different occasions. However, Dang, Piran, Han, Min, and Moon (2019) added that distributed computing provides the client with public and private data on a platform on the Internet. As technology advances, access to data becomes possible at less or no cost. Cloud computing is an approach that makes the storage and access of data easier (Mamatha & Kandewar, 2019). In fact, cloud computing is changing the way businesses work. Organizations have seen a move away from the old ways of manipulating data. Nowadays, transactions, presentations and advertising no longer require you to walk around with tons of files and devices in your hands. All you need is your internet. It is critical to know about cloud definition and architecture. According to Sharma and Trivedi (2014), cloud computing is a collection of resources that can be upgraded and downsized as needed. It is available over the Internet in a self-service model, requiring little or no interaction with the service provider. Edge computing is a version of cloud computing for processing time-sensitive data, offers application developers and service providers distributed computing capabilities at the edge of a system (Stefan & Liakat, 2015).

However, cloud computing is viewed as a body of water, called a dam, from which homes in the city draw water through a sewage system. Cloud computing is the host of multiple computers, allowing for location-independent and cost-effective sharing of resources on demand (Mamatha et al., 2019). In addition to location independence and cost efficiency, it also offers other advantages such as better scalability and improved flexibility (Mamatha et al., 2019). Cloud computing presents some security threats that delay the quick adoption of computing model, such as client and association susceptibility (Mathkunti, 2014). In order to achieve cloud computing, there must be a networked computer system consisting of software and hardware. According to a research, cloud computing has some disadvantages which are listed below

i. It may cause lock-in due to proprietary technology
ii. It may cause network latency by using internet to use some cloud applications
iii. In some cases cloud provider may cost more than on-premises systems
iv. It may be problematic while integrating on-premise system and cloud based system

The cloud is built on a large-scale distributed facility where resources are typically virtualized and offered services are distributed to users in the form of virtual machines, deployment environments, or software.

Idhammad, Afdel, and Belouch (2018) explained that attacking tools have made the existing sophisticated cloud intrusion detection systems increasingly sophisticated with large amounts of network traffic data, dynamic and complex characters, and type of attacks that are recent. Clearly, a cloud intrusion detection system should analyze huge amounts of network traffic data, efficiently identify fresh attack behaviours, and achieve high accuracy with low falsehood. However, pre-processing, analyzing, and identifying attacks in cloud platforms using conventional techniques have become very expensive in terms of computation, time, and budget. Therefore, efficient attack detection on cloud platforms requires the introduction of new distributed and intelligent methods such as machine learning techniques.

Machine learning is an area of computer science that uses methods of statistics in the creation of programs that get better performance over time and discover patterns in large data that humans are unlikely to find (Swamynathan, 2019). Machine learning is about studying and building algorithms will be able to be trained from data and make some predictions. Such algorithms work by building a model from sample inputs to make data-driven predictions or decisions, rather than following strictly static program instructions (Swamynathan, 2019). However, this study aims to build a Naive-Bayes -based model using to identify attacks on cloud computing platforms.

The methodology consists of the sequence of collecting CIDDs-001 data sets, pre-processing the data, selecting some improved data set functions using the filtered basic technique, identifying the cloud attacks, and then evaluating the performance of the model. Independent Individual variables that serve as input to the Naïve Bayes model are called features. The features can be understood as representations or characteristics that describe the data and help the models make prediction of the classes/labels. For instance, features in a structured dataset, kept in a CSV format, refer to each column representing a measurable piece of data that can be used for analysis.

According to Said, El Emary, Alyoubi and Alyoubi (2016), Cloud computing has a service models which are categorized into

i. Software as a Service,
ii. Platform as a Service (PaaS),
iii. Infrastructure as a Service (IaaS)

i. Software as a Service: Meanwhile, Mamatha et al. (2019) have pointed out that with SaaS, some online applications and software applications such as CRM, ERP used in managing the organizations are offered as a service. Santosh and Goudar (2012) explained that cloud consumers publish their applications in a hosting platform that application users can access over networks from different clients' devices (e.g. web browser, PDA, etc.). Examples of SaaS in real life are SalesForce.com, Gmail, Google Docs, etc.

ii. Platform as a Service (PaaS): The environments required to build up the applications are provided as a service, PaaS is a platform that supports it all. PaaS cloud enables users to develop cloud services and applications (e.g. SaaS) directly (Santosh et al., 2012). Hence, the dissimilarity between SaaS and PaaS is that SaaS only hosts completed cloud applications while PaaS provides a development platform that hosts both completed and running cloud applications (Santosh et al., 2012). An example of PaaS is Google AppEngine.

iii. Infrastructure as a Service (IaaS): IaaS is the basis of cloud services. This type of service provides storage, computing power, and management of the organization's database on-demand to the organization. Cloud customers clearly use IT frameworks in the IaaS cloud. In an IaaS cloud, virtualization is widely used to merge/decompose real resources in amazing ways. Cloud buyers have a specific opportunity to acquire or contract resource interests. The creation of free standalone virtual machines (VMs) independent of important data and other VMs is the most common method of virtualization (Arora et al., 2021).

As reported by Santosh et al.(2012), cloud end-users nonstop use cloud infrastructures provided in the IaaS cloud. Virtualization is being widely adopted in the IaaS cloud to merge physical hardware on an ad hoc basis to meet the demand of resource by the cloud users. The important tactic of virtualization is to set-up stand-alone virtual machines that are independent of both the other VMs and underlying hardware (Santosh et al., 2012). An example of IaaS is Amazon's Elastic Compute Cloud (EC2) including: AbiCloud, Eucalyptus, Nimbus and OpenNebula which are however the four popular cloud computing platforms have been identified by (Santosh et al., 2012).

AbiCloud: Peng, Zhang, Lei, Zhang, Zhang, and Li (2009) see Abicloud as a cloud computing platform is used in building, integrating and managing hybrid clouds in homogeneous platform where users can automatically and easily make provision and manage the resources and other virtual devices with Abicloud

Eucalyptus: Eucalyptus means "Elastic Utility Computing Architecture for Linking Your Programs to Useful Systems" (Peng et al., 2009) and it's mainly used for building an open-source privately

owned cloud platform. Eucalyptus can be used to link the user's programs to the needed systems. Being elastic, using the workstation implementations of Elastic Cloud Computing, and a famous computing standard are based on a protocol of service-level that allow users to lease networks for computing capacity. Presently, Amazon's EC2 is Eucalyptus compatible which can support other types of clients with minimal changes.

Nimbus: Nimbus is not a closed toolset and also a cloud computing solution that provides infrastructure as a service (Peng et al., 2009). It grants users the access to hire remote resources and build the needed computing platform through the use of virtual machines. Generally, all of these existing functional components can be classified into three types. One of them is client-based modules that support all types of cloud clients. EC2 customer module all belong to this type of component, Context customer module, cloud customer module, and reference customer module. The second type of components are the management of modules resource in the background, which are largely used to manage all types of physical resources on the cloud computing platform, including work service management modules, IaaS gateway modules, EC2 and other cloud platform support modules, and workspace pilot modules. The third type of components includes a web service resource framework module, a context agent module, an EC2 WSDL module, and a remote shim.

OpenNebula: OpenNebula is cloud service framework (Peng et al., 2009). It grants users to make provision and manage the virtual machines on some physical resources and it can set up user data clusters on an elastic virtual infrastructure that can adapt to service load changes automatically. The chief dissimilarity between OpenNebula and Nimbus is that OpenNebula does not implements a remote interface based on EC2 or WSRF through which the user can handle any security related issues while Nimbus does. OpenNebula is also an open and elastic virtual infrastructure management tool that allows storage, network and virtual technologies to be synchronized and allows users to dynamically deploy services on the distributed infrastructure according to data-centre and remote cloud resource allocation strategies. Through the internal interfaces and the OpenNebula data centre platform, users can easily set up all types of clouds.

OBJECTIVES

The specific objectives of this study are to:

i. collect and analyse cloud computing CIDDS-001 datasets;

ii. pre-process the datasets using data cleaning techniques to improve the quality;

iii. select promising features using Mutual-info filter-based technique; and
iv. classify cloud computing attacks with Naïve-Bayes algorithm and evaluate the model performance using accuracy, precision, recall and f1-score metrics.

SIGNIFICANCE OF THE STUDY

This study focuses on building a novel machine learning model to identify attacks in cloud computing environments. Organizations such as financial institutions that deal in storing large amounts of financial data, security researchers, government and general public will benefit enormously from the study. The threats which arise from the activities of attackers on cloud computing has increased processing time and slowed system performance which have put various organisations into some trouble moments, and that have been a cause for concern for many years. With a reliable and effective attack identification system, such concern will to a large extent be abolished. Cloud computing technology will therefore be reliable for the organisations to use without much worry.

RELATED WORKS

A few of related works were reviewed and research gaps were identified.

Kishore and Nagaraju (2022) evaluated the efficiency of classification methods in detecting attacks, showing that machine learning classifiers accurately detect DDoS attacks in a relatively short period of time. It was also noted that the optimal results for accuracy, F-score, and specificity in the trials were achieved by Logistic Regression, Gradient Boost, and Naive Bayes. Using logistic regression and Naive Bayes, the best precision value was found. It has been determined that AdaBoost provides the highest levels of accuracy and ROC AUC values. The Gradient enhancement algorithm's log loss value is the best. The research aim was to use machine learning classification algorithms to detect DDoS attacks on CIC-DDoS2019 dataset but it reported that naïve bayes produced good result but not best as expected.

To stop DDoS activities between virtual machines, a DDoS detection mechanism in the virtualization layer was suggested by (RadhikaP & KanimozhiP, 2021). Identifying that the traditional defence mechanism, such as firewalls that are unable to detect insider attach, RadhikaP et al., (2021) proposed detection method that combines the radial basis function (RBF) and particle swarm optimisation (PSO) to detect DDoS attacks and classify communication between virtual machines. The research aimed to propose DDoS detection approach in the hypervisor layer to discourage DDoS activities between virtual machines. Kushwah and Ranga (2020) proposed a new method to detect DDoS attacks in a cloud computing environment. The aim was to propose a new system for detecting DDoS attacks in cloud computing environment using voting extreme learning machine V-ELM artificial neural network.

Based on aggregated statistics from NetFlow, Majed, Noura, Salman, Malli and Chehab (2020) proposed a method for distinguishing DDoS traffic from ordinary traffic using Z-score variance metrics. for each of the aggregated flow, the obtained Z-scores of the three features captured were compared to the corresponding threshold values (CV). In the research work, using statistical thresholds is not practical because they have to be updated manually each time. Pham, Fang, Ha, Piran, Le, Le, Hwang, and Ding (2020) briefly examined the concepts, characteristics, security, and application of IoT-enhanced edge availability. In their information-driven society, the writers briefly examined the concepts, capabilities, safety, and applications of IoT-enabled edge processing, as well as those technologies' security implications. They described factors to take into account when developing a distributed, secure, and scalable computing architecture.

Bouyeddou, Kadri, Harrou, and Sun (2020) suggested a detection approach for DDOS attacks based on a statistical assessment of continually ranking likelihood scores and an exponential smoothing scheme. The aim of the research work was to introduce a reliable detection mechanism based on the continuous ranked probability score (CRPS) statistical metric and exponentially smoothing ES scheme for enabling efficient detection of DOS and DDoS attacks. They observed that although model development requires a lot of time, these statistical detection techniques had good detection fidelity. The identification of DDoS assaults by machine learning algorithms rapidly develops into a study topic to address the aforementioned issues.

Sarraf (2020) performed analysis and identification of DDoS attacks using ML methods. The CICIDS2017 dataset subset, which included 200,000 DDoS examples and benign classes, were used. One of the data's 84 categorical and numerical characteristics was deleted, leaving 83 features for machine learning modelling and feature engineering development. In feature engineering, correlation analysis and feature importance discovery utilising decision trees have been applied. The findings indicate that the most practical features were "Flow ID," "SYN Flag Cnt," and "Dst IP." 100% accuracy was achieved on CICIDS2017 datasets.

Roempluk and Surinta (2019) used KNN, MLP, and SVM machine learning algorithms to suggest an appropriate method to detect DDoS attacks. This method was evaluated on the KDD CUP 1999 and NSL-KDD datasets, the datasets were checked and deleted duplicated and decreased from 4million plus to 500 thousand plus records. The model was trained by K-nearest neighbour (KNN), multi-layer perception and support vector machine. The resource scheduling issue in cloud computing was resolved by El-Boghdadi and Rabie (2019) using DRL for Cloud Scheduling (DRLCS), one of the cutting-edge machine learning algorithms. The research stated that the subsequent approach used by DeepRM and DeepRM2, certain factors in cloud planning must be considered, such as processor, memory, task duration, and virtual machine load balancing that are necessary.

DeepRM and DeepRM2 only make use of the memory and CPU specs, though. A traditional arrangement approach was presented and put into practise by Nawrocki, 'Sniezy'nski, and Sojewski in 2019. The multi-level breakthrough in Internet application design enabled the general-purpose MCC condition.

By using Mobile Edge Computing (MEC), it will be possible to get around the low battery life and manageability of mobile devices as well as the significant latency of offloading apps to the cloud. It moves cloud processing capabilities closer to the mobile client at the cell edge (Xiao, Jia, Liu, Cheng, Yu, & Lv, 2019). Many observers claim that one of the current key issues is mobile subscriptions at the edge. It covers the distribution of resources and the offloading of computations. To suit client needs, asset allocation entails managing and storing assets. It depends on the property's accessibility and constraints. The service provider will distribute all assets to each customer in accordance with the length of each work.

Offloading of computation takes place at the external stage (edge server or cloud), and it is based on the device's processing power and capacity constraints. Due to the diverse business needs of these clients and their mobility, it's challenging to make provide for an appropriate solution to the executives a single setting. Artificial intelligence (AI) solutions are therefore suggested to address this optimisation challenge.

In order to alleviate issues and boost performance, resource management in mobile edge computing (MEC) (Zamzam, Tallal, &Mohamed, 2019). The heterogeneity, uncertainty, and cloud environment make it impossible to manage resource allocation using current policies. To safeguard data and enhance performance, Al-Janabi and Shehab (2019) explained secure edge computing in IoT. Nguyen et al., (2018) aimed to identify cyber attacks. They were able to show that the accuracy of their suggested model increased attack detection accuracy by up to 97.11%. The security concerns demand for optimal safety of information in the cloud, working to achieve improved result is encouraged; hence the research gap was identified. The researcher leveraged deep learning approach to detect cyber attacks in mobile cloud environment.

In a cloud computing setting, Idhammad M et al. (2018) investigated a novel technique to identify HTTP DDoS attacks. The information theory entropy (ITE) and RF learning algorithms constitute the foundation of the new detection technique that has been proposed. The incoming traffic signals' network header feature entropy was determined using a time-based sliding window technique. In a cloud infrastructure based on Open Stack, CIDDS-001 (Coburg Intrusion Detection Dataset) was used as an updated labelled stream-based dataset. To stop DDoS performance in virtual computers, Rawashdeh, Alkasassbeh, and Al-Hawawreh (2018) suggested an anomalous intrusion detection technique on the hypervisor layer. Evolutionary neural networks were used to build the detecting technique. The Scalable Neural Network has been integrated with Particle Swarm Optimization with the neural network for the DDoS attack detection and traffic data classification.

ANN is used in References (Saljoughi, Mehrdad, and Hamid, 2017) to identify intrusions and attacks. The NSL-KDD and KDD-CUP datasets were explored to test the model. The writers asserted that their suggested model discovered intrusions and attacks by unauthorised users. Using, Elzamly, Hussin, and Basari (2016) projected fundamental issues with security in distributed computing. The Delphi cloud procedure is used for informal social events and surveys. Information from reliable sources is gathered using the Delphi method. To forecast issues with distributed computing, the ANN algorithm has been employed as a quantifiable information model. To foresee cloud security incidents, the LMBP algorithm is employed. It has been discovered that LMBP algorithms are incredibly effective for test and preparation systems in terms of cloud security with banking organisations. Due to the diverse and scattered nature of cloud frameworks, advanced assaults are challenging to detect.

The importance of cyber attacks in cloud environments for preventing unauthorised access to cloud arrangements has been addressed by Sayantan, Stephen, and Arun-Balaji (2016). A technique has been put out to find digital intrusions on cloud platforms and distant processing equipment. The proposed technique involves the application of ANNs. ANNs were created utilising data on cloud stop trunk connections' system traffic. Because the ANN is calculated in great detail, a way to cut down on the number of structures collected from the system traffic data was developed and incorporated to this approach. This method is illustrated by ways for two sizable databases of system traffic data, showing superior outcomes compared to some methods of spotting digital attacks in cloud configurations.

The simplest ML method for recurrent and classification issues is probably K-NN. The K-NN algorithm uses fresh data and defines it using similarity metrics (like distance, for example). An overwhelming majority vote to its neighbours completes the ranking. Methods of information security cannot be used, these measures to be used legally, it is essential to comprehend the security requirements. A data categorization method founded on data secrecy was proposed by Zardari, Jung, and Zakaria (2014) in virtual and cloud environments, the K-NN information sorting method was used.

System and framework managers implement interrupts using the invention known as IDPS. The approved manager may get an email alert once the IDPS validates the interruption. Defences are improved by machine learning IDPS. A genetic algorithm ANN and ML are two unique ML techniques for enhancing network security. Selective new examples that the framework is unaware of are resolved by GA using prior examples. Cybercriminals are gradually learning how new devices that utilise AI to find flaws are designed. This enables online thieves to hide their goals when they test systems and distribute malware (Shamshirband & Chronopoulos, 2019).

In order to address the issue, Hussien and Sulaiman (2017) presented web prefetching strategies employing ML algorithms in mobile computing. One of the inventions utilised to lower the invisibility of traffic management on the Internet is prefetching. In order to address issues with inactivity in the context of data management, this study suggests using MCC conditions as an innovation. Because to prefetching, improper item information storage, and mobile phone capacity restrictions, excessive pre-fetching causes overhead and hinders frame execution. In (Arjunan and Modi, 2017), an updated security framework for CC's intrusion detection system is proposed. The writers combined a signature and anomaly-based technique to detect intrusions. They improved the effectiveness of the suggested strategy using Naive Bayes and other algorithms.

To lessen the threat posed by DDoS, Zekri, El Kafhali, Aboutabit, and Saadi (2017) developed a Distributed Denial of Service (DDoS) detection system based on the C4.5 algorithm. Hidden inventions and ingrained practises comprise flaws and security gaps that could be exploited by attackers. Any movement can be successfully grouped using ML methods based on predetermined layers. The computational intelligence network provides access to ML techniques.

Hanna, Bader, Ibrahim, and Adel (2016) employed Naive Bayes, a multilayer perceptron, SVM, and a decision tree (C4.5) to combine data from the access list of ML techniques. When it comes to tackling security concerns and attacks, these algorithms are highly helpful. When it comes to tackling security concerns and attacks, these algorithms are highly helpful. As a first step towards creating a safe and secure environment, Hanna and Associates (2016) reviewed and examined how to reduce the security concerns of cloud computing. Hou et al.'s (2019) made explanation of how to use machine learning to identify network security on edge computing devices to address the issue. An intelligent home framework created by Alibaba's ECS was artificially simulated as part of their analysis. The most cutting-edge IT innovation was considered during the creation of Device Architect. This study extends the usage of ML while recognising the need of system security in IoT frameworks.

Wani, Rana, Saxena, and Pandey (2019) used Tor Hammer as an attack tool to study the cloud environment and subsequently produced a dataset to identify intrusions. For classification, they tried a variety of Machine Learning algorithms; Support Vector Machine SVM, Naive Bayes, and Random Forest, and they demonstrated that SVM has the greatest accuracy, for instance 99.7%. The job involved dividing the dataset into positive and negative categories using a dataset vector. The outcomes show that the SVM RBF working strategy performs best in this task.

In their 2017 paper, Grusho, Zabezhailo, Zatsarinnyi, and Piskovskii addressed AI techniques and models for addressing information security issues. The ability to remotely access resource computers, active changes in the existing context of the virtual machines involved, vulnerabilities of idle VMs, insecure interfaces (unauthorised intrusion or hacking), misuse and deceitful use of cloud services, architectural restrictions on access to cloud infrastructure, and the dynamic nature of the cloud environment are essential for safety issues to the cloud computing environment. It is obvious that an IDS for the cloud needs to analyse massive amounts of network traffic data, identify new attack tactics effectively, and achieve high accuracy with few false positives.

However, utilising conventional techniques to pre-process, analyse, and identify intrusions in Cloud systems has become exceedingly expensive in terms of compute, time, and budget (Idhammad, Afdel, & Belouch, 2018). The proposed method calls for the employment of "instrumental" software systems that track incidents, compile comprehensive incident data, attempt to prevent incidents, and create incident reports for IS administrators as models for IDPSs. Identification of security policy issues, documentation of current threats, and prevention of policy violations by information exchange participants are all accomplished via IDPSs. The type of IDPS is determined by the particular events that should be tracked and the methods ("channels") via which these events should be implemented.

As a crucial step towards achieving a secure state for distributed computing, Hanna et al. (2016) described the accomplish moderation for security threat. The findings demonstrated that a simple decision tree model Chi-square Automatic Integration Detector (CHAID) algorithm security rating for ordering method is an effective strategy that enables the executive to evaluate the degree of the cloud making sure they give the necessary support.

According to the findings, a primary decision tree model, CHAID algorithm security score is a dependable technique that enables the chief to gauge the degree of cloud assurance and the types of help being offered.

Traditional methodologies have become increasingly expensive in terms of compute, time, and money for pre-processing, analysing, and detecting intrusions in Cloud systems (Idhammad, et al., 2018). The proposed method calls for the deployment of "instrumental" software systems that track occurrences, gather comprehensive incident data, attempt to thwart all, and IS administrators' incident reports as models for detection systems is created. Intrusion Detection and Prevention Systems (IDPSs) are employed to pinpoint security policy issues, compile a list of current dangers, and prevent information exchange participants from breaking security rules. The type of IDPS is determined by the particular events they must monitor and the channel (means) by which they must be implemented. Decisions that manage the accuracy of computer network designs, wireless technology that examines the behaviour of computer networks and analysis of computer processes are examples of these types of systems.

The system's drawback is that it depends heavily on the infrastructure and development of IS cloud security tools and technologies in place of specialised solutions. In order to achieve a secure state of distributed computing, Hana and associates (2016) studied methods to limit distributed computing security concerns. The findings demonstrate that a powerful tactic that enables the leader to measure the cloud level in choosing the type of support offered is the use of the fundamental decision tree model for the chi-square automatic interaction detector CHAID algorithm security score for the imperative approach. Numerous risks that are associated with distributed computing can seriously harm the authority and data that are backed by this innovation. The findings demonstrate that the CHAID algorithm for clustering's fundamental decision tree model for security scoring is a potent system that enables the leader to quantify the cloud's coverage and the quality of help proffered.

Based on the secrecy of the data, Zardari et al. (2014) suggested a data categorization approach. The Data was categorised based on several security needs using K-NN. The suggested method divides data into two categories, such as sensitive and non-sensitive data. The authors then encrypted sensitive data using the RSA technique for security. The required level of protection for various types of data can easily be determined using this type of strategy. This research was designed to determine security need of the system rather than to identify threats and attack it.

METHODOLOGY

In order to identify assaults in a cloud computing environment, machine learning-based models are built using a number of essential components. One is the sets of data that have been gathered, made accessible, and assessed. The platform that the datasets are executed on is another. In this sense, Python's functionality makes it the ideal target for this paradigm. Due to irregularities, noise, partial data, and missing values, real-world data can become highly dirty and damaged. In machine learning, it is generally accepted that the more data we have, the better models we can train. To create a new attack categorization model, the datasets must undergo specified procedures and we must follow the phases, which include data cleaning, data integration, data transformation, and data reduction. What is referred to as data pre-processing is what the aforementioned procedures amount to. Many environments can be used for data pre-processing, however Anaconda 2.40, which comes with a variety of programming environments, was used to launch the jupyter environment. When describing data analysis, Jupyter is an integrated interactive notebook environment that allows for editing and running human-readable papers.

ENTIRE ARCHITECTURE OF THE MODEL

This part provides an overview of the steps used by the machine learning-based model to recognise attacks in cloud computing. The first module always captures the incoming network traffic data to the cloud. The data are then pre-processed using a filter-based approach, and then passed to filtering. In addition to giving access to regular communications, this enables identification and prevention of suspicious attacks. Figure 1 shows the architecture of the full model as well as the identification process.

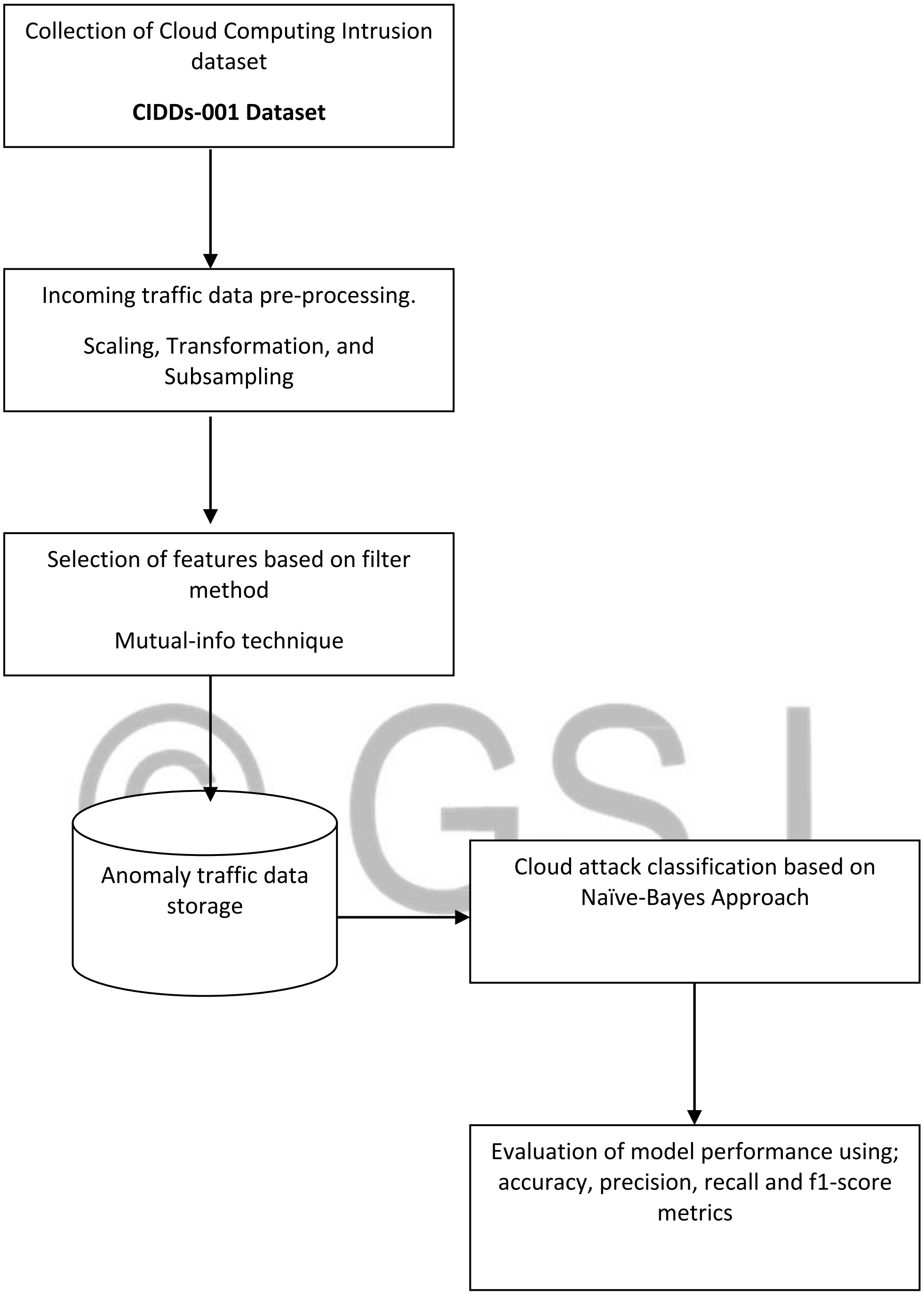
Collection of Cloud Computing Intrusion dataset
CIDDs-001 Dataset
Incoming traffic data pre-processing.
Scaling, Transformation, and Subsampling
Selection of features based on filter method
Mutual-info technique
Anomaly traffic data storage
Cloud attack classification based on Naïve-Bayes Approach
Evaluation of model performance using; accuracy, precision, recall and f1-score metrics

Figure 1: Architecture of the Naïve-Bayes Based model.

The figure 1 is the architecture that describes the various processes in the machine learning based intrusion detection system in cloud computing environment. The classification module was built in such a way that the attacks in the chosen cloud intrusion dataset can be efficiently classified. The final results of the classification/identification were evaluated based on accuracy metrics.

METRICS TO EVALUATE NAIVE BAYES ALGORITHM

The calculation for accuracy, precision, f1score and Recall were computed and there formulae are as follows:

$$ACCURACY = \frac{(TP+TN)}{P+N} \quad -- \quad 1.0$$

$$PRECISION = \frac{TP}{(TP+FP)} \quad -- \quad 1.1$$

$$RECALL = \frac{TP}{(\ TP+FN\ )} \quad -- \quad 1.2$$

$$F1-SCORE = \frac{2TP}{(\ 2TP+FP+FN)} \quad -- \quad 1.3$$

Where (TP) is number of True Positives, (TN) is True Negatives,(FP) is False Positives, and (FN) is False Negatives, P is number of Positives and N is number of Negatives.

RESULTS AND DISCUSSION

Implementation of an intrusion detective system requires deep understanding of machine learning and data processing. The dataset has 159373 rows × 16 columns and there was no any duplicate in the data set. As shown in the confusion matrix, the result showed high predictions by the model which implies that there's an improvement in this model compared to model proposed by Nguyen et al., 2018 as captured in the statement of problem. The classification report of this model further revealed that Attacker and Suspicious which represent 0 and 2 respectively have the highest precision and F1_score values with weighted avg of approximately 97.38% for the accuracy as against the 97.11% in the model proposed by (Nguyen et al., 2018)

Table 1 displays the outcomes.

Case/Metrics	Accuracy	Precision	Recall	F1_score	Data size(row by column)
Case_1	0.95	0.95	0.95	0.95	8098 by16
Case_2	0.95	0.95	0.95	0.94	159373 by 16
Case_3	0.97	0.97	0.97	0.97	153026 by 16
Case_4	0.97	0.97	0.97	0.97	186004 by 16

FINDINGS

Four metrics were utilised in the evaluation of the machine learning-based model: F1-score, accuracy, precision, and recall. The overall statistic, accuracy, provided a good indicator of how well the naïve-bayes performed. F1-score was derived using both recall and Precision. Recall determined the proportion of the total positive label instances that were properly tagged, while precision determined the proportion of positive predictions that were negative. The model for the four captures in case 1 to 4, the metrics show that the model performed above 95%. With mutual-info, promising features were identified in descending order.

Case 1 results produced 95% for accuracy, precision, recall and f1-score

Case 2 results produced 95% for accuracy, precision, recall while f1-score was 94%.

Case 3results produced 97% for accuracy, precision, recall and f1-score and finally,

Case 4 results produced 97% for accuracy, precision, recall and f1-score

Furthermore, accuracy is quite trustworthy because it is not biased towards the majority class while working with some unbalanced datasets, like CIDDS-001. Aside from that, the average setting has no impact on the accuracy result. It was observed that working on large dataset as CIDDS-001 requires a computer with a sizable quantity of memory and disc space to do the necessary computations.

SUMMARY, CONCLUSION AND RECOMMENDATION

Summary

As the model was being built through experimentation, several of the futures could not be trained to produce ideal outcomes due to some degree of imbalance in them; only attackTpye and class produced optimal results. Large data processing was difficult for systems with little memory. In order to partition the dataset, huge samples had to be chosen at random; this ensured that the outcome would not be impacted by the unbiased nature of the random technique employed. On a Jupyter computer with an Intel (inside) core i3 processor and 4 GB of RAM, the simulation is run. Seventy-five percent of each pre-processed dataset file is utilised to create a model or for training, while the remaining twenty-five percent is used for testing.

Conclusion

Due to raising demand for cloud computing, the vulnerability of the users also increases as they are prone to attacks by attackers who deploy sophisticated hacking systems. This therefore calls for developing an efficient machine-learning based model to identify and prevent attacks in a system. To successfully identify attacks, the model's final settings were fixed to size = 0.25 and random = 42. The multiclass algorithm naive-bayes requires technical expertise to handle. The average value therefore had an impact on configuring the metric parameters. The typical 'micro' was picked because of the characteristics of the data. The average setting was found to have an impact on the metric's outcome. Thus, it was deduced that the average setting depend on the kind of procedure used to develop a model. According to the evaluation's conclusions, Mutual-info fared better on CIDDS-001 in terms of metrics. As a result, it was possible to evaluate and calculate the average setting using the dataset. The evaluation's findings lead to the conclusion that, in terms of feature scores, ‘packets’ in the CIDDS-001’s features is the most promising. The model can therefore be used to identify intrusions in cloud computing systems and even more effectively against the accuracy 97.11% result gotten by Nguyen, Hoang, Niyato, Wang, Nguyen, and Dutkiewicz (2018), this model was able to achieve an accuracy of 97.14% and 97.38% in cases 3 and 4 of the captures.

Recommendation

Despite its complexity, the Naïve-Bayes algorithm for classification is the simplest and most user-friendly algorithm. And for the classification, it has generated some optimal performance indicators. Therefore, the model is advised for intrusion attacks, particularly those that share characteristics with the bruteforce and portscan hacking algorithms. Due to the enormous data computations, it is also advised to employ a supercomputer with unique specs while creating a model.

REFERENCES

[1] Arora, S., &Bal, J. S. (2021). *An Analysis Study on Architecture of Cloud Service Models In Cloud Computing. Journal Of Critical Reviews ISSN- 2394-5125 VOL 08, ISSUE 03, 2021*

[2] Al-Janabi S., and Shehab A. (2019). Edge Computing: Review and Future Directions. REVISTA, 368–380.

[3] Arjunan K., and Modi C. (2017). An enhanced intrusion detection framework for securing network layer of cloud computing. In Proceeding of the ISEA Asia Security and Privacy (ISEASP), Surat, India; pp. 1–10.

[4] Borylo, P., Tornatore, M., Jaglarz, P., Shahriar, N., Cholda, P., and Boutaba, R. (2020). Latency and energy-aware provisioning of network slices in cloud networks. Comput. Commun; 1–19.

[5] Bouyeddou, B., Kadri, B., Harrou, F., & Sun, Y. (2020). *DDOS-attacks detection using an efficient measurement-based statistical mechanism. Engineering Science and Technology, an International Journal, 23(4), 870–878. https://doi.org/10.1016/j.jestch.2020.05.002*

[6] Butt, U. A., Mehmood, M., Shah, S. B. H., Amin, R., Shaukat, M. W., Raza, S. M., &Piran, M. J. (2020). A review of machine learning algorithms for cloud computing security. Electronics, 9(9), 1379.

[7] Dang, L.M., Piran, M., Han, D., Min, K., and Moon, H. (2019). *A Survey on Internet of Things and Cloud Computing for Healthcare.; 8, 768.Deepali, Bhushan K. DDoS Attack Defense Framework for Cloud using Fog Computing. 2nd IEEE International Conference on Recent trends in Electronics, Information & Communication Technology (RTEICT). India. 2017: 534-538.*

[8] El-Boghdadi H., and Rabie A., (2019). Resource Scheduling for Offline Cloud Computing Using Deep Reinforcement Learning. Int. J. Comput. Sci. Netw., 342–356.

[9] Elzamly A., Hussin B., and Basari A.S., (2016). *Classification of Critical Cloud Computing Security Issues for Banking Organizations: A Cloud Delphi Study. Int. J. Grid Distrib. Comput., 137–158.*

[10] Grusho A., Zabezhailo M., Zatsarinnyi A., and Piskovskii V. (2017). On some artificial intelligence methods and technologies for cloud computing protection. Autom. Doc. Math. Linguist.; p62–74.

[11] Hanna M.S., Bader A.A., Ibrahim E.E., and Adel A.A. (2016). *Application of Intelligent Data Mining Approach in Securing the Cloud Computing. Int. J. Adv. Comput. Sci. Appl.; 151–159.*

[12] Hou S., and Xin H. (2019). *Use of machine learning in detecting network security of edge computing system. In Proceedings of the 4th International Conference on Big Data Analytics (ICBDA), Suzhou, China, pp. 252–256.*

[13] Hussien N., and Sulaiman S. (2017). *Web pre-fetching schemes using Machine Learning for Mobile Cloud Computing. Int. J. Adv. Soft Comput. Appl.; 154–187.*

[14] Idhammad, M., Afdel, K., &Belouch, M. (2018). Distributed intrusion detection system for cloud environments based on data mining techniques. Procedia Computer Science, 127, 35-41.

[15] Idhammad, M., Afdel, K., and Belouch, M.(2018). Detection system of HTTP DDoS attacks in a cloud environment based on information theoretic entropy and random forest, Security and Communication Networks, 2018

[16] Kishore BabuDasari, and NagarajuDevarakonda (2022). *Detection of DDoS Attacks Using Machine Learning Classification Algorithms. International Journal of Computer Network and Information Security(IJCNIS), Vol.14, No.6, pp.89-97, 2022. DOI:10.5815/ijcnis.2022.06.07*

[17] Kushwah, G.S., and Ranga, V. (2020). *Voting extreme learning machine based distributed denial of service attack detection in cloud computing. Journal of Information Security and Applications, 2020, 53, pp. 102532*

[18] Majed, H., Noura, H. N., Salman, O., Malli, M., &Chehab, A. (2020). *Efficient and secure statistical DDoS detection scheme. ICETE 2020 - Proceedings of the 17th International Joint Conference on e-Business and Telecommunications, 153–161. https://doi.org/10.5220/0009873801530161*

[19] Mamatha, K., &Kandewar, P.(2019). *Understanding Vulnerabilities and Data Security in Cloud Computing. Journal Of Architecture & Technology, IssnNo : 1006-7930*

[20] Nawrocki P., ´Snie ˙zy´ nski B., and Słojewski H. (2019). Adaptable mobile cloud computing environment with code transfer based on machine learning. Pervasive Mob. Comput., 49–63.

[21] Nguyen N., Hoang D., Niyato D., Wang P., Nguyen D., and Dutkiewicz E., (2018). *Cyberattack detection in mobile cloud computing: A deep learning approach, In Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC), Barcelona, Spain, pp. 1–6.*

[22] Peng J. J., Zhang,X. J. Lei Z., Zhang,B. Zhang F. W., and Li Q (2009). *Comparison of Several Cloud Computing Platforms, Second International Symposium on Information Science and Engineering (ISISE '09). IEEE Computer Society, Washington, DC, USA, pp. 23-27, DOI=10.1109/ISISE.2009.94.*

[23] Pham Q.V., Fang F., Ha V.N., Piran M.J., Le M., Le L.B., Hwang W.J., and Ding Z. A (2020). *Survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art. IEEE Access, 116974–117017.*

[24] RadhikaP, D., and KanimozhiP, P, S. (2021). *Detection of DDoS Attack Using Machine Learning Algorithm in Cloud Computing. International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-7, Issue-8, August 2021. ISSN: 2395-3470 www.ijseas.com 355*

[25] Rawashdeh, A., Alkasassbeh, M., and Al-Hawawreh, M.(2018). *An anomaly-based approach for DDoS attack detection in cloud environment. International Journal of Computer Applications in Technology, 2018, 57, (4), pp. 312-324*

[26] Roempluk T. and Surinta O. (2019). *A Machine Learning Approach for Detecting Distributed Denial of Service Attacks. Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical,*

Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), 2019, pp. 146-149, doi: 10.1109/ECTINCON.2019.8692243.

[27] Said, H. M., El Emary, I., Alyoubi, B. A., &Alyoubi, A. A. (2016). *Application of intelligent data mining approach in securing the cloud computing. International Journal of Advanced Computer Science and Applications, 7(9).*

[28] Saljoughi A., Mehrdad M., and Hamid M., (2017). *Attacks and intrusion detection in cloud computing using neural networks and particle swarm optimization algorithms. Emerg. Sci. J., (pp. 179–191).*

[29] Santosh Kumar and R. H. Goudar (2012). *Cloud Computing – Research Issues, Challenges, Architecture, Platforms and Applications: A Survey. International Journal of Future Computer and Communication, Vol. 1(4), December 2012. DOI: 10.7763/IJFCC.2012.V1.95*

[30] Sarraf, S. (2020). *Analysis and detection of ddos attacks using machine learning techniques. Am. Sci. Res. J. Eng. Technol. Sci, 66(1), (pp. 95-104).*

[31] Sayantan G., Stephen Y., and Arun-Balaji B., (2016). *Attack Detection in Cloud Infrastructures Using Artificial Neural Network with Genetic Feature Selection. In Proceedings of the IEEE 14th International Conference on Dependable, Autonomic and Secure Computing, Athens, Greece, (pp. 414–419).*

[32] Shamshirband S., and Chronopoulos A.T. (2019). *A new malware detection system using a high performance-ELM method. In Proceedings of the 23rd International Database Applications & Engineering Symposium, Athens, Greece; (pp. 1–10).*

[33] Sharma, R. & Trivedi, R. K. (2014). *Literature review: Cloud Computing –Security Issues, Solution and Technologies. International Journal of Engineering Research, Vol. 3(4), (pp. 221-225).*

[34] Stefan, H.; Liakat, M. (2015). *Cloud Computing Security Threats and Solutions. J. Cloud Comput.; 4, 1.*

[35] Swamynathan, M. (2019). Mastering machine learning with python in six steps: A practical implementation guide to predictive data analytics using python. Apress.

[36] Wani A., Rana Q., Saxena U., and Pandey N. (2019). *Analysis and Detection of DDoS Attacks on Cloud Computing Environment using Machine Learning Techniques. In Proceedings of the Amity International Conference on Artificial Intelligence (AICAI), Dubai, UAE; (pp. 870–875).*

[37] Xiao Y., Jia Y., Liu C., Cheng X., Yu J., and Lv W. (2019). Edge Computing Security: State of the Art and Challenges. Proc. IEEE 2019, 107, 1608–1631.

[38] Zamzam M., Tallal E., and Mohamed A. (2019). *Resource Management using Machine Learning in Mobile Edge Computing: A Survey. In Proceedings of the Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt; (pp. 112–117).*

[39] Zardari M.A., Jung L.T., and Zakaria N. (2014). *K-NN classifier for data confidentiality in cloud computing. In Proceedings of the International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, Malaysia; (pp. 1–6).*

[40] Zekri M., El Kafhali S., Aboutabit N., and Saadi Y. (2017). *DDoS attack detection using machine learning techniques in cloud computing environments. In Proceedings of the International Conference of Cloud Computing Technologies and Applications (CloudTech), Rabat, Morocco,(pp. 1–7).*