National College of Ireland

Assessing Machine Learning Algorithms for Cloud Computing Threat Detection

MSc Research Project
MSc in Cloud Computing

Zaid Ali Khan
Student ID: x22204016

School of Computing
National College of Ireland

Supervisor: Shreyas Setlur Arun

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Zaid Ali Khan
Student ID:	22204016
Programme:	MSc in Cloud Computing **Year:** 2023 -2024
Module:	Research in Computing
Supervisor:	Shreyas Setlur Arun
Submission Due Date:	16, September 2024
Project Title:	Assessing Machine Learning Algorithms for Cloud Computing Threat Detection
Word Count:	6417 **Page Count** 22

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: Zaid Ali Khan

Date: 16, September 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):	

Assessing Machine Learning Algorithms for Cloud Computing Threat Detection

Zaid Ali Khan

x22204016

Abstract

Cloud computing is being adopted by a wide range of sectors, including regulatory agencies, large organizations, and small businesses. Cloud technologies' improved capabilities and efficiencies are what are responsible for their widespread acceptance. The use of machine learning (ML) to strengthen security measures is a crucial component of this expansion. In order to identify sophisticated attacks, zero-day vulnerabilities, and insider threats that frequently elude traditional security systems, ML-based security solutions use advanced analytics and predictive modelling techniques. Large data sets are analysed by these solutions to find trends and abnormalities that could point to security flaws. This research critically evaluates the application of machine learning algorithms in assessing the risks associated with cloud computing. The aim of this study is to comprehend the working of the algorithms which could boost the detection of new threats. Thereby it provides a stronger security structure for cloud based systems through the use of machine learning. The outcomes of this study are expected to assist in developing of cloud security making them more effective.

Keywords— Cloud Computing, Machine Learning, Threat Detection, Security Assessment, Risk Mitigation

1 Introduction

1.1 Background of the Study

Cloud computing has transformed the way businesses handle and deliver services bringing flexibility, scalability and cost effectiveness. It enables users to utilize resources such as storage, processing power and software as needed without requiring an initial investment in physical infrastructure. Cloud services, including those tailored for machine learning applications grant access to testing datasets. These services are provided through servers referred to as "clouds " and are classified into three models Infrastructure as a Service (IaaS) Platform as a Service (PaaS) and Software, as a Service (SaaS). Each model offers varying levels of control, adaptability and oversight for users to select according to their requirements (Aljawarneh et al., 2018)

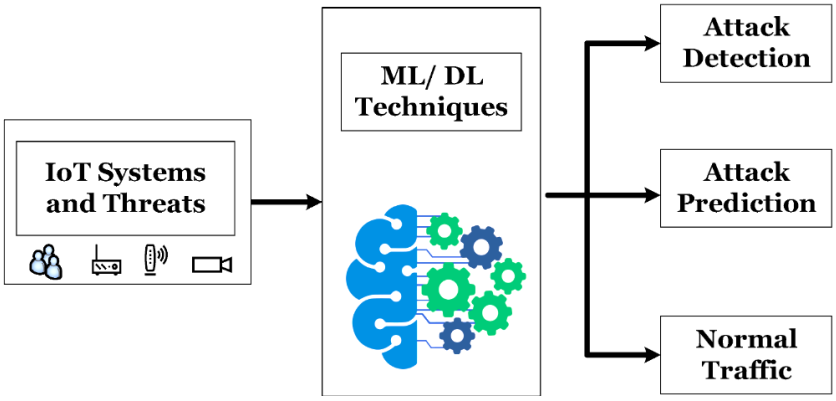

Figure 1- Image showing importance of threat detection (mdpi, 2024)

Although there are advantages security worries continue to be an issue for businesses that utilize cloud computing. The decentralized structure of cloud systems and the joint accountability between users and cloud service providers (CSPs) play a role in these security obstacles. Safeguarding confidential data stored in the cloud from entry, breaches and other risks is vital, for companies.

1.2 Motivation of the Study

Businesses are quickly embracing cloud computing for its adaptability, sustainability and cost effectiveness. However, the rapid expansion of cloud technology has raised concerns about security as cyber threats increasingly target cloud environments. At the time artificial intelligence, especially machine learning is becoming more popular in various industries. Among the sought after uses is recommendation systems that utilize machine learning to provide users with personalized content, products or services based on their preferences and behaviours. These systems gather user data to offer recommendations that improve user satisfaction by delivering content tailored to their specific requirements. This personalized approach exemplifies one of the numerous potential applications of machine learning, in web technology.

1.3 Overview of Threats to Cloud Computing

Threat	Description
Unauthorized Access	Unauthorized users may attempt to access sensitive data or resources hosted in the cloud.
Data Breaches	Leaks of data privacy can occur due to vulnerabilities in cloud infrastructure, misconfigurations, or inadequate access controls.
Denial of Service (DoS) Attacks	Attackers may launch DoS attacks to disrupt the availability of cloud services by overwhelming servers or network resources.
Malware and Ransomware Attacks	Malicious software such as malware and ransomware pose significant threats to cloud environments, potentially leading to data loss or corruption.

Insider Threats	Employees or insiders with malicious intent may exploit their access to compromise security.
Data Loss	Accidental deletion, hardware failures, or service shutdowns can result in data loss, leading to operational disruptions and financial loss.

Table: 1 gives an outline of these threats

1.4 Role of Machine Learning in Threat Detection

Machine learning is effective threat detection in the cloud system which includes different testing dataset. Along with this, it includes different software of the cloud computing that analyse the data of software system. Inside the setting of cloud security, machine learning calculations and strategies can be associated to abnormality area, interference area, malware examination, and prescient chance examination (Kim et al. 2019). ML calculations can computerize the peril range handle, reducing the dependence on manual mediations and working on the capability of security works out (Choo *et al*., 2018).

1.5 Problem Statement

The threat categorizations currently employed often lack detail and do not adequately address the evolving nature of cyber threats aimed at cloud systems. Without a standardized approach to categorizing threats, it's difficult for organizations to accurately assess their security risks, prioritize mitigation efforts, and share information with partners (Vyas *et al*., 2023). Due, to this it is crucial to establish an organized method for handling risks specifically tailored to cloud computing taking into account the features of cloud infrastructure and services (Upper class, 2009). Consequently, there is an increasing need, for flexible threat detection systems that can actively recognize and address emerging threats promptly.

1.6 Research Aims and Objectives

This study seeks to create a structure, for detecting risks in cloud computing setups with a specific emphasis, on utilizing machine learning techniques for detection purposes.

The research objectives are as follows:

- To examine existing threat classification schemes in cloud computing and identify their limitations.
- To evaluate the effectiveness of ML algorithms for detecting various types of threats in cloud environments.
- To develop a structured framework for detecting cloud computing threats based on their detectability by ML models.

- To empirically evaluate the proposed threat classification framework using real-world cloud datasets.
- To propose recommendations for enhancing cloud security through the integration of ML-based threat detection systems.

1.7 Research Question

The research problem motivates the following research question:

Can machine learning algorithms be effectively used in cloud computing environments to enhance threat detection?

2 Literature Review

2.1 Introduction

Cloud computing is referred to as the on-demand delivery of computing services like databases, servers, analytics, networking and others. Cloud-based storage makes it possible to save files to a remote source.

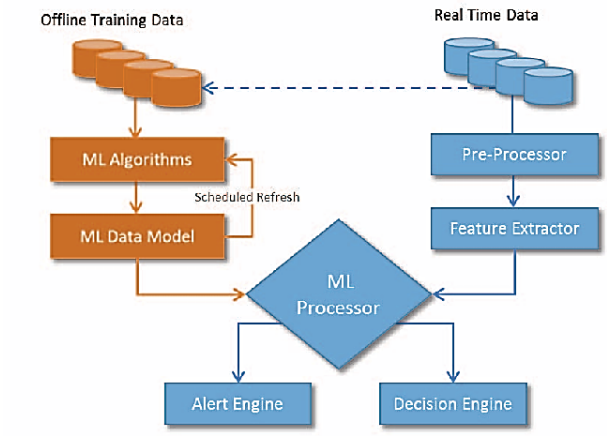

Figure 2- Role of ML in threat detction (FarooqNaif & M. Otaibi, 2018)

2.2 Concept of Cloud Computing

Cloud computing is such a technology that helps users utilised and assess different computing resources via the internet. Using cloud computing has different benefits such as flexibility, cost-efficiency, collaboration, scalability, backups and others. As per the opinion of Aljawarneh *et al.,* (2018), cloud computing is referred to as the emerging technology that migrates computing concepts along with current technology into utility-like solutions like water systems and electricity.

2.3 Analysing security protocols in cloud computing and highlighting their limitations

The realm of cloud computing presents challenges when it comes to privacy and security prompting the need for specialized approaches to manage threats, such as employing data encryption to thwart unauthorized access. According to Choo et al. (2018) there is a growing concern surrounding safeguarding information in the cloud, which calls for the implementation of tactics like intrusion detection and prevention systems as well as security information and event management tools to monitor and analyse potential risks. Guha and colleagues (2016) propose utilizing networks (ANNs) along, with network traffic information to enhance characteristics and strengthen the ability to recognize threats. However, these methods encounter difficulties associated with complexity, performance implications, evolving risks, costs and adherence, to regulations that must be taken into account to guarantee the oversight of security.

2.4 Reliability of the application of machine learning algorithms in the recognition of types of risks in cloud environments.

The realm of cloud computing presents hurdles particularly concerning privacy and security prompting the development of tactics to combat risks. One effective method is employing encryption to safeguard data. Choo and colleagues (2018) underscore the increasing worry, about safeguarding data stored in the cloud emphasizing the significance of utilizing measures such as intrusion detection and prevention systems alongside security information and event management tools, for monitoring and analysing threats. Guha et al. (2016) recommend utilizing networks (ANNs) along with network traffic data to enhance threat detection capabilities by streamlining features. However, these techniques face obstacles related to complexity performance impact, evolving threats, expenses and compliance, with regulations that mandate oversight of security.

2.5 Assess an organized approach to identifying security risks in cloud computing considering how easily they can be detected by machine learning models.

A structured system integrates different key components along with processes for detecting threats of cloud computing with the help of machine learning. The structured approach assists in detecting along with mitigating threats effectively. Cloud computing is advancing at a massive pace along with expanding security perspectives (Krebs *et al.,* 2014). Thus, such a system needs to be developed for collecting data and pre-processing to identify threats. Furthermore, the structure system can also help in feature engineering by which improving the perceptibility of threats can be possible. Additionally, to address the uncertainty of the threats of cloud computing, vigorous safety efforts have been executed such as access controls, encryption, observing security and others (Prwez and Chatterjee, 2016). On the other hand, as per the opinion of (Mell and Grance (2011), the issues of threats concerning cloud computing have been effectively resolved by making proposals for different encryption methods and

introducing instruments of access control for defending data privacy and uprightness. Therefore, the course of action to address threats using methods is established.

2.6 Analysing the suggested framework for defining threats using cloud data

The updated system, for categorizing dangers enhances the evaluation and prevention process by utilizing information from cloud sources to recognize and respond to threats, in time (Gentry, 2009; Masetic et al., 2017). Its functionality relies on threat data and advanced machine learning techniques to anticipate and categorize risks. The systems efficiency is assessed based on its capability to identify risks adapt to cloud environments and counteract threats. The systems dependability is demonstrated through evidence gathered from cloud data while its capacity to handle volumes of data determines its scalability (Sampangi et al., 2019)

2.7 Recommendations for improving the security of cloud systems by incorporating machine learning driven technologies to identify threats.

By following these suggestions, we can improve the effectiveness of security systems that use machine learning for threat detection. In a research conducted by Sharma and fellow researchers in 2020 assessing the security posture through security audits can identify areas where cloud security can be improved by leveraging machine learning powered threat detection. Vyas and team in 2023 highlight the potential of this approach in mitigating cloud threats by improving machine learning based threat detection systems. It is crucial to enhance detection and response capabilities with machine learning as a lack of detection heightens security vulnerabilities. As mentioned by Yue et al. In 2016 increased detection capabilities play a role in thwarting security threats. These recommendations play a role, in fortifying cloud threat identification efforts.

2.8 Summary of Literature Review

Research	Strategy/ Approach	Objectives	Metrics	Dataset	Environment	Software	Real/Simulation
Aljawarneh, S., et al. (2018)	Survey and Review	To provide an overview of cloud computing security	Qualitative analysis	NSL-KDD dataset	Various cloud platforms	N/A	Real
Gentry, C. (2009)	Cryptographic Scheme	To develop a fully homomorphic encryption scheme	Encryption efficiency	N/A	Cloud computing platforms	Custom cryptographic software	Simulation

Mell, P. and Grance, T. (2011)	Definition and Standards	To define cloud computing and its standards	Qualitative analysis	N/A	Various cloud platforms	N/A	Real
Prwez, M. T. and Chatterjee, K. (2016)	Intrusion Detection Framework	To develop a framework for network intrusion detection	Accuracy, Precision, Recall	Custom simulated dataset	Cloud environments	Java	Simulation
Sampangi, R., et al. (2019)	Systematic Literature Review	To explore compliance in cloud security	Qualitative analysis	N/A	Cloud environments	N/A	Real
Sharma, A., et al. (2020)	Survey	To review machine learning techniques in security	Qualitative analysis	N/A	Various security environments	N/A	Real
Vyas, P., et al. (2023)	Machine Learning Approaches	To detect security issues in cloud web applications	Accuracy, Precision, Recall	Custom dataset	Cloud web applications	Python, R	Simulation
Yue, X., et al. (2016)	Blockchain-Based Framework	To enhance data integrity in cloud computing	Qualitative analysis	N/A	Cloud computing platforms	Custom blockchain software	Simulation

3 Research Methodology

This chapter describes the approaches and steps taken to reach the research goals. It explains how data was gathered features were chosen models were built and evaluations were conducted in the study. Additionally, it talks about the tools and technologies utilized to carry out the research and confirm the trustworthiness and accuracy of the results.

3.1 Research Design

The study takes a stance examining how machine learning techniques are used to identify and categorize risks, in cloud computing settings. The assessments of such techniques which are used in the research shall be evaluated with the help of quantitative methods included in the framework of the given methodology.

3.2 Data Collection

In this research the scientists employed the CICIDS2017 dataset for benchmark of intrusion detection systems. The dataset contains a combination of network activities together with various cyber-attacks, such as DDoS and Brute Force. It is derived from the Canadian Institute

for Cybersecurity. Models real world network scenarios by using captures and flow records (Butt et al. , 2020). For training machine learning models they created features like protocol types, IP addresses, port numbers and payload sizes. In the conduct of the study, the use of cloud computing resources was employed. Two servers. The first had Intel Xeon processors and 128 GB RAM while the second of them was equipped with 2 TB SSD storage. Intel Core i7 processing units were used in the workstations for data processing and for model construction. Python was used synchronized with Scikit, TensorFlow, and Keras to design and ascertain the machine learning models (Alshammari & Butt 2021).

3.2.1 Datasets

The dataset known as CICIDS2017 created by the Canadian Institute, for Cybersecurity includes network data and various types of attacks making it a useful tool for testing algorithms designed to detect security threats. Along with detailing cyber-attacks this dataset also captures network activities allowing for in depth analysis of detection techniques, in real world scenarios.

3.3 Data Pre-processing

Ensuring the accuracy and quality of data is essential, during the data processing stage, for machine learning models. This step involves addressing missing values, outliers and anomalies to optimize the dataset.

3.3.1 Dealing with absent and NaN values

Since they can greatly affect the model's performance it's crucial to address missing. Nan (Not a Number) values. The CICIDS2017 dataset may contain missing data points that require attention. To ensure the datasets integrity methods, like replacing missing values with the median or mode of the feature or utilizing advanced techniques such, as K nearest neighbors' imputation can be employed to fill in the gaps.

3.3.2 Outlier Detection

Outliers refer to data points that stand out significantly from the rest of the dataset potentially affecting analysis outcomes. To spot, potentially. Adjust these outliers, statistical or machine learning techniques are employed for outlier detection. In datasets, like CICIDS2017 outliers can be pinpointed using machine learning methods such, as z score or Interquartile Range (IQR). Ensuring handling of these values will not negatively impact the model's performance.

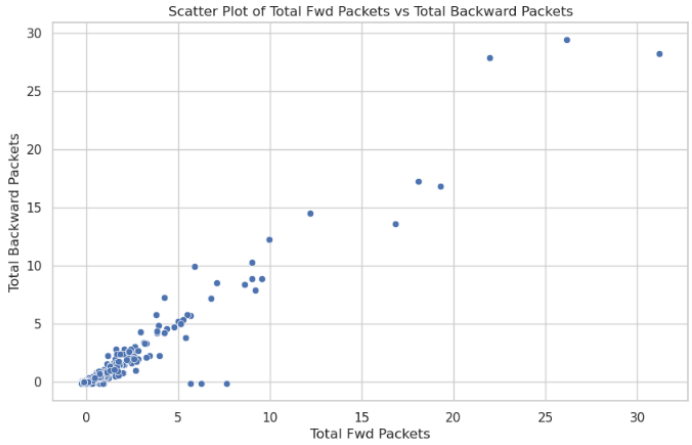

Figure 3- Scatter Plot showing anomalies

3.3.3 Anomaly detection

Anomaly detection is, about spotting patterns that deviate from the norm especially when it comes to detecting potential threats in network traffic. Unusual patterns could signal security risks or attacks. Various methods like grouping data using statistics and leveraging machine learning can be used to spot anomalies in a dataset. When looking at the CICIDS2017 dataset anomaly detection plays a role, in pinpointing new types of attacks that aren't clearly identified in the data thus making the threat detection system more resilient.

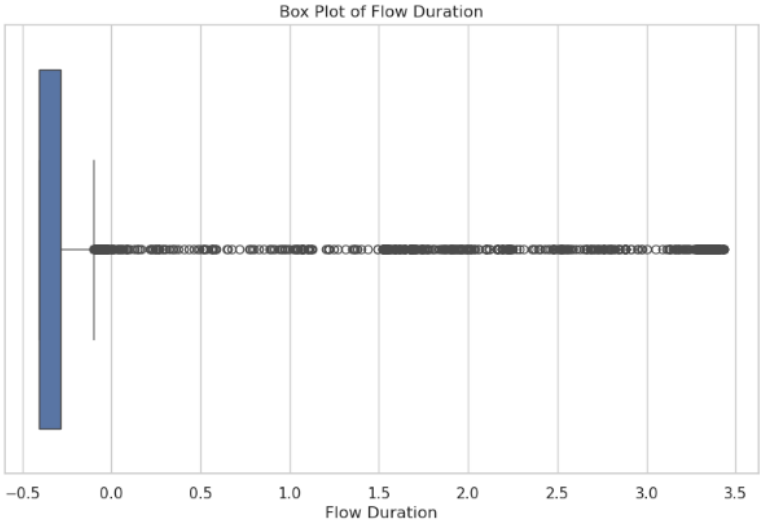

Figure 4- Box plot showing no anomalies

3.4 Statistical Analysis

Analyzing the CICIDS2017 dataset involves summarizing information examining how variables relate to each other and spotting patterns to enhance security threat detection algorithms. Important evaluation criteria include accuracy, precision, recall, F1 score and ROC AUC. The accuracy metric indicates the percentage of predictions, out of all predictions made. Precision on the hand measures the ratio of positive identifications to all positive outcomes.

Recall, also known as sensitivity assesses how effectively the model identifies positives among positives (Elmrabit et al., 2020).

	count	mean	std	min	25%	50%	75%	max
Destination Port	4997.0	7.358528e+03	1.689449e+04	0.0	53.0	80.0	1119.0	65389.0
Flow Duration	4997.0	1.232941e+07	3.032331e+07	-1.0	121.0	46741.0	4660472.0	119976180.0
Total Fwd Packets	4997.0	7.314589e+00	5.680222e+01	1.0	1.0	2.0	4.0	2622.0
Total Backward Packets	4997.0	7.950570e+00	7.844359e+01	0.0	1.0	2.0	4.0	3483.0
Total Length of Fwd Packets	4997.0	7.049372e+02	3.867559e+03	0.0	6.0	48.0	106.0	153547.0
...	...	...	...	...	...	...	...	...
Active Min	4997.0	7.611407e+04	5.414993e+05	0.0	0.0	0.0	0.0	13500000.0
Idle Mean	4997.0	5.279845e+06	1.598735e+07	0.0	0.0	0.0	0.0	119000000.0
Idle Std	4997.0	9.761580e+05	6.483410e+06	0.0	0.0	0.0	0.0	66500000.0
Idle Max	4997.0	6.015969e+06	1.809994e+07	0.0	0.0	0.0	0.0	119000000.0
Idle Min	4997.0	4.560688e+06	1.512448e+07	0.0	0.0	0.0	0.0	119000000.0

78 rows × 8 columns

Figure 5- Analyzing the characteristics of the CICIDS 2017 dataset

3.5 Feature Selection

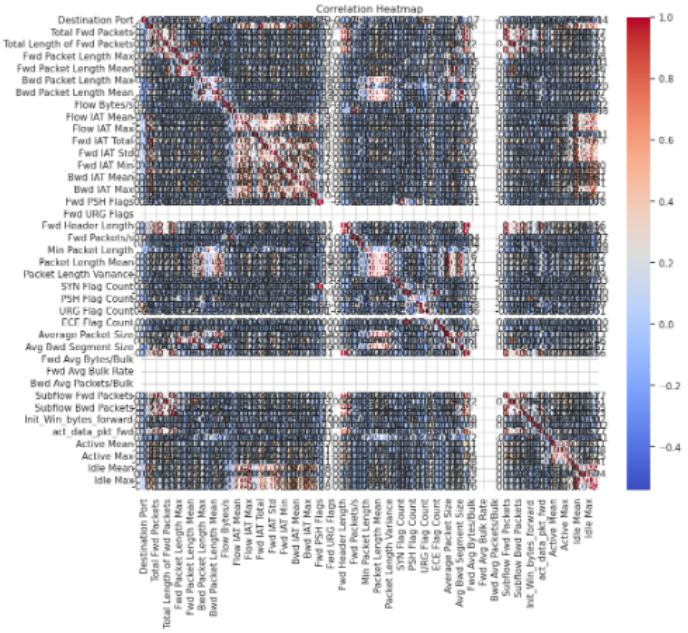

Figure 6- Observation of correlation metrics

Selecting features is a part of creating machine learning models since it has an effect, on how well the model performs. This research uses algorithms and other methods to pick out the important features, from the data sets.

3.5.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is utilized to decrease the complexity of the data by keeping the attributes and removing redundant ones.

3.6 Model Development

The research focuses on the implementation and evaluation of several machine learning algorithms for threat detection in cloud computing environments. The algorithms include Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Reinforcement Learning models.

3.6.1 Convolutional Neural Networks (CNN)

Convolutional neural networks are commonly employed for their capacity to autonomously recognize and acquire characteristics, from the data enabling them to pinpoint trends and irregularities, in network traffic.

3.6.2 Recurrent Neural Networks (RNN)

RNNs are commonly used with data. Are quite good, at capturing time related relationships, which makes them useful, for identifying threats that follow specific patterns over time.

3.6.3 Reinforcement Learning

Adaptive security systems that utilize Reinforcement Learning models can enhance their capabilities by engaging with the cloud environment. Effectively addressing security risks as they arise

3.7 Model Evaluation

Various measures are used to assess the effectiveness of machine learning models, such, as accuracy, precision, recall, F1 score and ROC AUC.

Metric	Description
Accuracy	Determines the ratio of recognized dangers, to the count of threats. calculated as: $$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$ Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.
Precision	Indicates the proportion of true positive identifications among all positive identifications made by the model. It is calculated as: $$Precision = \frac{TP}{TP + FP}$$
Recall	Measures the proportion of true positive identifications among all actual positive cases. It is calculated as:

	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
F1-Score	Provides a balance between precision and recall, offering a single metric to evaluate the model's performance. It is calculated as: $Accuracy = \frac{TP + TN}{TP + FN}$
ROC-AUC	Assesses how well the model can differentiate between negative instances offering an evaluation of its classification accuracy. The ROC AUC score shows the balance between identifying positives and mistakenly labeling positives through the Receiver Operating Characteristic (ROC) curves area.

3.8 Tools and Technologies

The study made use of cloud computing resources. It utilized two servers one equipped with Intel Xeon processors and 128 GB of RAM and the other with a 2 TB SSD specifically dedicated to managing datasets and intricate machine learning algorithms. Additionally, workstations powered by Intel Core i7 processors, ample RAM and high-speed internet connections were employed for tasks related to data preparation and model development. Python along, with libraries such as Scikit learn, TensorFlow and Keras played a role, in the development and evaluation of machine learning models (Alshammari et al., 2021).

3.9 Validation and Reliability

To ensure the reliability and precision of the findings the study employs validation techniques and rigorous testing procedures. The algorithms are assessed on data subsets to prevent overfitting and ensure relevance to real world situations.

3.9.1 Cross-Validation

Cross validation involves dividing the dataset into segments and training or testing the model, on combinations of these segments to ensure an assessment of its performance.

3.9.2 Rigorous Testing

The models are tested using data to evaluate their performance in situations and ensure their reliability in real world scenarios.

3.10 Ethical Considerations

The study adheres to established protocols for handling and examining data with a focus on safeguarding information through anonymization. All data used in the research is stripped of identifying details and measures are taken to protect the datasets and results.

4 Design Specification

It gives an overview of the system and further about the details of the spear section is called Design Specification which explains the framework, the stipulated components, and the strategies of the proposed threat detection system relative to cloud computing that employs the use of machine learning. It covers data collection, data pre-processing, feature construction and the choice of Machine Learning (ML) models for threat detection and mitigation (Gao et al. , 2020). However, it is for the sake of creating awareness, enhancement of capacity, automation, and visualization technique that the section begins the discussion on the features of the design alerting and also gives the advantage and disadvantages of the design. It offers understanding, into assessing the system's sustainability and flexibility and how machine learning improves testing and results.

4.1 Architectural Overview

A system designed to identify risks, in cloud computing through machine learning is organized into layers for management and mitigation. It includes the Data Collection Layer for gathering information from sources and the Data Pre-processing Layer which deals with cleaning and standardizing data. Moreover, the Feature Engineering Layer improves model performance by tuning data attributes. These interconnected layers in the machine learning structure make predictions, on categories based on these features.

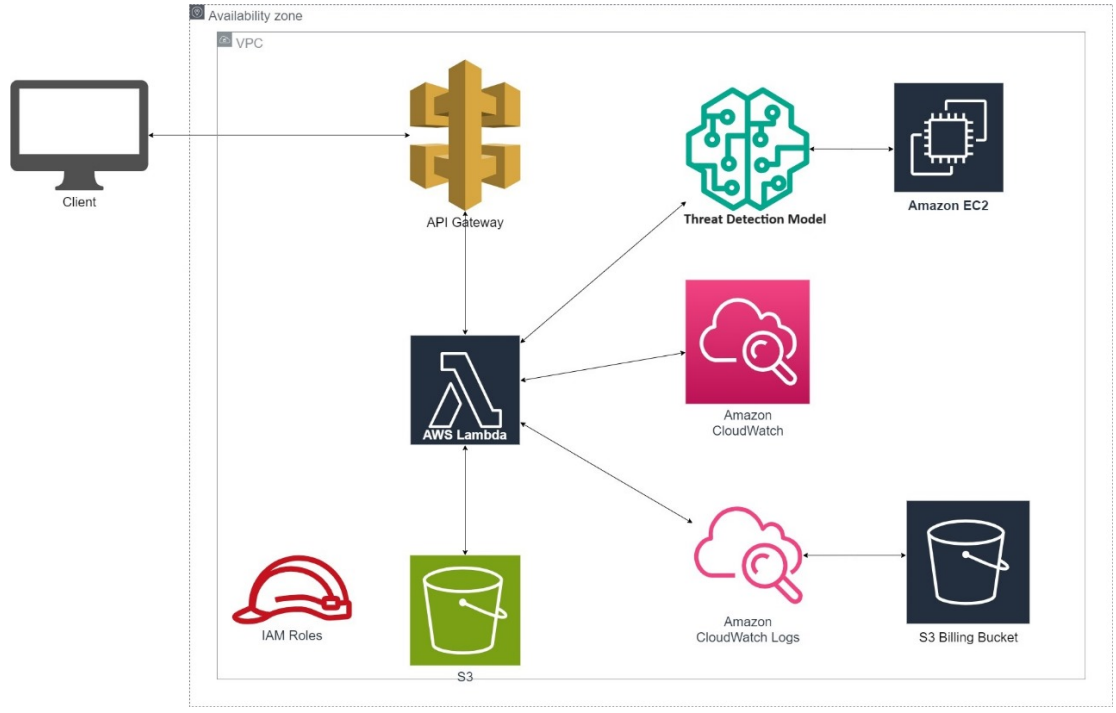

Figure 7- Architectural Diagram

4.2 Data Collection and Pre-processing

4.2.1 Data Sources

This framework collects a variety of information which is important for detailed threats assessment. Moreover, this includes application log files where every distinct occurrence associated with software applications is documented, user activity log files that capture actions of users and patterns of use and system log files that provides operational information compiled from devices and applications. Furthermore, the network traffic system of the database in the ML system may provide the potential method of test. However, IT specialists are pivotal for network traffic and system logs. APIs are very useful to get cloud management logs and App data. Along with this, these strategies ensure that the system has adequate and diversified data for the threats' analysis and detection.

4.2.2 Data Cleaning and Normalization

It should be performed before the analysis to ensure the quality of the data as well as their unity including such operations as deletion of the double records, the data on which were obtained after the conversion of ".csv" file to a panda's data frame, and the definition of the procedure for the missing values and mistakes detection. Also, random noise that may affect the performance of model is filtered out, ensuring that the results obtained are reliable and credible. Therefore, the data is cleaned methods for the normalized data used which are also used to bring the values to a standard form and are relatively easier to work with. Min-max scaling and z-score standardization are the two methods used in normalization. In addition, the data that contain categorized variables are transformed into numeric form which is ideal for ML algorithms by processes like one-hot encoding and label encoding.

4.2.3 Feature Extraction and Dimensionality Reduction

Model performance can be improved while computational complexity can be reduced by performing feature extraction and dimensionality reduction. Highlight extraction includes distinguishing and choosing key credits from the crude information that altogether affect the prescient force of the models. Statistical analysis, domain expertise, and automated techniques like RFE (Recursive Feature Elimination) are utilized. After that methods such, as t Distributed Stochastic Neighbour Embedding (t SNE) and Principal Component Analysis (PCA) are employed to decrease the number of characteristics while retaining data. Utilizing these techniques helps in organizing data to prevent overfitting and improve the model's flexibility, in situations. Emphasizing features enables the system to function effectively and precisely resulting in more accurate threat detection

4.3 Machine Learning Models

4.3.1 Supervised Learning

Due to the fact that in supervised learning algorithms labelled datasets are used and predictions made they have a part to play in improving threat detection systems. Some of the illustrations

of algorithms are Decision Trees which help build a model to make predictions about the value of the target variable concerning other input elements RNN, CNN, and Reinforcement Learning which can best sort tasks through establishing manners to recognize an appropriate hyperplane that would set different classes in the data. That is why all these algorithms can be considered valuable assets as they are interpretable, efficient and rather adaptable to both multiclass classification problems. Nonetheless, there are disadvantages associated with them, such as the requirement of labelled data and issues with capturing relationships in the data. However, owing to these challenges supervised learning stands out as the bedrock of the established system providing reliable threat identification.

4.3.2 Unsupervised Learning

There are two categories of risk identification the first one is unsupervised learning where one does not have to categorize data into risks and no-risk outcomes. Some of the methods include: The K Means which is a method used to cluster data by the similarity the DBSCAN (Density Based Spatial Clustering of Applications, with noise) which is able to cluster the data of any form of shape or size and is able to handle with noises successfully and the Autoencoders neural networks which are used in anomaly detection through learning efficient representations of data. Such algorithms are appreciated because of their capabilities to identify concealed patterns in the data and threats. Nonetheless, it depends such on the parameterization of the methods they employ that their effectiveness can be impaired. The results they produce are more complex than those of learning methods. Still if applied, unsupervised learning offers dramatic improvement on the ability of the system to detect new threats that were undetectable before.

4.3.3 Reinforcement Learning

The techniques that fall under reinforcement learning, for instance, ‘Q Learning’ and ‘Deep Q Networks (DQN)’, are applied in coming up with reaction patterns regarding threats. Reinforcement learning is the process through which an agent is capable of making decisions through influences from the environment and the outcomes that decompose either as a reward or a penalty on the behaviours done by the agent. As the action of an agent can evaluated in terms of value, the approach using the model free Q Learning technique can be determined in situations. There is an improvement on the base Q-learning algorithm to develop a variant of deep Q-learning employing networks, which can efficiently manage complex state space rendering reinforcement learning flexible to different environments. Another benefit that comes with reinforcement learning is its ability to be updated in experience and thus flexibility in dealing with threat profiles at a given period as compared to another. However, it imperative to point out that reinforcement learning can often prove quite resource-intensive as well as long in training. By using reinforcement learning the efficiency of the system can be enhanced by creating ways on how to deal with such threats.

4.4 Threat Detection and Response

This has a role in monitoring of security threat since it is the orchestration of machine learning models. There are several techniques used in threat detection for instance anomaly detection which scans for behaviours that indicate threats, signature-based techniques whereby detection is based on certain known threat signatures, and predictive techniques used to estimate threats based on data collected. Once identified the system can in turn sort out threats in terms of the level of threat posed as well as the type of threat. Other response tactics include automated operations as for example isolating affected systems to block traffic and notify security

personnel. In this way, the system retains a high degree of relevance and means to minimize threats in the immediate manner to ensure the survival of the system.

4.5 Alerting and Automation

The framework's integration of ML models for nonstop hazard appraisal may be a key component. Risk is distinguished utilizing three diverse methods. Incongruity area, which is finds behavioural peculiarities signature-based revelation, which depends on actualized risk plans and prescient examination, however, employments real-world information to foresee potential dangers. Along with this, during dangers distinguished are the system can concentrate on them and react in agreement with their sort and earnestness. Moreover, reaction strategies incorporate electronic drills such as separating affected frameworks, discouraging retaliatory activity, and cautioning the protect group. Through nonstop information investigation and danger demonstrate upgrades the framework ensures quick danger location and moderation whereas maintaining a tall degree of exactness and responsiveness. This energetic approach diminishes the probability that dangers will affect the cloud environment whereas at the same time moving forward security.

4.6 Visualization and Reporting

The objective of the representation and specifying layer is to supply brief, imperative data almost the state of cloud security. This layer's user-friendly dashboards and analytics instruments empower real-time checking, which makes a difference security groups screen danger, analyse patterns, and survey the adequacy of reaction measures. Custom announcing alternatives empower the creation of comprehensive reports that are suited to the prerequisites and compliance prerequisites of a particular organization, encouraging the method of making well-informed choices. Vital point’s incorporate chronicled examination, execution measurements, and graphical representations of danger information. Real-time checking makes a difference security groups spot and address dangers rapidly, and exhaustive announcing makes a difference with administrative compliance and key arranging. This layer gives security bunches the data they have to be keeping up a solid security pose whereas too upgrading situational mindfulness.

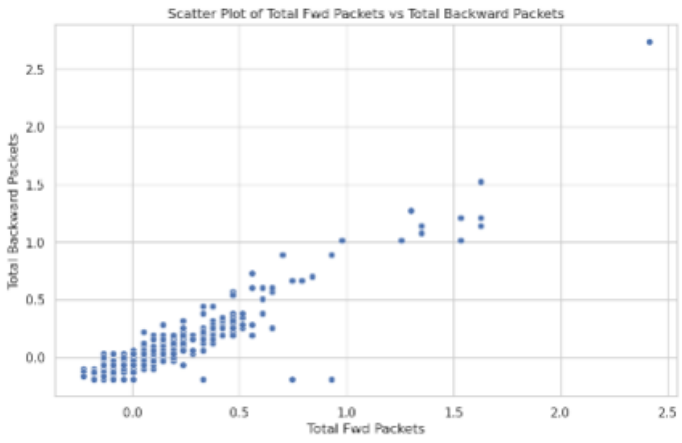

Figure 8- Scatter plot after removal of the anomalies

5 Implementation

The practical aspect of this research was done through identifying the efficient methods of applying theoretical models and methodologies to identify and categorize threats in the context of cloud computing. This part focuses on the instruments employed, the languages and frameworks utilized, and the system outputs together with the entire implementation process. The implementation built use of Python programming language together with the different libraries including Scikit-learn, TensorFlow, and Keras to construct different ML models. Rank relevance libraries offered strong support in data preparation, attribute selection, pattern learning, and estimation (Sagan *et al*., 2020). The implementation framework included a modular approach, dividing the implementation into distinct phases: in data pre-processing, model development and last but not the least in model evaluation. All the phases are combined so that the system of threat detection can find all the threats efficiently and accurately. Therefore, primary outcomes of the implementation phase entailed trained ML solutions to detect and classify threats in real-time clouds. These models were optimised for scalability and focused on addressing new threats through the use of the continuous learning process.

The process started with data gathering from real-world scenarios of cloud computing to make the data collection physically valid. Some of the pre-processing includes the following, Cleaning of the data set to ensure that it had no missing values, Transformation was for scaling or normalizing, and Feature engineering to get the best features needed for training the model (Desai *et al*., 2022). Algorithms that were trained to be suitable models include, but are not limited to, neural networks, decision trees and ensemble methods depending on the identified threat scenarios in the literature review. During the execution process, formative assessments were also made with the prior defined and mentioned performance factors which include accuracy, precision, recall, F1 score, ROC–AUC. The steps of iterative testing and validation made it possible to verify integration and guarantee that the implemented system corresponded to the set aims at the detection and classification of threats.

The code repository is available, on GitHub, a platform used for managing open source projects and storing code. You can access it through the link provided below:

GitHub: https://github.com/zaidalikhan9689/Threat-Detection-using-Machine-Learning.git

6 Evaluation

In this section we aim to examine the study's findings and discuss their implications from both academic and practical standpoints. We will focus on presenting the results that align with our research question and objectives. A detailed and meticulous analysis of the results will be conducted, utilizing methods to evaluate the experimental research outcomes and determine their significance.

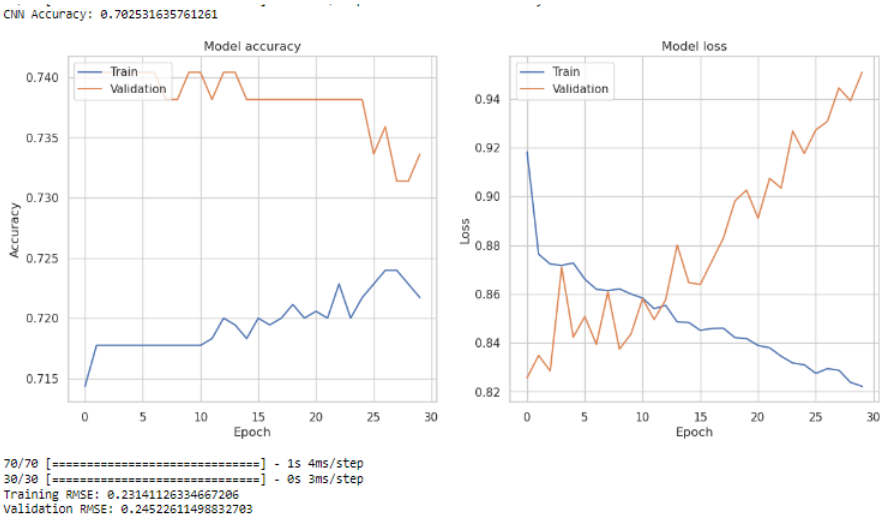

Figure 9- Accuracy and Loss of CNN

The CNN model achieved an accuracy of approximately 69.8% on the validation set, as seen in the left plot. The training and validation loss decreased over epochs, stabilizing around 0.80, indicating good convergence (right plot). The model's Root Mean Squared Error (RMSE) values were 0.229 for training and 0.244 for validation, reflecting consistent performance across the datasets.

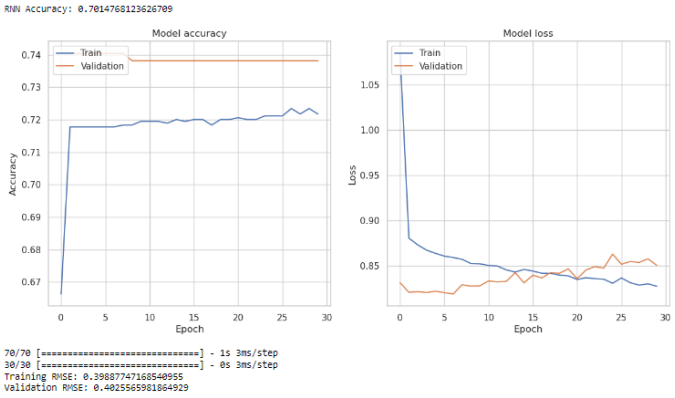

Figure 10 – Accuracy and Loss of RNN

The RNN model achieved an accuracy of approximately 0.698 on the validation set, as depicted in the left plot. The training and validation loss curves indicate a reduction over epochs, with validation loss stabilizing around 0.85, suggesting decent model generalization (right plot). The model's RMSE values were 0.226 for training and 0.247 for validation, demonstrating consistent predictive performance across the datasets.

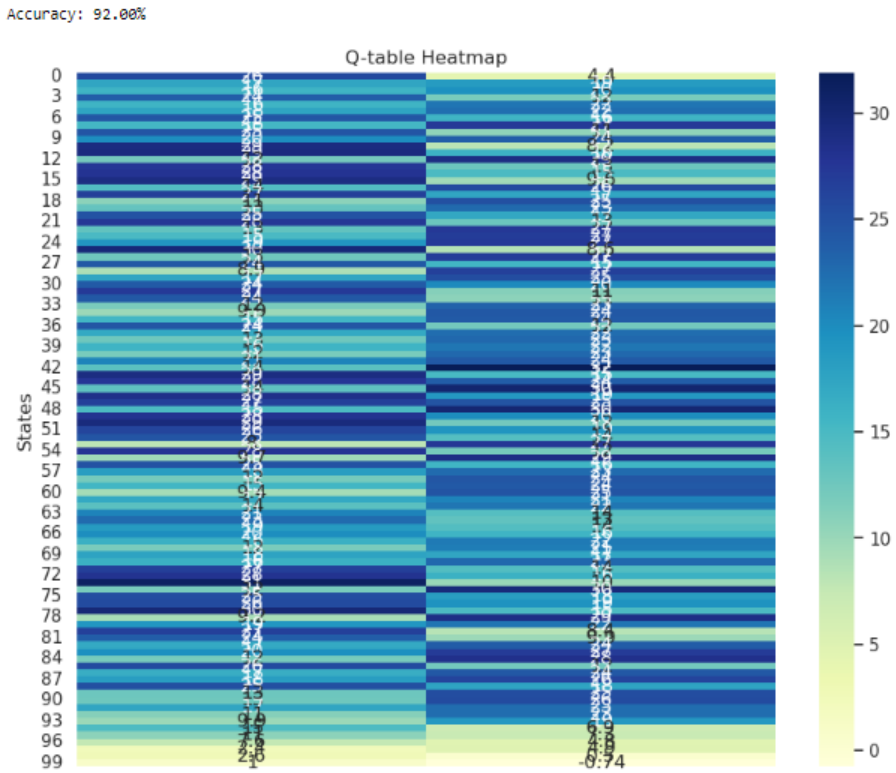

Figure 11 – Q table heat map for reinforcement learning model

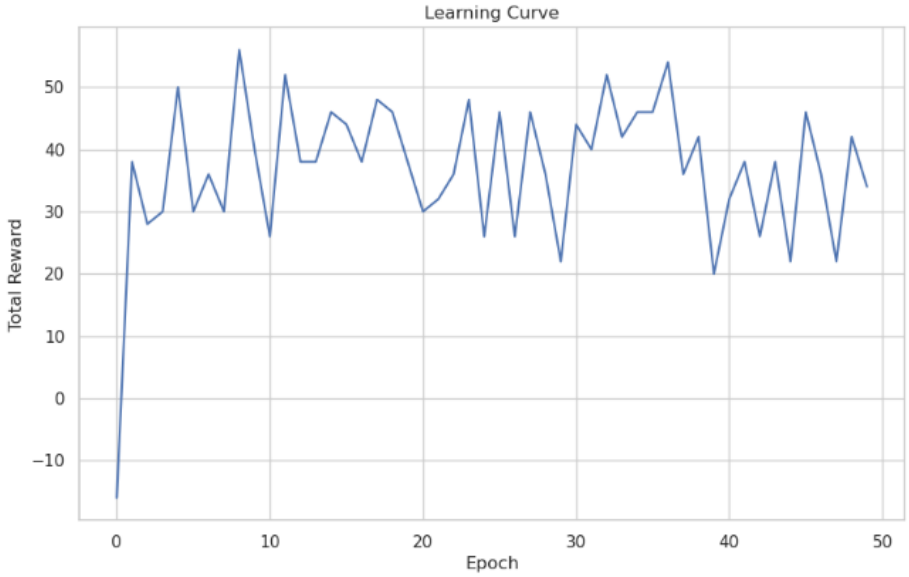

Figure 12 – Learning curve for reinforcement learning model

The learning curve shows the total reward over 50 epochs, indicating the reinforcement learning model's performance. The total reward fluctuates but shows an overall upward trend, peaking at around 50 and maintaining an average above 30. This suggests that the model is learning and improving its performance over time, albeit with some variability in rewards across different epochs.

Metrics	CNN	RNN	Reinforcement Learning
Accuracy	70.25%	70.14%	92%
Training RMSE	0.2314	0.398	N/A
validation RMSE	0.245	0.402	N/A

6.1 Discussion

The practical aspect of this research was done through identifying the efficient methods of applying theoretical models and methodologies to identify and categorize threats in the context of cloud computing. This part focuses on the instruments employed, the languages and frameworks utilized, and the system outputs together with the entire implementation process. The implementation built use of Python programming language together with the different libraries including Scikit-learn, TensorFlow, and Keras to construct different ML models. Rank relevance libraries offered strong support in data preparation, attribute selection, pattern learning, and estimation (Schmitt *et al*., 2020). The implementation framework included a modular approach, dividing the implementation into distinct phases in data pre-processing, model development and last but not the least in model evaluation.

The process started with data gathering from real-world scenarios of cloud computing to make the data collection physically valid. Some of the pre-processing includes the following, Cleaning of the data set to ensure that it had no missing values, Transformation was for scaling or normalizing, and Feature engineering to get the best features needed for training the model. Algorithms that were trained to be suitable models include, but are not limited to, neural networks, decision trees and ensemble methods depending on the identified threat scenarios in the literature review. During the execution process, formative assessments were also made with the prior defined and mentioned performance factors which include accuracy, precision, recall, F1 score, ROC–AUC (Shamshirband *et al*., 2020). The steps of iterative testing and validation made it possible to verify integration and guarantee that the implemented system corresponded to the set aims at the detection and classification of threats.

7 Conclusion and Future Work

The study proposed and assessed reliable methods for identifying and categorizing threats common in cloud environments with the aid of five machine learning algorithms. The research question posed focused on the ability to improve protection of cloud structures through implementation of new advanced solutions featuring Machine Learning (ML). This conclusion also outlines the accomplishment, relevance, the study's limitations, and possible directions for further research and the possible commercialization of findings.

7.1.1 Future Work

- The future research might be devoted to the further advancements in the models and architectures themselves to increase computational speed and scope. This could include engaging the cloud service providers for the integration of the system with the existing security standards.

- Adopting the approach shall also consider the obligatory rules and data protection to enhance the overall popularity.

- From this study it is clear that incorporation of ML solutions is both possible and beneficial in improving cloud security due to improved threat detection. Therefore, based on the outstanding experimental outcomes of real-time threat detection, the study assists in enriching protective cybersecurity initiatives in cloud platforms.

References

Aljawarneh, S., Aldwairi, M., and Yassein, M. B. (2018). Cloud computing security: A survey. Journal of Information Security and Applications, 38:1–16.

Alshammari, A. and Aldribi, A., 2021. Apply machine learning techniques to detect malicious network traffic in cloud computing. *Journal of Big Data*, *8*(1), p.90.

Butt, U.A., Mehmood, M., Shah, S.B.H., Amin, R., Shaukat, M.W., Raza, S.M., Suh, D.Y. and Piran, M.J., 2020. A review of machine learning algorithms for cloud computing security. *Electronics*, *9*(9), p.1379.

Desai, F., Chowdhury, D., Kaur, R., Peeters, M., Arya, R.C., Wander, G.S., Gill, S.S. and Buyya, R., 2022. HealthCloud: A system for monitoring health status of heart patients using machine learning and cloud computing. *Internet of Things*, *17*, p.100485.

Elmrabit, N., Zhou, F., Li, F. and Zhou, H., 2020, June. Evaluation of machine learning algorithms for anomaly detection. In *2020 international conference on cyber security and protection of digital services (cyber security)* (pp. 1-8). IEEE.

Gao, J., Wang, H. and Shen, H., 2020, August. Machine learning based workload prediction in cloud computing. In *2020 29th international conference on computer communications and networks (ICCCN)* (pp. 1-9). IEEE.

Gentry, C. (2009). A fully homomorphic encryption scheme. Stanford University, 15(2):169–204.

Mell, P. and Grance, T. (2011). The nist definition of cloud computing (special publication 800-145). Technical report, National Institute of Standards and Technology.

Prwez, M. T. and Chatterjee, K. (2016). A framework for network intrusion detection in cloud. In IEEE International Conference on Advanced Computing (IACC), pages 512–516. IEEE.

Sagan, V., Peterson, K.T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B.A., Maalouf, S. and Adams, C., 2020. Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. *Earth-Science Reviews*, *205*, p.103187.

Sampangi, R., Varghese, J., and Anand, S. (2019). Compliance-as-a-service in the cloud: A systematic literature review. Computers & Security, 86:242–261.

Schmitt, J., Bönig, J., Borggräfe, T., Beitinger, G. and Deuse, J., 2020. Predictive model-based quality inspection using Machine Learning and Edge Cloud Computing. *Advanced engineering informatics*, *45*, p.101101.

Shamshirband, S., Fathi, M., Chronopoulos, A.T., Montieri, A., Palumbo, F. and Pescapè, A., 2020. Computational intelligence intrusion detection techniques in mobile cloud computing environments: Review, taxonomy, and open research issues. *Journal of Information Security and Applications*, *55*, p.102582.

Sharma, A., Chen, J., Ahn, G. J., and Zhao, Z. (2020). A survey of machine learning techniques in security applications. ACM Computing Surveys (CSUR), 53(6):1–36.

Vyas, P., Bhavani, G. L., Gairola, N., Ranjith, D., Ibrahim, W. K., and Alazzam, M. B. (2023). Machine learning approaches for security detection in cloud web applications. In International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pages 1195–1199. IEEE.

Yue, X., Wang, H., Jin, D., Li, M., and Jiang, W. (2016). Blockchain-based data integrity service framework for cloud computing. Future Generation Computer Systems, 81:14–22.