A new approach to Cloud security posture management and anomaly detection in cloud traffic using gradient boosting classifier algorithm

Academic internship
MSc in Cybersecurity

Varun Gowda Doddakarade Nagendra
Student ID: 22171541

School of Computing
National College of Ireland

Supervisor: Prof. Vikas Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Varun Gowda Doddakarade Nagendra		
Student ID:	X22171541		
Programme:	MSc in Cyber Security	**Year:**	2023-2024
Module:	Academic Internship		
Supervisor:	Prof. Vikas Sahni		
Submission Due Date:	31/01/2024		
Project Title:	A new approach to Cloud security posture management and anomaly detection in cloud traffic using gradient boosting classifier algorithm		
Word Count:	6805	**Page Count**	20

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: Varun Gowda Doddakarade Nagendra

Date: 31/01/2024

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):	

A new approach to Cloud security posture management and anomaly detection in cloud traffic using gradient boosting classifier algorithm

Varun Gowda Doddakarade Nagendra
X22171541

Abstract

Cloud Security Posture Management is a set of practices and tools designed to ensure the consistent and effective security configuration of cloud resources that involves monitoring, assessing, and managing the security posture of an organization’s cloud infrastructure to prevent and address security misconfigurations, vulnerabilities, and compliance issues. This project navigates the area of cloud security with a dual focus on fortifying the security posture and detecting anomalies within cloud traffic. A conceptual security framework is devised by integrating features from prominent cloud security frameworks. This framework serves as a comprehensive guide for managing security postures in the cloud.

Simultaneously, the study delves into anomaly detection in cloud traffic, employing the Gradient Boosting Classifier algorithm to discern deviations from established norms. Rigorous preprocessing, feature selection, and model training characterize the methodology. Applying the proposed conceptual framework and the gradient boosting classifier aligns with the dynamic nature of cloud computing environments, ensuring robust security measures and efficient identification of anomalous activities. The model achieved an accuracy of 94.57% and 96% of the F1-score, which is better when compared to the SVM (89%) and LSTM (94%) models. These evaluation factors showcase the efficacy of the approach in enhancing cloud security resilience, offering a substantial contribution to the evolving landscape of cloud security management.

Keywords: Cloud Security, Anomaly Detection, Conceptual Framework, Gradient Boosting Classifier, Security Posture Management, Cloud Traffic, Feature Selection.

1 Introduction

In the ever-evolving landscape of information technology, cloud computing has emerged as a transformative force, reshaping the way organizations manage and deploy their IT resources. The advent of cloud platforms, including Microsoft Azure, Amazon Web Services, and Google Cloud Platform, has ushered in an era of unprecedented scalability, flexibility, and cost-efficiency. These platforms have become the bedrock upon which modern businesses build their digital infrastructure, enabling them to innovate, expand, and compete in a globalized world.

However, with the migration of critical business operations and sensitive data to the cloud, a paramount concern has arisen about security. Ensuring cloud-hosted resources’ safety, compliance, and resilience has become a top priority for organizations of all sizes and industries. The ever-evolving threat landscape, characterized by sophisticated cyberattacks and constant vulnerabilities, demands a proactive and comprehensive approach to cloud security.

In response to these challenges, Cloud Security Posture Management (CSPM) has emerged as a critical practice to mitigate security risks in cloud environments. CSPM involves the continuous assessment and enhancement of the security posture of cloud resources, to identify misconfigurations, vulnerabilities, and compliance violations. This practice empowers organizations to proactively detect and remediate security issues, reducing the risk of data breaches, compliance failures, and other security incidents.

As organizations embrace the cloud, they encounter a new set of security challenges. Cloud environments are dynamic, complex, and globally accessible, making them attractive targets for cyberattacks. Security breaches, data leaks, misconfigurations, and compliance violations are ever-present threats that demand vigilant attention. This section delves into the unique security challenges posed by cloud computing. It explores the implications of shared responsibility models, where cloud providers and customers share security responsibilities. The dynamic nature of cloud resources and the need for continuous security monitoring are also highlighted.

To address these security challenges, organizations have turned to CSPM as a pivotal practice. CSPM involves continuously assessing cloud configurations, security policies, and compliance posture. It encompasses tasks such as vulnerability scanning, configuration analysis, and policy enforcement. This section underscores the significance of CSPM in cloud security strategy. It outlines the objectives of CSPM, which include identifying misconfigurations, detecting vulnerabilities, enforcing security policies, and ensuring compliance with industry regulations and best practices.

In the rapidly advancing landscape of cloud computing, where data and processes transcend traditional boundaries, the security of cloud traffic emerges as a paramount concern. Anomalies in cloud traffic represent deviations from the expected and established patterns of behavior within these expansive and dynamic computing environments. As organizations increasingly migrate their critical operations to the cloud, the need to comprehend, detect, and mitigate anomalies becomes imperative for ensuring the integrity, confidentiality, and availability of data.

The study of anomalies in cloud traffic is motivated by several crucial factors. First and foremost is the sheer scale and complexity of cloud infrastructures, which make them susceptible to a diverse range of security threats. Understanding anomalies becomes essential for safeguarding sensitive information, preventing unauthorized access, and mitigating the potential impact of malicious activities. Moreover, the dynamic nature of cloud environments, characterized by virtualization, scalability, and resource sharing, introduces unique challenges in anomaly detection.

Research question

1. How can the diverse security features from standard cloud security frameworks be effectively integrated into a unified conceptual cloud framework, ensuring a comprehensive and cohesive approach to cloud security?
2. To what extent does the application of the Gradient Boosting Classifier algorithm enhance the efficacy of anomaly detection in cloud traffic?

1.1 Document Structure

The research project is organized as follows:

1.1.1 Section 1: Introduction

- Background to the research topic
- Research Problem and Context
- Potential Contributions and Benefits

1.1.2 Section 2: Literature Review

- Cloud Security Posture Management Frameworks
- Cloud security controls and threats
- Anomaly detection in cloud traffic
- Cloud Security Frameworks

1.1.3 Section 3: Research Methodology

- Research Design and Approach
- Data Collection Methods

1.1.4 Section 4: Design specifications

- Architecture and Components

1.1.5 Section 5: Implementation

1.1.6 Section 6: Evaluation

- Case study 1 - Evaluation of Gradient boosting classifier algorithm
- Case study 2 - Evaluation of the SVM algorithm
- Case study 3 - Evaluation of the LTSM algorithm
- Discussion

1.1.7 Section 7: Conclusion and Future Work

- Summary of Research Findings
- Final Remarks on the Proposed Solution's Significanc

2 Related Work

2.1 Cloud Security Posture Management (CSPM) Frameworks

In these papers, both researchers (Diogenes, 2019), (Rastogi, 2023) have studied CSPM frameworks other than RPA. Diogenes proposes a CSPM solution for monitoring Azure, AWS, and GCP security setups and compliance via the use of native cloud services and APIs.

Constant surveillance and prompt action in the event of a security breach are two of their main concerns. When it comes to CSPM, however, Rastogi has developed a cross-platform solution that provides real-time security analysis and granular control over security measures.

Research like this sheds light on the potential of CSPM frameworks in the absence of robotic process automation. While RPA offers advantages, determining whether integrating RPA will provide a net gain requires knowledge of current CSPM solutions.

The paper (Coppola, 2023) discusses the growing significance of cloud security, particularly in the context of organizations utilizing cloud services like Amazon Web Services (AWS) for handling substantial volumes of big data and deploying artificial intelligence (AI) technologies. The proposed Cloud Security Posture Management (CSPM) tool, designed with a focus on AWS and aligned with the NIST Cybersecurity Framework v1.1, addresses the challenges of maintaining security and compliance in a landscape where ubiquitous data access is desirable. Leveraging AWS services such as VPC traffic logs, GuardDuty, and CloudTrail, the CSPM tool integrates AI capabilities for continuous threat monitoring, misconfiguration alerting, risk identification, and remediation recommendations.

2.2 Cloud security controls and threats

This paper Sailakshmi (Sailakshmi, 2021) addresses the imperative need for comprehensive insights into securing business-critical applications in the dynamic landscape of cloud computing. By focusing on the Cloud Security Alliance’s (CSA) Top 20 Critical Controls, the author has tried to bridge information gaps and provide a foundation for securing cloud environments. Acknowledging the increasing adoption of cloud services, the research emphasizes the significance of understanding security controls and underscores the importance of cloud audits to ensure both security and compliance. A noteworthy aspect of the paper is its proposal for a comparative analysis across major cloud providers AWS, Google Cloud, and Azure against the Center for Internet Security (CIS) Top 20 Controls.

In this paper (Balachandra and Ramakrishna, 2019), the authors focus on addressing the security challenges associated with cloud computing, which has become an integral part of modern IT infrastructure. The authors highlight the transformative impact of cloud services, emphasizing the widespread adoption of resources delivered over the Internet. However, this growth is accompanied by significant security concerns that hinder the broader acceptance of cloud computing. The study identifies key issues such as data protection, scrutinizing the cloud utilization by vendors, and the need for secure information storage.

The paper (Singh, 2017) delves into the fundamental features of cloud computing and articulates the security issues and threats arising from this transformative technology. It addresses key topics such as cloud architecture framework, service and deployment models, and cloud technologies, and introduces cloud security concepts, threats, and attacks. Furthermore, the paper identifies and discusses several open research issues within the realm of cloud security, indicating a forward-looking approach to address evolving challenges in the field. However, a more in-depth critical analysis could involve a comparative examination of existing solutions to the identified security issues, practical case studies demonstrating the application of proposed solutions, and consideration of the evolving regulatory landscape governing cloud security.

2.3 Anomaly detection in cloud traffic

In this paper (Yang, 2019) the author Yang addresses the escalating threat of abnormal activities in network environments, emphasizing the critical need for timely anomaly detection to ensure network security. The proposed solution centers around a novel anomaly network traffic detection algorithm specifically tailored for cloud computing environments. The framework of the anomaly network traffic detection system is outlined, incorporating six crucial network traffic features such as the number of source and destination IP addresses, port numbers, packet types, and overall network packets.

The paper introduces a hybrid model that combines information entropy and Support Vector Machines (SVM) to effectively normalize network feature values and employ SVM for the detection of anomaly network behaviors.

In this paper, author (Alshammari, 2021) delves into the complex landscape of computer network security, where the ever-evolving nature of attacks poses significant risks. Notably, the study acknowledges the continual emergence of new attacks targeting open ports, necessitating advanced tools to combat network vulnerabilities effectively. In response to this dynamic threat landscape, machine learning (ML) has gained prominence as a powerful technique. The proposed research introduces a detection framework that leverages an ML model to enhance IDS capabilities in detecting network traffic anomalies.

The key challenges addressed in this research revolve around feature extraction, essential for training the ML model to differentiate between various types of attacks and regular traffic. The research utilizes the ISOT-CID network traffic dataset for model training, augmenting it with significant column features.

The paper (Girish, 2023) identifies a significant challenge in cloud environments the timely detection and prediction of anomalies. While traditional methods rely on manual anomaly detection through threshold levels and heartbeat monitoring, the paper underscores the growing trend in utilizing machine learning techniques for proactive anomaly detection. The proposed model introduces a novel approach for anomaly detection in the OpenStack cloud environment, employing Stacked and Bidirectional LSTM models to construct a neural network. Data for experimentation is collected from OpenStack using collected, encompassing 10 features and a class label. The use of LSTM neural networks demonstrates efficacy in detecting anomalies, with the proposed model achieving impressive detection accuracy rates of 94.61% for the training set and 93.98% for the test set.

While the paper provides valuable insights into anomaly detection using advanced neural network architectures in the OpenStack environment, a more comprehensive analysis could involve comparing the proposed model with existing anomaly detection methods, assessing its scalability, and addressing potential limitations or challenges in real-world deployments.

2.4 Cloud Security Frameworks

In this paper, the authors (Chauhan and Stavros, 2023) systematically examine and compare COBIT 5, NIST, ISO 27017, CSA Star, and AWS Well-Architected cloud security frameworks, evaluate their effectiveness, and address prevalent security concerns and threats within cloud computing, offering potential solutions and countermeasures to enhance the security of cloud-based systems.

The study conducts a comprehensive comparison of the above-mentioned frameworks, considering factors such as their focus, scope, approach, strengths, limitations, and the steps and tools necessary for implementation. This comparative analysis assists in understanding the differences and specific strengths of each framework, aiding decision-making in the selection and implementation of suitable security measures for cloud-based systems.

The emergence of new frameworks like FedRAMP in the US and C5 in Germany aims to enhance protection against threats and vulnerabilities specific to cloud environments. The paper (C. Di Giulio, 2017) provides a comprehensive overview of ISO/IEC 27001, C5, and FedRAMP, conducting a critical examination of their completeness and adequacy in addressing contemporary threats to cloud assurance. A comparative analysis is performed to question the level of protection offered by these certifications, identifying weaknesses in all three frameworks, and underscoring the need for improvements to align with the evolving security requirements of the current threat landscape.

While the paper sheds light on the convergence and shortcomings of these certifications, a more nuanced discussion could involve proposing specific enhancements or adaptations required in these frameworks to address identified weaknesses.

2.5 Summary

The literature review emphasizes the need for a unified conceptual cloud framework, integrating diverse security features to address dynamic challenges in cloud security. Simultaneously, the literature highlights the escalating threat of abnormal activities in cloud traffic, prompting the second question on the efficacy of applying the Gradient Boosting Classifier algorithm in anomaly detection. This inquiry aims to assess how advanced machine learning techniques contribute to enhancing the precision and efficiency of anomaly detection, thereby fortifying the security landscape of cloud environments.

Table 1: A summary of the literature review

Section	Key Findings/Contributions	Suggestions for Improvement
Cloud Security Posture Management Frameworks , Security controls and threats	Highlighting the potential of CSPM frameworks. The importance of constant surveillance and prompt action in security breaches. CSPM solutions for Azure, AWS, and GCP security setups	There is no single cloud security framework that covers all security controls. Hence there is a need to combine security controls from different cloud frameworks and the security controls identified through the literature review.

Anomaly Detection in Cloud Traffic	Novel anomaly network traffic detection algorithm for cloud environments. Importance of timely anomaly detection in ensuring network security. Incorporation of six crucial network traffic features in the detection system. Use of SVM and LTSM for detecting anomaly network behaviors.	Anomaly detection in the cloud using a Gradient boosting classifier algorithm and testing how this algorithm can increase the efficiency and accuracy of anomaly detection.

3 Research Methodology

3.1 Steps to prepare the conceptual cloud framework,

Stage 1: Identifying security controls.

Identifying security controls is the first step in establishing a security framework. This process involves systematically recognizing and documenting the specific measures and mechanisms that will be implemented to safeguard information, systems, and assets within an organization.

The Cloud Security Alliance (CSA) has introduced a valuable resource called the Cloud Controls Matrix (CCM), which serves as a comprehensive reference for both customers and vendors. This matrix presents a structured table outlining fundamental security principles, aiding in the assessment of overall security risks associated with cloud providers. The CCM delves into the intricate details of security principles across 13 domains within cloud security.

The CCM incorporates controls from various widely accepted industry security standards, such as NIST, ISO 27001/27002, ISACA, PCI, Jericho Forum, and NERC CIP. This integration ensures a holistic approach to cloud security by assimilating insights and best practices from diverse sources.

Stage 2: Combining security controls

Security controls from academic literature are compared with security controls collected from industry standards to remove duplicates.

3.2 Methodology for anomaly detection in cloud traffic

Dataset collection: Cloud traffic dataset CSE-CIC-IDS2018[1], for an AWS cloud service provider which contains network logs that provide information about the traffic within the cloud infrastructure. The dataset used here was collected in February 2018 for the entire

[1] https://registry.opendata.aws/cse-cic-ids2018

month. Another dataset, UNSW-NB 15[2] dataset was created by the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS).

Dataset pre-processing: Steps were taken to ensure the raw data's quality and suitability for anomaly detection. Duplicate records were identified and removed to avoid redundancy and maintain the dataset's integrity. Subsequently, missing values, if any, were handled to prevent biases in the analysis. Irrelevant or redundant information was filtered out to streamline the dataset, focusing on pertinent features for anomaly detection. Addressing potential outliers, a crucial step in anomaly detection was performed to mitigate their impact on the analysis and model performance.

To maintain data integrity, checks were implemented, verifying, for instance, the chronological order of timestamps and ensuring the consistency of certain fields.

Selection of algorithm model: The algorithm used in this project is the Gradient Boosting classifier, as it is an ensemble learning method that builds a strong predictive model by combining multiple weak learners, typically decision trees. This allows the model to capture complex relationships and patterns within the data, making it effective for detecting anomalies (M. K. Islam, 2020). As anomalies often manifest non-linear patterns in data, this algorithm is capable of handling non-linear relationships, allowing it to effectively identify complex patterns associated with anomalous behavior. The algorithm is versatile and can be applied to different types of data like structured and tabular data which are commonly found in anomaly detection scenarios.

Training and testing the selected model: The dataset UNSW-NB 15[2] is used to train the algorithm where 80% of the datasets are used to train the algorithm using training data (X_train, y_train) with the fit method. The remaining 20% is used to test the algorithm and to evaluate its performance.

Performance evaluation of the model: Evaluating a Gradient Boosting Classifier (GBC) for anomaly detection involves assessing its performance using relevant metrics and visualization techniques. The accuracy of the model is calculated by comparing the predicted labels (y_pred) with the actual labels from the test set (y_test) using the accuracy_score function. This metric indicates the overall correctness of the model's predictions. The classification report is used to evaluate precision, recall, and F1-score for each class as offers a more comprehensive understanding of the model's performance.

[2] https://www.kaggle.com/datasets/mrwellsdavid/unsw-nb15

4 Design Specification

The conceptual framework that is being proposed in this paper is prepared by combining the security controls that are implemented in standard frameworks and the security controls that are identified through examining real-world security challenges and academic literature review.

The framework development process is shown below,

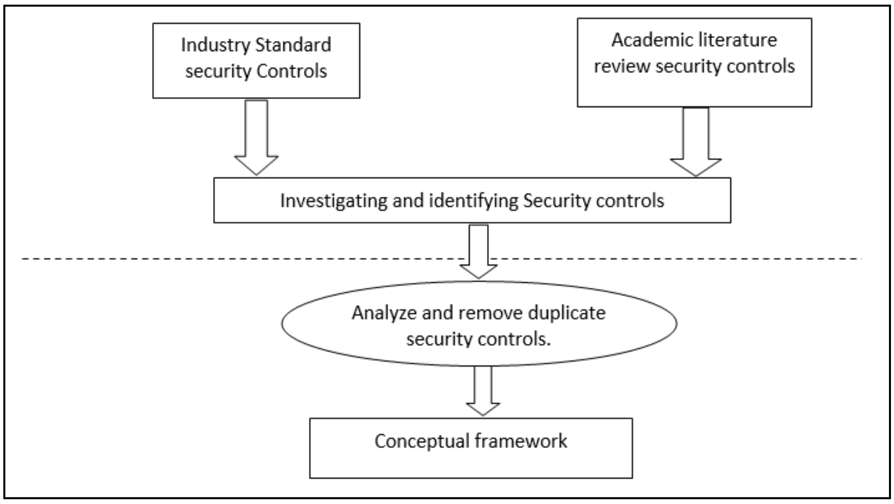
Industry Standard security Controls
Academic literature review security controls
Investigating and identifying Security controls
Analyze and remove duplicate security controls.
Conceptual framework

Figure 1: The conceptual framework development process

The below table illustrates the association between selected security controls and prominent industry-standard frameworks, namely the NIST[3] Cybersecurity Framework, ISO/IEC 27001/27002[4], ISACA COBIT[5], and PCI DSS[6]. The "Privacy" control, not explicitly addressed in NIST[3], ISO[4], and COBIT[4], is implicitly considered in PCI DSS[6]. "Availability," "Integrity," "Malware Protection," "Authentication," "Authorization," and "Encryption" controls find common ground across all four frameworks, emphasizing their fundamental importance.

"Leftover Data Removal" is not explicitly covered in any framework, while "Outdated Software Detection" is acknowledged by NIST[3] and PCI DSS[6]. "Trust" remains unaddressed across all frameworks. "Regulatory Compliance" is implicitly considered in NIST and COBIT and explicitly addressed in ISO/IEC 27001/27002[4] and PCI DSS[6]. This overview emphasizes the alignment of specific security controls with industry standards, aiding organizations in tailoring their security practices based on the guidelines provided by these frameworks.

[3] https://www.nist.gov/cyberframework

[4] https://www.iso.org/standard/27001

[5] https://www.isaca.org/resources/frameworks-standards-and-models

[6] https://www.pcisecuritystandards.org/standards/

Table 2: The summary of security controls from different cloud frameworks,

Security controls	NIST cybersecurity framework	ISO/IEC 27001/27002	ISACA COBIT	PCI DSS
Privacy				
Availability	✓	✓	✓	✓
Integrity	✓	✓	✓	✓
Trust		✓		
Leftover Data Removal				✓
Malware Protection	✓	✓	✓	✓
Outdated Software Det	✓			✓
Authentication	✓	✓	✓	✓
Authorization	✓	✓	✓	✓
Encryption	✓	✓	✓	
Regulatory Compliance	✓	✓		

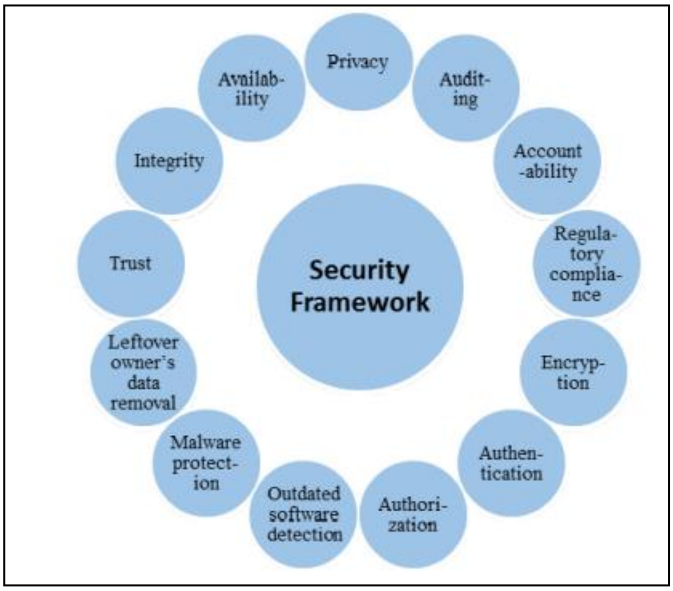

Fig 2: Security framework after combining security controls from standard frameworks.

Anomaly detection serves as a crucial layer of defense against evolving cyber threats, offering the capability to identify patterns and behaviours that deviate from the norm. In the context of cloud security, anomalies signify malicious activities, potential breaches, or system vulnerabilities that traditional security measures might overlook. Unlike rule-based approaches, which rely on predefined patterns, anomaly detection brings a proactive and adaptive dimension to security protocols. Detecting anomalous behaviour becomes not only a means of preventing unauthorized access or data breaches but also a strategy for ensuring the uninterrupted and secure functioning of cloud-based services.

The Gradient Boosting Classifier implemented for anomaly detection is a supervised machine learning algorithm that combines the strengths of multiple weak learners, typically decision trees, to create a robust and accurate predictive model (M. K. Islam, 2020). In the context of anomaly detection, the GBC is employed to discern patterns in the data that deviate from the norm, signaling potential anomalies.

4.1 Techniques and Frameworks Used:

4.1.1 Scikit-Learn: The implementation relies on the scikit-learn library, a versatile and widely adopted machine learning toolkit in Python. Scikit-learn provides a comprehensive set of tools for building and evaluating machine learning models, including a robust implementation of the Gradient Boosting algorithm. Scikit-learn provides an efficient and user-friendly interface for various machine learning algorithms, including Gradient Boosting.

4.1.2 Gradient Boosting Algorithm: The core of the anomaly detection model is the Gradient Boosting algorithm. This ensemble learning technique combines weak learners, typically decision trees, sequentially. Each tree corrects the errors of its predecessor, focusing on instances that are challenging to classify. The iterative refinement of the model enhances its ability to capture complex patterns in the data, making it well-suited for detecting anomalies that might exhibit subtle or non-linear characteristics.

4.2 Feature Selection: Before training the model, a careful selection of features relevant to anomaly detection is performed. This feature selection step is crucial for enhancing the model’s capacity to identify patterns associated with anomalous behaviour. The chosen features contribute significantly to the model’s understanding of the data and its ability to discriminate between normal and anomalous instances.

4.3 Data Preprocessing: The dataset undergoes preprocessing steps to ensure it is suitable for training the Gradient Boosting Classifier. This includes handling timestamp data, filling in missing values, and converting data to a format compatible with the algorithm. Proper data preprocessing is essential for preparing the input data in a standardized manner, ensuring the model can effectively learn from it.

4.4 Algorithm and Model Functionality:

4.4.1 Sequential Training: The Gradient Boosting Classifier employs a sequential training approach. Decision trees are trained one after another, with each tree focusing on instances where the model has previously struggled. This sequential learning process allows the model to progressively correct its mistakes, improving its ability to capture intricate patterns in the data. This is particularly valuable for anomaly detection tasks where anomalies may exhibit diverse and nuanced characteristics.

4.4.2 Ensemble Learning: The model utilizes ensemble learning by aggregating predictions from multiple decision trees. The ensemble approach enhances the model’s predictive power, enabling it to generalize well to unseen data and identify anomalies effectively. The collective decision-making process of the ensemble contributes to the model’s robustness and resilience to overfitting.

4.4.3 Predictive Accuracy: The primary goal of the Gradient Boosting Classifier is to achieve high predictive accuracy. In the context of anomaly detection, this involves correctly identifying instances as normal or anomalous. The accuracy metric provides a quantitative measure of the overall correctness of the model's predictions, offering insights into its reliability.

4.4.4 Evaluation Metrics: The model's performance is assessed using a variety of metrics. Accuracy measures the overall correctness of the model, while the confusion matrix breaks down predictions into true positives, true negatives, false positives, and false negatives. The Receiver Operating Characteristic (ROC) curve provides insights into the model's ability to discriminate between normal and anomalous instances, especially important in imbalanced datasets.

5 Implementation

The final stage of this model implementation involves developing the code to train and test the model. The steps followed here are given in detail,

Data preprocessing: In the dataset UNSW-NB 15[2], the data preprocessing steps involve addressing various aspects to ensure the dataset's suitability for machine learning models. Firstly, the 'dur' column, representing the duration of each record, is scaled to maintain consistent magnitudes. Categorical variables like 'proto' (protocol) and 'service' (service type) are encoded using one-hot encoding or label encoding to convert them into numerical representations. Similarly, the 'state' column, indicating the connection state, is encoded accordingly. The 'attack_cat' column, specifying the attack category, and the 'label' column, denoting normal or anomalous instances, may undergo encoding or transformation based on the analysis goals. Features such as 'sload' and 'dload' (source and destination load) are scaled to ensure consistent ranges. Additionally, any missing or null values in the dataset are appropriately handled through imputation or removal, to maintain data integrity. These preprocessing steps collectively prepare the dataset for training and evaluating machine learning models.

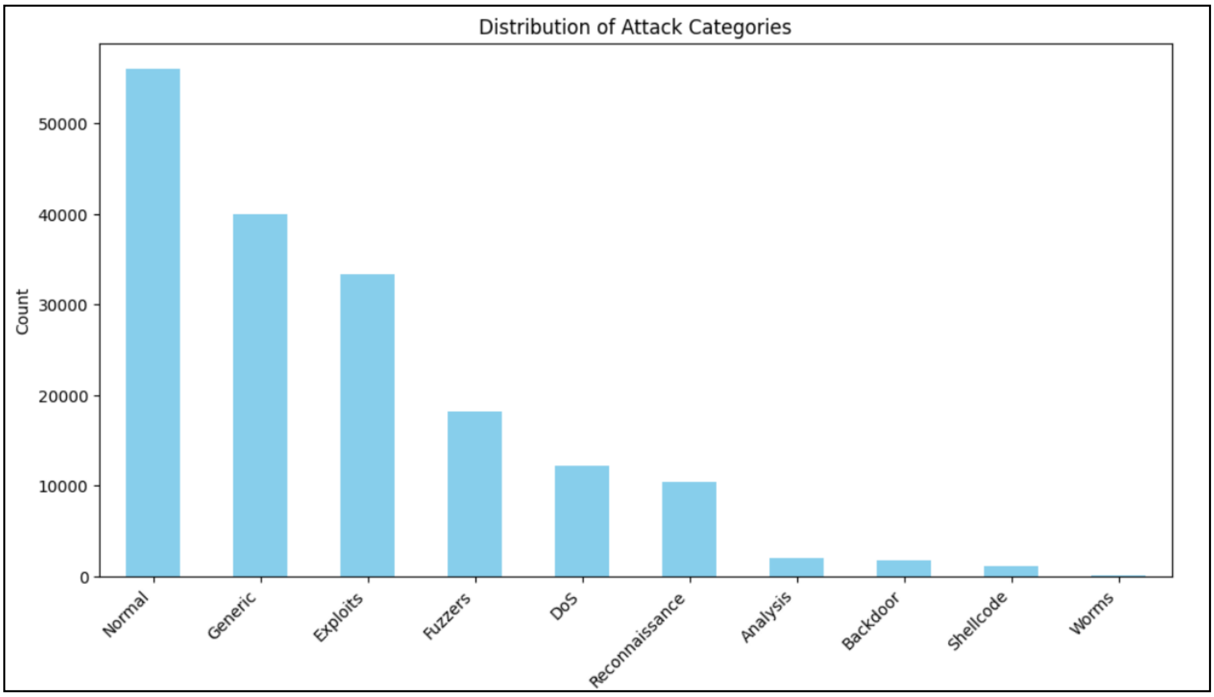

Figure 3: Various attack categories from the selected dataset

Outputs Produced: Trained Gradient Boosting Classifier Model: The primary output of the implementation is a well-trained Gradient Boosting Classifier model. This model has undergone training using the selected features from the cloud traffic dataset. It has learned to distinguish between normal and anomalous patterns in the data, making it capable of detecting anomalies effectively.

Evaluation Metrics and Visualizations: Comprehensive evaluation metrics have been generated to assess the performance of the Gradient Boosting Classifier. These metrics include accuracy, precision, recall, F1-score, false positive rate, and detection rate. Visualizations such as confusion matrices provide a detailed understanding of the model's ability to classify normal and anomalous instances correctly.

Codebase: The implementation involves a well-organized codebase written in Python, leveraging popular libraries such as pandas, scikit-learn, Matplotlib, and Seaborn. The code includes sections for loading and preprocessing the dataset, feature selection, model training using the Gradient Boosting Classifier, hyperparameter tuning, evaluation, and visualization of results. Proper comments within the code enhance its readability and usability for future reference.

5.1 Models Developed:

Gradient Boosting Classifier: The core machine learning model developed for anomaly detection is the Gradient Boosting Classifier. This model has been trained on the selected features of the cloud traffic dataset and is optimized to accurately classify instances as normal or anomalous. The trained model encapsulates the learned patterns and decision boundaries crucial for effective anomaly detection.

5.2 Tools and Languages Used

Programming Language: Python is a widely used and versatile programming language in the data science community. It offers extensive libraries and frameworks specifically designed for data analysis, machine learning, and visualization. Its syntax is clear and concise, making it easy to read and write, which enhances code readability and maintainability.

5.3 Libraries/Frameworks

pandas: It is a powerful data manipulation library in Python. It provides data structures like Data Frames that are well-suited for handling structured data. In this project, pandas are used for loading, cleaning, and preprocessing the cloud traffic datasets and its capability of handling tabular data makes it a natural choice for such tasks.

scikit-learn: It provides a vast array of tools for machine learning tasks, including classification, regression, clustering, and more. The Gradient Boosting Classifier used in this project is part of scikit-learn, offering a well-optimized and easy-to-use implementation of the algorithm.

Matplotlib, Seaborn: Matplotlib is a comprehensive 2D plotting library for Python, and Seaborn is built on top of Matplotlib, providing a high-level interface for statistical data visualization. In this project, these tools are utilized for creating visualizations of the data distribution, feature relationships, and evaluation metrics.

6 Evaluation

The evaluation section serves as a critical component of the study, offering a comprehensive analysis of the anomaly detection model’s performance based on the Gradient Boosting Classifier. This section aims to rigorously assess the effectiveness of the model in distinguishing between normal and anomalous instances within the network traffic data. The evaluation is guided by a set of carefully chosen metrics, including precision, recall, F1-score, and accuracy, which collectively provide insights into the model’s predictive capabilities.

The primary objective of the evaluation is to demonstrate the model’s ability to identify anomalies with a high degree of accuracy while minimizing both false positives and false negatives. This is particularly crucial in the context of anomaly detection, where the consequences of overlooking genuine anomalies or misclassifying normal instances as anomalies can have significant implications for network security.

6.1 Case study 1: Evaluation of Gradient boosting classifier algorithm

In this case study, a gradient-boosting algorithm was employed to detect network intrusions using a dataset UNSW-NB 15² consisting of 119,341 instances. The model achieved an impressive accuracy of 94.57%, demonstrating its capability to accurately classify network activities as normal or malicious. Despite a class imbalance with a higher number of instances in the malicious class, the model exhibited high precision, recall, and F1-Score for both normal and malicious instances. With a precision of 96% for normal instances and 94% for malicious instances, the model effectively minimized false positives and false negatives. This balanced performance reflected in the F1-Scores of 91% for normal instances and 96% for malicious instances, makes the Gradient Boosting algorithm well-suited for real-world network security applications. Recommendations include addressing class imbalance, continuous monitoring of model performance, and adaptation to changing network patterns for sustained effectiveness in identifying network intrusions.

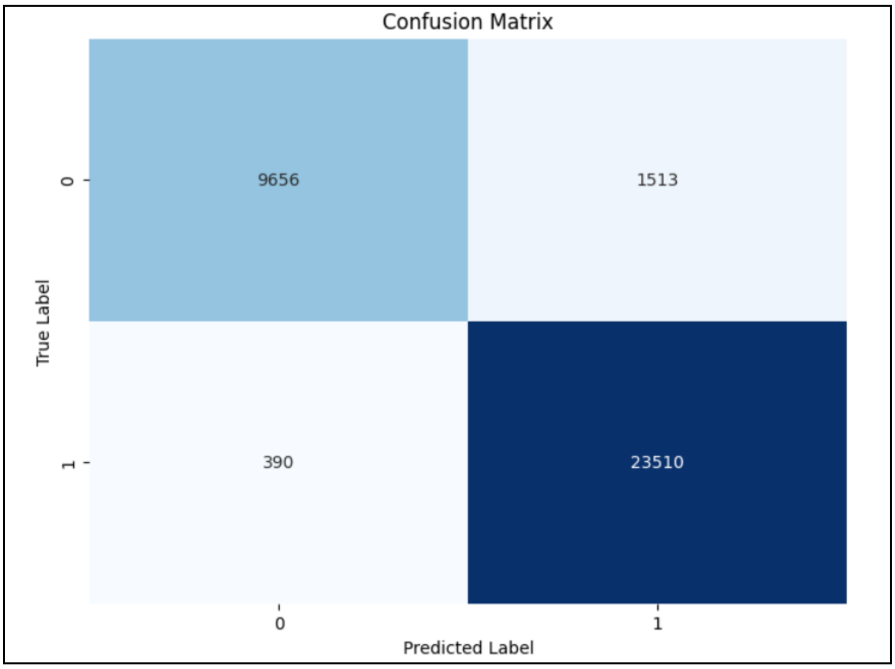

Figure 4: Confusion matrix of Gradient boosting classifier model

6.2 Case study 2: Evaluation of SVM model

In this case study, the Support Vector Machine algorithm was applied to detect network intrusions using a dataset UNSW-NB 152 comprising 119,341 instances. The SVM model demonstrated commendable accuracy at 89.53%, effectively classifying network activities as normal or malicious. Despite a class imbalance, the model exhibited balanced precision, recall, and F1-Scores for both normal and malicious instances. With a precision of 90% for normal instances and 89% for malicious instances, the model effectively minimized false positives and false negatives. The balanced performance reflected in the F1-Scores of 82% for normal instances and 93% for malicious instances, makes the SVM algorithm suitable for real-world applications in network security.

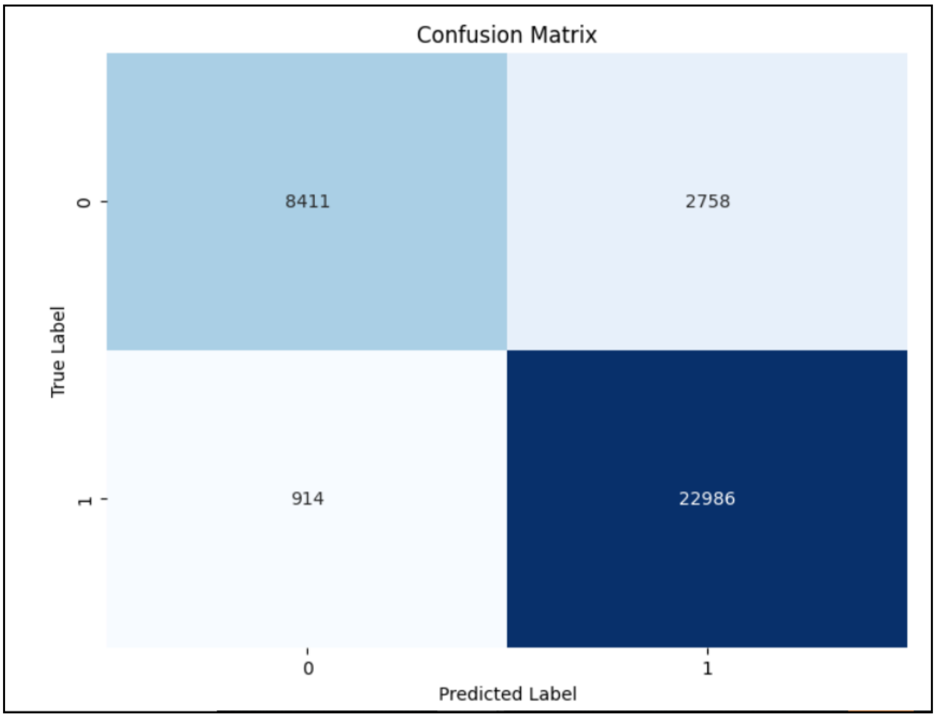

Figure 5: Confusion matrix of Support Vector machine model

6.3 Case study 3: Evaluation of the LSTM model

In this case study, a Long Short-Term Memory (LSTM) neural network was employed for network intrusion detection using a dataset UNSW-NB 15[2] of 119,341 instances. The LSTM model exhibited outstanding performance, achieving an accuracy of 94.32%. Its ability to effectively distinguish between normal and malicious network activities is reflected in the precision, recall, and F1-Score metrics. With a precision of 95% for normal instances and 94% for malicious instances, coupled with high recall values of 87% and 98% respectively, the model demonstrated robust capabilities in minimizing both false positives and false negatives. The balanced F1 scores of 91% for normal instances and 96% for malicious instances underline the model's reliability.

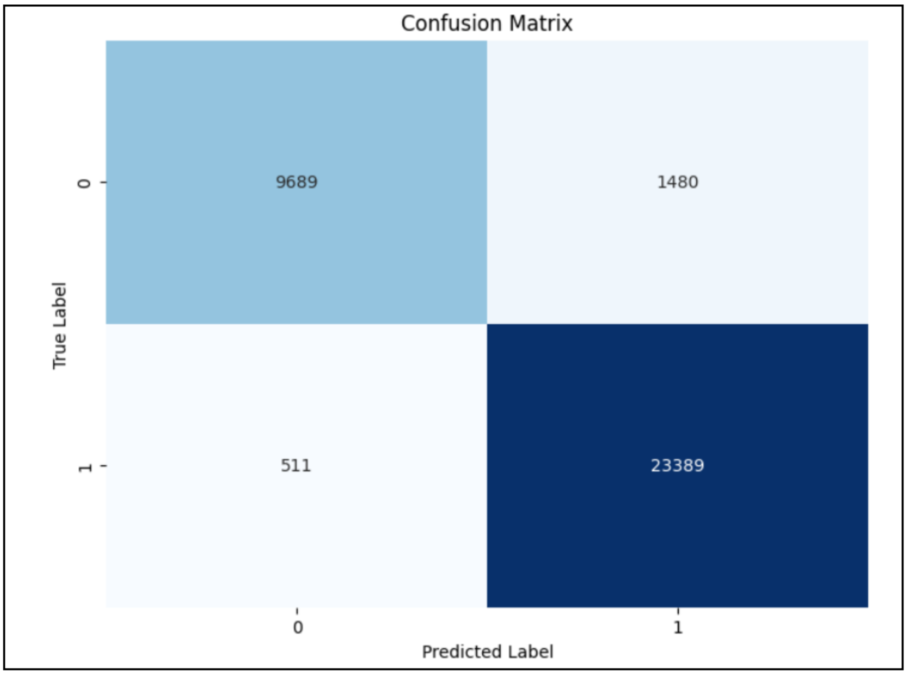

Figure 6: Confusion matrix of Long Short-Term memory

6.4 Discussion

Based on the comprehensive evaluation of performance metrics, the Gradient Boosting model emerges as the most effective for network intrusion detection. It achieved the highest accuracy of 94.57% and demonstrated a balanced performance with commendable precision, recall, and F1 scores for both normal and malicious instances. The model's ability to minimize false positives and false negatives, along with its overall strong predictive capability, makes Gradient Boosting the preferred choice among the evaluated models. The below table summarizes the evaluation of the Gradient boosting classifier algorithm against other models in anomaly detection.

Table 3: Evaluation summary of all models

Model	Accuracy	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
Gradient Boost	94.57%	95%	86%	90%	94%	98%	96%
SVM	89.53%	90%	75%	82%	89%	96%	93%
LSTM	94.32%	95%	87%	91%	94%	98%	96%

The comprehensive evaluation of the Gradient Boosting Classifier for anomaly detection after running the dataset containing various types of attacks and network logs provides valuable insights into the model's performance. This detailed discussion aims to critically analyse the findings, highlight strengths, address limitations, and suggest avenues for improvement, all within the context of existing literature.

Effective Handling of Class Imbalance: While the overall accuracy is impressive, it is crucial to acknowledge the potential impact of class imbalance, especially in scenarios where the instances of anomalies are substantially lower than normal instances. The datasets that are considered here are the real-time traffic captured and hence they have the majority of traffic as normal, if the datasets have more balanced network traffic the performance of the model can be improved. The provided dataset comprises a total of 119,341 instances, with a notable class imbalance. Class 0 (Normal) accounts for 56,000 instances, while Class 1 (Malicious) consists of 63,341 instances. The Gradient Boosting algorithm's ability to handle class imbalance contributes to its success in detecting various attacks without being skewed towards the majority class. This class distribution indicates a substantial prevalence of normal instances compared to malicious ones. The impact of class imbalance on model development and evaluation should be carefully considered, as it can influence the model's ability to generalize and detect instances of the minority class effectively.

Comparison with Existing Models: The findings should be contextualized within the broader landscape of anomaly detection models. Comparisons with other state-of-the-art models, such as Support Vector Machines (Yang, 2019) and the Long short-term memory algorithm model (Girish, 2023) provide a benchmark for the model's effectiveness. The

Gradient Boosting algorithm emerged as a robust and versatile choice, demonstrating high accuracy compared to other models and balanced performance across various attack types in the context of network intrusion detection.

7 Conclusion and Future Work

The two critical aspects of cloud security are addressed in this project, the integration of diverse security features into a unified conceptual cloud framework and the application of the Gradient Boosting Classifier algorithm for anomaly detection in cloud traffic.

A conceptual security framework was successfully devised by integrating features from prominent cloud security frameworks. This framework provides a comprehensive guide for managing security postures in the cloud, ensuring a cohesive and robust approach to cloud security. The incorporation of multiple perspectives enhances the framework’s versatility and applicability across different cloud platforms. The second research question focused on the efficacy of the Gradient Boosting Classifier algorithm in enhancing anomaly detection in cloud traffic. Rigorous preprocessing, feature selection, and model training were performed, resulting in a model capable of discerning deviations from established norms. Evaluation metrics such as accuracy, recall, F1-score, and precision demonstrated highly effective performance in capturing and identifying anomalies within cloud traffic.

In the future, the conceptual security framework devised for cloud security posture management can be extended and implemented on major cloud platforms such as AWS, Microsoft Azure and GCP to ascertain its adaptability and effectiveness in diverse environments. Further exploration of alternative anomaly detection algorithms presents an avenue for assessing their performance across various datasets, allowing for a comparative analysis with the Gradient Boosting Classifier utilized in the initial study. Additionally, there is an opportunity to investigate the real-time implementation of the anomaly detection system, exploring possibilities for seamless integration with existing cloud security architectures to facilitate proactive threat identification. In the future, we can intend to increase the performance of the model through multiclass classification.

References

Coppola, G., Varde, A.S. and Shang, J. (2023) ‘Enhancing cloud security posture for ubiquitous data access with a cybersecurity framework based management tool’, 2023 IEEE 14th Annual Ubiquitous Computing, Electronics &amp;amp; Mobile Communication Conference (UEMCON) [Preprint]. doi:10.1109/uemcon59035.2023.10316003.

Diogenes, Y., 2019. The Quest for Visibility and Control in the Cloud. ISSA Journal, 17(3).

Rastogi, S., 2021. Cloud Computing Simplified: Explore Application of Cloud, Cloud Deployment Models, Service Models and Mobile Cloud Computing (English Edition). BPB Publications.

Sailakshmi, Vyshnavi, "Analysis of Cloud Security Controls in AWS, Azure, and Google Cloud" (2021). Culminating Projects in Information Assurance. 112. https://repository.stcloudstate.edu/msia_etds/112

B. R. Kandukuri, R. P. V. and A. Rakshit, "Cloud Security Issues," 2019 IEEE International Conference on Services Computing, Bangalore, India, 2019, pp. 517-520, doi: 10.1109/SCC.2009.84.

Singh, A., & Chatterjee, K. (2017). Cloud security issues and challenges: A survey. Journal of Network and Computer Applications, 79, 88-115. https://doi.org/10.1016/j.jnca.2016.11.027

Yang, C. Anomaly network traffic detection algorithm based on information entropy measurement under the cloud computing environment. Cluster Comput 22 (Suppl 4), 8309–8317 (2019). https://doi.org/10.1007/s10586-018-1755-5

Alshammari, A., Aldribi, A. Apply machine learning techniques to detect malicious network traffic in cloud computing. J Big Data 8, 90 (2021). https://doi.org/10.1186/s40537-021-00475-1

Girish, L., Rao, S.K.N. Anomaly detection in cloud environment using artificial intelligence techniques. Computing 105, 675–688 (2023). https://doi.org/10.1007/s00607-021-00941-x

Chauhan, M. and Shiaeles, S. (2023) ‘An analysis of cloud security frameworks, problems and proposed solutions’, Network, 3(3), pp. 422–450. doi:10.3390/network3030018.

C. Di Giulio, R. Sprabery, C. Kamhoua, K. Kwiat, R. H. Campbell and M. N. Bashir, "Cloud Standards in Comparison: Are New Security Frameworks Improving Cloud Security?," 2017 IEEE 10th International Conference on Cloud Computing (CLOUD), Honololu, HI, USA, 2017, pp. 50-57, doi: 10.1109/CLOUD.2017.16.

Haber, M.J., Chappell, B. and Hills, C., 2022. Mitigation Strategies. In Cloud Attack Vectors: Building Effective Cyber-Defense Strategies to Protect Cloud Resources (pp. 221-296). Berkeley, CA: Apress.

C. Tunc et al., "Cloud Security Automation Framework," 2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W), Tucson, AZ, USA, 2017, pp. 307-312, doi: 10.1109/FAS-W.2017.164.

Rupra, S.S. and Omamo, A. (2020) A Cloud Computing Security Assessment Framework for Small and Medium Enterprises. Journal of Information Security, 11, 201-224.

J. Surbiryala, C. Li and C. Rong, "A framework for improving security in cloud computing," 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 2017, pp. 260-264, doi: 10.1109/ICCCBDA.2017.7951921.

Kumar, R. and Goyal, R. (2019) ‘On cloud security requirements, threats, vulnerabilities and countermeasures: A survey’, Computer Science Review, 33, pp. 1–48. doi:10.1016/j.cosrev.2019.05.002.

G. SAWHNEY, G. KAUR and R. Deorari, "CSPM: A secure Cloud Computing Performance Management Model," 2022 International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, 2022, pp. 1-5, doi: 10.1109/ICCR56254.2022.9995865.

Luigi Coppolino, Salvatore D’Antonio, Giovanni Mazzeo, Luigi Romano, "Cloud security: Emerging threats and current solutions", Computers & Electrical Engineering, Volume 59, 2017

Sari, A., 2015. A review of anomaly detection systems in cloud networks and survey of cloud security measures in cloud storage applications. Journal of Information Security, 6(02), p.142.

Paulikas, G., Sandonavičius, D., Stasiukaitis, E., Vilutis, G. and Vaitkunas, M., 2022, October. Survey of Cloud Traffic Anomaly Detection Algorithms. In International Conference on Information and Software Technologies (pp. 19-32). Cham: Springer International Publishing.

Zhao, C., Qian, Q., Xue, J., Pu, J., Yu, G. and Zhang, N., 2023, April. Design and Implementation of Enterprise Sensitive Information Monitoring System Based on RPA. In 2023 5th International Conference on Communications, Information System and Computer Engineering (CISCE) (pp. 263-267). IEEE.

Axmann, B. and Harmoko, H., 2022. Process & Software Selection for Robotic Process Automation (RPA). Tehnički glasnik, 16(3), pp.412-419.

Martínez-Rojas, A., Sánchez-Oliva, J., López-Carnicer, J.M. and Jiménez-Ramírez, A., 2021, August. Airpa: An architecture to support the execution and maintenance of AI-powered RPA robots. In International Conference on Business Process Management (pp. 38-48). Cham: Springer International Publishing.

M. K. Islam, P. Hridi, M. S. Hossain and H. S. Narman, "Network Anomaly Detection Using LightGBM: A Gradient Boosting Classifier," 2020 30th International Telecommunication Networks and Applications Conference (ITNAC), Melbourne, VIC, Australia, 2020, pp. 1-7, doi: 10.1109/ITNAC50341.2020.9315049.

Hagemann, T. and Katsarou, K., 2020, December. A systematic review on anomaly detection for cloud computing environments. In Proceedings of the 2020 3rd Artificial Intelligence and Cloud Computing Conference (pp. 83-96).