National College of Ireland

Anomaly Detection in Cloud System using Novel Aspect of SMOTE Sampling and Machine Learning Classifiers

MSc Research Project
MSc in Cloud Computing

Gauri Misra
Student ID: x20259611

School of Computing
National College of Ireland

Supervisor: Rashid Mijumbi

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Gauri Misra
Student ID:	x20259611
Programme:	MSc in Cloud Computing
Year:	2023
Module:	MSc Research Project
Supervisor:	Rashid Mijumbi
Submission Due Date:	14/12/2023
Project Title:	Anomaly Detection in Cloud System using Novel Aspect of SMOTE Sampling and Machine Learning Classifiers
Word Count:	7563
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:	Gauri Misra
Date:	14th December 2023

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):	

Anomaly Detection in Cloud System using Novel Aspect of SMOTE Sampling and Machine Learning Classifiers

Gauri Misra
x20259611

Abstract

The main problem of class imbalance in machine learning (ML) models is addressed in this study, which presents a novel approach to improve anomaly detection in cloud systems. To increase the detection accuracy of uncommon anomalies which are usually underrepresented in cloud datasets—the main contribution is the combination of powerful machine learning classifiers with the Synthetic Minority Oversampling Technique (SMOTE). Various models, including Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), Gradient Boosting Machine (GBM) and Random Forest (RF) were thoroughly evaluated as part of the research. Findings demonstrated that Deep Neural Network and Gradient Boosting Machines were capable to identify outliers with pinpoint accuracy. In terms of cloud security, the significance of this research is the comprehensive methodology of employing the SMOTE sampling techniques. The proposed research assesses the various machine learning models that will detect anomalies in cloud computing with higher accuracy. New avenues for research include developing and testing data balancing algorithms with more sophisticated features, and study hybrid models that combine the best features of different approaches. The essential quality in the ever-evolving world of cyber threats, these endeavours may produce innovative and highly adaptable security solutions for cloud computing. The four classification models that have been created for the detection of the anomaly in cloud environment, the Random Forest and the Deep Neural Network models achieved reliable accuracy of 0.98 and 0.99 respectively. However, the RNN classifier model achieved a poor accuracy of prediction which is 0.14. An excellent result of 100% accuracy is achieved by the Gradient Boosting Model (GBM). All the performance parameter values of GBM Model are equal to 1.

Keywords: Machine Learning, Anomaly Detection, Deep Learning, Recurrent Neural Networks, Deep Neural Networks, SMOTE, Random Forests, Gradient Boosting Machines.

1 Introduction

Access and management of digital resources has been completely transformed by cloud computing. Its efficiency, scalability, and adaptability have made it a popular technology, opening up new possibilities for data processing, storage, and delivery. Cloud computing is becoming increasingly vulnerable to security breaches as it grows in importance within

IT systems. Due to the sensitive nature of the data stored in the cloud, its security is of the utmost importance. Unauthorized access, data breaches, and different types of cyber assault are still potential risks to cloud systems, even though security mechanisms have improved Garg et al. (2019). More robust and adaptable security solutions are needed because traditional security measures are ineffective when it comes to cloud computing due to their dynamic nature and complicated infrastructure.

A. Problem Statement: Identifying outliers, which can be signs of security breaches or other systematic issues, is a major obstacle in cloud security. Unusual access patterns or systems behavior are two examples of anomalies in cloud systems that could indicate either operation or security risks. Despite their success in exporting known dangers, existing cloud anomaly detection technology suffers from some limitations:

- **Class Imbalance:** The problem of class imbalance arises in many anomaly detection machine learning models because the number of normal examples greatly exceeds the number of abnormal ones. Because of this disparity, models may be skewed towards normalcy in their predictions, leading to an excessive number of false negatives when looking for outliers.
- **Zero-Day Attacks:** There is still a long way to go before one can reliably identify zero-day attacks. Since they take advantage of undiscovered vulnerabilities, these attacks pose a significant threat and are difficult to detect using conventional methods that depend on known attack signatures.

B. Research Aims and Objectives: This research aims to enhance anomaly detection in cloud systems by integrating advanced machine learning techniques. The primary objectives include:

- Developing Advanced Machine Learning Models: Exploring and integrating state-of-the-art class balancing methods for Random Forests (RF) and Gradient Boosting Machines (GBM) to address the issue of class imbalance in anomaly detection.
- Leveraging Deep Learning Techniques: Investigating the potential of Deep Neural Networks (DNNs) and Recurrent Neural Networks (RNNs) to improve the identification and prediction of anomalies, particularly focusing on their ability to detect zero-day threats.
- Enhancing Cloud Security: Providing actionable insights and recommendations to cloud service providers and users, aiming to strengthen the security and resilience of cloud infrastructures against emerging vulnerabilities.

Research Question: When compared to a novel class balancing approach developed for Gradient Boosting Machines (GBM) and Random Forest (RF), how effective are Recurrent Neural Networks (RNNs) and Deep Neural Networks (DNNs) in detecting anomalies, specifically zero-day threats in high-dimensional, active cloud data?

C. Significance of the Research: This research is significant as it contributes to the evolving field of cloud computing security. By addressing key challenges in anomaly detection, the study aims to:

- Improve the efficiency and accuracy of anomaly detection in cloud environments.

- Offer new insights into the application of advanced machine learning and deep learning techniques in the context of cloud security.
- Provide valuable recommendations to enhance the security protocols of cloud systems, benefiting a wide range of stakeholders including cloud service providers, businesses, and individual users.

D. Structure of the Report: The first section of this report is Introduction which emphasis on the problem it aims to address and why this study is important and provides a reason for conducting the research. The second section is Related Work which looks into the existing studies to find gaps in the research and summarizes whatever has been done before in the field of anomaly detection in cloud systems using machine learning approaches. The third section is Research Methodology which mentions the procedures and approaches that will be followed to achieve the aim of the research and objective. Then the fourth section is the Design Specification, the methods and structure supporting the implementation and the related specifications are covered in this section. The fifth section is Implementation which investigates the proposed solution's implementation. Then the sixth section is Evaluation, where the detailed assessment of results have been presented and also the major findings of this research are mentioned here. The last section is Conclusion which concludes the complete study and also describes the future work.

2 Related Work

2.1 Anomaly Detection in a Large-Scale Cloud Platform

Anomaly detection is a key part of making the cloud more reliable because it watches how things work and finds possible systems problems. Girish and Rao (2023) came up with a new way to find problems in cloud systems that use both supervised and unsupervised machine learning algorithms. They used real-world data searches to show that their method worked. In the same way, Islam et al. (2021) stressed how important it is to watch quality service in a cloud platform like the IBM cloud platform. They created an automated system using deep learning networks this system is very good at finding problems in real time and improving working efficiency by cutting down on the amount of work that needs to be done by hand.

He et al. (2023) looked at CloudShield, also made to find complex threats in cloud computing settings, such as speculative execution attacks like specter and meltdown. This study shows how useful AI-driven systems like these are. To keep systems from breaking down and reduce false alarm fatigue, dear work stresses how important is to tell the difference between harmless and dangerous animals.

Another example is Wang et al. (2021), who created a self-evolving lonely detection tool to help cloud computing systems be safe by finding problems on their own. Their self-evolving method changes based on how the system works, sewing promise for making cloud systems more reliable. Muneer et al. (2023), did more studies on finding cyber security events in cloud environments using a machine learning best method that combines supervised and supervised techniques. According to the author, their review of a real-world datasheet showed that the method could quickly find threats, which improved the security of cloud environments.

Using such advanced machine learning methods not only Max's operation run more smoothly but also makes customers much happier by finding and stopping cloud outages

before they happen. The work of Islam et al. (2021), gives a thorough look at the structure and use of AI in cloud monitoring. this is a big step forward for cloud computing and a plan for how AI-based detection systems can be added to large-scale cloud infrastructures in the future.

2.2 Application of Machine Learning in Detecting Malicious Network Traffic

When it comes to improving network security in a cloud environment, machine learning is becoming dispensable for detecting hostile actions. To strengthen the security of cloud-based networks, Alshammari and Aldribi (2021) used a variety of machine learning models, such as support vector machine, DT, and neural network, applied to a heterogeneous dataset. Both the decision tree and rend of forest model achieved an accuracy of 100% which is due to the reason of Novel feature selection. Also, the implemented neural network model has an accuracy of 90%, and with cross-validation, the accuracy is also achieved at 96% in a split validation. It is also found that the Naive Bayes model is not the best on the support vector machine model and achieved an accuracy of 81% during aspect validation. Machine learning and Big data analytics can strengthen cloud security systems in the face of ever-changing threats, according to the research of Mohammad and Pradhan (2021). They propose a method that strategically you just advanced analytics and machine learning models to improve cloud security.

Rana et al. (2022) provide a comprehensive analysis of cloud intrusion detection systems, demonstrating how machine learning is becoming increasingly important for cloud security. The study provides useful insights into the present state of cloud security by reviewing different intrusion detection systems that are based on machine learning. These machine learning designs are effective however there are still problems like diverse data sets and no real-time data processing. To extend the applicability of the models to different cloud environments and dynamic network conditions, and to provide strong, real-time cloud security solutions, continued research is necessary to address these constraints.

2.3 Enhancing Anomaly Detection in Cloud Systems

Lin et al. (2019) study on cyber security looks at how machine learning and deep learning can be used to fight cyber security risks that are already at an edge. The main focus of their study was on using LSTM networks with long-term and short-term memory along with an attention method to sort network traps. It was able to reach an amazing 96.2% accuracy by using the SMOTE and an improved loss function to deal with the class mismatch problems that come up a lot in anomaly detection. 96% of the time, both the Precision readings and the call, were right. The recall rates were significantly improved after using a SMOTE and the upgraded loss function. For online assault samples, the recall rate increased from 0% to 98%, and for infiltration samples, it showed modest gains from 11% to 17%, despite the initial low precision. Accuracy and recall rates of over 93% were also achieved using more conventional machine learning methods including decision trees, KNN, and RF classifiers. SVM and Gaussian Naive Bayes models are performing worse. With bigger data sets in particular, the research demonstrated that deep learning outperforms traditional learning in terms of training time.

A state-of-the-art multimodal deep learning method for anomaly detection utilizing dispersed tracing data and LSTM networks was introduced by Nedelkoski et al. (2019).

By using this method, it became much easier to find problems with the way system parts were being run. In 2020, Rabbani et al. proved that advanced methods work well in cloud-based IT operations. They also underlined the significance of deep learning in identifying issues with cloud infrastructures. According to considerable research, LSTM networks appear to be a promising method for finding small anomalies in high-dimensional datasets. Until these ideas are put into practice and put to the test in real-world situations, their effectiveness cannot be clearly assessed. Future research is required to ascertain the suitability of such strategies for use in other computer context.

2.4 Cloud Computing Security Optimization through Algorithm Implementation

According to Dhabliya (2021), One of the top priorities is to develop strategies that can improve the security of cloud. Possible security approaches included intrusion detection systems. Encryption and access control systems. The requirement of using access control system to restrict the privileges of user and reduce the risk of data breaches was highlighted in the study result. The privacy of the data and incomprehensibility should be assured via security measures. The result of the study indicates that intrusion monitoring system are critical for determining and removing the risk. The study yielded significant finding but it did not provide enough information about how actual cloud systems works. Testing the algorithms with a simulated settings can be helpful to make the complexity of cloud architecture seem more than it really is. Future studies can be focused on these areas to strengthen the security and make them more reliable.

2.5 Class Imbalance Issues in Anomaly Detection for Cloud Systems

The imbalance of the class makes it difficult to determine the anomalous occurrences in cloud system Ochani et al. (2019). A well-balanced distribution of the classes can be achieved by a creative strategy that consists of over sampling and under sampling tactics. The research looked at deep learning techniques for handling large amount of data that uniquely distributes across different classes. Because of the major issue with the previous study, they will be better prepared to plan in next research. To further investigate how deep learning could help level the playing field in anomaly detection Johnson and Khoshgoftaar (2019), conducted additional research. According to the authors, Researchers and practitioners working with imbalanced data sets in cloud systems might greatly benefit from deep learning ability to reveal intricate patterns in data. Rogić and Kašćelan (2021), broadened the discussion to include the classification of customer segments using & learning and support vector rule extraction. Even though their research goes beyond the usual scope of anomaly detection, it does show how successful methods like & learning can be in tagline class in balance problems; these findings are relevant to cloud anomaly detection as well. Understanding class in balance and security challenges is crucial in the domain of mobile cloud computing, according to Saran et al. (2022). Since mobile cloud systems have their own set of security challenges, their research shows that tailored anomaly detection methods are essential in these environments. In a way that sidesteps the problem of class in balance, Aldallal and Alisa (2021), addressed the importance of intrusion detection systems in protecting data stored in the cloud. The need for precise and only detection in an efficient intrusion detection system was highlighted by

their demonstration of machine learning practical uses to improve cloud platform security protocol.

2.6 Systematic Review of Anomaly Detection in Cloud Computing

In their comprehensive systematic review of anomaly detection in cloud computing, Hagemann and Katsarou (2020) examined 215 papers. Machine learning, deep learning, and statistical methods were all covered in this review, which focused on how they were used in different cloud-based network situations. The discussion on anomaly detection techniques in the article explored a wide range of subjects including failure detection performance monitoring and intrusion detection. A thorough explanation of evolutionary research is focused on the significance of confirming approaches with the help of existing datasets Researchers will find this study useful as it summarizes the existing situation of a nominal detection techniques while showing how things have progressed recently. Cloud system resilience and security must be bribery goals of this area of study in future with an emphasis on constant improvement and new developments.

2.7 Deep Learning Techniques for Anomaly Detection in Cloud Systems

Considered as a reliable method for the detection of outliers within the cloud environment, Deep learning algorithms are gaining popularity among people for the ability to uncover complex trends and patterns in the information.Jauro et al. (2020) Analyze the impact of deep learning models on cloud computing and also provided guidance to experts about how to improve the workflow. Based on the analysis of the sequential data in cloud, Chkirbene et al. (2020) concluded That deep learning and LSTM is a powerful method for debunking urban legends.

Dhabliya (2021) investigated that how the machine learning algorithms can significantly improve the security of cloud computing. The results of the analysis highlighted the critical requirement of bolstering the computer system security with the help of State-of-the-art methods such as deep running. The results also proved that the mathematical deep learning model has the ability to effectively improve the security.

Muhamad et al. (2023) examined and compared few procedures of class balancing. Some deep learning algorithms such as LSTM might be helpful for cloud systems while focusing on categorization instead of anomaly recognition when addressing the issues. It becomes simpler to examine issues over time when LSTM can organize things in a progressive manner. SMOTE is proved to be a good method for a more legitimate sorting and identification of outliers. Samriya and Kumar (2020) showed a new attack detection system for cloud computing that uses a mixed clustering optimization method. Deep learning could be used in these devices. LSTM with an attention method can help the system find more complicated intrusion patterns.

2.8 Novelty and Motivation of the Work

This study offers a new way to find problems in cloud systems by testing how well Recurrent Neural Networks (RNNs) and Deep Neural Networks (DNNs) can find oddities, especially zero-day threats, in high-dimensional, dynamic cloud data. This study also

Model/Algorithm	Study (Author, Year)	Performance Metrics
ANN	Alshammari & Aldribi (2021)	94% accuracy in cross-validation, 96% in split-validation (90% training data)
KNN	Alshammari & Aldribi (2021)	High accuracy, some classification errors in split-validation
SVM	Alshammari & Aldribi (2021)	81% accuracy in split-validation
Deep Learning Neural Networks	Islam et al.	Outperforms traditional methods by identifying issues up to 20 minutes earlier
Naïve Bayes	Alshammari & Aldribi (2021)	Unreliable results in cross-validation
LSTM with AM	Lin et al. (2019)	Classification accuracy of 96.2%, Precision and Recall rates around 96%
GaussianNB, SVM	Lin et al. (2019)	Classification results that are worse than those obtained with deep learning methods
Multimodal Deep Learning Solution	Nedelkoski et al. (2019)	Enhanced anomaly identification capabilities (specific metrics not provided)
Advanced Techniques	Rabbani et al. (2020)	Validation of effectiveness in cloud-based IT operations (specific metrics not provided)
Cryptographic Algorithms	Dhabliya (2021)	Protecting the security and privacy of data effectively (specific measures not given)
Intrusion Detection Systems	Dhabliya (2021)	important for identifying threats in real time (particular metrics not supplied)

Figure 1: Summary of previous work

looks at class division, which most of the reports have not consider. Based on the machine learning techniques RF and GBM, a new way of class balance is called for. This project aims to use machine learning and deep learning to make cloud systems better at finding strange behavior. A model or method to recognize these problems and avert their detrimental impacts beforehand is required due to the increase in fraud, security problems, and other technical roadblocks. The study looks at deep learning techniques like RNNs and DNNs to build new anomaly detection systems and also tackles the issue of class imbalance. This paper fills a gap in the research by exploring new ideas, noting the flaws in conventional methods, and adding to earlier work on finding different things in cloud systems.

3 Methodology

This section will be demonstrating the methodology adopted for detecting the anomalies in cloud data. The methodology focuses on utilizing advanced machine learning algorithms and presents a new approach for detecting anomalies in cloud data. This approach, called SMOTE (Synthetic Minority Over-sampling Technique) sampling, is employed. This novel methodology will not only enhance the detection of anomalies but it also plays a significant role in enhancing the security and dependability of cloud systems.

3.1 Phase 1: Data Preparation

For this research the KDD99 dataset is used for this project. Numerous network links are required to find security flaws in the dataset, which frequently serves as a benchmark for the field of cyber security. Eliminating any questions, filling in any gaps, and improving the accuracy of the information were the first steps in handling the data. This stage was made in an effort to establish a direct correlation between the system's capacity for problem detection and the accuracy of the provided data.

3.2 Phase 2: Addressing Class Inequality

This complete part is about coming up with a new way to balance classes because of the problem of class imbalance, which happens when abnormalities are less common than normal ones Peterson et al. (2020). Random forest as well as gradient-boosting machines excel in categorising data into distinct categories, making them ideal for this task.

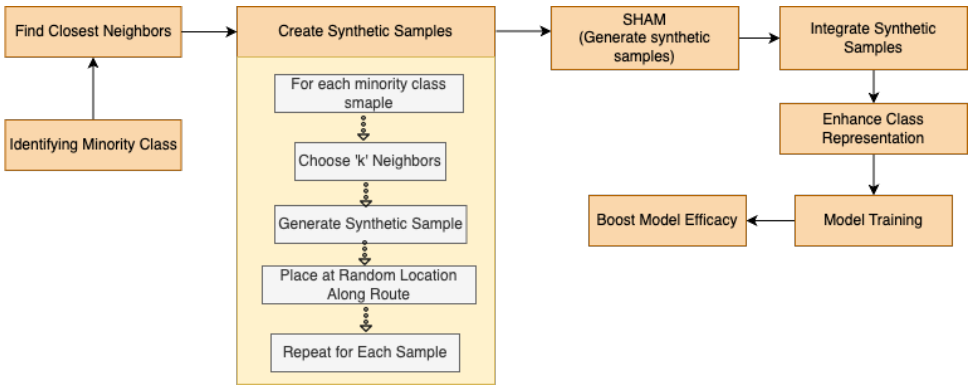
Class Balancing Techniques:

- **Weight Modification:** The model's weight was adjusted throughout the training process to better indicate the minority class. With this approach, errors with the model are much more likely to occur.
- **SMOTE Implementation:** To rectify the data-set imbalance issue, synthetic samples were generated for the minority class using SMOTE.

Novel approach - SMOTE Sampling, A proposed solution to address the problem of class imbalance involves the application of Synthetic Minority Over-sampling Technique (SMOTE) sampling within the context of cloud system data. The Synthetic Minority Over-Sampling Technique (SMOTE) is a novel method used for oversampling in which synthetic samples are created for the minority class, specifically the anomalies, within the feature space. The creation of a balanced dataset through the use of SMOTE improves the learning capabilities of machine learning classifiers. This allows for effective learning from both normal and anomalous instances, resulting in enhanced detection performance.

3.2.1 Steps in SMOTE Sampling:

Finding the minority group in the data is the first phase. The identification is vital since a SMOTE is striving to enhance the representation of the collection of the minority community. Finding the closest neighbours inside the minority class is the further step after identifying the class itself. At least five individuals who live in the same neighbourhood are chosen as k. The process for creating synthetic samples is established in this crucial stage, which lays the framework for the following steps.



The SMOTE procedure relies heavily on the creation of SHAM samples. For each minority class sample, one of the closest 'k' neighbours of the method is chosen. Using the minority class sample to locate its nearest neighbours in the future space allow one to create a synthetic sample. The phoney sample is placed at a random location along this route. The resulting synthetic samples will be more representative of the minority

class's feature space and distributed more uniformly if one applies this technique. The final step is to include the created examples into the initial dataset. To make the dataset more suitable for training models in which class balance is crucial, it is necessary to increase the representation of the minority class. The distribution of classes is now more evenly distributed thanks to this modification. Using this approach, SMOTE ensures that underrepresented groups are adequately considered and included in the training of prediction models. When given unfair datasets, this makes the models even more useful.

3.2.2 Mathematical Explanation

Algorithm: Enhanced Anomaly Detection using SMOTE and ML Classifiers

Input: Dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ where x_i are features and y_i are labels; $y_i \in \{0,1, \ldots, C-1\}$ for a C-class classification problem.

Output: Optimized ML models for anomaly detection.

Steps:

1 Data Pre-processing:

- Normalize numerical features: $x_{\text{norm}} = \frac{x-\mu}{\sigma}$
- Encode categorical features: $x_{\text{encoded}} = \text{OneHot}(x_{\text{cat}})$

2 Class Balancing using SMOTE:

- For each minority class sample x, generate synthetic samples by the following process:
- Find k nearest neighbors for x in the minority class.
- For each neighbor x_{nn}, create a synthetic sample x_{new} as:

$$x_{\text{new}} = x + \lambda \cdot (x_{\text{nn}} - x)$$

- Where λ is a random number between 0 and 1

3 Model Training:

- For each model M in {DNN, RNN, RandomForest, GBM}:
- Train M on balanced dataset D_{SMOTE}.
- Evaluate M using metrics like accuracy, F1-score, ROC-AUC.

4 Model Evaluation and Selection:

- Select the model M^* with the best performance based on evaluation metrics.
  Return: Optimized model M^* for anomaly detection.

Return: Optimized model M^* for anomaly detection.

3.2.3 SMOTE Implementation Proposed in the Present Case:

In a cloud computer setting, SMOTE is used after the first steps of preparing the data. To avoid problems with class imbalance, SMOTE tries to make fake samples that correctly represent the minority class in the dataset. This is very important because unfair class representation in areas like cyber security can lead to biased models. When the model is taught using SMOTE on a data set that more closely matches how classes are spread out in the real world, the results are more accurate and reliable. As you can see, this

way shows how a cloud-based model can find trends and handle huge amounts of data like KDD99"

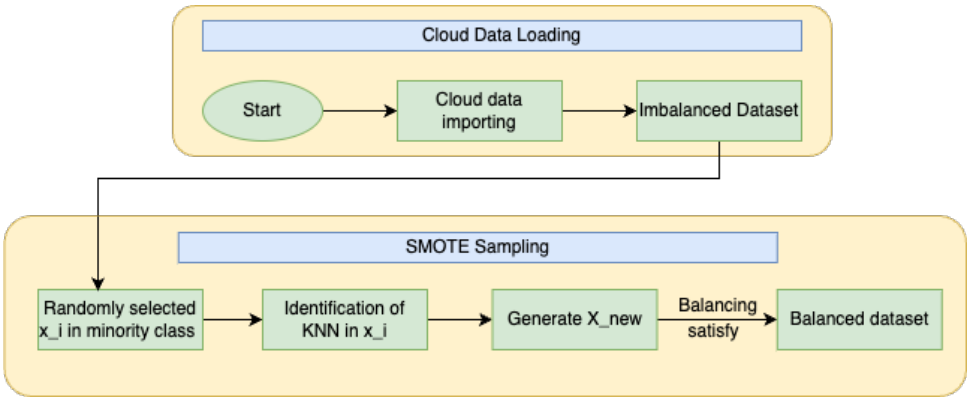


Figure 2: SMOTE Implementation

3.3 Phase 3: Deep Learning Integration

At this point, Deep Neural Networks (DNNs) and Recurrent Neural Networks (RNNs) were added to the structure for finding anomalies.

Deep Learning Models:

More accurate finding of anomalies required training and using Recurrent Neural Networks (RNNs) and Deep Neural Networks (DNNs). For even the smallest mistakes to be found, these models were carefully made. Their ability to learn and spot trends naturally made the detection system much more accurate and reliable. They were able to gracefully change and respond to constantly changing data trends because they were so good at learning.

The deep learning models performed great not only in finding mistakes in the data but also spotting complex attack trends and **zero-day flaws**. No previous data is available to find zero-day bugs, which are security holes that no one has ever seen before due to their capability to spot trends, DNNs and RNs performed exceptionally well on this task. By searching for trends and peculiarities that typical models would overlook, they might be able to identify these flaws and provide a more thorough and trustworthy security response. This element of the deep learning model is essential for cloud security because it requires ongoing monitoring to protect systems and data from new risks.

3.4 Phase 4: Model Evaluation and Comparison

To determine what each model could accomplish beyond ensuring that the numbers were accurate, a thorough review procedure was used. below mentioned are the evaluation metrics:

- **Precision, Recall, F1 Score, and AUC-ROC Curve:** These dimensions provide a complete picture of how the models operate, particularly in terms of their ability to distinguish between good and bad actions.
- **Comparison of Machine Learning Models:** A comparative analysis between traditional machine learning models and advanced DNNs and RNNs was performed.

This comparison highlighted the strengths and potential limitations of each model, offering an evidence-based assessment of their real-world applicability.

3.5 Phase 5: Practical Application and Recommendations

The research extended into the realm of practical utility, applying the findings to actual cloud environments.

Application in Cloud Settings

- **Integration into Cloud Security Infrastructures:** The research examined how the proposed anomaly detection systems could be incorporated into existing cloud security frameworks.
- **Customized Recommendations:** Tailored suggestions were compiled for cloud service providers, focusing on enhancing their security measures based on the variety of cloud services and their unique security challenges.

The study culminated in a significant contribution to the knowledge in cloud security, emphasizing the role of machine learning and deep learning in transforming anomaly detection.

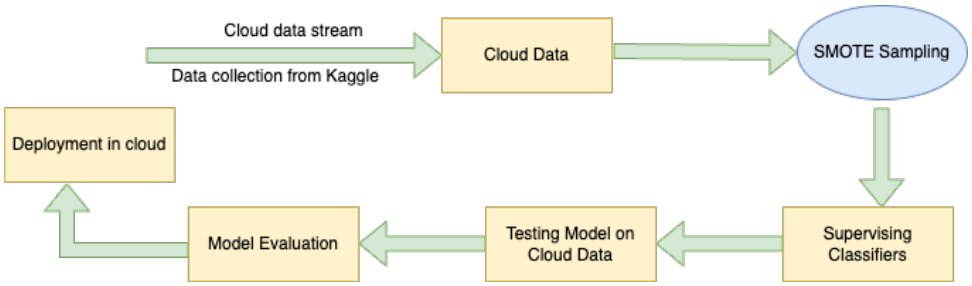


Figure 3: Framework for Research Methodology

4 Design Specification

An intricate architecture for the detection of anomaly in cloud environment including advanced machine learning algorithms and deep learning is the central focus of this study. The functional descriptions, underlying architecture and prerequisites of the suggested model are detailed in this design specification chapter.

4.1 Underlying Architecture

The preparation of data, Balancing class training the model and assessing the model are all the stages that are included in the multi-tier architecture approach. The preparation using the KDD99 dataset is used in first layer. The cleaning, normalization and conversion of categorical variables to numerical one is included in this stage.

The process of detecting anomaly with an imbalanced class is a challenge that is addressed in the second tier. For creating a balanced training data set it uses a class balancing method that is suitable for GBM and RF. It also makes use of SMOTE that is Synthetic Minority Over-sampling technique.

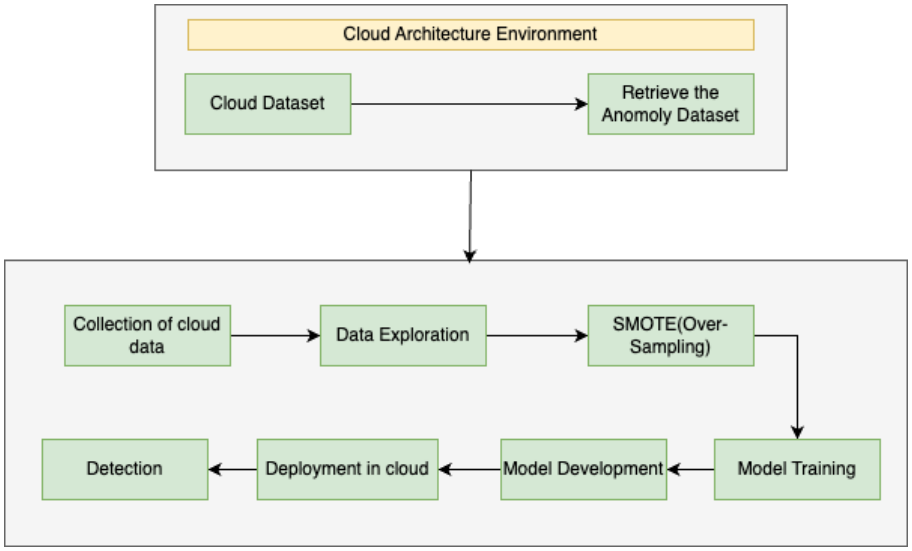


Figure 4: Architecture of the Proposed System

4.2 Advanced Machine Learning Integration

A combination of RNN and DNN is used in the subsequent tier. High dimensional and complex patterns of the data are common in the cloud computing environment. These models are suitable in decoding such patterns. RNNs are great at detecting patterns in sequential data that makes them suitable for the real time detection of anomalies. On the other hand, DNNs are suitable for handling large amounts of data which is common in cloud.

4.3 Algorithm/Model Functionality

- Deep Neural Networks: Deep Neural Network of this architecture is trained for the identification of complex correlations and patterns in the data. These will significantly contribute to an extremely precise detection system with the help of multiple layers for extracting and analysing information at different levels of abstraction.
- Recurrent Neural Networks: The system is based on RNN that is optimized for sequential data. These are suitable for the analysis of time series data that is usually found in cloud system because of the unique design that allows them to save the information over the time.
- Random Forests and Gradient Boosting Machines: These algorithms work by merging the predictions of multiple models into a single but more accurate and reliable forecast. These are fine tuned to deal with the unbalanced data set in more effective manner for improvising the capability of anomaly detection.

5 Implementation

Multiple important procedures are included in the final stage of execution of the proposed model for enhancing the security of cloud system. The phase includes model development common data transformation and using multiple tools and programming language for

the same. Each step of this process is carefully planned to be aligned with these specific requirements of the cloud computing.

5.1 Final Model Development and Training

Advanced machine learning models are used for detecting anomalies in cloud environment. The implementation is stage is based on development and training of these models. The models used in this are Random Forests (RF), Gradient Boosting Machines (GBM), Deep Neural Networks (DNNs), and Recurrent Neural Networks (RNNs).

5.2 Data Transformation and Preparation

The transformation of raw data into a format that is understandable by the machine learning models was an important part of the implementation phase. This included following stages:

- Data Cleaning and Normalization: Extensive cleaning of the data was performed on KDD99 dataset for the elimination of inconsistencies and errors. It is also ensured that all the numerical values were of the same size to make it suitable for the machine learning model.
- Feature Engineering: Ability of the machine learning models for the identification of outliers was improved by creating new features from the existed data. This include developing the features that capture regular cloud system behavior patterns. This is done to help in recognizing changes and variations in abnormalities.
- Encoding Categorical Features: For the machine learning algorithms to work with categorical features, these features were converted into numerical features with the help of OneHotEncoding.

5.3 Model Training and Optimization

- Random Forests and Gradient Boosting Machines: The balanced data set that is created after the application of preprocessing techniques was used for training the machine learning models. The adjustment in the hyperparameters helped in increasing the accuracy of the models for detecting anomalies. The parameters included the number of trees, depth of the tree, and rate of learning. The goal was to minimize the issue of overfitting while improving the accuracy.
- Deep Neural Networks: There are multiple hidden layers that created the DNN model. Multiple layers have different purpose such as Learning multiple levels of data representation and data structure. The optimization methods (Adam and RMSprop) are used to fine tune the weight of network. These method help do successfully understand the integrate patterns available in the cloud.
- Recurrent Neural Networks: For training the RNN model using the LSTM unit, Sequential cloud data is used for capturing the temporal linkages. Fine tuning of the number of LSTM unit and learning rate is used for enhancing the pattern recognition ability of the model.

5.4 Tools and Languages Used

A wide variety of tools and programming languages used for designing the proposed solution. These are selected based on their ability to handle the large amount of data and relevance with the project title.

- Python: Python is used in this work as the primary programming language because of the comprehensive collection of libraries and frameworks that are particularly designed for the facilitation of data science and machine learning activities. Some of these libraries are tensorFlow, and Keras for deep learning and Pandas for manipulating the data.
- Machine Learning and Deep Learning Libraries
- Scikit-learn: Traditional machine learning models were implemented using this library. Tools for preparing the data, training the models, tuning the hyperparameters and evaluation of the model were all made efficient by it.
- TensorFlow and Keras: Each RNN and DNN was constructed and trained with the help of deep learning frameworks. They offered a full array of tools for training the models on big data set and also allowed for flexibility in creating unique neural network designs.
- Data Handling and Visualization Tools
- Pandas: Cleaning of the data transformation and preparation were all accomplished with the help of this package. Handling massive amount of data in datasets become an easy task with its robust structure of data and operation.
- Matplotlib and Seaborn: Visualization of the data and results using these tools was critical for the comprehension of distribution of data, feature correlations and performance indicators for the model.

5.5 Outputs Produced

Several critical results were produced during the phase of implementation:

- Transformed Dataset: A cleaned, normalized and feature engineered version of the data set was created for training the machine learning models.
- Trained Models: RF, GBM, DNNs, and RNNs, models are used and trained for the anomaly detection within the cloud environment.
- Model Evaluation Reports: The results of all the models including the recall, accuracy, F1 score and area under the curve are documented in reports. The effectiveness of anomaly detection for every model is detailed in the report.

6 Evaluation

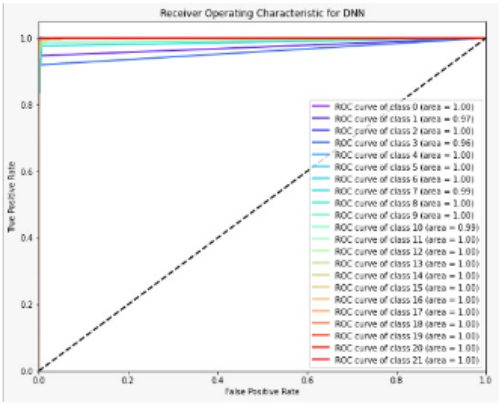
The outcomes of the anomaly detection are examined in detail in this section with an emphasis on cloud computing Xu et al. (2019). Gradient Boosting Machines (GBM), Random Forest Classifiers (RF), Recurrent Neural Networks (RNN), and Deep Neural Networks (DNN) are the machine learning models that are used in the investigation stop the accuracy of every model in terms of anomaly detection ability within the cloud environment is thoroughly examined.

6.1 Experiment 1: Deep Neural Network (DNN)

- **Performance Metrics:** The DNN model performed exceptionally well in identification issues within the cloud computing environment as it showed impressive success rate of 98.17%. Depending on the class values the model performed better or worse in terms of memory, accuracy and F1score.
- **ROC Curve Analysis:** Receiver operating characteristic (ROC) and Area under the garbage showed their exceptional abilities of problem finding. The DNN model distinguished between the abnormal and normal situations in classes with an AUC that is close to 1.00.

	precision	recall	f1-score	support
0	0.65	0.65	0.65	17
1	0.48	0.68	0.57	19
2	0.65	0.76	0.70	17
3	0.96	0.72	0.86	25
4	0.72	0.76	0.74	28
5	0.58	0.58	0.58	26
6	0.61	0.70	0.65	27
7	0.72	0.65	0.68	45
8	0.80	0.65	0.75	57
9	0.82	0.96	0.86	51
10	0.88	0.92	0.91	70
11	0.99	0.95	0.97	212
12	0.98	0.94	0.96	265
13	0.91	0.88	0.90	136
14	0.95	0.92	0.94	228
15	0.99	0.99	0.99	1184
16	0.99	0.98	0.99	742
17	0.98	0.99	0.99	892
18	1.00	1.00	1.00	6194
19	1.00	1.00	1.00	3186
20	1.00	1.00	1.00	5574
21	1.00	1.00	1.00	18772
accuracy			0.99	37792
macro avg	0.85	0.85	0.85	37792
weighted avg	0.99	0.99	0.99	37792

Performance of DNN Model in Cloud Anomaly Detection



ROC Curve for DNN Model

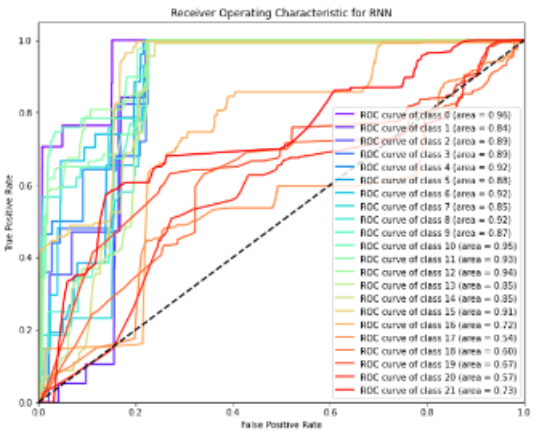
- **Implications:** The DNN has the potential to be used in real life cloud security scenarios because of the excellent AUC score and accuracy. This is particularly true in the situations when identifying cyber threats cannot be effectively done with the traditional methods.

6.2 Experiment 2: Recurrent Neural Network (RNN)

- **Performance Metrics:** When compared to RNN model that achieved an overall accuracy of 13.95% the DNN model has better performance. This means it might be hard to find time connections in the information, which is important in cloud computing settings.
- **Classification Challenges:** The sorting report showed that most of the classes had low scores for accuracy and memory, and some even had no scores at all. The RNN model might have trouble learning from the uneven information, even if SMOTE is used to make it more even.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	17
1	0.00	0.00	0.00	19
2	0.00	0.00	0.00	17
3	0.03	0.36	0.06	25
4	0.00	0.00	0.00	28
5	0.00	0.00	0.00	26
6	0.00	0.00	0.00	27
7	0.00	0.00	0.00	43
8	0.00	0.00	0.00	37
9	0.00	0.00	0.00	52
10	0.00	0.00	0.00	79
11	0.19	0.51	0.27	212
12	0.00	0.00	0.00	209
13	0.00	0.00	0.00	130
14	0.00	0.00	0.00	225
15	0.14	0.82	0.24	1184
16	0.05	0.58	0.09	742
17	0.00	0.00	0.00	892
18	0.18	0.29	0.22	6194
19	0.18	0.59	0.27	3186
20	0.00	0.00	0.00	5676
21	0.00	0.00	0.00	18772
accuracy			0.14	37792
macro avg	0.03	0.14	0.05	37792
weighted avg	0.05	0.14	0.07	37792

Performance of DNN Model in Cloud Anomaly Detection



ROC Curve for RNN Model

- **Implications:** The results show that the RNN design needs to be improved even more for cloud anomaly detection, especially when it comes to handling uneven data and getting important time features.

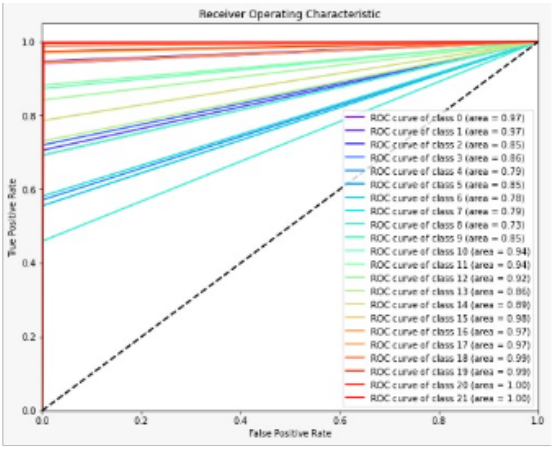
6.3 Experiment 3: Random Forest Classifier (RF)

- **Performance Metrics:** With a score of 97.84%, the RF model was pretty accurate. The classification report showed that most classes had good accuracy, memory, and F1 scores. However, some classes had lower scores, which meant they could do better.

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.94	0.97	17
1	0.72	0.95	0.82	19
2	0.92	0.71	0.80	17
3	0.82	0.72	0.77	25
4	0.67	0.57	0.62	28
5	0.60	0.69	0.64	26
6	0.47	0.56	0.51	27
7	0.62	0.58	0.60	43
8	0.71	0.46	0.56	37
9	0.64	0.69	0.67	52
10	0.80	0.87	0.84	79
11	0.90	0.88	0.89	212
12	0.88	0.84	0.86	209
13	0.75	0.73	0.74	130
14	0.80	0.79	0.79	225
15	0.97	0.97	0.97	1184
16	0.95	0.94	0.94	742
17	0.94	0.95	0.94	892
18	0.99	0.99	0.99	6194
19	0.98	0.97	0.98	3186
20	0.98	1.00	0.99	5676
21	1.00	1.00	1.00	18772
accuracy			0.98	37792
macro avg	0.82	0.81	0.81	37792
weighted avg	0.98	0.98	0.98	37792

Performance of RF Model in Cloud Anomaly Detection



ROC Curve for RF Model

- **ROC Curve Analysis:** The ROC graphs showed that the RF model could effectively tell the difference between different classes, even though performance differed between them.

- **Implications:** The fact that RF works well to find problems in cloud computing shows that it works well. The variation in the results in different classes showed that more tuning is required of the parameters particularly in the categories where the model underperformed.

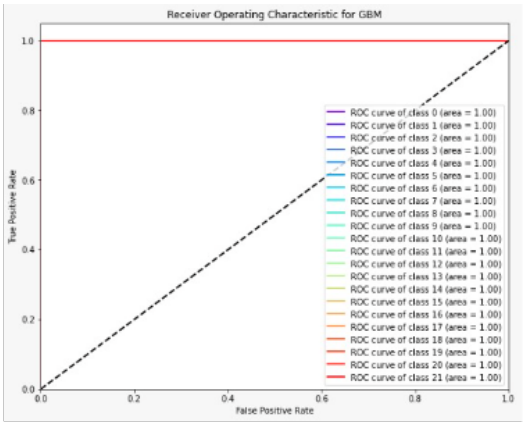
6.4 Experiment 4: Gradient Boosting Machines (GBM)

- **Performance Metrics:** GBM effectively addressed all the classes with remarkable Percent accuracy rate and other evaluation matrices.

GBM Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	17
1	1.00	1.00	1.00	19
2	1.00	1.00	1.00	17
3	1.00	1.00	1.00	25
4	1.00	1.00	1.00	28
5	1.00	1.00	1.00	26
6	1.00	1.00	1.00	27
7	1.00	1.00	1.00	43
8	1.00	1.00	1.00	37
9	1.00	1.00	1.00	52
10	1.00	1.00	1.00	79
11	1.00	1.00	1.00	212
12	1.00	1.00	1.00	209
13	1.00	1.00	1.00	130
14	1.00	1.00	1.00	225
15	1.00	1.00	1.00	1184
16	1.00	1.00	1.00	742
17	1.00	1.00	1.00	892
18	1.00	1.00	1.00	6194
19	1.00	1.00	1.00	3186
20	1.00	1.00	1.00	5676
21	1.00	1.00	1.00	18772
accuracy			1.00	37792
macro avg	1.00	1.00	1.00	37792
weighted avg	1.00	1.00	1.00	37792

Performance of GBM Model in Cloud Anomaly Detection



ROC Curve for GBM Model

- **ROC Curve Analysis:** With an AUC score of 1, all classes performed well in the roc analysis. This indicated that these can effectively distinguish between the typical and out of the order cloud behavior.

- **Implications:** Determining the issues within the cloud environment can never be easier Than using GBM as shown by this research. Additional testing can be performed on a wider range of data for determining the suitability and reliability of this strategy. The 100 percent accuracy of the model raises the concern about overfitting.

Model	Accuracy	Precision, Recall, F1-Score	ROC Curve Analysis	Implications and Challenges
Deep Neural Network (DNN)	98.17%	Varied across different classes	High sensitivity and specificity (AUC close to 1.00)	Applicability in real-world scenarios; effective in certain categories
Recurrent Neural Network (RNN)	13.95%	Poor precision and recall for most classes	-	Needs optimization for cloud anomaly detection; challenges with imbalanced data
Random Forest Classifier (RF)	97.84%	Good across most classes, some lower scores	The effective distinction between classes; varied performance	Reliable in anomaly detection; requires tuning in underperforming classes
Gradient Boosting Machines (GBM)	100%	Perfect across all classes	AUC of 1.00 for all classes	Highly suitable for anomaly detection; potential overfitting concerns

6.5 Overall Implications and Future Directions

- Academic Perspective: This research significantly contributes to the existing research by demonstrating the ability of using machine learning methods for the detection of anomalies within the cloud environment. This also demonstrates that

not all the machine learning models can work effectively in all scenarios and it is important to consider the disadvantages and advantages of each model as per the scenario.

- Practitioner Perspective: The choice of suitable machine learning models for the cloud security can be improved with the help of this study. GBM and RF have a lot of promise, but it's important to be aware of the risk of overfitting and the need to tune carefully.

- Future Research: Future studies should focus on optimizing RNN models for cloud data, exploring hybrid models that combine the strengths of different algorithms, and testing the models on more diverse and complex datasets to ensure the generalizability of the findings.

6.6 Discussion

The result obtained by experimenting on open-source cloud-based data to investigate anomaly detection in cloud computing shows some substantial results that also address the value obtained for different parameters of machine learning and deep learning models. A notable performance by Deep neural networks and gradient-boosting machine indicates their ability to detect anomalies efficiently. But the GBM model's almost flawless accuracy raises the question of whether it will be overfitted, limiting its usefulness (Dattakavi, 2022). The recurrent neural network model is performing not so good with an imbalanced data set. Classifier trends using random forest demonstrated potential, but their efficacy differed, suggesting that model adjustment was necessary. Constantly improving models is a reasonable approach, as previous researches have demonstrated. Hybrid models and sophisticated data harmonization strategies should be investigated in future studies . Validating the applicability of these models in the expanding field of cloud security requires testing with varied data sets.

With remarkably high accuracy rates of 97.84% for the random forest model and also perfect accuracy for the GBM model, the work represents a major step forward in cloud security by successfully tackling class imbalance with random forest and GBM models. Out-performing previous findings from LSTM networks DNN and RNN demonstrated improved detection capabilities, particularly for zero-day threats. The superiority of deep learning over iron and model in only detection is demonstrated by their relative accuracy of 98.17% and 13.95%.

Cloud computing security has never been better, according to this study that addresses the ever-changing problems posed by cyber security threats. An important step towards creating stable cloud computing environments, the study improves detection efficiency and accuracy.

7 Conclusion and Future Work

With the development of sophisticated anomaly detection models the obtained results demonstrated that cloud computing security has been significantly improved. Using the combined form of machine learning and deep learning models they set out the pressing need for robust security measures in cloud settings. The findings of this research

demonstrated that DNN and GBM models might safeguard cloud infrastructures by detecting anomalies with exceptional accuracy. The importance of this research lies in the comprehensive methodology it allows to assess the appropriateness of several machine learning models in the context of cloud security. This research has some limitations as it says that more diverse data set are needed for testing these models. Exploring the hybrid models that consist of multiple aspects of different methods and creating more advanced algorithms for balancing the data are the two main areas that can be used in future research. The dynamic nature of the cyber threats requires flexible and new cloud computing security solutions. This research however lays the framework for the future studies and emphasis the requirement of the continuous enhancement and modification in the security approaches for cloud system.

References

Aldallal, A. and Alisa, F. (2021). Effective intrusion detection system to secure data in cloud using machine learning, *Symmetry* **13**(12): 2306.

Alshammari, A. and Aldribi, A. (2021). Apply machine learning techniques to detect malicious network traffic in cloud computing, *Journal of Big Data* **8**(1): 1–24.

Chkirbene, Z., Erbad, A., Hamila, R., Gouissem, A., Mohamed, A. and Hamdi, M. (2020). Machine learning based cloud computing anomalies detection, *IEEE Network* **34**(6): 178–183.

Dhabliya, M. D. (2021). Cloud computing security optimization via algorithm implementation, *International Journal of New Practices in Management and Engineering* **10**(01): 22–24.

Garg, S., Kaur, K., Kumar, N., Kaddoum, G., Zomaya, A. Y. and Ranjan, R. (2019). A hybrid deep learning-based model for anomaly detection in cloud datacenter networks, *IEEE Transactions on Network and Service Management* **16**(3): 924–935.

Girish, L. and Rao, S. K. (2023). Anomaly detection in cloud environment using artificial intelligence techniques, *Computing* **105**(3): 675–688.

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

He, Z., Hu, G. and Lee, R. B. (2023). Cloudshield: Real-time anomaly detection in the cloud, *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy*, pp. 91–102.

Islam, M. S., Pourmajidi, W., Zhang, L., Steinbacher, J., Erwin, T. and Miranskyy, A. (2021). Anomaly detection in a large-scale cloud platform, *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, IEEE, pp. 150–159.

Jauro, F., Chiroma, H., Gital, A. Y., Almutairi, M., Shafi'i, M. A. and Abawajy, J. H. (2020). Deep learning architectures in emerging cloud computing architectures: Recent development, challenges and next research trend, *Applied Soft Computing* **96**: 106582.

Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance, *Journal of Big Data* **6**(1): 1–54.

Lin, P., Ye, K. and Xu, C.-Z. (2019). Dynamic network anomaly detection system by using deep learning techniques, *Cloud Computing–CLOUD 2019: 12th International Conference, Held as Part of the Services Conference Federation, SCF 2019, San Diego, CA, USA, June 25–30, 2019, Proceedings 12*, Springer, pp. 161–176.

Mohammad, A. S. and Pradhan, M. R. (2021). Machine learning with big data analytics for cloud security, *Computers & Electrical Engineering* **96**: 107527.

Muhamad, F. P. B., Mulyani, E., Bunga, M. S. and Mushafa, A. F. (2023). Class balancing methods comparison for software requirements classification on support vector machines, *Sinkron: jurnal dan penelitian teknik informatika* **8**(2): 1196–1208.

Muneer, S. M., Alvi, M. B. and Farrakh, A. (2023). Cyber security event detection using machine learning technique, *International Journal of Computational and Innovative Sciences* **2**(2): 42–46.

Nedelkoski, S., Cardoso, J. and Kao, O. (2019). Anomaly detection from system tracing data using multimodal deep learning, *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*, pp. 179–186.

Ochani, M., Sawarkar, S. and Narwane, S. (2019). A novel approach to handle class imbalance: a survey, *Int J Eng Dev Res (IJEDR)* **7**(2): 1–9.

Peterson, K. T., Sagan, V. and Sloan, J. J. (2020). Deep learning-based water quality estimation and anomaly detection using landsat-8/sentinel-2 virtual constellation and cloud computing, *GIScience & Remote Sensing* **57**(4): 510–525.

Rana, P., Batra, I., Malik, A., Imoize, A. L., Kim, Y., Pani, S. K., Goyal, N., Kumar, A. and Rho, S. (2022). Intrusion detection systems in cloud computing paradigm: Analysis and overview, *Complexity* **2022**.

Rogić, S. and Kašćelan, L. (2021). Class balancing in customer segments classification using support vector machine rule extraction and ensemble learning, *Computer Science and Information Systems* **18**(3): 893–925.

Samriya, J. K. and Kumar, N. (2020). A novel intrusion detection system using hybrid clustering-optimization approach in cloud computing, *Materials Today: Proceedings*, Vol. 2, pp. 23–54.

Saran, M., Yadav, R. K. and Tripathi, U. N. (2022). Machine learning based security for cloud computing: A survey, *International Journal of Applied Engineering Research* **17**(4): 332–337.

Wang, H., Guo, J., Ma, X., Fu, S., Yang, Q. and Xu, Y. (2021). Online self-evolving anomaly detection in cloud computing environments, *arXiv preprint arXiv:2111.08232* .

Xu, S., Qian, Y. and Hu, R. Q. (2019). Data-driven edge intelligence for robust network anomaly detection, *IEEE Transactions on Network Science and Engineering* **7**(3): 1481–1492.