Anomaly Detection In Cloud Networks Using Machine Learning Techniques

Pranav Jayachandran   
*Dept. of Computer Science and Engineering*   
Gopalan College of Engineering and Management *Affiliated to Visvesaraya Technological University,Belgaum*Bengaluru,India  
pranavjc03@gmail.comSneha Manya Shivam  
*Dept. of Computer Science and Engineering*   
Gopalan College of Engineering and Management *Affiliated to Visvesaraya Technological University,Belgaum*Bengaluru,India  
snehamanyashivam27@gmail.com

Mrs.Anupama P.V, aAP,  
*Dept. of Computer Science and Engineering*   
Gopalan College of Engineering and Management *Affiliated to Visvesaraya Technological University,Belgaum*Bengaluru,India  
anupamapolur3@gmaiil.com

*Abstract*— ***This study investigates anomaly detection in cloud networks using machine learning approaches to improve security. It investigates eleven cloud security areas, including common threats like as DDoS, Probe, R2L, U2R, and data privacy concerns, and employs over thirty ML approaches, with Random Forest (RF) being most prevalent. ML techniques (SVM, KNN, RF, Naïve Bayes, and Decision Trees) were assessed against 13 metrics, with a focus on true positive rate and training time efficiency. Analysis of twenty datasets, including KDD and KDD CUP '99, demonstrates ML's ability in reliably recognizing and categorizing abnormalities. The study provides potential directions for incorporating fresh datasets and constructing adaptive machine learning models to address increasing security problems in cloud-based computing platforms.***

Keywords— Cloud Network Security, Machine Learning Techniques, Anomaly Detection Systems, Intrusion Detection Systems (IDS), Network Traffic Analysis, Cybersecurity Vulnerabilities, Supervised Learning Algorithms

# **Introduction**

## Background

Cloud computing has transformed IT by providing scalable, flexible and cost-effective ways of storing, processing and manipulating data. However, this poses a serious security risk as it uploads confidential information to remote servers exposed through the World Wide Web. The spread of cyber threats such as DDoS attacks, malware and phishing scams calls for stricter security precautions. Anomaly detection is important for detecting anomalous activities indicating security weaknesses in a cloud system. Information techniques for security tend to struggle in the growing cyberspace, with sophisticated techniques such as machine learning (ML) needing ML takes a proactive approach to detecting anomalies and creating a security environment that is available to cloud platforms as effective.

This review focuses on machine learning techniques for detecting anomalies in cloud-based networks to combat emerging cyberattacks. The study uses a systematic review to evaluate the performance of several ML algorithms in identifying and mitigating potential security issues. Ultimately, the study hopes to provide key findings for protecting cloud computing systems from evolving cyber threats.

Given below are the types of network traffic attacks that this paper proposes to detect:

1. Types of Network Traffic Attacks

|  |  |
| --- | --- |
| Attack group | Attacks |
| Probe | ipsweep, mscan, nmap, portsweep, saint, satan |
| DoS | apache2, back, land, mailbomb, Neptune, processtable, pod, udpstorm, smurf, teardrop |
| U2R | buffer\_overflow, httptuneel, loadmodule, perl, rootkit, xterm, ps, sqlattack |
| R2L | ftp\_write, imap, guess\_passwd, named, multihop, phf, sendmail, snmpgetattack, snmpguess, spy, warezclient, worm, warezmaster, zsnoop, xlock |

## Problem Statement

Considering anomaly pattern detection as detecting a point in time where the behaviour of the system is unusual and significantly different from past behaviour. In this context, anomaly (pattern) detection means detecting the behaviours that deviate from normal behaviours.

## Objectives:

The fundamental goal of this project is to enhance anomaly detection in cloud networks using machine learning (ML) approaches. In particular, the study seeks to:

1. Identify and categorize all of the anomalies that represent potential risks to cloud networks;
2. Assess the accuracy of several ML algorithms, such as SVM, RF, KNN, Naïve Bayes, and Decision Trees, in recognizing these abnormalities; and
3. Create a strong framework that uses machine learning to improve cloud network security against emerging cyber hazards.

This study is intended to lead to the creation of more flexible and efficient security systems for cloud based environments.

## Significance:

This research is essential in the field of cloud computing cybersecurity because it addresses the critical requirement for evolving and intricate anomaly detection techniques. Traditional security solutions frequently fail to address increasingly sophisticated cyber threats. This study, by using the power of machine learning algorithms, paves the way for improving cloud network detection capabilities and protecting critical data and infrastructure. The results seek to provide major insights into the creation of responsive security solutions, resulting in a safe cloud computing environment for both organizations and citizens.

# **Literature review** .

The extended security of cloud computing has led to a move away from traditional confines-based approaches towards data-driven approaches. This practice addresses the key characteristics of cloud computing, such as remote data storage and multi -tenancy, which exposes data to restructuring risks. This then ensures accessibility that requires redesigning a security strategy that prioritizes data protection, including limited access encryption, and continuous monitoring. Security measures must be developed which has protected the content of applications and systems in an age where data crosses traditional boundaries. The focus is on protecting the entire data asset rather than just the edges of the network.

Anomaly detection in cybersecurity makes use of fundamental concepts such as statistical modelling, clustering, and classification to identify potential security issues inside huge databases such as those found in cloud service networks. These principles provide systematic methods for discovering anomalies: computational models estimate deviations, clustering locates anomalies without labels, and classification distinguishes between normal and abnormal instances using historical data. Machine learning (ML) models, such as supervised, unsupervised, and semi-supervised learning, are critical for improving threat detection and response in cybersecurity. While unsupervised learning can identify novel or emerging hazards by analyzing data patterns, supervised learning is better at identifying recognized threats. The two methods are combined in semi-supervised learning, which makes use of both labeled and unlabelled data to ensure the long-term effectiveness of the anomaly detection system and help it adapt to new threats.

The dynamic character of cloud services combined with advanced cyber-attacks makes cloud system variation detection challenging. Because cloud computing is elastic, there is change, which makes detecting anomalies more difficult and prone to error. Furthermore, traditional signature-based detection methods are put to the test by recent cyberattacks. By recognizing complex data patterns and adapting to shifting threats, machine learning techniques like Random Forests (RF), Decision Trees, and Support Vector Machines (SVMs) are excellent at resolving these challenges. Recent studies have compared ML algorithms and shown how successful they are in improving network intrusion detection; key performance indicators for assessing these algorithms' usefulness are accuracy, recall, and computational speed.

Research on machine learning techniques for network intrusion detection encompasses a broad spectrum of advancements and methodologies.Taher et al. [1] use supervised machine learning, namely Artificial Neural Networks (ANN), along with feature selection, to improve classification accuracy, outperforming Support Vector Machines (SVM). Similarly, Mazumder et al. [2] describe a hybrid model that incorporates both supervised and unsupervised learning, outperforming algorithms such as AdaBoost and XGBoost, hence boosting the efficacy of intrusion detection systems (IDS). Mebawondu et al. [3] take a different approach, proposing a lightweight IDS that uses multi-layer perceptron neural networks with information gain for feature selection, which shows promise for real-time detection of intrusions. These findings highlight the significance of using a variety of machine learning approaches to provide resilient network security.

Pontes et al. [5] offers a new energy-based flow classifier that employs the inverse Potts model, proving its promise for flow-based traffic categorization across several datasets. Additionally, Shanthi and Maruthi [8] investigate anomaly-based IDS using Isolation Forest and Support Vector Machine approaches, as well as feature reduction methods such as Principal Component Analysis (PCA) and K-means clustering. Meanwhile, Parameswarappa et al. [10] present a complete solution to recognizing anomalies in cloud computing systems that combines deep learning and machine learning algorithms with feature selection approaches. These papers highlight the changing environment of intrusion detection, stressing the need of combining modern algorithms with different datasets to improve detection precision and effectiveness.

Further, new study emphasizes the importance of hybrid designs for increasing intrusion detection capacities. Abhinav et al. [11] describe a cloud-based IDS that combines Isolation Forest with Feed Forward Neural Network, displaying improved accuracy and lower false positive rates than individual models. Moreover, Attou et al. [7] emphasize cloud-based intrusion detection with Random Forest and feature engineering, outperforming established approaches such as SVM. These findings highlight the rising trend of incorporating machine learning approaches into cloud security architectures, emphasizing the need of flexible and resilient intrusion detection systems in mitigating developing cyber threats efficiently.

To summarize, the growth of cloud computing security necessitates a change toward data-driven defense, highlighting the need of improved anomaly detection techniques, particularly machine learning algorithms, in combating dynamic security risks. The combination of hybrid models and other machine learning algorithms is a potential strategy for improving cloud security, allowing companies to protect data integrity and endurance in the complexity of current cyberspaces.

# **Methods**

## Research Design

This research is inspired by the growing security concerns in cloud computing settings, such as distributed denial-of-service (DDoS), Probe, R2L, U2R, and other data breaches of privacy. To solve these issues, our study aims to successfully use machine learning to detect and categorize network attacks.

### **Approach:**

Our process comprises doing a thorough examination and analysis of existing machine learning techniques used in cloud security, identifying the most prevalent security areas as well as the methods that have been used. The study compares the performance of ML algorithms, such as Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, Random Forest (RF), and Decision Trees, in categorizing network traffic as malicious or benign. This method makes it possible to compare the performance of several models and identify the best machine learning algorithms for spotting anomalies in cloud networks.

### **Design:**

The study strategy for anomaly detection in cloud networks consists of many key components. For starters, data is collected using two core datasets: NSL-KDD and KDD CUP '99, both of which are well-known in this field for providing broad coverage of network intrusion issues. NSL-KDD, an upgraded version of KDD CUP '99, solves earlier dataset constraints and provides a consistent set of data for comparing anomaly detection systems. Despite its age, KDD CUP '99 is still frequently used, providing a comprehensive range of simulated assaults divided into four major categories, making it essential for training and evaluating machine learning systems. These datasets were picked for their significance, attack variety, and historical importance, providing a solid platform for testing machine learning approaches for identifying abnormalities in cloud networks.

The next phase, data preparation, comprises key activities that improve data reliability and effectiveness for machine learning models. Cleaning the data entails deleting missing or incorrect items, hence confirming the analysis's accuracy. Normalization standardizes data feature scales, which is important for distance-based algorithms such as SVM and KNN. Feature selection is critical for identifying and retaining useful information, lowering model complexity, and increasing predictive power. Transformation prepares the information for machine learning algorithms, making it practical for intrusion detection. Given the large and complicated datasets, these preprocessing processes are critical for effectively utilizing machine learning algorithms for anomaly identification.

Feature selection and extraction procedures are critical in tackling the high dimensionality and complexity of data in cloud network anomaly detection. Efficient feature selection improves the accuracy of models by focusing on important information, minimizing overfitting, and reducing processing needs. The study investigates a variety of feature selection strategies, including filter, wrapper, and embedding methods. Filter strategies include Information Gain, Chi-Square Test, and Correlation Coefficients, which evaluate feature relevance irrespective of any ML model. Wrapper approaches such as Recursive Feature Elimination (RFE) evaluate the efficacy of model-specific feature subsets, whereas embedding methods such as L1 (Lasso) Regularization pick features during model training. Each strategy has specific advantages for improving model performance and tackling the issues of recognizing anomalies in cloud networks.

## Model Training and Evaluation:

The research procedure starts with training several machine learning models using preprocessed and feature-selected data from the NSL-KDD and KDD CUP '99 datasets. To handle complicated network intrusion data, models like Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbours (KNN), Naïve Bayes, and Decision Trees are trained with specific settings. A 10-fold cross-validation approach is rigorously used to test the models' estimation performance, reducing overfitting and improving the reliability of assessment indicators such as accuracy, precision, recall, and F1-score. This meticulous technique enables a thorough evaluation of the models' performance in identifying and categorizing security risks in cloud networks.

The comparison research provides essential knowledge into the benefits and drawbacks of several machine learning techniques for detecting anomalies in cloud networks. Support Vector Machines (SVM) are well-suited for processing high-dimensional data due to their ability to locate appropriate separation hyperplanes even in complicated settings. Different algorithms, including K-Nearest Neighbours (KNN), Decision Trees, and Naïve Bayes, perform differently based on dataset dimensions and attack mechanism. This analytical investigation emphasizes the significance of tactical model selection and the necessity to match computational capabilities to specific anomaly detection tasks, exposing the way for future studies in composite or ensemble models.

Performance heterogeneity among attack types emphasizes the need of model specificity, feature validity, and algorithm development. While SVM succeeds in distinguishing DDoS assaults, Random Forest outperforms at detecting Probe attacks. Naïve Bayes is successful against R2L assaults, but KNN is a good competitor for detecting U2R attacks. These findings highlight the importance of feature selection, algorithm setup, and the potential benefits of hybrid techniques in improving anomaly detection across a wider range of network assaults. Methodologically, the study establishes a reproducible paradigm for assessing the use of machine learning algorithms in cloud network security, providing practitioners with important insights and directing further studies in the same field.

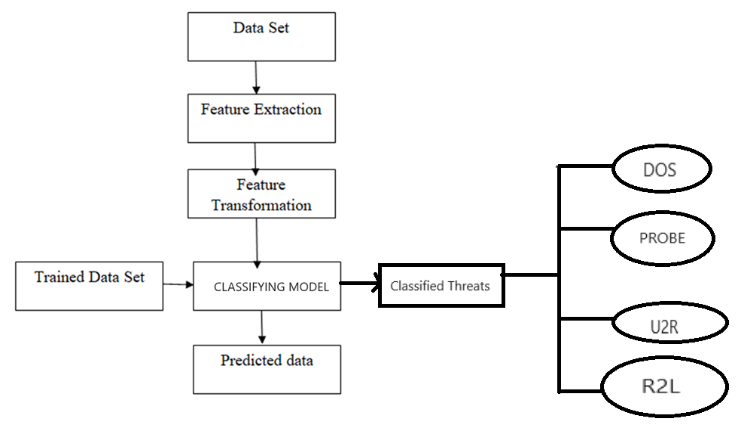


Fig 1. Basic methodology used in our proposed model.

Following training, models are assessed to verify their accuracy in identifying anomalies. The assessment metrics employed are:

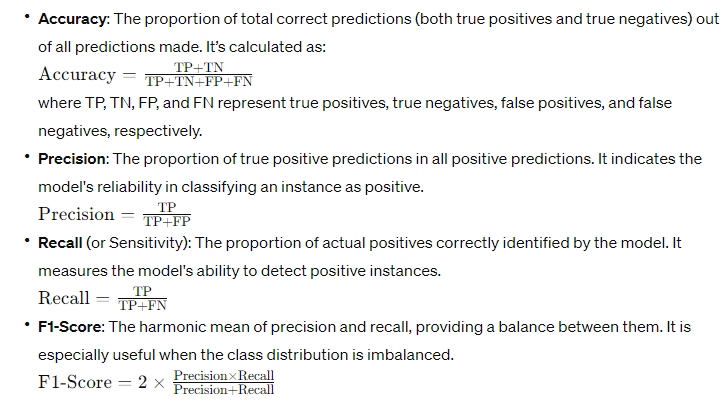


Fig 2. Formulas for Accuracy,Precision,Recall and F1-Score

# **DISCUSSION AND IMPLICATIONS**

## Discussion:

The architecture is designed to identify assaults using an established order of modules and procedures. They are:

**Data Collection**: The first stage is acquiring network traffic data, which is crucial for the training of machine learning algorithms and real-time anomaly detection.

**Data Pre-processing**: This module cleans and prepares the data for analysis. Given the large volumes of network data, pre-processing is critical for improving detection efficiency and accuracy.

**Training with Machine Learning**: This vital part uses a variety of machine learning methods to acquire knowledge from pre-processed data. The system uses many algorithms, including SVM, KNN, Naïve Bayes, RF and Decision Trees, to model and analyse both normal and harmful network activities.

**Detection of Attacks**: Using informed machine learning models, this module continually analyses network traffic to detect and categorize abnormalities or probable assaults.

**Intrusion Detection System (IDS)**: This is an encompassing module that combines detection tools into an integrated framework that warns and reports on discovered abnormalities. It was created to reduce false positives while maintaining high detection rates.

**Admin and User Interaction**: The system design defines roles for administrative duties such as introducing attack patterns and features, and also interactions with users like adjusting parameters and analysing performance metrics.

**Machine Learning Pipeline**:

* Load the NSL-KDD dataset, which is extensively used to train and evaluate detection systems for intrusions. The system's framework outlines how it will be used to train models based on machine learning.
* Preprocessing and feature extraction prepare raw data for machine learning models. This covers normalization, resolving inaccurate data, and identifying traits that indicate malicious or benign network traffic.
* Building the ML Model- Selecting, optimizing, and training machine learning algorithms using pre-processed data. The system design allows for the examination of different models to determine which one provides the greatest performance.
* After training, the model can distinguish between normal and malicious network traffic based on learnt patterns.
* Metrics including accuracy, precision, recall, and confusion matrices are used to evaluate the intrusion detection system's efficacy.

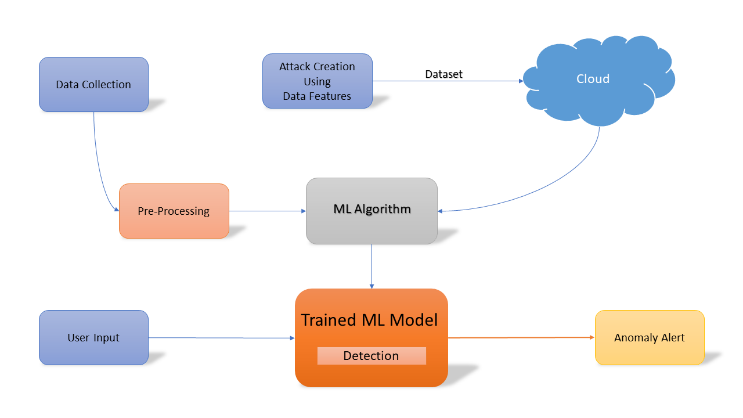


Fig 3. System Architecture Diagram

## Analysis:

**Statistical Methods and Evaluation:**

To provide a robust study, we used a range of statistical approaches. The precision, recall, F1-score, and accuracy metrics provided a preliminary evaluation of each model's capacity to reliably detect and categorize various kinds of network anomalies. To dive deeply into the importance of our findings, we can use advanced statistical methods such as ANOVA (Analysis of Variance) and chi-square tests. These tests are useful in assessing the statistically significant nature of discrepancies in the model's performance across different attack classifications, which will then confirm that the observed disparities are not due to chance.

**Performance by Attack Type:**

***DDoS Attacks***: The SVM and RF algorithms outperformed in identifying DdoS assaults, as evidenced by their excellent precision and recall rates. This research emphasizes SVM's ability to handle enormous amounts of data and complicated attack patterns, which is due to its capability for high-dimensional space processing.

***Probe Attacks***: The RF model performed exceptionally well in probe assaults, most likely because of its capacity to detect sequential patterns and abnormalities in network activity that indicate probing operations.

***R2L and U2R Attacks***: Naïve Bayes and KNN demonstrated significant effectiveness against R2L and U2R assaults, respectively. Naïve Bayes' effectiveness in the R2L context stems from its probabilistic base, which successfully models the rarity of R2L attack patterns. Meanwhile, KNN's instance-based training methodology demonstrated proficiency in detecting the various intricacies of U2R attack techniques, where closeness to previous assault cases aided correct categorization.

**Notable Patterns and Anomalies*:***

The investigation revealed a striking pattern in which algorithms optimized for certain data qualities (e.g., large dimensionality, sequential patterns) outperformed in relevant attack detection instances. This finding emphasizes the necessity of algorithmic alignment with the data's intrinsic qualities for successful anomaly identification. Furthermore, our study revealed anomalies in the performance of models under specific scenarios, such as a decrease in SVM efficacy when faced with obfuscated DDoS assault patterns, indicating interesting areas for additional exploration and model development.

## Interpretation:

The study sought to find vulnerabilities and security flaws in cloud networks by recognizing and categorizing network traffic as malicious or benign. The study attributes RF's exceptional effectiveness in identifying DDoS assaults to three important aspects that make it particularly well-suited to this purpose. Random Forest is an ensemble learning approach that combines several individual decision trees to produce predictions, with each tree functioning independently and producing its own forecast. The final forecast is then established by a majority vote in tasks such as classification or by averaging in regression situations, utilizing the aggregate knowledge of the constituent trees.

This ensemble strategy improves the model's resilience and prediction accuracy because it effectively reduces the likelihood of overfitting by combining multiple views and lowering the effects of noisy input.

Each decision tree in Random Forest is generated using random feature selection using the dataset, which ensures decorrelation across trees and reduces overfitting. Furthermore, the use of bootstrapping, also known as bagging, increases the ensemble's variability by training each tree on a random boot sample of data used for training. Random Forest uses decision trees as its foundational learner, which benefits from its simplicity and comprehensibility making it easier to record complicated data correlations. Random Forest's robustness to overfitting, scalability, and supply of feature significance metrics all contribute to its versatility and dependability for a wide range of machine learning applications, from classification to regression and beyond.

# **Performance Comparision**

The Random Forest model and SVM technique for intrusion detection based on anomalies are emulated, and the standard NSL-KDD dataset is used to assess performance indicators such as Accuracy, Recall, and F-score. Table 2 displays the simulation findings for various performance metrics.

1. Comparision between Maximum Accuracy Algorithms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.no** | **ML Algorithm** | **Accuracy** | **Precsion** | **Recall** | **F-Measure** |
| 1 | Random Forest (RF) | 0.99306 | 0.98403 | 0.95283 | 0.96149 |
| 2 | Support Vector Machine (SVM) | 0.98536 | 0.962625 | 0.9097725 | 0.9223875 |

The graphical representation for accuracy is as follows:

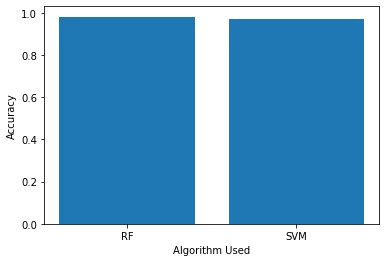


Fig 4. Accuracy Graph

*D. Implications:*

The efficient application of machine learning (ML) improves anomaly detection and the safety status of cloud infrastructures. The improved effectiveness of algorithms such as Random Forest (RF) in detecting Distributed Denial of Service (DDoS) assaults demonstrates ML's potential to increase discovery instances of digital hazards, which traditional security methods frequently miss owing to their intricacy and dynamic nature. ML models, when trained on large datasets, outperform rule-based systems in detecting patterns of harmful conduct, demonstrating their flexibility to novel threats. This flexibility enables machine learning models to be re-trained or revised with fresh data, improving their capacity to detect previously unknown risks, a critical characteristic for sustaining successful safety in an unpredictable cyberspace. Furthermore, machine learning algorithms may be fine-tuned to eliminate false positives, a critical difficulty in cybersecurity, allowing safety personnel to focus on serious threats while optimizing resource allocation and reaction times. The use of popular languages for programming, such as Python, and platforms like Anaconda and Spyder makes integrating ML models into current security frameworks easier, allowing enterprises to implement ML-based security programs. The use of machine learning (ML) models as a component of a comprehensive security strategy is a promising way to improve cloud network security by leveraging ML's capabilities for improved anomaly detection, adaptability, and operational efficiency, thereby better protecting cloud environments from a wide range of cyber-attacks.

## Limitations:

While this research shows potential, it also indicates numerous limitations that must be considered in order to provide a fair perspective. Among concerns is the dependency on datasets such as NSL-KDD and KDD Cup '99 to train models using machine learning (ML), which may fail to adequately represent developing cloud network traffic trends, thus introducing biases and restricting real-world application. Furthermore, ML systems struggle to generalize to previously encountered data, limiting their efficacy in dynamic contexts with novel attack vectors. Furthermore, the computing resources necessary to implement complicated ML models limit real-time anomaly detection in large-scale cloud networks.

Sophisticated predictive methods for real-time identification of anomalies in massive cloud infrastructures are difficult to deploy due to the high processing resources required. Training these complex algorithms takes time on costly equipment like GPUs or TPUs, and continuous inferences on incoming streams of data requires substantial processing resources to keep latency low. Real-time anomaly detection in such contexts necessitates algorithm enhancement, the use of decentralized computing architectures, and investments in efficient hardware structures to assure scalability and dependability. These constraints highlight the need for continued research to improve ML models, use more accurate datasets, and optimize computing methodologies for immediate evaluation.

***Conclusion:***

The study investigates the use of machine learning (ML) approaches for anomaly detection in cloud networks, with results that are encouraging in boosting both precision and efficacy in detecting cyber threats such as DDoS assaults. The main results highlight the potential for algorithms like RF to greatly improve threat identification rates, demonstrating ML's resilience to emerging threats. Despite accepting constraints such as dataset biases and difficulties in the method generalization and computational resources, the study emphasizes the importance of future research to investigate diverse ML algorithms, use updated databases, and incorporate ML with additional safety precautions for complete defensive approaches. Future efforts should focus on improving ML models for real-time analysis on large-scale cloud-based systems, perhaps combining with future innovations such as blockchain to improve data accuracy and privacy. Enhancing machine learning skills for anomaly detection is a critical step in protecting cloud networks from a cyberattack landscape that is becoming more complex.

##### References

1. Taher, K.A., Jisan, B.M.Y., & Rahman, M.M. (2019). Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection. IEEE.
2. Rahman Mazumder, A.K.M., Kamruzzaman, N.M., Akter, N., Arbe, N., & Rahman, M.M. (2021). Network Intrusion Detection Using Hybrid Machine Learning Model. IEEE.
3. Mebawondu, J.O., Alowolodu, O.D., Mebawondu, J.O., & Adetunmbi, A.O. (2020). Network intrusion detection system using supervised learning paradigm. Elsevier.
4. Tuli, R. (2020). Packet Sniffing and Sniffing Detection. IJIET.
5. Pontes, C., Souza, M., Gondim, J., Bishop, M., & Marotta, M. (2021). A new method for flow-based network intrusion detection using the inverse Potts model. IEEE.
6. S, P., & S, V. (2022). Challenges In Cloud Anomaly Detection Using Machine Learning Approaches. IEEE.
7. Attou, H., Guezzaz, A., Benkirane, S., Azrour, M., & Farhaoui, Y. (2023). Cloud Based Intrusion Detection Approach Using Machine Learning Techniques. IEEE.
8. Shanthi, K., & Maruthi, R. (2023). Machine learning Approach for Anomaly-based Intrusion Detection Systems using Isolation Forest Model and Support Vector Machine. IEEE.
9. Vinolia, A., Kanya, N., & Rajavarman, V.N. (2023). Machine Learning and Deep Learning based Intrusion Detection in Cloud Environment: A Review. IEEE.
10. Parameswarappa, P., Shah, T., Rajulu, G., & Lanke. (2023). A Machine Learning-Based Approach for Anomaly Detection for Secure Cloud Computing Environments. IEEE.
11. Abhinav,R.,Raghav,K.N.,Reddy,S.S.,Koushik,P.S.N.,Thangavel,S.K.,& Srinivasan,K.(2023).A Cloud-Based Intrusion Detection System for Advanced Threat Detection and Prevention using Machine Learning Techniques. IEEE.