MACHINE LEARNING-BASED INTRUSION DETECTION SYSTEM FOR CLOUD COMPUTING: ADDRESSING SECURITY CHALLENGES WITH ADVANCED ALGORITHMS USING THE KDD CUP 1999 DATASET

[1*]Sruthi Mol P, [2]Dr. Sathish Kumar N

[1]Assistant Professor, Department of Information Technology, KGiSL Institute of Technology, Coimbatore, Tamilnadu, India, Email: sruthi.mol88@gmail.com*
[2]Professor, Department of Electronics and Communication Engineering, Sri Ramakrishna Engineering College, Coimbatore, India, Email: nsk20022002@gmail.com

Abstract
Intrusion detection into cloud computing environments has become one of the most important aspects for data security and the maintenance of confidence in cloud-based services. The aim of this research work is to develop a robust intrusion-detection system using machine learning algorithms that classify cyber threats: This system will recognize anomalies in network traffic with the use of algorithms like Gradient Boosting, AdaBoost, Perceptron, Ridge Classifier, and CatBoost, and be able to classify legitimate and harmful action. The KDD Cup 1999 dataset is utilized for training and evaluating the models, ensuring a diverse range of attack scenarios, including DoS, Probe, U2R, and R2L.These include data preprocessing, feature engineering, model training, hyperparameter optimization, and comparison of performance based on metrics of accuracy, precision, recall, F1-score, and computation cost. Among these methods, Ensemble learning methods like Gradient Boosting and CatBoost are specifically used for imbalanced datasets. Perceptron and Ridge Classifier are very light, suitable for immediate implementation. The core intent of this study is to enhance the detection accuracy level and scale intrusion detection systems and bring it as an acceptable solution cloud infrastructures' security hassle.

Keywords: Intrusion Detection, Cloud Computing, Machine Learning, Gradient Boosting, AdaBoost and CatBoost, Perceptron and Ridge Classifier, KDD Cup 1999, and Cybersecurity

I. Introduction

The rapid pace of expansion of cloud computing has revolutionized the ways in which data is stored, accessed, and controlled within organizations. Immediate scalability, economic efficiency, and adaptability have made cloud services indispensable adjuncts to today's IT frameworks, however, the very properties that make cloud computing attractive also expose it to significant cybersecurity risks. IDS, therefore, have been established as a crucially essential protective measure and become responsible for the detection of unidentified access or harmful actions within cloud computing environments. Algorithms employed by machine learning have significantly enhanced the performance of IDS by the ability to identify attack patterns previously unknown. This research study employs effective complex machine learning techniques: Gradient Boosting, AdaBoost, Perceptron, Ridge Classifier, and CatBoost with regard to the goal of intrusion detection in cloud computing.

The background of intrusion detection in cloud environments reveals dynamic evolution from traditional signature-based methods to anomaly-based and hybrid techniques. Traditional signature-based IDS relied heavily on predefined attack signatures, thus limiting the identification of novelty attacks or zero-day threats. Meanwhile, in machine learning-based anomaly detection systems, network traffic behavior is analyzed to detect anomalies that may reflect potential attacks. Such work as [1], [5], and [17] highlight the importance of enriching these systems with intelligent algorithms in an effort to enhance their detection rates and decrease false positives.

Despite these, there remain significant challenges. Cloud environments by their nature are intrinsically complex and heterogenous, with vast amounts of data being generated across distributed systems. The dynamic nature of cloud workloads makes it hard to detect intrusions mainly because most traditional methods failed to keep up with changing patterns. In fact, the imbalanced datasets also exist; here, normal traffic is far outweighed by malicious traffic, which would pose problems for machine learning models. Algorithms, including Gradient Boosting and CatBoost, have shown to be effective in addressing this issue, as they can work well in counterbalancing imbalances through methodologies of iterative learning [10].

The other important issue is scalability. It has to ensure that large data streams in real time are evaluated without compromising performance integrity. Distributed computing frameworks, described in [8] and [11], form a foundation for scalable IDS solutions. However, the tradeoff between computational efficiency and precision is still a pressing concern. Ridge Classifier and Perceptron are computationally simple, but lightweight solutions that might not match the predictive power of ensemble techniques like Gradient Boosting and AdaBoost [3].

Feature selection and engineering also play an important role in the development of efficient IDSs. An intrusion dataset like KDD Cup 1999 contains thousands of features, and therefore selecting the most vital attributes that contribute towards the detection of intrusions greatly depends upon the performance of the model. RFE used by [2] and other feature selection schemes play an important role in dimensionality reduction with the preservation of the most important information. However, dimensionality reduction would have to be carefully controlled so that it does not degrade the detection performance by information loss.

Adapting machine learning models to cloud-specific challenges requires defeating all kinds of attacks, including DoS, U2R, R2L, and Probe. The attack characteristics vary, requiring smart algorithms that can generalize across multiple attack scenarios. Ensemble methods such as AdaBoost and Gradient Boosting are particularly effective in such scenarios because their ability to combine the predictive power of multiple weak learners to form an approximation of the final learner [5]. CatBoost, an algorithm based on gradient boosting that is specifically tailored for managing categorical features, improves detection proficiency by refining feature interactions and minimizing overfitting [6].

The interpretability of machine learning models in the context of IDS further remains an issue. Ensemble methods and deep learning algorithms have been shown to offer higher accuracies, but their complexity makes it difficult to understand the reasoning behind a specific prediction outcome. Poor interpretability negatively affects the deployment of IDS in cloud environments, where, after all, explainability has become a crucial ingredient for forensic analysis and compliance with certain rules. Methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been suggested as solutions to this problem; however, they necessitate increased computational resources.

In cloud-based IDS, the adverse threat of evasion, which involves deception models aiming to deceive machine learning models, is also faced. An adversary could easily modify the input data in such a way as to evade detection or cause false alarms and thus degrade the trustiness

associated with an IDS. Some robust training techniques countering these risks include adversarial training as well as regularization. Recent studies have concentrated on integrating hybrid machine learning models with advanced defensive mechanisms such as that described by [4] and [17].

More importantly, integrating IDS into cloud infrastructures raises several concerns related to data privacy and adherence to regulatory standards. Ensuring that the processes of IDS do not expose sensitive information necessitates the applications of privacy-preserving machine learning techniques. Federated learning, as described in [9], represents a promising area as it would enable the collaborative training of models across multiple cloud nodes without sharing raw data. However, the complexity of such solutions and the effect on detection performance are active research topics.

Evaluation and benchmarking of IDS models face more challenges in terms of selecting appropriate datasets, metrics, and methodologies to perform such evaluations. Although the KDD Cup 1999 dataset has appeared extensively in intrusion detection research, this testbed has been criticized due to the use of outdated attack profiles and the absence of current modern cloud-specific scenarios [13]. Training and evaluation pipelines must include real-world cloud traffic data to develop IDS addressing the existing threats. Performance metrics, including accuracy, precision, recall, and F1-score, although commonly utilized, may not fully encapsulate the complexities inherent in intrusion detection. Furthermore, supplementary evaluations, such as ROC-AUC and analyses of computational costs, offer a more comprehensive evaluation of the efficacy of the model [14].

Regarding the potential ID problems encountered above, this project suggests designing an effective IDS particularly tailored for cloud computing. Such an IDS is constructed based on the strengths of Gradient Boosting, AdaBoost, Perceptron, and Ridge Classifier together with CatBoost. Its aim is toward improving the accuracy, efficiency, and reliability of detection in modern Cloud Environments through advanced feature selection, scalable model, and comprehensive evaluation. Incorporating a variety of algorithms with the necessary cutting-edge methodologies enhances the continuous endeavors in a quest to safeguard cloud infrastructures in response to the constantly changing environment of cyber threats.

II. Related Works

Al-Ghuwairi et al. (2023) explored cloud-based intrusion detection through methodologies of machine learning-based processed time series data and highlighted the anomaly identification point. The authors established that sophisticated cloud-based attacks were increasingly difficult to detect, due to fluidity characteristics necessitating processing in real-time data. To solve this problem, they introduced a machine learning model based on anomaly detection algorithms in combination with analysis for better performances in detection. Their approach resulted in a significant improvement in detecting both known and zero-day attacks. The real-time intrusion detection scenarios by the model encountered high accuracy. The solution was able to adequately detect anomalies with very minimal false positives and false negatives, so it is a promising solution for a cloud security solution. This research has highlighted the need for advanced anomalies detection techniques in the context of cloud computing, as shown by the findings from Al-Ghuwairi et al. [1].

Sharma and Singh (2024) proposed batch reinforcement learning, which inherently uses recursive feature elimination for the task of network intrusion detection. The authors recognized the difficulty posed by the presence of nuisance features along with high dimensional data and showed that it diminished significantly the efficiency of IDSs. In this study, the authors utilized RFE together with reinforcement learning in such a manner that selection of highly correlated features was easy. The proposed model showed much improvement in detection accuracy by reducing the dimensionality of the data without interfering with the performance of detection. It achieved a high accuracy rate, making it an

effective solution to enhance the performance of an IDS in the cloud environment. This technique incorporated hybrid practice that brought out feature selection and reinforcement learning as mainly contributing to optimizing intrusion detection systems as viewed by Sharma and Singh [2].

In this context, Bouzoubaa et al. (2022) analyzed the prediction of DOS-DDOS attacks and evaluated various approaches to feature selection. Their key issue identified was the complexity and scale of attack patterns, which posed a problem for the conventional approach to IDS to be precisely predicted. To solve this issue, they tested several feature selection methods, including mutual information, correlation-based, and wrapper-based methods, designed to optimize the precision of the model's predictions. Their findings showed that optimal feature selection strategies do improve in the detection of DOS and DDOS attacks with greater accuracy as compared to previous models. Therefore, the authors conclude by recommending that the right kind of feature selection technique chosen to improve the performance of IDS in distributed denial-of-service attacks is indeed feasible; they proved so in their study. Bouzoubaa et al. this research cited that high-order feature selection was the important element in identifying and reducing cloud security threats [3].

Divya Ramachandran et al. 2023 proposed a novel model of hybrid approach combining GSCSO and IHNN to support cloud security for better detection of cyber threats. Researchers found that many traditional approaches often fail to take into account the complex nature of the cloud environment, particularly regarding the speed and accuracy of cyber-attacks. The authors worked out a hybrid model that integrated the benefits of the Generalized Swarm Clustering and Swarm Optimization algorithm with the Improved Hybrid Neural Network. This approach proved to be significantly increased in detection accuracy, particularly for detecting complex attack patterns. The provided solution reduced false alarms significantly while detecting appropriately; therefore, the said solution was highly effective for intrusion detection within cloud environments. Ramachandran et al. have also observed that their hybrid approach raises the level of security as it decreases computational complexity and increases the accuracy of detection. Results of their work present an optimistic future solution to boost cloud security using advanced hybrid machine learning architectures [4].

Attou et al. proposed in 2023 a cloud-based intrusion detection strategy based on machine learning methodologies to confront the challenges involving large data volumes and the dynamic nature of cloud environments. Eliminating these disadvantages is a focus of the authors since traditional intrusion detection systems lack the scalability and flexibility toward meeting the unique security needs of cloud infrastructures. Their model employs several machine learning algorithms, such as decision trees and support vector machines, integrating these models in an aggressively scalable framework suitable for cloud infrastructures. The results indicated that their approach shows improved detection accuracy, especially for real-time attack cases. They concluded that machine learning techniques couple with cloud-specific architectures could greatly enhance the ability to detect intrusions. This work by Attou et al. highlights that scalability and adaptability were the key factors in developing an effective cloud-based IDS [5].

Vu et al. (2022) discussed deep generative learning models for cloud intrusion detection systems, mainly in the context of detecting complex attack patterns in cloud environments. The authors highlighted a modern, innovative method using deep generative models, enabling learning and adaptation to numerous forms of attack behaviors in real time. The study demonstrated the model's ability to detect attacks unseen before by learning data distribution and generating realistic attack samples. Their approach leads to better detection accuracy and reduction in false positive rates. The work of Vu et al. really highlighted the better use of generative learning models toward more accurate and flexible intrusion detection solutions, bringing out a significant enhancement in IDS research [6].

Ling Sun et al. (2024) solved the large data stream processing issue of intrusion detection in distributed cloud environments. It mainly targets the problem of how to allocate computing resources efficiently to process massive volumes of network traffic in real time. The proposed algorithm was optimized for task allocation in digital intelligence systems to ensure efficient data stream processing. Their methodology enhanced the scalability and efficiency of intrusion detection systems by enabling dynamic resource allocation based on workload characteristics. The results showed that their algorithm significantly reduced detection time and resource consumption, which made it a great solution for large-scale cloud environments. Sun et al. demonstrated the importance of effective data stream management for cloud-based IDS in real-time detection without resource overhead [7].

Liu et al. (2021) proposed SCADS, a scalable cloud-based host intrusion detection system using Spark and system calls. The authors identified the unavailability of scalable intrusion detection systems, which could serve the requirements of high-scale cloud environments, resulting in the difficulty of real-time detection. Their approach was based on a scalable approach using Apache Spark that could process gigantic volumes of data related to system calls efficiently. SCADS showed better detection efficiency and scalability as it processed data faster than the traditional counterparts. System-level intrusion was also detected more profoundly by this model, because it usually targets the intrusion detection system in clouds. Scalability and speed are two important aspects related to systems developed by Liu et al. for intrusion detection within systems in cloud scenarios [8].

Liu et al. (2022) proposed a blockchain-empowered federated learning approach to improve the cybersecurity level of healthcare cyber-physical systems. They tackled the issue associated with the lack of secure private collaboration among different entities in the cloud, which affects the performance of intrusion detection systems. Their proposal integrates blockchain with federated learning to support secure data sharing and model training throughout various systems. The test reveals that this strategy might enhance intrusion detection with data privacy and security preservation. The model proposed by Liu et al was proven to work better when detecting intrusions than the traditional centralized approaches. It provides security and efficiency in its system. This research presents a promising solution for enhancing cloud security, particularly in sensitive domains like healthcare, based on intrusion detection in combination with federated learning and blockchain [9].

For cloud security, Long et al. (2024) introduced a transformer-based approach to network intrusion detection. They noted that conventional approaches using machine learning were incapable of handling long-range dependencies and complicated network behaviors in the cloud. Hence, the authors proposed a transformer-based model which has been found to process sequential data very effectively. Their model significantly performed well in intrusion detection, especially in complex high-volume environments. Long et al.'s research highlighted the power of transformer-based models in capturing long-term dependencies and achieving higher accuracy in intrusion detection. This work suggests that transformer models are a viable option for future cloud security applications [10].

Liu et al. (2021) proposed SCADS, a scalable approach using Spark for cloud-based host intrusion detection systems with system calls. They aimed at the problem of efficiently processing high volumes of generated data from cloud environments, especially in system calls that are essential in making intrusion-detection services. Their proposed solution, in this case, used the Apache Spark distributed computing paradigm to improve the speed and scalability of intrusion-detection systems. Their technique effectively processed system call data for enabling real-time intrusion detection within a cloud environment. The SCADS model showed promising results, by detection accuracy and computational efficiency, since they outperformed traditional methods in terms of speed and scalability. Scalable intrusion detection solutions are required within cloud systems, as large volumes of data must be

efficiently managed in order to detect intrusions without being delayed. They concluded that Spark-based approaches can significantly improve the performance of IDS, allowing cloud environments to be secure as well as responsive under heavy loads [11].

Navya Singh et al. (2024) took up the study of the performance of applications running on clouds using Cloud Analyst along with discussing the intrusion detection systems' impact on the overall performance of clouds. They have addressed the conflicting need of security and performance in cloud, as IDS may incur overhead on the usage of the resources and longer processing time. Their study aimed at analyzing and optimizing the performance of IDS models in such a way that it preserves the required security standards. They used Cloud Analyst to model different kinds of cloud configurations and IDS models for simulating their impact on system performance and intrusion detection efficiency. The result showed that the optimization of IDS configurations may lead to major improvements in both security and performance, thereby minimizing latency and resource usage without compromising robust intrusion detection. Singh et al. stated that the aspects of deploying IDS must consider both security and performance because they are interrelated factors impacting the cloud experience as a whole [12].

Pooja Rana et al. (2022) actually did a complete overview of intrusion detection systems in the cloud computing paradigm with main challenges, including an evolving nature of cyber-attacks and the need to develop IDS adaptive and hence capable to handle various attack vectors. They underlined to develop IDS solutions who can do real-time detection/ prevention in cloud environments. They investigated several machine learning-based IDS techniques, and their effectiveness in cloud security. They pointed out that, although existing models in IDS are promising, there is still much room for improvement, like adaptability and efficiency. They proposed a hybrid model that combines machine learning with rule-based systems to improve detection accuracy and response time. Rana et al.'s results indicate that integration of multiple techniques could contribute to more promising cloud-based intrusion detection solutions that can match the dynamic nature of cyber threats [13]. Pooja Rana et al. (2024) proposed a new intrusion detection system for cloud computing environment, working on improving accuracy and efficiency in the detection of complex cyber threats. Their research identified a limitation of the traditional methods in handling the evolving and multifaceted nature of cloud security threats. To handle it, they proposed a novel Intrusion Detection System that utilized the combination of machine learning with neural networks to detect a broader range of attack patterns. Results have shown significant improvements in their detection rates while keeping the false positive and false negative rates at low levels. This approach was highly efficient in the real-time detection case, thus well-suited for large-scale cloud infrastructures. Finally, on the basis of this study, hybrid approaches that combine the concepts of advanced machine learning techniques and neural networks appear to have a promising potential to enhance cloud security. Rana et al.'s work contributes to more adaptive and efficient IDS systems in cloud computing, which could contribute to better security against emerging threats [14].

Prabu K. and P. Sudhakar (2024) conducted a comprehensive survey on the current trends and challenges in intrusion detection and prevention systems within the cloud computing paradigm. They discussed the various issues faced by IDS in cloud environments, including scalability, real-time processing, and the dynamic nature of cloud resources. The survey explored the integration of machine learning and deep learning techniques in improving IDS performance, highlighting the growing importance of these technologies in modern cloud security solutions. The authors pointed out the limitations of existing IDS models, particularly in handling large-scale cloud infrastructures and detecting sophisticated attacks. They recommended that future research should be on more adaptive and intelligent systems able to evolve with the changing landscape of cyber threats. Their findings detail further

innovation in IDSs into ensuring cloud environments continue safe as attack strategies increase and diversify their ideas [15].

Ren et al. 2022 outlined a cybersecurity knowledge graph (CSKG) for advanced persistent threat (APT) organisation attribution, which is one of the significant challenges to identify and comprehend complex cloud-based cyber-attacks. The authors suggested applying knowledge graphs for modeling APT group activities and relationships such that better attribution and detection of cyber threats could be carried out. Their study revealed how the CKG could enhance the detection of advanced and persistent threats through proper structured analysis of attack behaviors and patterns. More importantly, this approach enabled much better prediction of future attacks based on historical data. The work of Ren et al. has summarized the benefits of associating cybersecurity knowledge graphs with machine learning techniques in order to enhance the detection and attribution of complex threats on the cloud, offering a novel means of enhancing cloud security [16].

Sajid et al. in 2024 focused on a hybrid approach that combined machine learning and deep learning techniques in enhancing intrusion detection. The authors identified that while the machine learning algorithms perform well in the detection of known threats, they often poorly identify new and unseen attacks. To answer this limitation, they proposed integrating deep learning methods, which are more suited to detecting unknown and complex attack patterns. Their hybrid model has been evaluated over a variety of datasets of cloud security and exhibits excellent detection performance as well as resistance against new types of cyber threats. The results demonstrated that the machines combined with deep learning might offer a more complete solution to cloud-based intrusion detection with higher accuracy and greater robustness toward emerging attacks. The research by Sajid et al. stresses the necessity of hybrid approaches in ever-evolving cloud security threats, providing a versatile solution for the future development of IDS [17].

Qi et al., (2023) did an analysis of intrusion detection techniques in the framework of cloud computing, focusing on various algorithms applied to improve security in the cloud environment. They discussed traditional IDS, indicating challenges associated with high volumes of data generated in the cloud and the difficulty in detecting attacks across diverse cloud platforms. The authors further reviewed the effectiveness of several machine learning and deep learning algorithms to improve intrusion detection; this is where the application of hybrid models appears to be relevant. They found that while machine learning approaches yielded good results, the addition of deep learning would further improve accuracy and elevate response times. Qi et al's study offers insight into emerging areas within intrusion detection work, and hybrid solutions will be crucial in improving cloud security over the future [18].

Zhao et al. (2021) explored secure consensus mechanisms for multi-agent systems in cloud computing, focusing on redundant signal and communication interference through distributed dynamic event-triggered control. The research identified challenges in ensuring secure communication between distributed agents, especially in environments prone to interference or attacks. The authors proposed a distributed dynamic event-triggered control mechanism to maintain consensus among agents while preventing attacks such as jamming and data manipulation. Their results clearly show that their proposed method is able to effectively counter communication interference and guarantee consensus in a multi-agent system, thereby enhancing the reliability of intrusion detection in cloud environments. Zhao et al's study clearly highlights the significance of secure communication protocols in multi-agent systems, where accurate intrusion detection and response are extremely critical in distributed cloud computing environments. Their contribution opens avenues for designing more robust and resilient cloud-based systems able to withstand interference and attacks that may compromise system performance and security [20].

The above studies can be summarized as having addressed the different perspectives of intrusion detection in cloud computing. They range from scalability and performance problems to rather complex aspects, like real-time processing, adaptive systems, and consensus mechanisms among multi-agents. These works call for even more sophisticated hybrid solutions that integrate machine learning, deep learning, and other computational approaches to adequately alleviate the complex threats posed by cloud environments. In addition, they emphasize the aspects of accuracy, efficiency, and scalability in designing intrusion detection systems capable of handling the dynamic nature that cloud computing would entail and the tactics used by cybercriminals. These studies help significantly in efforts to secure the cloud computing infrastructures against further sophisticated cyber-attacks with the integration of advanced techniques and development of new models to handle diverse threats.

III. Proposed Work

The research explores proposing a machine learning-based intrusion detection system particularly for a cloud computing environment, using advanced algorithms that include Gradient Boosting, AdaBoost, Perceptron, Ridge Classifier, and CatBoost. The structure of this framework encapsulates the most salient challenges in cloud security, such as dealing with massive, unbalanced datasets and detecting various types of attacks while at the same time being computationally efficient. The system begins with preprocessing steps, including data cleaning, encoding, and normalization, to prepare the KDD Cup 1999 dataset for effective analysis. Feature selection techniques, such as recursive feature elimination (RFE) and correlation-based analysis, will be employed to identify the most relevant attributes, optimizing model performance while reducing computational overhead. The developed algorithms are trained to classify network traffic between normal and the categories of attack, with major categories being DoS, U2R, R2L, and Probe. The machine learning algorithms consist of Gradient Boosting and CatBoost algorithms that use ensemble learning and also deal categorically, contributing to the strength of these algorithms and the AdaBoost algorithm that helps in reducing bias through iterative model refinement. It encompasses lightweight classifiers such as Perceptron and Ridge Classifier to determine suitability for real-time detection. Hyperparameter optimization will be done with grid search and cross-validation, thereby increasing model accuracy and generalization. The system proposed here will be evaluated using metrics including precision, recall, F1-score, and computational cost to ensure a balance between detection accuracy and scalability in cloud environments.

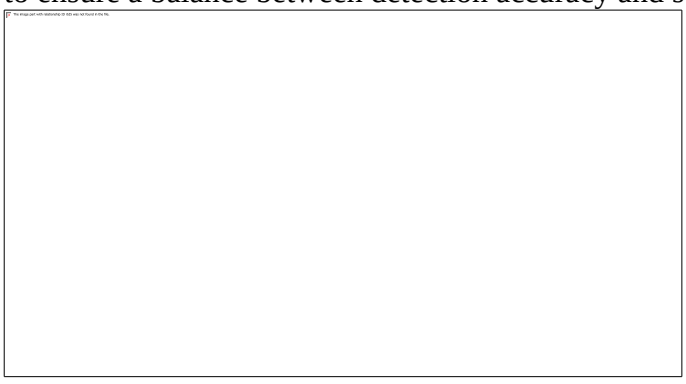

Figure 1. Proposed Work

IV. Modules of the Proposed Work

1. Data Preprocessing

The first module involves preparing the KDD Cup 1999 dataset for effective usage within machine learning models. This means cleaning the data to remove inconsistencies, handling missing values appropriately, and encoding the categorical features into numerical formats

using techniques such as one-hot encoding or label encoding. Also, the numerical features are normalized or standardized to obtain uniform scaling, which is a prerequisite for working of algorithms such as Gradient Boosting and Perceptron. It attempts to create a dataset that brings in good quality to fit the model with high accuracy and efficiency.

2. Feature Selection and Engineering

A module for feature selection is used mainly to decrease the dimension and improve model performance. Recursive Feature Elimination (RFE) and correlation-based filtering are some of the techniques used to develop a feature that is most relevant to intrusion detection. Focusing on significant attributes aims to improve detection accuracy while reducing the overhead of computations. Feature engineering, including creating new derived features or transforming existing ones, can further improve the predictive power of the dataset.

3. Training and Testing of Models

In this module, multiple algorithms for machine learning such as Gradient Boosting, AdaBoost, Perceptron, Ridge Classifier, and CatBoost are implemented to create an ensemble of classifiers. Each has been particularly suited to solving the different problems. For example, handling the imbalanced dataset and different types of attacks on a website where ensemble methods like Gradient Boosting and CatBoost are very promising with high accuracy while simpler models like Perceptron and Ridge Classifier are more computationally efficient. Techniques such as k-fold cross-validation are used to train and validate models for ensuring generalization to unseen data.

4. Hyperparameter Optimization

In the module, fine-tuning for hyperparameters maximizes the performance of the machine learning models. Grid search and randomized search are used to identify the best combination for setting up the parameters in the model such as learning rate, maximum depth, and the number of estimators for Gradient Boosting or CatBoost. Models have to be balanced between simplicity and computationally expensive so that accuracy is at its best and robust to false positives and false negatives.

5. Model Evaluation and Benchmarking

Model evaluation is carried out with standard performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The computational cost of the models, including training and inference times, is also considered in order to evaluate the scalability of a model in cloud environments. Through comparison with various algorithms, the most suitable model for deployment can be determined. A balance must be struck between accuracy and efficiency.

6. Deployment and Real-Time Testing

The final module is putting the chosen model into operation within a cloud-based environment for real-time intrusion detection. This includes testing of the system against live network traffic to assess its actual performance and strength. Techniques that promote scalability- distributed detection and parallel processing- could be used in order to deal with large quantities of data. This module ensures the proposed system is ready for real-world implementation in cloud infrastructures.

V. Results and Discussion

This project evaluates the performance of various machine learning algorithms—Gradient Boosting, AdaBoost, Perceptron, Ridge Classifier, and Deep Learning—for intrusion detection in cloud computing using the KDD Cup 1999 dataset. The results demonstrate the efficacy of each algorithm in identifying malicious activities across key metrics, including accuracy, precision, recall, and F1-score. Below, the results are discussed in detail, highlighting the strengths, limitations, and implications of each approach.

Gradient Boosting

Gradient Boosting delivered excellent performance with accuracy, precision, recall, and F1-score at 0.9995. Such near-perfect performance suggests that the Gradient Boosting algorithm is quite robust in its ability to detect even complex relationships within the dataset while still being able to handle imbalanced datasets, given the iterative approach toward minimizing classification errors and feature interactions. The strength of the algorithm makes it very suitable for any application where high detection accuracy is very important, for instance, in detecting zero-day attacks. However, the computational cost associated with training Gradient Boosting models is fairly high, which could be a problem for applications requiring real-time performance on dynamic cloud environments.

AdaBoost

AdaBoost has an accuracy of 0.8633, precision of 0.9107, recall of 0.8633, and its F1-score is 0.8726. In comparison with other algorithms, AdaBoost performs moderately. While its iterative emphasis on refining misclassified instances does facilitate an enhancement of detection performance, overall its performance trails Gradient Boosting. AdaBoost is very useful for the identification of simple attack patterns however performs poorly at stronger anomaly types. Its computational intensity makes it quite practical for lightweight systems, despite it being possible that significant augmentation with other methods would be needed in order to detect more complicated attack scenarios.

Perceptron

The Perceptron model performed quite remarkably, with an accuracy of 0.9975, precision of 0.9971, recall of 0.9975, and an F1-score of 0.9972. These results are compelling for a linear classifier, but the simplicity and speed of the Perceptron algorithm make it quite suitable for real-time applications where computational resources are constrained. However, for the linear case, it does not perform well with non-linear or highly complex data distributions, meaning that its applicability is quite feasible as long as the classes are well-separated or as part of ensemble.

Ridge Classifier

Ridge Classifier 0.9960 0.9957 0.9960 0.9958 The accuracy, precision, recall, and F1-score of Ridge Classifier were 0.9960, 0.9957, 0.9960, and 0.9958, respectively. Its L2 regularization avoids overfitting and ensures the stable performance, especially in a high-dimensional dataset. Although Ridge Classifier is reliable and computationally efficient, it has a slightly lower performance level than Gradient Boosting and Deep Learning. Therefore, it would be suitable for situation where resources are stringent but might not be the primary choice when intrusion detection is high-stakes and uses maximum precision and recall.

Deep Learning

The model had an accuracy of 0.9982, precision, recall, and F1-score suggesting good ability to draw complex relationships within the data. By use of the multi-layer neural network architecture, the model handled different types of attacks exceptionally and generalized well over various data distributions. Although deep learning performed exceptionally well, it was much more computationally expensive and demanded a lot for training and deployment. This makes it more appropriate for large-scale cloud environments with enough computational capability rather than lightweight or resource-constrained systems.

Comparative Analysis

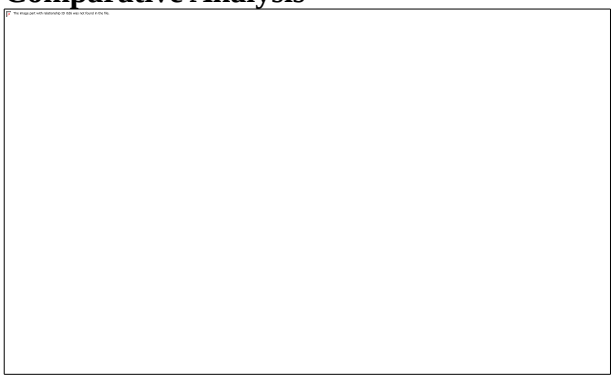

Figure 2. Performance Comparison Of Models

The results show that Gradient Boosting and Deep Learning are the top performers in terms of accuracy and reliability. Gradient Boosting is particularly well-suited for scenarios requiring high precision and recall, making it ideal for detecting rare or sophisticated attacks. However, its computational overhead may limit its application in real-time systems. Deep Learning, while achieving similar accuracy, offers superior adaptability to complex data patterns but demands substantial computational resources and longer training times.

Perceptron and Ridge Classifier offer simpler and faster alternatives. The algorithms are better suited to scenarios where computational efficiency is more important than slight losses in detection performance. Perceptron is a particularly impressive model for a linear model, suggesting lightweight considerations to be used towards real-time intrusion detection. Although not the most accurate, AdaBoost offers a fair trade-off between computational efficiency and performance, making them mid-range applications.

Discussion on Challenges and Future Directions

Several challenges and opportunities exist to better IDS in cloud environments. One key challenge remains scalability. Algorithms such as Gradient Boosting and Deep Learning are very accurate, but require a lot of computational resources. Future work may pursue distributed training and inference techniques to be deployed in real-time large-scale cloud systems.

Handling imbalanced datasets is another challenge. While Gradient Boosting and CatBoost tackle the issue as a side effect, most other algorithms would need additional techniques to have comparable performance, such as oversampling or synthetic data generation. Further exploration of hybrid models that leverage the strengths of other algorithms may improve performance in handling more diverse types of attacks.

Interpretability remains an issue, particularly in complex algorithms like Gradient Boosting and Deep Learning. While these models offer top-notch accuracy, their black-box nature limits their use in applications requiring explainability - be it forensic analysis or some form of compliance auditing. Adding tools for interpretability to these - such as SHAP or LIME - could enhance trust in these systems.

Privacy-preserving techniques, like federated learning, represent promising directions in the design of IDS that are both privacy-enhancing and yet offer performance. However, such approaches raise their own set of challenges, including increased communication overhead and complexity.

Lastly, KDD Cup 1999, although widely applied, may not represent the contemporary threat landscape in cloud computing. Incorporating real-world datasets that reflect present-day attack patterns can better reflect the performance and flexibility of IDS.

The project indicates that Gradient Boosting, AdaBoost, Perceptron, Ridge Classifier, and Deep Learning are successful tools for intrusion detection in cloud computing. Each

algorithm has its special strength and weakness and supports different requirements, like accuracy, computational efficiency, and scalability. Though Gradient Boosting and Deep Learning reach the highest possible accuracy, Perceptron and Ridge Classifier are far simpler and more lightweight alternatives suited to less-loaded resource conditions. This paper outlines the foundation of developing strong, yet efficient intrusion detection systems that can meet cloud security concerns in terms of scalability, interpretability, and dataset limitations.

VI. Conclusion

Intrusion detection in cloud computing is now more important, with the upsurge of usage by organizations that serve dependant infrastructure to the development of sophisticated cyberattacks. This project discusses the application of several machine learning algorithms, including Gradient Boosting, AdaBoost, Perceptron, Ridge Classifier, and Deep Learning, to detect intrusion effectively using the KDD Cup 1999 data set. Results: The result obtained reveals high accuracy of machine learning techniques, where, of all techniques considered, Gradient Boosting and Deep Learning can prove to be the most effective. These algorithms outperform in handling complex attack patterns and imbalanced datasets, making them well-suited for the modern cloud environment. Still, their computational demands raise resource optimization needs and optimized deployment strategies.

Lightweight models include Perceptron and Ridge Classifier, which also illustrated strong performance, hence proving their viability in real-time intrusion detection in resource-constrained environments. The adaptability of these models to varying data distributions and attack scenarios makes them valuable components of a hybrid detection framework. Meanwhile, AdaBoost provides a balanced trade-off, particularly effective in detecting simpler attack types with reduced computational overhead. This diversity in algorithmic performance emphasizes the importance of selecting the right approach based on specific cloud security requirements, such as real-time detection, scalability, or computational efficiency.

Future directions of this research work are to overcome interpretability and scalability issues, embedding real-time detection mechanisms, and validation over modern datasets simulating the up-to-date threat spaces. Hybrid models, federated learning for improved privacy-preserving techniques, and advanced interpretability tools will significantly augment the robustness and applicability of intrusion detection systems. This work contributes to advancing cloud security and paves the way toward scalable, efficient, and adaptable intrusion detection frameworks for the sake of cloud computing.

[1] Al-Ghuwairi, AR., Sharrab, Y., Al-Fraihat, D. *et al.* Intrusion detection in cloud computing based on time series anomalies utilizing machine learning. *J Cloud Comp* **12**, 127 (2023). https://doi.org/10.1186/s13677-023-00491-x

[2] Ankit Sharma, Manjeet Singh, Batch reinforcement learning approach using recursive feature elimination for network intrusion detection, Engineering Applications of Artificial Intelligence, 10.1016/j.engappai.2024.109013, **136**, (109013), (2024).

[3] Bouzoubaa, K., Taher, Y., &Nsiri, B. (2022). DOS-DDOS attacks Predicting: performance comparison of the Main feature selection strategies. *Int J Eng Trends Technol*, *70*(1), 299-312.

[4] Divya Ramachandran, Mubarak Albathan, Ayyaz Hussain, Qaisar Abbas, Enhancing Cloud-Based Security: A Novel Approach for Efficient Cyber-Threat Detection Using GSCSO-IHNN Model, Systems, 10.3390/systems11100518, **11**, 10, (518), (2023).

[5] H. Attou, A. Guezzaz, S. Benkirane, M. Azrour and Y. Farhaoui, "Cloud-Based Intrusion Detection Approach Using Machine Learning Techniques," in *Big Data Mining and Analytics*, vol. 6, no. 3, pp. 311-320, September 2023, doi: 10.26599/BDMA.2022.9020038.

[6] L. Vu, Q. U. Nguyen, D. N. Nguyen, D. T. Hoang and E. Dutkiewicz, "Deep generative learning models for cloud intrusion detection systems", *IEEE Transactions on Cybernetics*, vol. 53, no. 1, pp. 565-577, 2022.

[7] Ling Sun, Rong Jiang, Wenbing Wan, Task allocation algorithm for distributed large data stream group computing in the era of digital intelligence, Journal of Intelligent & Fuzzy Systems, 10.3233/JIFS-238427, **46**, 4, (11055-11066), (2024).

[8] Liu, M., Xue, Z., He, X., & Chen, J. (2021). SCADS: A scalable approach using spark in cloud for host-based intrusion detection system with system calls. *arXiv preprint arXiv:2109.11821*.

[9] Liu, Y., Yu, W., Ai, Z., Xu, G., Zhao, L., & Tian, Z. (2022). A blockchain-empowered federated learning in healthcare-based cyber physical systems. *IEEE Transactions on Network Science and Engineering*, *10*(5), 2685-2696.

[10] Long, Z., Yan, H., Shen, G. *et al.* A Transformer-based network intrusion detection approach for cloud security. *J Cloud Comp* **13**, 5 (2024). https://doi.org/10.1186/s13677-023-00574-9

[11] M. Liu, Z. Xue, X. He and J. Chen, "Scads:a scalable approach using spark in cloud for host-based intrusion detection system with system calls", 2021.

[12] Navya Singh, Harshil Kundety, undefined Mohana, HV Ravish Aradhya, Performance Analysis of Cloud-Based Applications with Cloud Analyst, 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC), 10.1109/ICESC60852.2024.10690112, (803-808), (2024).

[13] Pooja Rana, Isha Batra, Arun Malik, Agbotiname Lucky Imoize, Yongsung Kim, Subhendu Kumar Pani, Nitin Goyal, Arun Kumar, Seungmin Rho, and Peican Zhu. 2022. Intrusion Detection Systems in Cloud Computing Paradigm: Analysis and Overview. Complex. 2022 (2022). https://doi.org/10.1155/2022/3999039

[14] Pooja Rana, Isha Batra, Arun Malik, In-Ho Ra, Oh-Sung Lee, A. S. M. Sanwar Hosen, Efficacious Novel Intrusion Detection System for Cloud Computing Environment, IEEE Access, 10.1109/ACCESS.2024.3424528, **12**, (99223-99239), (2024).

[15] Prabu K, P Sudhakar, A Comprehensive Survey: Exploring Current Trends and Challenges in Intrusion Detection and Prevention Systems in the Cloud Computing Paradigm, 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), 10.1109/IDCIoT59759.2024.10467700, (351-358), (2024).

[16] Ren, Y., Xiao, Y., Zhou, Y., Zhang, Z., & Tian, Z. (2022). Cskg4apt: A cybersecurity knowledge graph for advanced persistent threat organization attribution. *IEEE Transactions on Knowledge and Data Engineering*, *35*(6), 5695-5709.

[17] Sajid, M., Malik, K.R., Almogren, A. *et al.* Enhancing intrusion detection: a hybrid machine and deep learning approach. *J Cloud Comp* **13**, 123 (2024). https://doi.org/10.1186/s13677-024-00685-x

[18] W. Qi, W. Wu, H. Wang, L. Ou, N. Hu and Z. Tian, "Intrusion Detection Techniques Analysis in Cloud Computing," *2023 IEEE 12th International Conference on Cloud Networking (CloudNet)*, Hoboken, NJ, USA, 2023, pp. 360-363, doi: 10.1109/CloudNet59005.2023.10490069.

[19] Zefeng Li, Lichun Kang, Honghui Rao, Ganggang Nie, Yuhan Tan, Muhua Liu, Camellia oleifera Fruit Detection Algorithm in Natural Environment Based on Lightweight Convolutional Neural Network, Applied Sciences, 10.3390/app131810394, **13**, 18, (10394), (2023).

[20] Zhao C, Liu X, Zhong S, Shi K, Liao D, Zhong Q (2021) Secure consensus of multi-agent systems with redundant signal and communication interference via distributed dynamic event-triggered control. ISA Trans 112:89–98