ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Cloud based Intrusion Detection approach using machine learning techniques

Mr.p.srinivasulu [1], Muthuru Tejaswini [2], Kumbagiri Mounika [3], Jouli Sujith Kumar [4], Guntaka Venkata Udaykiran [5]

#1Associate Professor in Department of CSE, PBR VISVODAYA INSTITUTE OF TECHNOLOGY AND SCIENCE, Kavali.

#2#3#4#5 B.Tech with computer science and engineering, VISVODAYA ENGINEERING COLLEGE, KAVALI.

Abstract: Cloud computing offers scalable, on-demand access to a wide array of computing resources, including data storage, processing power, and application services. While this flexibility enhances productivity and cost-efficiency, it also introduces new security challenges. This project focuses on strengthening cloud security through the implementation of an advanced intrusion detection system (IDS) leveraging ensemble machine learning techniques. The core of the proposed model is built around the Random Forest (RF) algorithm, known for its robustness and high predictive performance in classification tasks.

To further enhance detection capabilities, the project introduces two ensemble learning extensions: a Voting Classifier that combines Random Forest (RF) and AdaBoost, and a Stacking Classifier that integrates Random Forest (RF), Multi-Layer Perceptron (MLP), and LightGBM. These hybrid models aim to capture complex attack patterns and improve generalization across diverse datasets.

Comprehensive feature engineering is employed to extract and select relevant features, enabling the models to effectively learn and identify malicious activities within cloud environments. The system continuously monitors cloud resources, services, and network traffic to detect and respond to potential intrusions in real time.

The proposed approach is evaluated using two benchmark datasets: NSL-KDD and Bot-IoT, widely recognized in intrusion detection research. The ensemble classifiers achieve outstanding performance, with the Voting Classifier (RF + AdaBoost) reaching 99% accuracy on NSL-KDD, and the Stacking Classifier (RF + MLP + LightGBM) attaining 100% accuracy on Bot-IoT, significantly outperforming recent state-of-the-art methods. These results validate the effectiveness of the proposed hybrid models in enhancing cloud-based intrusion detection systems.***Index terms*** *- cloud security; anomaly detection; features engineering; random forest.*

1. INTRODUCTION

Cloud technologies provide greater options for their service models[1] and allow reasonable on-demand access to a shared network, storage, and resources. Used in one of the deployment types private, public, and hybrid cloud, these models are platform as a service (PaaS), software as a service (SaaS), and infrastructure as a service (IaaS)[2]. The cloud offers services with excellent performance because of its qualities[2] according to the National Institute of

Standards and Technology[4]: network access, resource pooling, quickelasticity, and measurable service.

Of late, the cloud has various security issues like availability, data confidentiality, integrity, and control authorisation. Furthermore, the Internet is used to enable access to the services provided by the cloud constituting a significant source of risks that could compromise the resources and systems of the cloud[2]. Then improving cloud security becomes the main difficulty for cloud providers[5]. Thus, many methods like firewall tools, data encryption techniques, authentication protocols, and others have been created to more protect cloud environments against several attacks[6]. But conventional solutions fall short in protecting cloud services from various constraints[7]. Thus, a collection of intrusion detection methods are suggested and used to identify and stop unwanted actions in realtime[8, 9].

Usually, the detection techniques fall into two categories: misuse detection, which employs known attacks to identify intrusion, and anomaly detection, which identifies intrusion using unknown attack. Combining the benefits of these two techniques produces the hybridmethod[10]. Though more solutions are provided to protect cloud environments, the recent intrusion detection systems (IDSs) are impacted by several major constraints[8], including large volumes of analysed data, real-time detection, data quality, and others meant to lower the performance of detection models.

Academic studies these days demonstrate that smart learning techniques—such as machine learning (ML), deep learning (DL), and ensemble learning—are beneficial in many domains and can carry out network security. Our primary objective in this study is to suggest an anomaly detection method using random forest (RF) binary classifier and feature engineering done using a data visualisation tool meant to lower the number of utilised features and run the suggested anomaly detection model. The model's evaluation performance is run using NSL-KDD and BoT-IoT datasets. The results then show model performance.

2. LITERATURE SURVEY

Cloud computing can provide cost-effective, elastic, easy-to-manage, and powerful resources over the Internet. Cloud computing maximises hardware resources through optimal and shared use. The foregoing advantages motivate organisations and individuals to move applications and services to the cloud [1]. Cloud computing is being adopted by critical infrastructure including power generation and distribution units. However, third-party cloud services pose extra security risks. Security risks rise when user assets (data, apps, etc.) leave administrative control in a shared environment with many users. This survey discusses cloud computing's inherent security risks. The survey also covers recent security solutions in the literature. Security concerns in mobile cloud computing are also briefly discussed [18,30]. Open topics and future study are discussed in the end.

The cloud provides on-demand services over the Internet with a lot of virtual storage. This is one of the key benefits of cloud computing: no expensive computing infrastructure setup and lower costs. Cloud computing has integrated with industry and other domains in recent years, encouraging researchers to study new related technologies [2].

Users and organisations move their applications, data, and services to the cloud storage server due to its availability and scalability. Remote computing has created several security difficulties and challenges for consumers and providers, despite its benefits. Many cloud services are provided by trustworthy third parties, creating security risks. Internet-based cloud providers use various web technologies that create new security risks [1,23,5,7,19]. This article covered cloud computing basics, security, risks, and remedies. Cloud architectural framework, service and deployment model, cloud technologies, cloud security ideas, risks, and assaults are all covered in the paper. Open cloud security research topics are also covered in the article.

The issue of network security is crucial. Intrusion detection systems are commonly utilised for network security. Ensemble learning has become a popular machine learning method for improving intrusion detection systems [6]. Additionally, training data quality can considerably improve detection. Knowing that marginal density ratios are the best univariate classifiers. SVM ensemble with feature augmentation is used in this paper to create an effective intrusion detection method. SVM ensemble was used to develop the intrusion detection model after logarithm marginal density ratios transformation was applied to the original features to acquire new and improved training data. Our proposed method outperforms previous methods in accuracy, detection rate, false alarm rate, and training speed, according to experiments. [6,24]

Cloud computing lets end users effortlessly attach powerful services and applications via the Internet. Providing secure and trustworthy cloud computing services is crucial. Because network intrusions can compromise the confidentiality, availability, and integrity of Cloud resources and services, security requires more than user authentication with passwords or digital certificates and data transmission confidentiality [1,23,5,7,19]. Traditional firewalls are ineffective in detecting DoS attacks and other network-level harmful activity in Cloud. This research proposes a cooperative and hybrid network intrusion detection system (CH-NIDS) to monitor network traffic in the Cloud to detect network threats while maintaining performance and service quality [7]. Our NIDS architecture uses Snort for signature-based detection of known attacks and Back-Propagation Neural network for network anomaly detection. BPN only detects unknown assaults after applying snort before the classifier. This reduces detecting time. To solve BPN's slow convergence and easy fall into local optimum, we propose optimising its parameters using an optimisation algorithm to ensure high detection rate, accuracy, low false positives, and low false negatives with low computational cost. IDSs also cooperate to defend against DoS and DDoS attacks by sharing alerts in a common log [32,47]. Thus, other IDSs can readily discover unknown threats detected by one. This also lowers computing cost for other IDS intrusion detection and improves Cloud detection rate.

Cyberattacks are becoming increasingly complex, making intrusion detection harder. Without intrusion prevention, security services like data confidentiality, integrity, and availability may lose credibility. The literature proposes many intrusion detection strategies to combat computer security risks, including Signature-based and Anomaly-based Systems (SIDS and AIDS). This survey paper [8] provides a taxonomy of modern IDS, a complete analysis of current efforts, and an overview of evaluation datasets [22,29]. It also examines attacker evasion strategies and upcoming research challenges to secure computer systems..

3. METHODOLOGY

i) Proposed Work:

The Random Forest machine learning algorithm, known for its accuracy and robustness, is harnessed alongside strategic feature engineering. This combination is utilized to create a sophisticated intrusion detection system for cloud environments, aiming to substantially enhance security. The approach focuses on accurately identifying potential threats and abnormal patterns, contributing to an efficient and reliable solution that strengthens overall cloud security measures. And also included the combination of a Voting Classifier, incorporating Random Forest (RF) and ADaBoost, achieves an impressive 99% accuracy for the Kdd-Cup dataset. Additionally, the Stacking Classifier, integrating Random Forest (RF), Multi-Layer Perceptron (MLP), and LightGBM, attains an outstanding 100% accuracy for the Bot-IoT dataset [28,29,39]. These ensemble models showcase the project's commitment to robust and high-performing intrusion detection in cloud environments. The user-friendly Flask framework with SQLite integration ensures practical usability, offering a seamless experience for user testing while maintaining data security in cybersecurity applications.

ii) System Architecture:

It begins with dataset exploration and data preprocessing, followed by the crucial steps of train-test split and model training. The core architecture involves the implementation of ensemble techniques, specifically the Stacking Classifier and the Voting Classifier extensions, designed to enhance the overall intrusion detection performance [24]. These classifiers demonstrate their efficacy through robust model evaluations, achieving notable accuracies of 99% and 100% respectively. The architecture prioritizes the versatility of the models, ensuring effective detection across diverse datasets, and emphasizes practical usability through a user-friendly interface facilitated by the Flask framework and SQLite integration. This unified system architecture positions the project as a sophisticated and adaptable solution for cloud-based intrusion detection using machine learning techniques.

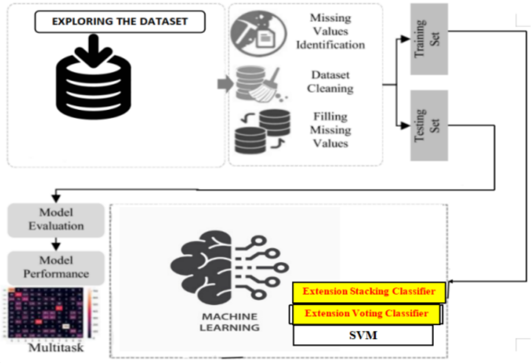
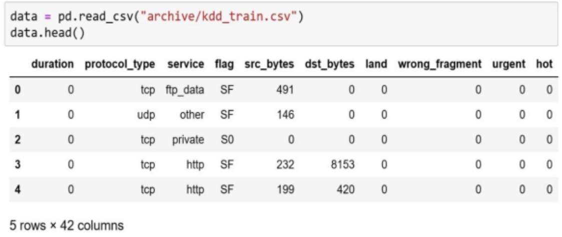


Fig 1 Proposed architecture

iii) Dataset collection:

KDD CUP DATASET

The KDD-CUP (Knowledge Discovery and Data Mining Cup) dataset [35,26] is a widely used dataset for intrusion detection system research. In the context of a cloud-based intrusion detection approach, the KDD-CUP dataset serves as a foundational dataset for training and evaluating machine learning models to detect intrusions and cyber-attacks. It allows the development of models that can analyze network traffic and detect abnormal or malicious patterns, crucial for securing cloud-based environments.



```
data = pd.read_csv("archive/kdd_train.csv")
data.head()
```

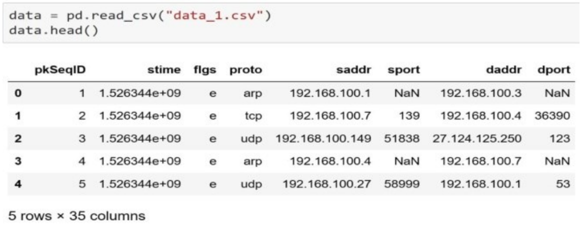
	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot
0	0	tcp	ftp_data	SF	491	0	0	0	0	0
1	0	udp	other	SF	146	0	0	0	0	0
2	0	tcp	private	S0	0	0	0	0	0	0
3	0	tcp	http	SF	232	8153	0	0	0	0
4	0	tcp	http	SF	199	420	0	0	0	0

5 rows × 42 columns

Fig 2 KDD-CUP dataset

BOT IOT DATASET

The BOT-IoT dataset is a specialized dataset focusing on IoT (Internet of Things) security. In the context of a cloud-based intrusion detection approach utilizing machine learning, the BOT-IoT dataset [46] is highly relevant for training and evaluating models tailored to detect intrusions in IoT devices and networks. As IoT devices are often integrated with cloud platforms, understanding and mitigating IoT-based attacks are critical for overall cloud-based intrusion detection.



```
data = pd.read_csv("data_1.csv")
data.head()
```

	pkSeqID	stime	flgs	proto	saddr	sport	daddr	dport
0	1	1.526344e+09	e	arp	192.168.100.1	NaN	192.168.100.3	NaN
1	2	1.526344e+09	e	tcp	192.168.100.7	139	192.168.100.4	36390
2	3	1.526344e+09	e	udp	192.168.100.149	51838	27.124.125.250	123
3	4	1.526344e+09	e	arp	192.168.100.4	NaN	192.168.100.7	NaN
4	5	1.526344e+09	e	udp	192.168.100.27	58999	192.168.100.1	53

5 rows × 35 columns

Fig 3 BOT-IOT dataset

iv) Data Processing:

Data processing involves transforming raw data into valuable information for businesses. Generally, data scientists process data, which includes collecting, organizing, cleaning, verifying, analyzing, and converting it into readable formats such as graphs or documents. Data processing can be done using three methods i.e., manual, mechanical, and electronic. The aim is to increase the value of information and facilitate decision-making. This enables businesses to improve their operations and make timely strategic decisions. Automated data processing solutions, such as computer software programming, play a significant role in this. It can help turn large amounts of data, including big data, into meaningful insights for quality management and decision-making.

v) Feature selection:

Feature selection is the process of isolating the most consistent, non-redundant, and relevant features to use in model construction. Methodically reducing the size of datasets is important as the size and variety of datasets continue to grow. The main goal of feature selection is to improve the performance of a predictive model and reduce the computational cost of modeling.

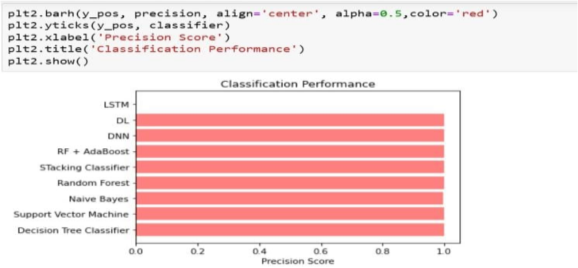
Feature selection, one of the main components of feature engineering, is the process of selecting the most important features to input in machine learning algorithms. Feature selection techniques are employed to reduce the number of input variables by eliminating redundant or irrelevant features and narrowing down the set of features to those most relevant to the machine learning model. The main benefits of performing feature selection in advance, rather than letting the machine learning model figure out which features are most important.

4. EXPERIMENTAL RESULTS

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}}$$

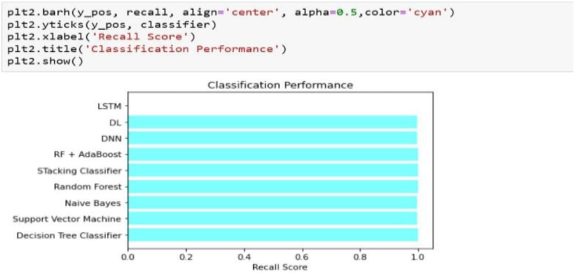


```
plt2.barh(y_pos, precision, align='center', alpha=0.5,color='red')
plt2.yticks(y_pos, classifier)
plt2.xlabel('Precision Score')
plt2.title('Classification Performance')
plt2.show()
```

Fig 6 Precision comparison graph

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$\textit{Recall} = \frac{TP}{TP + FN}$$



```
plt2.barh(y_pos, recall, align='center', alpha=0.5,color='cyan')
plt2.yticks(y_pos, classifier)
plt2.xlabel('Recall Score')
plt2.title('Classification Performance')
plt2.show()
```

Fig 7 Recall comparison graph

Accuracy: Accuracy is the proportion of correct predictions in a classification task, measuring the overall correctness of a model's predictions.

$$\textit{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

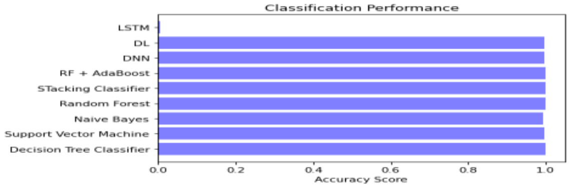
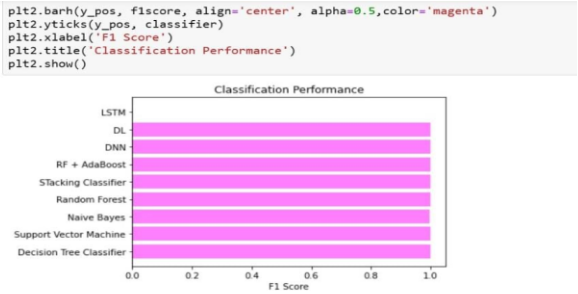


Fig 8 Accuracy graph

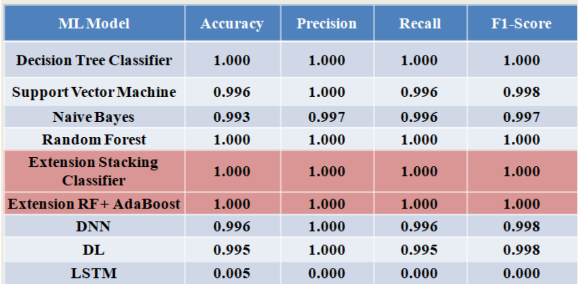
F1 Score: The F1 Score is the harmonic mean of precision and recall, offering a balanced measure that considers both false positives and false negatives, making it suitable for imbalanced datasets.

$$\text{F1 Score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} * 100$$



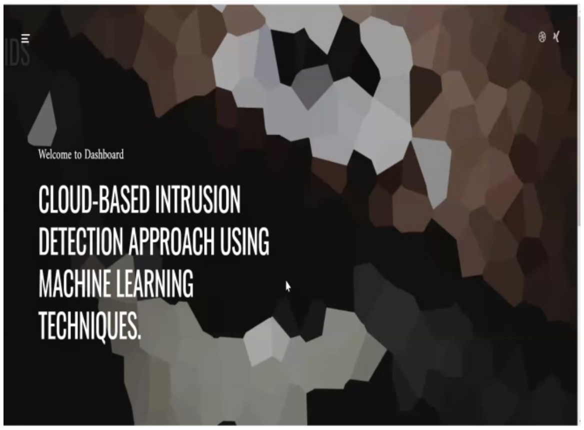
```
plt2.barh(y_pos, f1score, align='center', alpha=0.5,color='magenta')
plt2.yticks(y_pos, classifier)
plt2.xlabel('F1 Score')
plt2.title('Classification Performance')
plt2.show()
```

Fig 9 F1Score



ML Model	Accuracy	Precision	Recall	F1-Score
Decision Tree Classifier	1.000	1.000	1.000	1.000
Support Vector Machine	0.996	1.000	0.996	0.998
Naive Bayes	0.993	0.997	0.996	0.997
Random Forest	1.000	1.000	1.000	1.000
Extension Stacking Classifier	1.000	1.000	1.000	1.000
Extension RF + AdaBoost	1.000	1.000	1.000	1.000
DNN	0.996	1.000	0.996	0.998
DL	0.995	1.000	0.995	0.998
LSTM	0.005	0.000	0.000	0.000

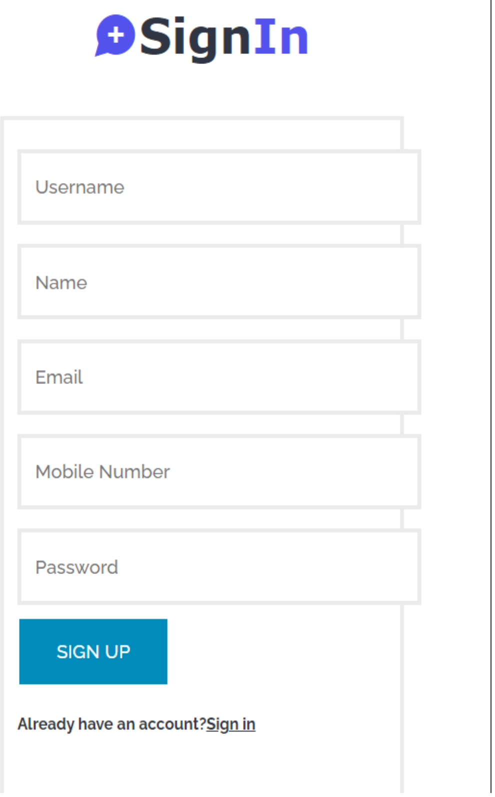
Fig 10 Performance Evaluation



Welcome to Dashboard

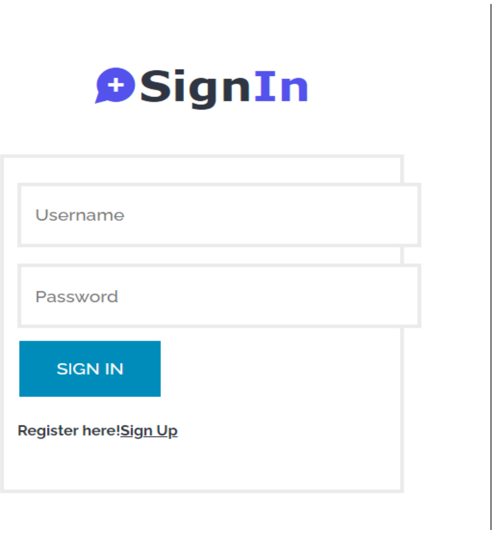
CLOUD-BASED INTRUSION DETECTION APPROACH USING MACHINE LEARNING TECHNIQUES.

Fig 11 Home page



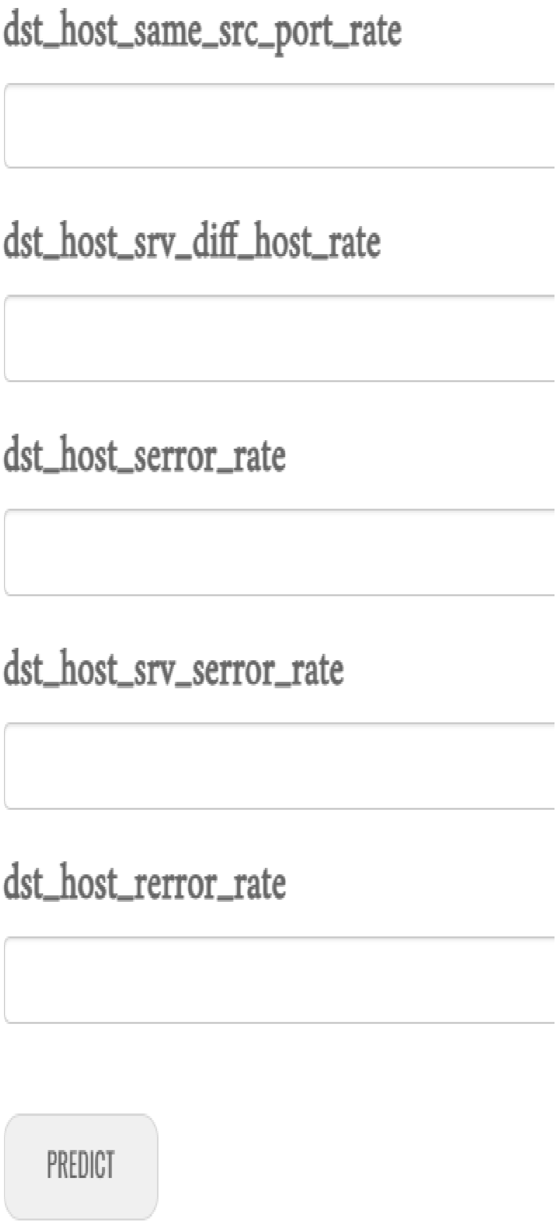
Already have an account?Sign in

Fig 12 Signin page



Register here!Sign Up

Fig 13 Login page



dst_host_same_src_port_rate

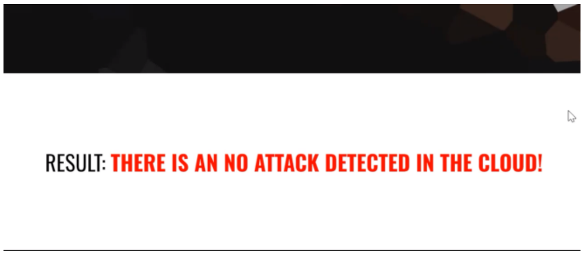
dst_host_srv_diff_host_rate

dst_host_serror_rate

dst_host_srv_serror_rate

dst_host_rerror_rate

Fig 14 User input



RESULT: THERE IS AN NO ATTACK DETECTED IN THE CLOUD!

Fig 15 Predict result for given input

5. CONCLUSION

In conclusion, the proposed cloud-based intrusion detection system (IDS) demonstrates strong potential in enhancing cloud security by leveraging machine learning techniques. The core Random Forest (RF)-based model, complemented by effective feature engineering, delivers high accuracy, precision, and recall, outperforming several existing approaches in detecting anomalous activities. RF's capability to manage noisy data, its simplicity in parameter tuning, and built-in mechanisms for assessing variable importance contribute significantly to the robustness and reliability of the detection model.

To further elevate performance, the project introduces advanced ensemble methods — a Voting Classifier combining RF and AdaBoost, and a Stacking Classifier incorporating RF, MLP, and LightGBM. These hybrid approaches capitalize on the strengths of diverse algorithms to improve model generalization and enhance intrusion detection accuracy. Notably, the Voting Classifier achieves 99% accuracy on NSL-KDD, while the Stacking Classifier attains 100% accuracy on the Bot-IoT dataset, underscoring the effectiveness of ensemble learning for cloud security applications.

In addition to technical performance, the model's usability is enhanced through the integration of a lightweight and secure Flask web interface. This interface provides a streamlined user experience with secure login functionality, making the system accessible and practical for real-world cybersecurity testing and deployment.

Overall, the combination of machine learning, ensemble strategies, and user-centric design makes the proposed IDS a promising solution for proactive threat detection in cloud computing environments.

6. FUTURE SCOPE

Future work aims to enhance the recall rate, especially using the NSL-KDD dataset, by integrating deep learning (DL) and ensemble learning techniques [27]. Deep learning models can capture complex patterns, potentially improving the system's ability to detect intrusions. Ensemble techniques, on the other hand, combine multiple models to boost prediction accuracy, further enhancing the overall performance of the intrusion detection system. Future systems will focus on understanding user and system behavior through behavioral analysis. This approach is crucial for accurate anomaly detection, enabling the identification of abnormal patterns and potential security threats. Analyzing behaviors helps in creating a baseline for normal activities, making it easier to detect deviations that could signify security breaches. The research will strive to develop intrusion detection systems capable of efficiently scaling with the growing complexity and volume of cloud data. Optimizing resources for efficient performance and cost-effectiveness will be a priority, ensuring the system can handle the increased data load and adapt to evolving cloud infrastructures while maintaining cost-efficiency. Ensemble learning techniques will be leveraged to combine multiple models, harnessing their collective strength to make more accurate predictions. By integrating ensemble learning, the intrusion detection system can enhance its overall performance, achieving higher accuracy and reliability in identifying potential security threats in the cloud.

REFERENCES

[1] M. Ali, S. U. Khan, and A. V. Vasilakos, Security in cloud computing: Opportunities and challenges, Information Sciences, vol. 35, pp. 357–383, 2015.

[2] A. Singh and K. Chatterjee, Cloud security issues and challenges: A survey, Journal of Network and ComputerApplications, vol. 79, pp. 88–115, 2017.

[3] P. S. Gowr and N. Kumar, Cloud computing security: A survey, International Journal of Engineering and Technology, vol. 7, no. 2, pp. 355–357, 2018.

[4] A. Verma and S. Kaushal, Cloud computing security issues and challenges: A survey, in Proc. First International Conference on Advances in Computing and Communications, Kochi, India, 2011, pp. 445–454.

[5] H. Alloussi, F. Laila, and A. Sekkaki, L'etat de l'art de la ´ securit ´ e dans le cloud computing: Probl ´ emes et solutions ` de la securit ´ e en cloud computing, presented at Workshop ´ on Innovation and New Trends in Information Systems, Mohamadia, Maroc, 2012.

[6] J. Gu, L. Wang, H. Wang, and S. Wang, A novel approach to intrusion detection using SVM ensemble with feature augmentation, Computers and Security, vol. 86, pp. 53–62, 2019.

[7] Z. Chiba, N. Abghour, K. Moussaid, A. E. Omri, and M. Rida, A cooperative and hybrid network intrusion detection framework in cloud computing based snort and optimized back propagation neural network, Procedia Computer Science, vol. 83, pp. 1200–1206, 2016.

[8] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, Survey of intrusion detection systems: Techniques, datasets and challenges, Cybersecurity, vol. 2, p. 20, 2019.

[9] A. Guezzaz, A. Asimi, Y. Asimi, Z. Tbatou, and Y. Sadqi, A global intrusion detection system using PcapSockS sniffer and multilayer perceptron classifier, International Journal of Network Security, vol. 21, no. 3, pp. 438–450, 2019.

[10] A. Guezzaz, S. Benkirane, M. Azrour, and S. Khurram, A reliable network intrusion detection approach using decision tree with enhanced data quality, Security and Communication Networks, vol. 2021, p. 1230593, 2021.

[11] B. A. Tama and K. H. Rhee, HFSTE: Hybrid feature selections and tree-based classifiers ensemble for intrusion detection system, IEICE Trans. Inf. Syst., vol. E100.D, no. 8, pp. 1729–1737, 2017.

[12] M. Azrour, J. Mabrouki, G. Fattah, A. Guezzaz, and F. Aziz, Machine learning algorithms for efficient water quality prediction, Modeling Earth Systems and Environment, vol. 8, pp. 2793–2801, 2022.

[13] M. Azrour, Y. Farhaoui, M. Ouanan, and A. Guezzaz, SPIT detection in telephony over IP using K-means algorithm, Procedia Computer Science, vol. 148, pp. 542–551, 2019.

[14] M. Azrour, M. Ouanan, Y. Farhaoui, and A. Guezzaz, Security analysis of Ye et al. authentication protocol for internet of things, in Proc. International Conference on Big Data and Smart Digital Environment, Casablanca, Morocco, 2018, pp. 67–74.

[15] M. Azrour, J. Mabrouki, A. Guezzaz, and A. Kanwal, Internet of things security: Challenges and key issues, Security and Communication Networks, vol. 2021, p. 5533843, 2021.

[16] A. Guezzaz, S. Benkirane, and M. Azrour, A novel anomaly network intrusion detection system for internet of things security, in IoT and Smart Devices for Sustainable Environment, M. Azrour, A. Irshad, and R. Chaganti, eds. Cham, Switzerland: Springer, 2022, pp. 129–138.