CLOUD-BASED DATA INTRUSION DETECTION SYSTEM USING ML TECHNIQUES

Manikrao Mulge[1], Namratha[2], Neha[3], Prerana[4], Sandhya Rani[5],
GNDEC, Bidar, Karnataka, INDIA.
msmulgi@gmail.com[1], namratha5858@gmail.com[2].

Abstract-Cloud computing (CC) stands as a groundbreaking technology, streamlining access to network and computer resources while offering a plethora of services, such as storage and data management, all with the aim of optimizing system functionality. Despite its array of advantages, cloud providers grapple with notable security challenges, particularly concerning the safeguarding of resources and services. Addressing these concerns and bolstering security measures necessitates vigilant monitoring of resources, services, and networks to promptly detect and respond to potential attacks. Central to this effort is the implementation of an advanced mechanism known as an intrusion detection system (IDS), which plays a pivotal role in regulating network traffic and identifying anomalous activities. This paper introduces an innovative cloud-based intrusion detection model that harnesses the power of the random forest (RF) algorithm alongside cutting-edge feature engineering techniques. Specifically, the integration of the RF classifier aims to enhance the accuracy (ACC) of the detection model significantly. The efficacy of the proposed model is rigorously evaluated using the NSL-KDD dataset, demonstrating a remarkable 99.99% ACC. This performance surpasses that of existing methodologies, underscoring the effectiveness and robustness of the proposed approach in addressing security challenges within cloud environments.

Keywords —Accuracy, Anomaly detection, Cloud security, Feature Engineering, Intrusion Detection System (IDS), Random forest (RF)

I. INTRODUCTION

Cloud technologies enable convenient access to shared networks, storage, and resources, offering a wide range of service models[1] such as platform as a service (PaaS), software as a service (SaaS), and infrastructure as a service (IaaS)[2][19]. These services can be deployed in private, public, or hybrid cloud[3] environments. The National Institute of Standards and Technology[4] identifies key characteristics of cloud computing, including network accessibility, resource pooling, scalability, and metered services, which contribute to high performance.

Despite its advantages, the cloud faces numerous security challenges, including threats to availability, data confidentiality, integrity, and access control. The Internet serves as a significant avenue for potential threats to cloud systems and resources[2]. Consequently, enhancing cloud security has become a critical priority for providers[5]. Various security measures, such as firewall tools, encryption algorithms, and authentication protocols, have been developed to mitigate these risks[6]. However, traditional security systems alone are insufficient to protect against evolving threats[7].

Intrusion detection approaches play a vital role in identifying and preventing unauthorized activities in real time[8][9]. These methods typically fall into two categories: misuse detection, which identifies known attacks, and anomaly detection, which detects unusual behavior indicative of unknown attacks. A hybrid approach combining the strengths of both methods has been proposed to improve detection accuracy[10].

Despite advancements in security solutions, current intrusion detection systems (IDSs) face significant limitations[8], including processing large volumes of data, realtime detection capabilities, and ensuring data quality, all of which impact detection performance. To address these challenges, researchers are increasingly turning to intelligent learning methods[6][11] such as machine learning (ML), deep learning (DL), and ensemble learning[12-18].

This research aims to propose an anomaly detection approach utilizing a random forest (RF) binary classifier. Feature engineering techniques are employed to streamline the feature set, enhancing the efficiency of the anomaly detection model. The performance of the proposed model is evaluated using NSL-KDD and BoT-IoT datasets, demonstrating its effectiveness.

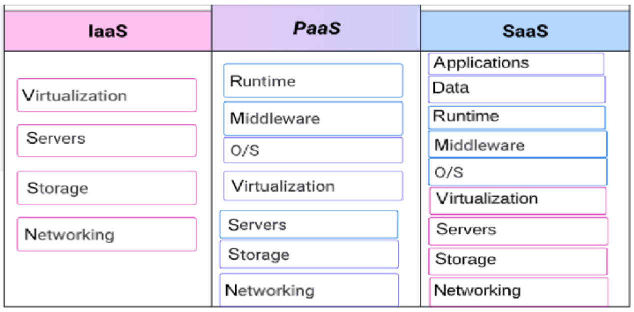


Fig. 1 Cloud computing models.

II. LITERATURE SURVEY

Numerous studies have delved into the realm of cloudbased intrusion detection, leveraging a variety of machine learning (ML) techniques to enhance security measures[22]. These studies have contributed significantly to the development of robust intrusion detection systems (IDS) capable of identifying and mitigating potential threats in cloud environments.

Kanimozhi and Jacob[28] conducted a comprehensive evaluation of different classifier methods for detecting botnet attacks. Their study employed a calibration curve to assess the performance of classifiers such as K-nearest neighbors (KNN), Naïve Bayes (NB), Adaboost with decision trees (DT), support vector machine (SVM), and random forest (RF) on the CSECIC-IDS2018 dataset. Through rigorous experimentation, they demonstrated the effectiveness of these classifiers in detecting various botnet attacks, providing insights into optimal detection strategies.

Zhou et al.[29] proposed a deep neural network (DNN)-based IDS tailored for cloud environments. Their system underwent three phases: data acquisition, preprocessing, and DNN classification. By employing support vector machines (SVM) alongside the DNN model, they achieved an impressive accuracy rate of 96.3%, showcasing the potential of deep learning techniques in bolstering intrusion detection capabilities.

Tang et al.[30] introduced a software-defined networking IDS harnessing deep learning methodologies. By leveraging a twostage deep learning approach, they aimed to detect and mitigate malicious assaults in cloud networks. Their system demonstrated promising results, achieving an accuracy rate of 96.93% based on attack detection performance, highlighting the efficacy of deep learning-based IDS frameworks.

Mishra et al.[32] proposed a classification-based ML approach for detecting distributed denial of service (DDoS) attacks in cloud computing environments. Utilizing techniques such as KNN, RF, and NB, their model exhibited a remarkable accuracy rate of 99.76%, with random forest (RF) emerging as the top-performing classifier. Their study underscored the importance of ML techniques in fortifying cloud security against evolving cyber threats.

Alshammari and Aldribi[22] applied ML methodologies to detect malicious network traffic in cloud computing environments. Leveraging the ISOT-CID dataset for evaluation, their study showcased the effectiveness of ML techniques in identifying and mitigating potential security breaches. By employing supervised learning algorithms, they demonstrated significant improvements in detection accuracy, reaffirming the role of ML in enhancing cloud security measures.

These studies collectively highlight the pivotal role of ML techniques in augmenting intrusion detection capabilities in cloud environments. By leveraging advanced ML algorithms and frameworks, researchers have made significant strides in fortifying cloud security against emerging threats, paving the way for more resilient and adaptive security solutions in the future[22-28].

III. METHODOLOGY

The first step in our methodology involves collecting relevant datasets for training and testing our intrusion detection model. We primarily utilize the NSLKDD dataset, which offers a comprehensive collection of network traffic data suitable for intrusion detection research.

Data preprocessing is crucial for ensuring the quality and integrity of the datasets before feeding them into the intrusion detection model. This step involves several tasks, including data cleaning, normalization, and feature engineering. Categorical variables are transformed into numerical values using techniques such as one-hot encoding or label encoding. Additionally, outliers and inconsistencies are identified and addressed to enhance the robustness of the model.

Feature selection plays a vital role in reducing the dimensionality of the dataset and improving the efficiency of the intrusion detection model. We employ various techniques, including graphical data visualization and statistical analysis, to identify the most relevant features for detection. By selecting a subset of informative features, we aim to enhance the model's predictive accuracy while reducing computational complexity.

Once the dataset is preprocessed and features are selected, we proceed to develop the intrusion detection model. Our approach leverages the Random Forest (RF) classifier, known for its effectiveness in handling complex datasets and mitigating issues such as overfitting. The RF algorithm is trained on the preprocessed dataset to learn patterns indicative of normal and abnormal network behavior.

To assess the performance of our intrusion detection model, we employ various evaluation metrics, including accuracy (ACC), precision, recall, and F1-score. These metrics provide insights into the model's ability to correctly classify instances of normal and malicious network activity. Additionally, we conduct cross-validation and utilize techniques such as confusion matrices to analyze the model's performance across different datasets and scenarios.

Our experiments are conducted in a controlled environment using a computer equipped with standard hardware specifications. We utilize Python programming language and relevant libraries such as scikit-learn for model development and evaluation. The experiments are designed to compare the performance of our proposed intrusion detection framework with existing models and assess its effectiveness in detecting various types of intrusions.

To validate the robustness and generalization capability of our intrusion detection model, we conduct rigorous testing on independent datasets and real-world scenarios. By evaluating the model's performance under different conditions and settings, we aim to demonstrate its reliability and effectiveness in detecting intrusions in diverse environments.

Finally, we analyze the results obtained from our experiments, examining the model's performance metrics and identifying areas for improvement. Through comprehensive analysis and interpretation of the results, we aim to validate the efficacy of our proposed intrusion detection framework and provide insights for future research directions.

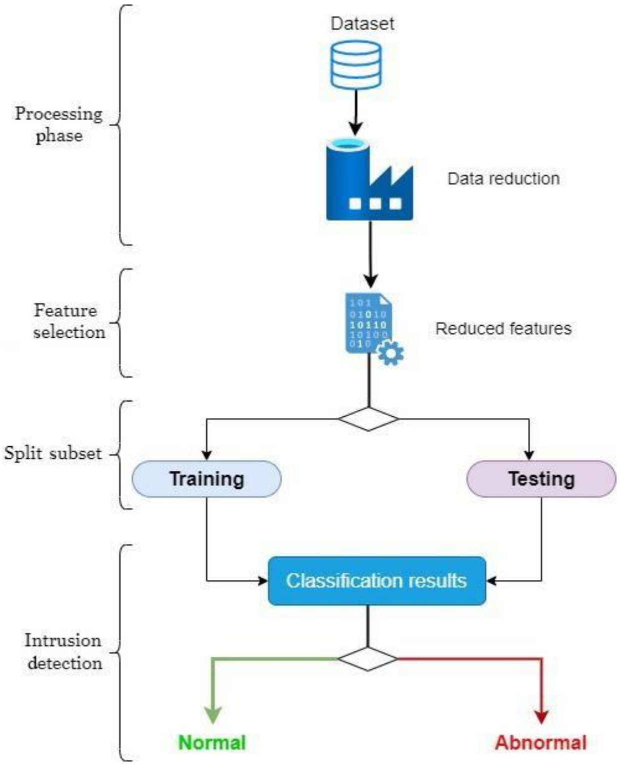


Fig. 2 Proposed model architecture.

Reference	Year	Method used	ACC (%)	Dataset
Chiba et al.[7]	2016	BPN	–	–
Alshammari And Aldribi[22]	2021	ANN KNN DT SVM NB RF	92.00 100.00 100.00 81.00 60.00 100.00	ISOTCID
Kanimozhi and Jacob[28]	2019	ANN RF KNN SVN Adaboost NB	99.90 99.80 99.73 99.80 99.90 99.20	CSE-CIC-IDS 2018
Zhou et al.[29]	2018	DNN	96.30	–
Tang et al.[30]	2016	DNN	75.75	NSL-KDD
Zhang et al.[31]	2018	DNN	–	–
Mishra et al.[32]	2021	RF, KNN, NB	99.76	–
Jiang et al.[33]	2018	LSTM	98.94	NSL-KDD
Khan et al.[34]	2019	ANNs	–	–
Potluri and Diedrich[35]	2016	DL	97.50	NSL-KDD
Kim et al.[36]	2016	LSTM	96.93	KDD CUP'99
Ahmad et al.[38]	2022	RF, NB, SVM, KNN	92.00	–
Mubarakali et al.[39]	2020	SVM	96.23	–

Table 1 Comparison of various current IDS models.

Our research was conducted and evaluated within a controlled experimental environment, utilizing a computer equipped with a Core i5 8250U CPU running at 1.8 GHz and 12 GB of RAM, operating on Windows 10 Professional 64-bit. Python 3 served as the primary programming language for implementing the Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM) models, following feature reduction through graphical visualization.

To validate our proposed model, we assessed its performance using the Accuracy (ACC) metric and compared it against other existing models. We partitioned the entire dataset randomly, allocating 70% for training purposes and reserving the remaining portion for testing. The optimal parameters for each classifier's performance were determined based on the dataset employed in both the training and testing phases.

Our research utilized two primary datasets: NSL-KDD and Bot-IoT. The NSL-KDD dataset, developed as an enhancement to the KDD 1999 dataset, addresses issues of redundancy and duplication, offering an appropriate number of records arranged in a standardized format (80% eKDDTrain + 20% ARFF). It comprises 41 initial features, including key properties such as:

Duration: representing the connection duration in seconds. Protocol Type: indicating the type of protocol used (TCP, UDP, ICMP).

Src Bytes: denoting data bytes sent from the source to the destination.

Dst Bytes: indicating the number of data bytes sent between source and destination.

Count: representing the number of connections made to the same host in the previous two seconds.

Srv Count: indicating the number of prior two-second connections to the same service as the current connection.

The categorical variable "Protocol Type" was transformed into numerical values using the dummies function. Through graphical visualization, we determined that the "Class" variable, representing anomaly detection, was minimally influenced by protocol types.

Furthermore, graphical visualization revealed that the "Class" variable could be predicted based on specific conditions of the "Count," "Duration," and "Dst Host Srv Count" variables. For instance, instances where the "Duration" variable exceeded 1500 seconds indicated potential anomalies. Additionally, if the "Src Bytes" variable was greater than 0 or the "Dst Bytes" variable exceeded 50,000, anomalies could be detected.

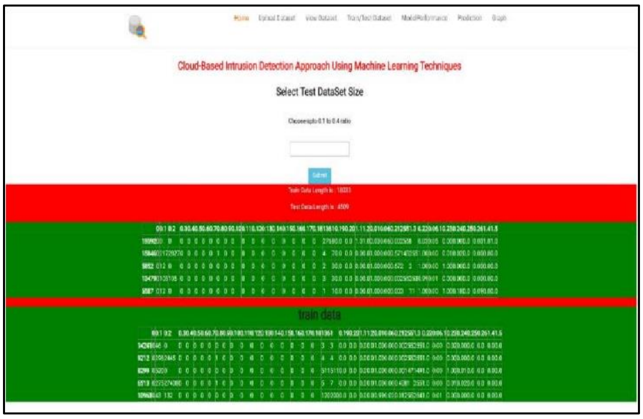
Following feature selection based on visualization insights, we narrowed down the number of features from 41 to two: "Src Bytes" and "Dst Bytes." Subsequently, we developed a Random Forest (RF) model to detect intrusions, utilizing these selected variables. The Bot-IoT dataset, encompassing various IoT traffic forms, was also considered for its enriched features and comprehensive data on IoT device behaviors. Shafiq et al. identified the top five variables with improved characteristics using Machine Learning (ML) approaches, further enhancing our dataset's predictive capabilities.

IV.RESULTS AND DISCUSSION

Our intrusion detection approach was evaluated on the NSL-KDD , showcasing its effectiveness in identifying intrusions within cloud and IoT environments.

Our approach demonstrates competitive performance compared to existing systems, with high accuracy . This validates its efficacy in detecting intrusions across diverse network traffic scenarios.

Further improvements could focus on enhancing feature engineering techniques and exploring ensemble methods for even more robust intrusion detection capabilities.



Cloud-Based Intrusion Detection Approach Using Machine Learning Techniques

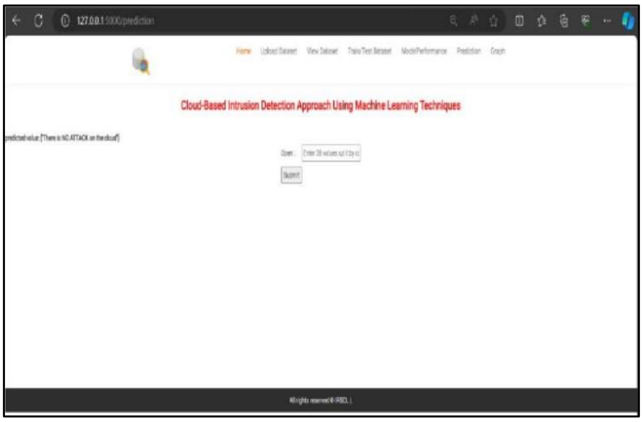
Select Test DataSet Size

Choose upto 0.1 to 0.4 ratio

Train Data Length is : 18033

Test Data Length is : 4509

Fig.3 Training/Testing data.



Cloud-Based Intrusion Detection Approach Using Machine Learning Techniques

predicted value: ['There is NO ATTACK on the cloud']

Enter 38 values split by c

Fig.4 Predicting the attack

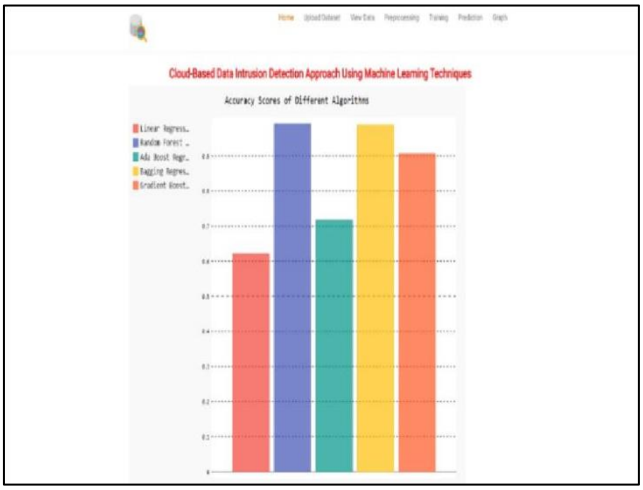


Fig.5 Accuracy score of different algorithms

V. CONCLUSION

Intrusion discovery is a new technology that has bettered the security of the Cloud. Lately, ML algorithms have been used to develop this fashion because they are veritably helpful to secure and cover systems. In this paper, we present an approach for detecting intrusions by combining graphic visualization and RF for cloud security. Also, the first one is used for features engineering and the alternate one is used to predict and detect intrusions. Before the training of the model, we reduced the number of features to two. Grounded on the attained results, the RF classifier is a remarkably more accurate system to predict and classify the attack type than DNN, DT, and SVM. We've demonstrated the eventuality of using a small number of features by differing the results with those of other classifiers. But recall is still not well enough using NSL- KDD, so in unborn work, we will concentrate on this point by using DL and ensemble literacy ways to improve our model.

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 Computer Applications*, 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.

[17] A. Guezzaz, A. Asimi, M. Azrour, Z. Tbatou, and Y. Asimi, A multilayer perceptron classifier for monitoring network traffic, in *Proc. 3rd International Conference on Big Data and Networks Technologies*, Leuven, Belgium, 2019, pp. 262–270.

[18] S. Benkirane, Road safety against sybil attacks based on RSU collaboration in VANET environment, in *Proc. 5th International Conference on Mobile, Secure, and Programmable Networking*, Mohammedia, Morocco, 2019, pp. 163–172.

[19] Q. Zhang, L. Cheng, and R. Boutaba, Cloud computing: State-of-the-art and research challenges, *J. Internet Serv. Appl.*, vol. 1, pp. 7–18, 2010.

[20] M. K. Srinivasan, K. Sarukesi, P. Rodrigues, M. S. Manoj, and P. Revathy, State-of-the-art cloud computing security taxonomies: A classification of security challenges in the present cloud computing environment, in *Proc. 2012 International Conference on Advances in Computing, Communications and Informatics*, Chennai, India, 2012, pp. 470–476.

[21] A. L. Buczak and E. Guven, A survey of data mining and machine learning methods for cyber security intrusion detection, *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[22] A. Alshammari and A. Aldribi, Apply machine learning techniques to detect malicious network traffic in cloud computing, *Journal of Big Data*, vol. 8, p. 90, 2021.

[23] A. Geron, *Hands-On Machine Learning with Scikit-Learn & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol, CA, USA: O’Reilly Media, Inc., 2017.

[24] N. Chand, P. Mishra, C. R. Krishna, E. S. Pilli, and M. C. Govil, A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection, in *Proc. 2016 International Conference on Advances in Computing, Communication, & Automation (ICACCA)*, Dehradun, India, 2016, pp. 1–6.

[25] A. B. Nassif, M. A. Talib, Q. Nasir, H. Albadani, and F. M. Dakalbab, Machine learning for cloud security: A systematic review, *IEEE Access*, vol. 9, pp. 20717–20735, 2021.

[26] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, A survey of deep learning-based network anomaly detection, *Cluster Comput.*, vol. 22, pp. 949–961, 2017.

[27] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study, *Journal of Information Security and Applications*, vol. 50, p. 102419, 2020.

[28] V. Kanimozhi and T. P. Jacob, Calibration of various optimized machine learning classifiers in network intrusion detection system on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing, *International Journal of Engineering Applied Sciences and Technology*, vol. 4, no. 6, pp. 209–213, 2019.

[29] L. Zhou, X. Ouyang, H. Ying, L. Han, Y. Cheng, and T. Zhang, Cyber-attack classification in smart grid via deep neural network, in *Proc. 2nd International Conference on Computer Science and Application Engineering*, Hohhot, China, 2018, pp. 1–5.

[30] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, and M. Ghogho, Deep learning approach for network intrusion detection in software defined networking, in *Proc. 2016 International Conference on Wireless Networks and Mobile Communications*, Fez, Morocco, 2016, pp. 258–263.

[31] L. Zhang, L. Shi, N. Kaja, and D. Ma, A two-stage deep learning approach for can intrusion detection, in *Proc. 2018 Ground Vehicle Syst. Eng. Technol. Symp. (GVSETS)*, Novi, MI, USA, 2018, pp. 1–11.

[32] A. Mishra, B. B. Gupta, D. Perakovic, F. J. G. Penalvo, and C. H. Hsu, Classification based machine learning for detection of DDoS attack in cloud computing, in *Proc. 2021 IEEE International Conference on Consumer Electronics*, Las Vegas, NV, USA, 2021, pp. 1–4.

[33] F. Jiang, Y. Fu, B. B. Gupta, Y. Liang, S. Rho, F. Lou, F. Meng, and Z. Tian, Deep learning based multi-channel intelligent attack detection for data security, *IEEE Transactions on Sustainable Computing*, vol. 5, no. 2, pp. 204–212, 2018.

[34] A. N. Khan, M. Y. Fan, A. Malik, and R. A. Memon, Learning from privacy preserved encrypted data on cloud through supervised and unsupervised machine learning, in *Proc. 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies*, Sukkur, Pakistan, 2019, pp. 1–5.

[35] S. Potluri and C. Diedrich, Accelerated deep neural networks for enhanced intrusion detection system, in *Proc. 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation*, Berlin, Germany, 2016, pp. 1–8.

[36] J. Kim, J. Kim, H. L. T. Thu, and H. Kim, Long short term memory recurrent neural network classifier for intrusion detection, in *Proc. 2016 International Conference on Plateform Technology and Service*, Jeju, Republic of Korea, 2016, pp. 1–5.

[37] J. Zhang, Anomaly detecting and ranking of the cloud computing platform by multi-view learning, *Multimedia Tools and Applications*, vol. 78, pp. 30923–30942, 2019.

[38] F. B. Ahmad, A. Nawaz, T. Ali, A. A. Kiani, and G. Mustafa, Securing cloud data: A machine learning based data categorization approach for cloud computing, *http://doi.org/10.21203/rs.3.rs-1315557/v1, 2022.*

[39] A. Mubarakali, K. Srinivasan, R. Mukhalid, S. C. Jaganathan, and N. Marina, Security challenges in Internet of things: Distributed denial of service attack detection using support vector machine-based expert systems, *Computational Intelligence*, vol. 36, no. 4, pp. 1580–1592, 2020.