International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Securing the Cloud Storage Using Novel Machine Learning-based Intrusion Detection System

Ms. Deeksha Chaudhary[1], Gowrishankar Jayaraman[2], Vishal Sharma[3], Dr. Sadaf Hashmi[4], Deepak Minhas[5]

Submitted: 03/02/2024 **Revised**: 11/03/2024 **Accepted**: 17/03/2024

Abstract: Intrusion detection is an approach for detecting unauthorized access or malicious behavior inside a system. When applied to cloud storage, it entails monitoring and analyzingnetwork activity, system records and consumer activity to identify any possible security vulnerabilities or abnormalities. This proactive method assists in securing sensitive data and the integrity of cloud-based infrastructure, ensuring persistent security against cyber-attacks. In this study, we produced an intelligent intrusion detection algorithm named Gorilla Troops fine-tuned Modified Random Forest (GTO-MRF) for enhancing security in cloud storage. Initially, we collected a dataset comprising simulated cloud-based intrusion scenarios and a variety of attack types. The proposed model was evaluated using this diverse dataset to enhance security measures. UnitVector Transformation (UVT) algorithm is employed to pre-process the gathered raw data. We extracted primary features from the pre-processed data using Kernel Principal Component Analysis (KPCA). The Gorilla Troops Optimization (GTO) approach improves the approach by adjusting the tree architectures and feature importance weights. We implemented the proposed model in software. The result evaluation phase is performed with multiple metrics such as training time, False alarm rate and encryption time to evaluate the suggested GT-MRF approach. We conducted a comparison analysis with other conventional methodologies. The experimental results illustrate that the proposed GT-MRF approach performed better than other conventional approaches for enhanced intrusion detection models in cloud storage.

Keywords: *Cloud Security Cloud Storage,Gorilla Troops fine-tuned Modified Random Forest (GT-MRF), Machine Learning, Intrusion Detection*

1. Introduction

An intrusion detection system (IDS), often known as IDSS, is a software application or hardware part that keeps an eye on a network to quickly catch any behavior that would indicate an attack. Platform, software and architecture of three different cloud computing service models are in danger of data security attacks [1].The greatest service development in the IT industry is cloud computing. The main benefit of cloud computing is that it enables access despite location or time constraints. Lower costs, more control over storage space, support for mobile as well as group applications and services are all benefits of cloud computing. In addition, cloud services are multisource, allowing customers to select different service providers based on their needs [2]. Several companies, banks and governments have embraced cloud computing services as they become more commonplace. Strong security measures are necessary because this transformation also exposed these systems to various dangers by hackers and intruders. An instance is the Amazon Web Services (AWS) platform, which offers services with expiration dates and restricted validity determined by the duration of the service license [3]. The volume of data is increasing, particularly crucial in the field of medicine, therefore regular backups and updates are necessary. Because it is private, healthcare data is a prime target for hackers looking to acquire information or modify it for illegal goals like gaining political or financial advantage. Medical records and health data could contain particular patient histories, details on prescription medications and equipment, along with other private patient data[4]. Research that has been conducted in foreign hospitals focuses on users' awareness of the import of cyber security, including the use of strong passwords, the removal or filtering of unsolicited emails, data encryption, the private handling of credentials, cautious information access and prompt reporting of any security breaches. While having IDS in place, the health industry can have issues with its local infrastructure [5].Depending on the services provided by the cloud service provider, the cloud offers several kinds of security. Highly sensitive data could not always be stored in the cloud, yet some

[1]*Assistant Professor, Maharishi School of Engineering and Technolgy, Maharishi University of Information Technology, Lucknow, India - 226036, Email Id- deeksha.choudhary@muit.in, Orcid Id- 0009-0001-8298-6082*

[2]*Assistant Professor, , Department of Computer Science Engineering, , Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Karnataka - 562112, India, Email Id- gowrishankar.j@jainuniversity.ac.in, Orcid Id- 0000-0002-4320-8683*

[3]*Assistant Professor, Department of Mechanical Engineering, Vivekananda Global University, Jaipur vishal_sharma@vgu.ac.in, Orcid Id- 0000-0003-3145-3654*

[4]*Associate Professor, Department of ISME, ATLAS SkillTech University, Mumbai, Maharashtra, India, Email Id- sadaf.hashmi@atlasuniversity.edu.in, Orcid Id- 0009-0008-8729-8423*

[5]*Centre of Research Impact and Outcome, Chitkara University, Rajpura-140417, Punjab, India, Email Id- deepak.minhas.orp@chitkara.edu.in, Orcid Id- https://orcid.org/0009-0002-0585-6046*

operational data can be saved online. When compared to cloud security, the data stored on internal servers needs to be secured with a similar level of security. Enhancing cloud resource security takes a lot of work, which helps to build security for the cloud deployment of private and sensitive medical data[6].Cybercrime has become a serious worldwide danger to individuals, organizations and governments. Hackers and cybercriminals are developing new and advanced methods that allow access to computer networks, steal personal data and impair regular corporate operations [7].Even with a lot of research on IDSs, a number of important problems remain to be resolved. More accuracy is required from IDSs to detect a larger range of intrusions with fewer false alarms and other problems, such as missing data in the cloud [8].To provide ongoing protection against cyber attacks, this proactive approach helps to safeguard critical data and the integrity of cloud-based infrastructure using GTO-MRF approach.

The remainder of the paper is split up into parts. Section 2 displays related studies that are based on objectives. Section 3 displays the data-collecting procedure along with their suggested techniques. The performance analysis and their discussion are given in sections 4 and 5. Section 6 signified the paper's conclusion.

2. Related work

The study [9] suggestedthata primary finding of the research was how few network traces were available for training contemporary machine-learning models to keep off intrusions strange to the Internet of Things (IoT). One dataset that was taken specifically was the Knowledge Mining in Databases (KDD) Cup.Thestudy [10] developed the research of an exciting universal intrusion detection framework that consists of five modules: preprocessing, auto-encoder, database, classification and feedback. The goal of the framework was to increase the intrusion detection system performance. The study [11] proposed the HIIDS method based on machine learning and met heuristic algorithms was explained in the study for IOT-based applications like healthcare. To identify and stop harmful traffic, the best hybrid Genetic Algorithm Digital Transformation (GA-DT) version-based HIIDS was finally used in the design of an IOT-based healthcare architecture.The study [12]oversaw the creation of an advanced, large-scale intrusion detection system (CIDS) on the internet to guard against attacks.The primary functions of the five major modules of the proposed CIDS were to perform network monitoring, traffic flow capture, feature extraction, flow analysis, intrusion detection along with its response and action recording. It has been demonstrated that the suggested method resolved every issue relevant to cloud attacks brought up in the literature. The study [13] suggested thathaving IDS can adapt to these kinds of barriers. In the past, proposed IDS that were cross-layer based, had two detection layers.Malicious behavior used in the study included the black hole and Distributed Denial of Service (DDoS) attacks. In the study [14] a framework for better communication was provided by the cloud which specifically suggested the Energy Aware Smart Home (EASH) architecture. The paper analyzed the issue of network attack types and communication breakdowns in EASH. The article [15] organized on developing network-based intrusion detection systems (NIDS), offered a categorization founded on significant Machine Learning (ML) and Deep Learning (DL) methodologies, created the term. Along with the potential to use the flaws of the suggested methodologies to improve ML and DL-based NIDS in the future, a number of research topics were noted.The study [16] employed the Random Forest(RF) classifier to present an effective approach with a uniform detection system based on supervised machine learning technology. The article [17] suggested the research project that develops a Network Intrusion Detection System (NIDS) using the deep learning idea as its base. The study[18]explained thatan improved Long Short-Term Memory (LSTM) system was used by the network based Secured Automatic Two-level Intrusion Detection System (SATIDS).

3. Method

In this study, first, we collect the data and preprocess the data using the Unit Vector Transformation(UVT) algorithm. After we preprocess the data, the Kernel Principal Component Analysis(KPCA) method is employed for feature extraction. In this study, we proposed Gorilla Troops optimized fine-tuned Modified Random Forest (GT-MRF) for enhancing security in cloud storage. Fig .1 shows the flow of methodology.

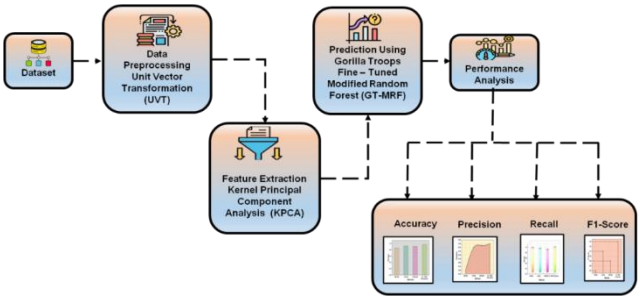

Fig.1 Flow of Methodology

3.1 Dataset

The dataset has [19] rows with data Transaction Identification (TID) totaling over 80,000, with the number of items [Item no] in each transaction displayed in the columns. Due to its size, the dataset is divided into four separate databases. Minimal support is used for creating frequent item sets. After loading the complete information, the algorithm uses the transaction of data which is displayed in Table 1 to determine the frequent item-sets.

Table 1. Transaction Database

TID	Transactions
1	B1,B3, A1,A2,A4
3	B2,B4, A2,A4
4	B3,B4, A3
8	B1, A2
9	B3,A 4

3.2. Preprocessing using Unit Vector Transformation

The vector's magnitude, or hypotenuse, is found using this method by using the Pythagorean Theorem($v_x{}^2 + v_y{}^2 = v^2$).

$$||\vec{u}|| = \sqrt{v_x * v_x + v_y * v_y} \quad (1)$$

Divide the value of every element by the vector to normalize the vector and scale it down to 1. When split by 10, a vector with 10 entries would have a value of 1. Scaling down to vector size 1 requires dividing each other by the same amount of 10. With this method, the direction of the vector can be modified without altering its lengthto reduce the size in cloud. While performing unit vector transformations, a new variable called $x'$with a range of $[0, 1]$could be established.

$$x' = \frac{x}{||x||} \quad (2$$

3.3.Feature Extraction using Kernel Principal Component Analysis (KPCA)

This kernel function, called KPCA, improves the PCA. The kernel methoddevelops the kernel Hilbert space via a non-linear mapping in place of the initial linear PCA process in cloud storage. For the feature space$(w_l \ldots . w_l)$.

For the KPCA, the analogous system is used in equation (3).

$$\tau(\rho(w_l).U) = (\rho(w_l).DU) forall\ l = 1 \ldots . k \quad (3)$$

Whereτ depicts the Eigen vector and U is the Eigen values. The stages data that follow outline the KPCA technique. A vector of S, where the Eigenvalues should be included in the feature space D and the Eigenvectors should be greater than 0.

- The eigenvalue that corresponds to the non-zero eigenvalues that correspond to the vectors in D should be normalizedin equation (4).
- $D = \frac{1}{k}\sum_{i=1}^{k} \quad \rho(w_l)\rho(w_l)^S$ (4)
- Where the coefficient is$s$.
- Calculate the eigenvalues and Eigenvectors of S, with the condition that the Eigenvalue be a part of the feature space D and the Eigenvector be greater than 0.
- The Eigen value that pertains to the non-zero Eigen values that relate to the vectors in D should be adjusted.

$$1 = \sum_{j,i=1}^{k} \quad \sigma_j^l \sigma_i^l \left(\rho(w_l).\rho(w_i)\right) = (\sigma^l.L\sigma^l) = \tau_l(\sigma^l.\sigma^l) \quad (5)$$

Where the σ coefficient.In the end, the kernel function is used for dot products without executing the map (ρ).

$$\left(U^l.\rho(w)\right) = \sum_{j=1}^{k} \quad \sigma_j^l\ (\rho(w_l).\rho(\rho w)) \quad (6)$$

3.4. Prediction of using Gorilla Troops fine-tuned Modified Random Forest (GT-MRF)

The GTO algorithm put forward that depends on the social intelligence of GTO natively. Here, novel methods forthe stage of exploration and exploitation are used to arithmetically represent gorilla behavior. In addition, it is led by a silverback gorilla, which makes decisions for the troop as a whole and goes food hunting in the cloud data. According to this method, the weaker gorilla in the set is represented by the population weaker solutions. Additionally, a different gorilla tries to distance itself to find the best ways to improve each gorilla's position. Additionally, the following provides a brief discussion and summary of the GT-MRF exploration and exploitation stages.

The three different models that are described by using Equation (7) during the exploration step of the GTO algorithm's flow chart include going to another gorilla, moving to an unfamiliar site and migrating to a known spot.

$$HW(s+1) = \{(V-K)q_1 + Kifqand < o(q_2 - C)W_q(s) + T, Uifqand \geq 0.5\, W(s) - T\left(T\left(W(s) - HW_q(s)\right) + q_3\left(W(s) - HW_q(s)\right)\right) ifqand < 0.5 \quad (7)$$

While $HW(s+1)$symbolizes the position vector in the next s iteration, $K$and Vd stand for the variable's lower and upper limits, respectively. The current vector of gorilla location is denoted byWs, a randomly picked group of gorillas is indicated by Wq and the position of a random

gorilla is indicated by $HWq(s)$. Moreover, random parameters from 0 to 1 are represented by rand, q_1, q_2 and q_3. The option to select the migration to an unidentified point is indicated by the parameter$o$. Moreover, the mobility of the silverback is indicated by the$T$ and U variables. The following equation (8) is used to calculate the variable C:

$$C = M \times (1 - \frac{s}{Maxls}) \quad (8)$$

M denotes the population's magnitude and $Maxls$ for the maximum number of iterations. Two different models are using for cloud storage during the exploitation stage. They rely on a comparison between the$C$values that is calculated using a parameter W. The gorilla complies with every order from the silverback to go to various locations to look for food supplies. It is expressed in Equation (9) andit is only applicable if$\geq W$.

$$HW(s+1) = T \times I(W(s) - W_{silverback}) + W(s) \quad (9)$$

This $W_{silverback}$ is used to show the silverback's location vector. When juvenile gorillas reach older age, they compete with adult females for resources. The following equation (10) expresses it and it can be utilized if$< W$.

$$HW(j) = W_{silverback} - (W_{silverback} \times Q - W - W(s) \times Q) \times S \quad (10)$$

The modified random forest (RF) method, as its name suggests, is a group of classification trees, each of which votes once for the class that is given to the input securing the cloud storage data the most frequently. The quantity of classification trees and the MRF classifier are merged. Forty prominent characteristics are chosen based on the feature importance standards. Equation (11) is used to compute the grouping result.

$$D(S) = F_s \sum_{j=1}^{L} (d_j(S) = O) \quad (11)$$

Where S is the original dataset's training set. The S dataset's subsets are denoted by $T$and$L$. Using a random vector, the technique automatically creates L decision trees for each subgroup. The classification result is given by $D(S)$, while the classification result of the ith decision tree is shown by$d_j(S)$. The target category, in this case, is O. Nevertheless, several random forest hyper-parameters are employed to either improve the model's prediction accuracy or speed up the method. When handling high-dimensional data, MRF can achieve superior performance by executing an implicit FS procedure. Within MRF, feature significance can be determined by using the Gini importance as a measuring criterion. These relevance scores, which are seen as an overgrowth, aid in identifying the decision trees are important to the classifier.

$$j(s) = 1 - e_1^2 - e_0^2 \quad (12)$$

To calculate the Gini impurity, use equation (12). In this case, any node of an RF is represented by the symbol t. The optimal split, which is a measurement of entropy, is found using the Gini impurity. In addition, Equation e_i(13), which finds the ratio of m_i samples to the total number of samples (n), indicates the class,with $j = 0,1$.

$$e_i \frac{m_i}{m} \quad (13)$$

Reducing δj can be performed by dividing and sending products to two distinct sub-notes $(s_o \& s_r)$ based on a threshold on variable Θ. Equation (14) shows the process.

$$\delta j(s) = j(s) - e_o j(s_o) - e_r j(s_r) \quad (14)$$

Then, using all-inclusive values which are available in the node overall thresholds, an exhaustive search is carried out. Equation (15) is used to store the decreases in Gini impurity values for each variable independently while considering all nodes s. When considering a specific problem at hand, IG considers how many times feature Θ is chosen during a split and significant inside the classifier.

$$J_H(\ominus) = \sum_q \quad \sum_r \quad \delta j_{\ominus}(s, S) \quad (15)$$

4. Result

In this section, we compared our suggested technique to other existing methods [20] like Support Vector Machine (SVM) [20], LogisticRegression (LR) [20] and Naive Bays (NB)[20] for the prediction of IDS in cloud storage, which is described in detail as shown in table 2.

Table 2. Outcomes of parameters

Methods	Accuracy	Precision	Recall	F1-Score
NB [20]	0.86	0.82	0.92	0.87
LR [20]	0.93	0.94	0.89	0.91
SVM [20]	0.92	0.95	0.85	0.89
GT - MRF [Proposed]	**0.97**	**0.96**	**0.94**	**0.93**

4.1. Accuracy

Scalability and accessibility are combined with distant data storage through cloud storage. Network traffic is monitored by IDS, which spot and stop unwanted activity. When it comes to ensuring data availability, confidentiality, and integrity with a high degree of reliability and few false positives,accuracy is the capacity to precisely identify unauthorized access or security breaches inside the cloud storage. Fig.2 shows the proposed system's Accuracy. The recommended method GT-MRF attained 0.97 of accuracy. When Compared to other existing approaches, our proposed method GT-MRF achieved a highest Accuracy rate, while NB is 0.86, LR is 0.93, and SVM is 0.92.

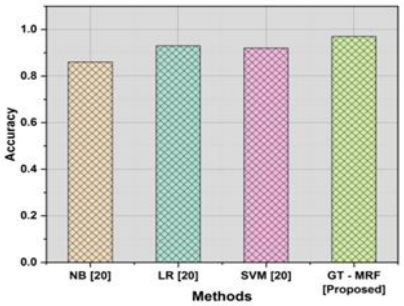

Fig. 2 Result of accuracy

4.2. Precision

In the context of IDS and cloud storage, precision is defined as the percentage of accurately noticed security risks and unauthorized access events to all detected events. It makes sure that the IDS's notifications about possible intrusion into the cloud storage environment are highly accurate, reducing the number of alarms and useless actions. Fig.3 shows the proposed system's Precision. The recommended method GT-MRF attained a precision (0.96). While compared to other existing approaches, NB reached 0.82, LR attained 0.94, and SVM achieved 0.95.

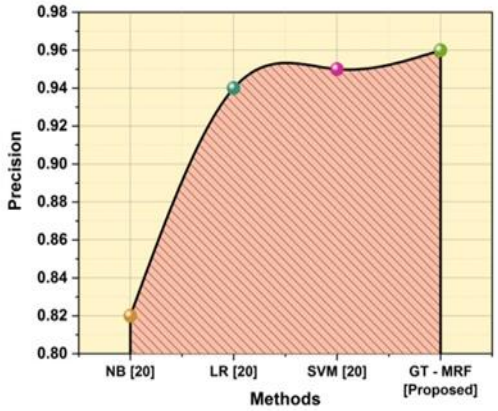

Fig. 3 Result of precision

4.3. Recall

In the context of IDS and Cloud Storage, recall refers to the capacity to recognize and detect any kind of security breaches and events involving unauthorized access inside the cloud environment. By ensuring thorough coverage and identifying possible threats, reduces the possibility of unnoticed attacks and data breaches. Fig.4 shows the proposed system's recall. The recommended method's GT-MRF attained a recall value (0.93). While compared to other approaches, NB achieved 0.92, LR reached 0.89, and SVM attained 0.85.

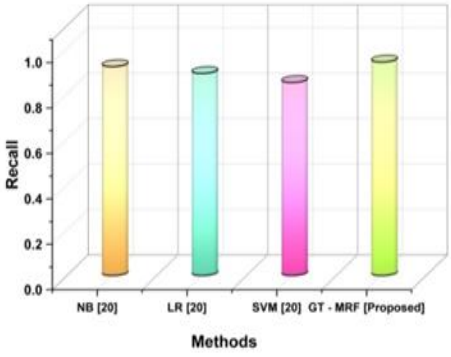

Fig. 4 Result of recall

4.4. F1-score

To provide a fair assessment of detection accuracy and completeness, the F1-score for Cloud Storage and (IDS) calculates the harmonic mean of precision and recall. It evaluates how well the system can detect and address security threats while reducing false positives and unnoticed attempts in cloud storage. Fig.5 shows the proposed system's F1 score. The recommended method' GT - MRF attained the F1-score (0.93). When compared to other approaches, NB performed 0.87, LR reached 0.91, and SVM attained 0.89.

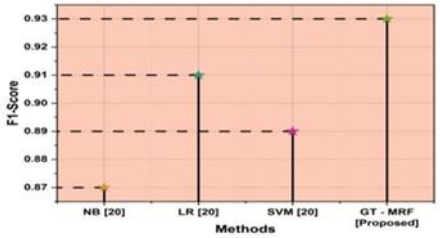

Fig 5. Result of F1-score

5. Discussion

The study consists of drawbacks to current methods like NB,LR and SVM. SVMs' high computational cost and sensitivity to parameter adjustment could cause them to perform poorly on huge datasets. Considering linear connections, LR can perform poorly when dealing with non-linear data. In addition to simplifying feature dependencies, Naive Bays can be sensitive to unwanted features. However GT-MRF has benefits. It handles complicated interactions, minimizes overfitting and

supports a variety of data formats by utilizing ensemble learning. The process of fine-tuning improves the resilience of the model and the MRF technique reduces the risk of overfitting and increases prediction accuracy, which makes it perfect for several real-world applications, such as intrusion detection in cloud storage systems.

6. Conclusion

The security of the cloud has been enhanced by a proposed technique called intrusion detection. To improve intrusion detection capabilities in cloud storage systems, the GT-MRF approach created to employ modern methods of machine learning. With the use of machine learning, GT-MRF improves its capacity to detect and respond to harmful activity as well as unauthorized access in cloud-based systems by improving the structure of trees along with the feature importance of weights utilizing the GTO techniquebecause this greatly helps in system monitoring and security.The innovation lay in the integration of GTO, which enhanced the model by adapting tree architectures and feature weights. The performance of proposed for precision,F1-score,accuracy,recall (0.96, 0.93, 0.97 and 0.94) indicate that the suggested GT-MRF strategy performed better than other traditional approaches. This demonstrated the model efficiency in achieving superior intrusion detection in cloud storage. Despite these successes, limitation exists and further research is required to address potential challenges and improve the algorithm's adaptability to evolving cyber threats. The proposed GT-MRF model proved to be a valuable asset in advancing the domain of intrusion detection for cloud-based systems, showcasing its significance in fortifying security measure and the way for future developments in this critical domain.

References

[1] Attou, H., Mohy-eddine, M., Guezzaz, A., Benkirane, S., Azrour, M., Alabdultif, A., &Almusallam, N. (2023). Towards an intelligent intrusion detection system to detect malicious activities in cloud computing. Applied Sciences, 13(17), 9588.

[2] Horchulhack, P., Viegas, E. K., &Santin, A. O. (2022). Toward feasible machine learning model updates in network-based intrusion detection. Computer Networks, 202, 108618.

[3] Dina, A. S., &Manivannan, D. (2021). Intrusion detection based on machine learning techniques in computer networks. Internet of Things, 16, 100462.

[4] Meryem, A., &Ouahidi, B. E. (2020). Hybrid intrusion detection system using machine learning. Network Security, 2020(5), 8-19.

[5] Hossain, M. A., & Islam, M. S. (2023). Ensuring network security with a robust intrusion detection system using ensemble-based machine learning. Array, 19, 100306.

[6] Seth, S., Singh, G., &KaurChahal, K. (2021). A novel time-efficient learning-based approach for smart intrusion detection system. Journal of Big Data, 8(1), 1-28.

[7] Khraisat, A., Gondal, I., Vamplew, P., &Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity, 2(1), 1-22.

[8] Khan, M. A., & Kim, Y. (2021). Deep Learning-Based Hybrid Intelligent Intrusion Detection System. Computers, Materials & Continua, 68(1).

[9] Ponnusamy, V., Humayun, M., Jhanjhi, N. Z., Yichiet, A., &Almufareh, M. F. (2022). Intrusion Detection Systems in Internet of Things and Mobile Ad-Hoc Networks. Computer Systems Science & Engineering, 40(3).

[10] Saif, S., Das, P., Biswas, S., Khari, M., &Shanmuganathan, V. (2022). HIIDS: Hybrid intelligent intrusion detection system empowered with machine learning and metaheuristic algorithms for application in IoT based healthcare. Microprocessors and Microsystems, 104622.

[11] Elmasry, W., Akbulut, A., &Zaim, A. H. (2021). A design of an integrated cloud-based intrusion detection system with third party cloud service. Open Computer Science, 11(1), 365-379.

[12] Amouri, A., Alaparthy, V. T., &Morgera, S. D. (2020). A machine learning based intrusion detection system for mobile Internet of Things. Sensors, 20(2), 461.

[13] Atul, D. J., Kamalraj, R., Ramesh, G., Sankaran, K. S., Sharma, S., &Khasim, S. (2021). A machine learning based IoT for providing an intrusion detection system for security. Microprocess. Microsystems, 82, 103741.

[14] Ahmad, Z., Shahid Khan, A., WaiShiang, C., Abdullah, J., & Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Transactions on Emerging Telecommunications Technologies, 32(1), e4150.

[15] Rani, D., &Kaushal, N. C. (2020, July). Supervised machine learning based network intrusion detection system for Internet of Things. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

[16] Sahar, N., Mishra, R., &Kalam, S. (2021). Deep learning approach-based network intrusion detection

system for fog-assisted iot. In Proceedings of international conference on big data, machine learning and their applications: ICBMA 2019 (pp. 39-50). Springer Singapore.

[17] Elsayed, R. A., Hamada, R. A., Abdalla, M. I., &Elsaid, S. A. (2023). Securing IoT and SDN systems using deep-learning based automatic intrusion detection. Ain Shams Engineering Journal, 14(10), 102211.

[18] Kasongo, S. M., & Sun, Y. (2020). Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset. Journal of Big Data, 7, 1-20.

[19] Suthanthiramani, P., Muthurajkumar, S., Sannasi, G., &Arputharaj, K. (2021). Secured data storage and retrieval using elliptic curve cryptography in cloud. Int. Arab J. Inf. Technol., 18(1), 56-66.

[20] Islam, U., Al-Atawi, A., Alwageed, H. S., Ahsan, M., Awwad, F. A., &Abonazel, M. R. (2023). Real-Time Detection Schemes for Memory DoS (M-DoS) Attacks on Cloud Computing Applications. IEEE Access.