Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 8, pp. 2285–2297
DOI : 10.47974/JIOS-1806

Machine learning-based network intrusion detection : Performance evaluation and comparative analysis

Barkha Kumari [§]
Vinay Singh [†]
Faculty of Computing and Information Technology
Usha Martin University
Ranchi 835103
Jharkhand
India

Mohit Kumar *
Department of Information Technology
School of Computing
MIT Art, Design and Technology University
Pune 412201
Maharashtra
India

Abstract

Intrusion detection systems are used to investigate malicious behavior that takes place within a system or network. Programming or equipment utilized for interruption identification scans a framework or organization for dubious exercises. The rising interconnectedness of PCs has made interruption discovery crucial for network security. Internet of Things (IoT) gadgets with their organization administrations are frequently helpless against assaults since they are not intended for security. This is designated by malevolent clients to take advantage of weaknesses or impede numerous weakness assaults. Thusly, manage this weakness; an interruption discovery framework that includes AI methods is required. Interruption Recognition Framework (IRF) is designated to get interruption in a correspondence framework by taking a gander at the IDS types and techniques. To bring down misleading problems and raise discovery rates, interruption location precision should be gotten to the next level. In ongoing works, numerous techniques have been utilized to upgrade the

[§] *E-mail:* **singhvarsha35@gmail.com**
[†] *E-mail:* **vinaysinghumu@gmail.com**
* *E-mail:* **mohitsmailbox13@gmail.com (Corresponding Author)**

presentation. An interruption location framework's essential assignment is to examine a lot of organization traffic information. To tackle this issue, a very much organized characterization framework is required. This issue is moved toward in the recommended way. Support Vector Machine (SVM) and Naive Bayes AI calculations are utilized. These techniques are many times used to address arrangement issues. An evaluation of the interruption recognition framework is directed utilizing the NSL-KDD information revelation dataset. The outcomes show that SVM_ outflanks Gullible Bayes. Successful arrangement strategies like Help Vector Machine and Credulous Bayes are utilized in relative examination, and their precision and misclassification rate are figured.

Subject Classification: *18M35, 32U35.*

Keywords: *Intrusion, Hazards of penetration, Attack detection, Remedial techniques.*

1. Introduction

In the realm of computer and organizational security, Intrusion Detection Systems (IDS) play a vital role in identifying irregular activities. These systems are broadly categorized into misuse-based and anomaly-based IDS. While misuse-based IDS effectively detect known attacks with a lower false alarm rate, they struggle with identifying novel attacks that deviate from established patterns. On the other hand, anomaly-based IDS, after establishing a baseline of normal behavior, flag any significant deviations from it as potential intrusions, capable of identifying both known and unknown threats but often accompanied by a higher false alarm rate. To mitigate this issue, various AI approaches are integrated. With the proliferation of computing devices like smartphones, tablets, laptops, and PCs, data security becomes paramount, and IDS are deployed to safeguard information transmitted over networks. The effectiveness of IDS depends on the characteristics of the network and the reference dataset used. Commonly utilized datasets include KDD Cup 99, NSL KDD, and attack datasets derived from specific environments. In the context of IoT devices, AI techniques, including supervised, unsupervised, or reinforcement learning, are employed to develop IDS. These systems, whether hardware or software-based, continuously monitor network or system activities, generating reports for management. Measurable classifiers are known as Bayesian classifiers. They have the ability to foresee the probability that a given model will squeeze into. There are three classes of Calculations in AI as follows: Managed learning Calculation, Solo Learning, and Support Learning [1]. In this study, we propose an IDS approach aimed at accurately and efficiently identifying

malicious network traffic. The subsequent sections outline the dataset utilized, the classification algorithms employed, and conclude with reflections on further research avenues.

2. Literature Survey

Since spilled data can bring about critical misfortunes, safeguarding an association's or a singular's PC and organization data has become pivotal. Consequently, interruption discovery frameworks are utilized to stop this damage. To work on the usefulness of IDS, different AI methods are made. The essential objective [2] is to address the issue of the Interruption Discovery Framework's (IDS) absence of adaptability. The proposed interruption location framework is equipped for distinguishing both known and obscure dangers. The Grouping Administrator (CM), Leader (DM), and Update Chief (UM) are the three principal instruments that include the proposed IDS. The proposed IDS's functional productivity is assessed utilizing the NSL-KDD dataset. Strategies that were directed as well as unaided were supported [3]. In fact, associations and individuals face various dangers while trading individual and expert information. Since the web has turned into an essential piece of our regular daily existences, the security of our information is generally helpless. IDS is essential in shielding web clients against unsafe organization assaults on account of this [4]. An interruption identification framework, or IDS, watches out for network traffic and conveys warnings when it sees any problematic action. The three classifications that will be the subject of this paper are AI, calculations NB, SVM, and KNN. In the initial step, these strategies will be used to decide the ideal precision utilizing the USNW NB 15 DATASET. Our data set is handled involving the best calculation in the subsequent stage, which depends on the result of the first [5]. To survey the model's presentation, we will run our tests on two unmistakable datasets. To guarantee the adequacy of the recommended strategy, execution is estimated utilizing the NSL-KDD and UNSWNB15 datasets.A mixture model is made by consolidating AI techniques, for example, Outrageous Learning Machine (ELM) and Backing Vector Machines (SVM) [3]. To make top notch datasets, changed K-implies is utilized. It makes little datasets that address the aggregate sum of unique preparation datasets. The classifier's preparation time is abbreviated by this step. KDDCUP 1999 is the execution rendition. It shows a 95.75 rate exactness. To report this issue, a few AI strategies like SVM, Irregular Woodland (RF), and ELM are explored. When contrasted with different procedures,

ELM yields more exact outcomes. One-fourth of the datasets are tests, a big part of the datasets are full datasets, etc [6]. It is recommended to utilize another mixture grouping technique on Fake Fish Swams (AFS) and Counterfeit Honey bee States (ABC) [7]. Relationship based Element Determination (CFS) and Fluffy C-Means Bunching (FCM) are utilized [8] to isolate preparing datasets and eliminate pointless highlights. The qualities are applied to datasets that have major areas of strength for a with the class however no relationship with the other datasets [9]. This approach accomplishes a 99 percent location pace of inconsistencies and a 0.01 percent bogus positive rate by utilizing the NSL-KDD and UNSW-NB15 datasets. A counterfeit honey bee province and AdaBoost calculations are utilized in a mixture A-NIDS way to deal with produce a high identification rate (DR) and low misleading positive rate (FPR) [10].

3. Exiting Methodology

The KDD Cup dataset is utilized to evaluate the presentation of the proposed model. Countless occasions from the 10% KDD preparing dataset are utilized to prepare classifiers like SVM and ELM. Rather than utilizing the total KDD dataset, 10% of it is utilized in light of the fact that doing so will bring about various issues. Emblematic components like banner, administration, and convention can be added, modified, or erased [11]. At last, the occurrences are ordered into four gatherings: R2L, DoS, Test, and Ordinary. They utilized the dataset to prepare SVM and ELM. They utilized a staggered model with remedied KDD dataset for the testing strategy. Utilizing the KDD Cup 1999 dataset, the recommended model's exactness has arrived at up to 95.75 rates and a deception pace of 1.87 percentages.

4. Problem Statement

The amount of data that is too much is causing a rise in false alarm reports of network intrusions and a decline in detection accuracy. When an unknown assault occurs on the system, this is one of the main problems. Reducing the false alarm rate and raising accuracy rates are the prmary goals. Machine learning algorithms such as SVM and Naïve Bayes have been employed to address the mentioned difficulties.

5. Proposed Approach

The three vital stages in the proposed worldview are result assessment, order and dataset pre-handling. Each phase of the proposed framework is vital and essentially affects how well it works [12]. The main pieces of this paper are inspecting how SVM and Innocent Bayes classifiers work.

5.1 *Initial Processing*

The classifier can't handle representative elements found in the dataset. All non-numeric or representative perspectives are altered or disposed of during this stage. During the pre-handling stage, non-numeric or emblematic highlights are taken out or supplanted. Non-numeric or representative components are eliminated or modified as a feature of the pre-handling process, which is urgent generally in light of the fact that they don't assume a critical part in interruption detection [13]. Representative components like banner, administration, and convention can be added, adjusted, or erased. At last, the occurrences are ordered into four gatherings: R2L, DoS, Test, and Ordinary.

5.2 *Techniques*

A relative report looking at the exactness and misclassification pace of SVM and Guileless Bayes for dataset characterization was led. At first, the natural dataset is acquired, and the class property involves 24 unmistakable attack types that are sorted into 4 gatherings. They are Dos, Test, r2l, and ordinary. Following, marking to change an ostensible property to a twofold characteristic, pre-handling is utilized. Non-numeric highlights are disposed of from the interruption recognition framework to expand its presentation. Highlight decrease is completed utilizing strategies like Cfs_Subset_Eval to acquire assorted results and upgrade the dataset's performance [14]. Following preprocessing, the gave dataset is exposed to highlight decrease and standardization. Cfs_Subset_Eval is a property determination method. It decides the worth of qualities by considering each component's one of a kind expectation gauge as well as the level of overt repetitiveness among them. Order under SVM falls under the regulated learning approach, which prepares different information sources from various subjects. The Help Vector Machine produces at least one hyper-planes in a high-layered space that is provided. SVM produces at least one hyper-planes. The best hyper-plane is the one that separates the furnished information into various classes with the chief parcel as

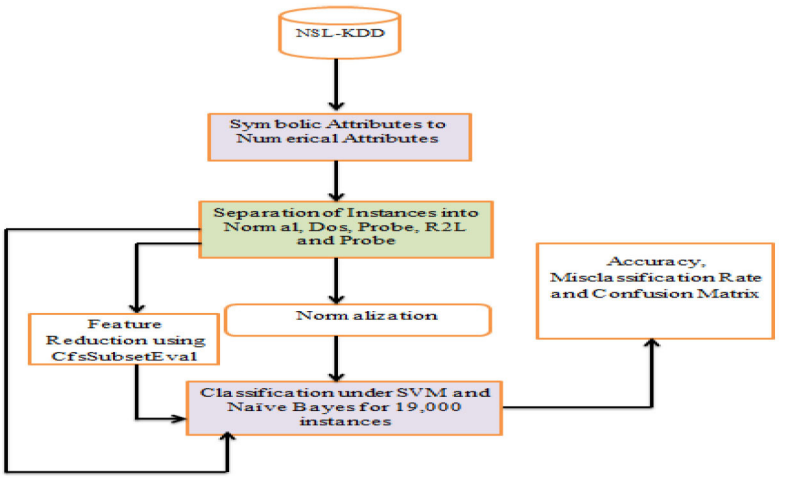

Figure 1

Block Diagram for different instances

productively as could be expected. Different portion capabilities are applied by a non-direct classifier to survey the edges between hyper-planes [15]. Likewise, Gullible Bayes is utilized in the strategy. Measurable classifiers are known as Bayesian classifiers. They have the ability to foresee the probability that a given model will squeeze into a particular class. The Bayes hypothesis is its basis [16]. It works under the reason that the characteristic incentive for a specific class is free of the trait values. It is known as class contingent freedom hypothesis. The guileless Bayes classifier works as such: T addresses the preparation set of tests, each with a class mark. A sum of k classes exist: X1, X2, X3, X(k-1), Xk. While An addresses a n-layered vector, A = {a1,a2,...,an} shows n estimated upsides of the n qualities, m1,m2,...,mn. b) The classifier will decide A, which fits to the class with the biggest posteriori likelihood, molded, for a given example A.This approach's block chart is shown in Figure 1. In light of the determined exactness and the obtained discoveries, a diagram has been made. We can concentrate on the diagram and see that SVM performs better compared to Straightforward Bayes

INPUT: Exercise Program Variable T Predictor P. A collection of test datasets is the output.

STEPS:

1. Start by reading Training Set T.

2. Determine each class's conditional probability, P. H ←> class variable dependant class X P(X | H) = P(H | X)*P(X)P(H)
3. Determine which class has the highest likelihood.
4. Create a matrix of confusion
5. Determine Misclassification Rate and Accuracy

OUTPUT: Classes of Output

1. Determine the Target Function T.
2. Capability of interest = min|| w1 ||(y + 2 where W means the weight vector, λ is the regularization boundary, xi is the information test, and yi is the result name.
3. Utilize weighted inclination drop learning
4. Update the weight rule for the mistakenly classified yield, w=w+η(yixi - 2λw1).
5. Update the weight rule w=w1+η(i - 2λw1) for the result that was effectively characterized, where η is the learning vector.
6. Offer back T
7. End capability

6. Performance Calculations

As shown in the table 1 exactness rate and misclassification rate were utilized to survey the SVM and Naive Bayes calculation execution across in excess of 18,700 models [17]. The data from the disarray network is utilized to assess these calculations' exhibition measures.

Table 1

Algorithms' calculation results

Method Used	Rate of Accuracy	Rate of Misclassification
SVM Method	95.73	3.237
Naïve Bayes Method	68.37	36.43
SVM-Normalization Method	94.84	3.534
Naïve Bayes-Normalization Method	86.008	27.353

6.1 *Evaluation*

The NSL-KDD dataset is utilized to assess the model after pre-handling and randomization are applied. There are in excess of 18,000 examples complete in the dataset [18]. As assessment measures, exactness rate and misclassification rate are utilized.

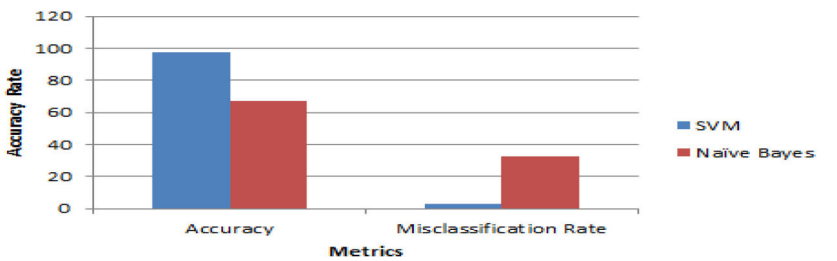

Figure 2

Metrics

As shown in Figure 2, portrays the examination of order exactness and Misclassification pace of the first dataset after preprocessing [19]. From the diagram it tends to be deducing that SVM achieves exactness of 97.29 rates and Naive Bayes accomplishes precision pace of 67.26 rates for in excess of 18700 examples. Innocent Bayes has high Misclassification rate than SVM and for in excess of 18700 examples.

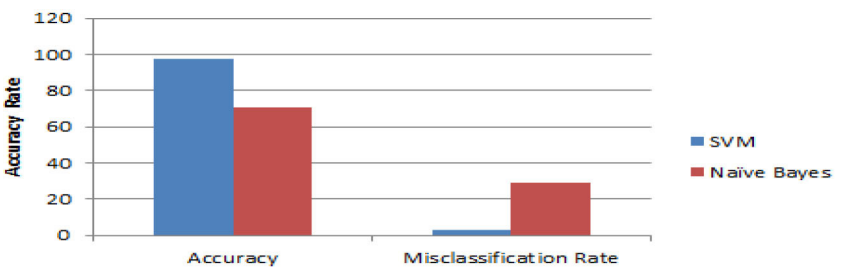

Figure 3

Normalization result

The correlation of the dataset's misclassification rate and grouping precision following standardization is displayed in the Figure 3 [20]. In light of the chart, it very well may be reasoned that SVM accomplishes a precision % of 93.85, though Gullible Bayes accomplishes an exactness pace of 71.001 for a sum of in excess of 18000 cases. Contrasted with SVM,

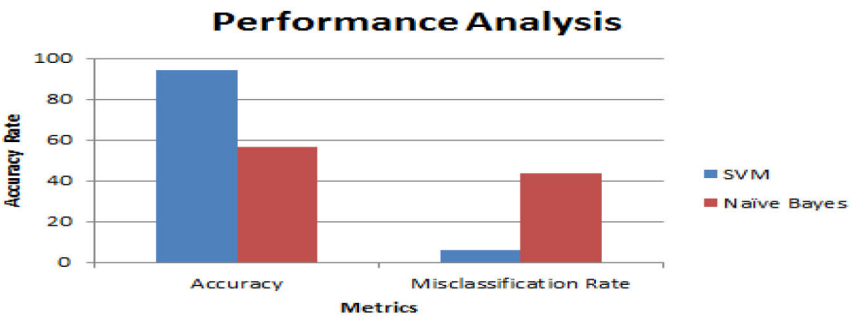

Figure 4

Accuracy and Misclassification rate after feature reduction

Gullible Bayes has a higher misclassification rate for up to in excess of 18,000 cases. For in excess of 18000 cases, the precision rate for Naive Bayes has dropped.

The dataset's examination of misclassification rate and arrangement exactness following component decrease is displayed in the Figure 4 [21]. The main goal of SVM implies to generate a hyperplane line based on a feature of set in regulated for classification of data points [22].

For each new attacks, new rules for detecting attacks needs to be defined [23].

Inview of the chart, it very well may be reasoned that SVM accomplishes an exactness % of 91.95, while Gullible Bayes accomplishes a precision pace of 66.54 for a sum of in excess of 18700 cases. For in excess of 18700 cases, Credulous Bayes has a higher misclassification rate than SVM [24]. In recent years, AI and machine learning (ML) have gained significant popularity in effectively addressing cybersecurity challenges [25-28]. So, numbers of researchers are trying hard to design state of art methods for preventing data from theft or modification or even entering it in one way or another [29]. Insider threats are very difficult to detect comparatively external threats [30]

7. Conclusion

The current patterns call for interruption identification and avoidance. Interruption identification and counteraction are fundamental since organizations and data frameworks assume a significant part in our day-to-day exercises. Interruption identification frameworks have utilized different procedures. AI is one of the most significant of them. SVM and Naive Bayes are two instances of the AI calculations shrouded in this

assessment. It recommends while tending to more than 18,700 cases. SVM works better compared to Naive Bayes. To increment precision, a half and half staggered model will be worked for future works those arrangements with enormous volumes of information.

References

[1] H. Wang, J. Gu, and S. Wang, "An effective intrusion detection framework based on SVM with feature augmentation," *Knowledge-Based Systems*, vol. 136, pp. 130-139 (2017).

[2] S. Roshan, Y. Miche, A. Akusok, and A. Lendasse, "Adaptive and online network intrusion detection system using clustering and extreme learning machines," *Journal of the Franklin Institute*, vol. 355, no. 4, pp. 1752-1779 (2018).

[3] W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system," *Expert Systems with Applications*, vol. 67, pp. 296-303 (2017).

[4] Y. K. Mali and A. Mohanpurkar, "Advanced pin entry method by resisting shoulder surfing attacks," in 2015 International Conference on Information Processing (ICIP), pp. 37-42, IEEE, Dec. (2015).

[5] B. G. Atli, Y. Miche, A. Kalliola, I. Oliver, S. Holtmanns, and A. Lendasse, "Anomaly-based intrusion detection using extreme learning machine and aggregation of network traffic statistics in probability space," *Cognitive Computation*, vol. 10, no. 5, pp. 848-863 (2018).

[6] P. He, J. Zhu, S. He, J. Li, and M. R. Lyu, "A feature reduced intrusion detection system using ANN classifier," *Expert Systems with Applications*, vol. 88, pp. 249-247, Dec. (2017).

[7] T. S. Ruprah, V. S. Kore, and Y. K. Mali, "Secure data transfer in android using elliptical curve cryptography," in 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), pp. 1-4, IEEE, Feb. (2017).

[8] K. A. Garcia, "Analyzing log files for postmortem intrusion detection," in 2012 IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), pp. 1690-1704, Jun. (2012).

[9] R. M. Elbasiony, E. A. Sallam, T. E. Eltobely, and M. M. Fahmy, "A hybrid network intrusion detection framework based on random forests

and weighted k-means," *Ain Shams Engineering Journal*, vol. 4, no. 4, pp. 753-762 (2013).

[10] M. Kumar, D. Kumar, and M. A. K. Akhtar, "Mathematical model for sink mobility (MMSM) in wireless sensor networks to improve network lifetime," in International Conference on Communication, Networks and Computing, pp. 133-141, Singapore: Springer Singapore, Mar. (2018).

[11] Y. Mali, V. U. Rathod, M. M. Kulkarni, P. Mokal, S. Patil, V. Dhamdhere, and D. R. Birari, "A comparative analysis of machine learning models for soil health prediction and crop selection," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 10s, pp. 811-828 (2023).

[12] P. Wasnik and N. Chavhan, "A review paper on designing intelligent intrusion detection system using deep learning," in 2023 11th International Conference on Emerging Trends in Engineering & Technology-Signal and Information Processing (ICETET-SIP), pp. 1-6, *IEEE*, Apr. (2023).

[13] M. D. T. Bennet, M. P. S. Bennet, and D. Anitha, "Securing smart city networks—intelligent detection of DDoS cyber attacks," in 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), pp. 1575-1580, *IEEE*, Dec. (2022).

[14] G. J. Pandeeswari and S. Jeyanthi, "Analysis of intrusion detection using machine learning techniques," in 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), pp. 1-5, *IEEE*, Dec. (2022).

[15] A. Oseni, N. Moustafa, G. Creech, N. Sohrabi, A. Strelzoff, Z. Tari, and I. Linkov, "An explainable deep learning framework for resilient intrusion detection in IoT-enabled transportation networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 1, pp. 1000-1014 (2022).

[16] A. Kumar, M. Kumar, R. P. Mahapatra, P. Bhattacharya, T. T. H. Le, S. Verma, and K. Mohiuddin, "Flamingo-optimization-based deep convolutional neural network for IoT-based arrhythmia classification," *Sensors*, vol. 23, no. 9, p. 4353 (2023).

[17] N. Janu, M. Tiwari, N. Prabha, M. K. Kumari, P. B. Sharma, A. Thakur, and D. S. Rajput, "Development of an efficient real-time H.264/AVC advanced video compression encryption scheme," *Journal of Discrete*

Mathematical Sciences and Cryptography, vol. 24, no. 8, pp. 2245-2255 (2021).

[18] D. Li, L. Deng, W. Liu, and Q. Su, "Improving communication precision of IoT through behavior-based learning in smart city environment," *Future Generation Computer Systems*, vol. 108, pp. 512-520 (2020).

[19] A. Kumar, R. S. Soni, R. R. Prajapati, S. K. Gupta, N. J. Rajput, and R. Jain, "An enhanced quantum key distribution protocol for security authentication," *Journal of Discrete Mathematical Sciences and Cryptography,* vol. 22, no. 4, pp. 499-507 (2019).

[20] A. A. Diro and N. Chilamkurti, "Distributed attack detection scheme using deep learning approach for Internet of Things," *Future Generation Computer Systems*, vol. 82, pp. 761-768 (2018).

[21] B. K. Mohanta, U. Satapathy, and D. Jena, "Addressing security and computation challenges in IoT using machine learning," in Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2020, pp. 67-74, Springer Singapore (2021).

[22] P. Malhotra and S. K. Malik, "Fake news detection using supervised machine learning techniques," *Journal of Information and Optimization Sciences*, vol. 43, no. 1, pp. 7-15 (2022).

[23] M. Malik, M. Kumar, V. Kumar, A. K. Gautam, S. Verma, S. Kumar, and D. Goyal, "High level browser security in cloud computing services from cross-site scripting attacks," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 25, no. 4, pp. 1073-1081 (2022).

[24] A. T. Siahmarzkooh, J. Karimpour, and S. Lotfi, "A cluster-based approach towards detecting and modeling network dictionary attacks," *Engineering, Technology & Applied Science Research*, vol. 6, no. 6, pp. 1227-1234 (2016).

[25] A. Kumar, R. S. Soni, R. R. Prajapati, S. K. Gupta, N. J. Rajput, and R. Jain, "An improved quantum key distribution protocol for verification," *Journal of Discrete Mathematical Sciences and Cryptography,* vol. 22, no. 4, pp. 491-498 (2019).

[26] A. Pratap, A. Kumar, and M. Kumar, "Analyzing the need of edge computing for Internet of Things (IoT)," in Proceedings of Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020, pp. 203-212, Singapore: Springer Singapore, May (2021).

[27] V. Singh, V. K. Yadav, P. Kumari, M. R. Singh, A. Singh, and A. S. Soni, "Source redundancy management and host intrusion detection in wireless sensor networks," Recent Advances in Computer Science and Communications (formerly Recent Patents on Computer Science), vol. 14, no. 1, pp. 43-47 (2021).

[28] S. Upadhyay, M. Kumar, A. Upadhyay, S. Verma, K. Kavita, M. Kaur, and P. A. Castillo, "Challenges and limitation analysis of an IoT-Dependent system for deployment in smart healthcare using communication standards features," *Sensors*, vol. 23, no. 11, p. 5155 (2023).

[29] K. K. Jabbar, H. A. Hailal, and A. T. Naser, "Information security based on sub-system keys generator by utilizing polynomials method and logic gate," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 26, no. 4, pp. 1135–1143 (2023), DOI: 10.47974/JDMSC-1549.

[30] S. K. Henge and A. K. Saini, "Analysis and detection of insider attacks using behaviour rule based architecture in enterprise multitenancy," *Journal of Discrete Mathematical Sciences & Cryptography*, vol. 707, pp. 718 (2023).

Received July, 2024