RESEARCH ARTICLE **ISSN 2738-9898 (Print) ISSN 2738-9901 (Online)**

Hybrid Method for Network Anomaly Detection: Combining Clustering and Classification with Feature Selection

Ashish KC Khatri [1] and Rammani Adhikari [2,*]

[1] Department of Information and Communication Technology, Pokhara University, Nepal
[2] School of Engineering, Faculty of Science and Technology, Pokhara University, Nepal
* Author to whom correspondence should be addressed; E-Mail: *rma.tbi@gmail.com*

Received : 15 February 2025; Received in revised form : 12 April 2025; Accepted : 18 April 2025; Published: 04 July 2025

Abstract

In this study, we have devised an innovative method for automating the classification of network attacks, leveraging a hybrid approach to feature selection. By employing this technique, we were able to dynamically reduce the feature set from an initial 47 to a more manageable 15, streamlining the analysis process while retaining crucial information for the accurate identification of attack patterns. The method also integrates a clustering classification framework, where K-means clustering serves as the foundation for grouping similar data points. To determine the optimal number of clusters, we employed the elbow method, resulting in the selection of a value for k that maximizes cluster homogeneity. Through rigorous validation, we ensured the robustness of our clustering approach, achieving a silhouette coefficient of 0.7048, indicative of well-defined and distinct clusters. Subsequently, we trained and evaluated an XGBOOST algorithm on our refined dataset. The XGBOOST algorithm, renowned for its effectiveness in handling structured data and classification tasks, exhibited remarkable performance. Our model achieved an outstanding overall accuracy rate of 0.9991, underscoring its proficiency in accurately classifying network attacks with a high degree of precision and reliability.

Keywords

Cyber Security, Distributed Denial of Service (DDoS), Extreme Gradient Boosting (XGBOOST), Intrusion Detection System (IDS), One-Class Support Vector Machine (OCSVM), Sub Space Clustering

1. Introduction

The improvement of the performance of the intrusion detection for new and unknown attacks by classifying them based on similar behaviour of the cluster is a challenge in the Internet of Things (IoT) based networks. It is anticipated that by year 2025, 41.6 billion IoT devices will be interconnected, which poses many challenges for the practical realization of IoT. For example, there are issues with data integrity and confidentiality in massive IoT networks. There are now more security issues to be concerned about, such as zero-day attacks on internet users [1].

Intrusion Detection System (IDS) are crucial in addressing these security challenges. IDS can be classified as misuse (also called signature-based) detection and anomaly (also called behaviour-based) detection. Misuse detection approaches are designed to detect attacks by using a database of predefined attack patterns. They can detect known attacks but are unable

to defend the system against unknown attacks because such attacks do not exist in the predefined pattern list [2]. Among the numerous threats that exist, the Distributed Denial of Service (DDoS) attack stands out as a relatively straightforward yet highly potent technique for targeting both intranet and internet resources. Typically, in such attacks, legitimate users cannot access web-based services due to a substantial number of compromised machines. DDoS attacks can be executed at various levels, including the network, transport, and application layers, employing diverse protocols like TCP, UDP, ICMP, and HTTP [3].

A study proposing an ensemble-based intrusion detection model has been conducted. This model integrates logistic regression, naive Bayes, and decision trees, utilizing a voting classifier to enhance performance. The model's efficacy was compared against several prominent state-of-the-art techniques using the CICIDS2017 dataset. The findings demonstrate a notable improvement in accuracy over existing models for both binary and multi-class classification scenarios. The result obtained was 88.92% for binary classification and 88.96% for multi-class classification [1].

Likewise, another machine learning technique for unsupervised anomaly detection has been introduced, combining Sub-Space Clustering (SSC) with One-Class Support Vector Machine (OCSVM) to detect attacks without prior knowledge. This approach was evaluated using the well-known NSL-KDD dataset. Experimental results showed that their method outperforms several existing techniques. The 89% test accuracy was obtained from the study [2].

Similarly, the Convolutional Neural Network's capability to extract spatial features and the Long Short-Term Memory Network's strength in capturing temporal features was used to develop a hybrid intrusion detection system. To enhance the model's performance, the system incorporated batch normalization and dropout layers. The model was trained on three datasets—CIC-IDS 2017, UNSW-NB15, and WSN-DS—using both binary and multiclass classification approaches. An accuracy of 93.5% was obtained for binary classification and 82% accuracy for multi-class classification [4]. The research with implementation of hybrid and ensemble methods are getting popular. This research implemented two feature dimensionality reduction techniques: (i) Auto-Encoder (AE), a deep learning method, and (ii) Principal Component Analysis (PCA). The low-dimensional features obtained from these techniques are then used to construct various classifiers, including Random Forest (RF), Bayesian Network, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA), for designing an intrusion detection system (IDS). The experimental results demonstrate that using low-dimensional features in both binary and multi-class classification improves Detection Rate (DR), F-measure, False Alarm Rate (FAR), and Accuracy. This research successfully reduces the CICIDS2017 dataset's feature dimensions from 81 to 10, while maintaining a high accuracy of 99.6% in binary classification [5].

Moreover, a new hybrid model that combines machine learning and deep learning to enhance detection rates while ensuring dependability was introduced. This method optimizes pre-processing by integrating SMOTE for data balancing and Extreme Gradient Boosting (XGBoost) for feature selection. They compared our approach to various machine learning and deep learning algorithms to identify the most efficient one for implementation in the pipeline. Additionally, they selected the most effective model for network intrusion based on a set of benchmarked performance analysis criteria. Their method demonstrated outstanding results when tested on the KDDCUP'99 and CIC-MalMem-2022 datasets, achieving accuracies of 99.99% and 100%, respectively, without any overfitting or Type-1 and Type-2

errors [6]. Furthermore, an Enhanced Intrusion Detection System using Agent Clustering and Classification based on Outlier Detection (EIDS-ACC-OD) is proposed. Initially, preprocessing is conducted to eliminate unwanted data through outlier detection. Subsequently, a modified K-means clustering algorithm is developed for data segmentation. Finally, K-Nearest Neighbour (KNN) is employed to categorize the attacks. The result obtained from KNN was 92.2% accuracy [7].

Additionally, an unsupervised machine learning model using k-means clustering was proposed to create an Intrusion Detection System (IDS) with high efficiency and low false positive and false negative rates. The model was tested on the NSL-KDD dataset, which contains 25,192 entries across 22 different data types. The study evaluated the model using 11, 22, 44, 66, and 88 clusters, resulting in efficiency rates of 70.75%, 81.61%, 65.40%, 61.30%, and 55.43% respectively; false positive rates of 0.74%, 4.03%, 15.55%, 21.47%, and 31.91% respectively; and false negative rates of 99.82%, 98.14%, 97.76%, 96.32%, and 95.70%, respectively. Notably, the best performance was achieved when the number of clusters matched the number of data types in the dataset. Based on these findings, it is recommended to explore other data mining techniques. Additionally, a follow-up study using the k-means algorithm combined with a signature-based approach is proposed to reduce the false negative rate, and the development of a system to automatically determine the optimal number of clusters is suggested. [8]

The other study tried to identify which model fits better with each kind of attack to define a set of reasoner modules. In addition, this research work proposes to organize these modules to feed a selection system, that is, a dynamic classifier. Finally, the study shows that when using the proposed dynamic classifier model, the detection range increases, improving the detection by each model in terms of accuracy. [9] Additionally, the research on ensemble learning is widely used for detection. a network intrusion detection system (NIDSE) is designed for Smart Homes using an ensemble model to identify attacks on smart home devices. The task of classifying these attacks is approached as a predictive classification problem, utilizing XGBoost. This ensemble method adds models sequentially to correct errors until no further improvements can be achieved. The performance of NIDSE is evaluated on the IoT network intrusion (IoT-NI) dataset, which includes various types of network attacks such as host discovery, synchronized sequence number (SYN), acknowledgment (ACK), and hypertext transfer protocol (HTTP) flooding. Cross-validation results indicate that the XGBoost classifier achieves a micro-average precision of 94% and a macro-average precision of 85% in classifying the nine different types of attacks [10]. This study presents AttackNet, a deep learning model using an adaptive CNN-GRU architecture for detecting botnet attacks in IIoT environments. Tested on recent datasets, including N-BaIoT, it achieved 99.75% accuracy, 0.0063 loss, and precision/recall of 99.75% and 99.74%, respectively. AttackNet outperforms existing methods by 3.2%–16.07%, offering superior real-time threat detection and classification in IIoT networks [11].

This study proposes an improved IDS for IoT security using multimodal big data representation and transfer learning. PCAP files are processed with Spark-based optimization, and semantic features are extracted using word2vec. Network bytes are converted to images, with texture features captured via an attention-based ResNet. Text and texture features are

combined for attack classification, achieving 98.2% accuracy on CIC-IoT 2022, CIC-IoT 2023, and Edge-IIoT datasets [12].

This study proposes a hybrid intrusion detection system combining SVM and Grey Wolf Optimization (GWO). SVM handles classification, while GWO optimizes kernel functions, feature selection, and parameters. Tested on NSL-KDD and TON_IoT datasets, the model outperforms others in accuracy, precision, recall, and F-score [13].

Despite advancements, there remains scope for improvement through novel approaches. A key limitation in existing studies is the manual assignment and aggregation of attack types into general classes. As technological evolution leads to an increase in feature dimensionality, relying on a single feature selection method may be insufficient. Therefore, integrating two distinct feature selection techniques provides a more robust foundation for identifying the most relevant features, thereby enhancing detection performance. Motivated by prior research employing various machine learning and deep learning models, this study proposes a hybrid feature selection method combined with the XGBoost algorithm. The primary contribution is the development of an automated attack classification framework leveraging clustering techniques to enhance detection and classification accuracy.

2. Methodology

The proposed approach integrates feature selection, clustering, and classification to enhance network anomaly detection. The workflow, illustrated in Figure 1, consists of data preprocessing, feature selection, clustering, and classification using XGBoost.

2.1 Data collection

The study utilized the CICIOT2023 dataset, comprising 1,382,658 records with 47 features. Preprocessing involved:

a) Memory optimization by converting float64 data to float32, reducing memory usage from 458 MB to 253 MB.
b) Standard scaling to normalize numerical features between -1 and 1 using z-score normalization. Standard scaling uses mean and standard deviation to compute the standard score (also called as z score) as follows:

$$\text{z score} = \frac{\text{original value}(x) - \text{mean}\ (\mu)}{\text{Standard Deviation}\ (\sigma)} \quad (1)$$

The Standard Scaler in Scikit-learn is a powerful tool for pre-processing numerical data, particularly when your data is susceptible to outliers or skewness. The numerical data were converted into values ranging from -1 to 1.

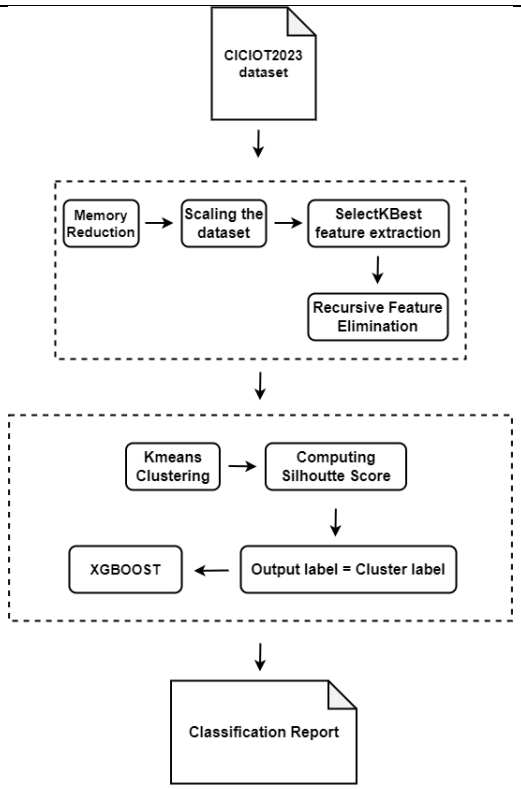

Figure 1: Overall Methodology

2.2. Feature Selection

To improve model efficiency, a hybrid feature selection approach (combination of SelectKBest and Recursive Feature algorithm) was used:

1. SelectKBest identified the top 30 features based on statistical significance.
2. Recursive Feature Elimination (RFE) further refined the selection to 15 features using a Random Forest-based ranking. The data are depicted as shown in Table 1.

Table 1: Feature selection

Description	No. of features
Original features	47
After SelectKBest	30
After RFE	15

Based on relevant research [4], SelectKBest was used with $k = 30$ to reduce the features from 47 to the top 30, listed in Table 2, which were then passed to the RFE algorithm.

Table 2: Selected features after applying SelectKBest

Selected Features	**Selected Features**
flow_duration	SSH
Header_Length	TCP
Protocol Type	UDP
fin_flag_number	ARP
syn_flag_number	ICMP

rst_flag_number	Tot sum
psh_flag_number	Min
ack_flag_number	Max
ack_count	AVG
syn_count	Std
fin_count	Tot size
urg_count	Magnitude
rst_count	Radius
HTTP	Covariance
HTTPS	Variance

The final set of 15 features used for clustering, as shown in Table 3, was selected using RFE with a Random Forest classifier. RFE iteratively ranks features based on importance weights and eliminates the least important ones, retraining the model at each step, until the optimal subset is identified [14].

Table 3: Selected Feature after RFE algorithm

Selected Feature	**Selected Feature**
flow_duration	urg_count
Header_Length	Tot sum
Protocol Type	Min
fin_flag_number	Max
syn_flag_number	AVG
psh_flag_number	Tot size
ack_count	Magnitue
syn_count	

2.3. Clustering with K-Means

K-Means clustering was employed to group similar network behaviors. The optimal number of clusters was determined as k = 8 using the elbow method (Figure 2), consistent with [15]. Cluster validity was assessed using the Silhouette Score, yielding a value of 0.7048

2.4. Silhouette Score

The Silhouette Score evaluates clustering quality without requiring a training set, making it well-suited for unsupervised tasks. It is defined as:

$$S(x_i) = \frac{b(x_i) - a(x_i)}{\max\{b(x_i), a(x_i)\}} \quad (2)$$

Where, x_i is an element in cluster πk, $a(x_i)$ is the average distance of x_i to all other elements in the cluster πk (within dissimilarity), $b(x_i) = min\{dl(x_i)\}$, among all clusters $l \neq k$.

The clustering algorithm's performance will be evaluated using the Silhouette score, which ranges from -1 to 1. A score closer to 1 indicates better clustering quality.

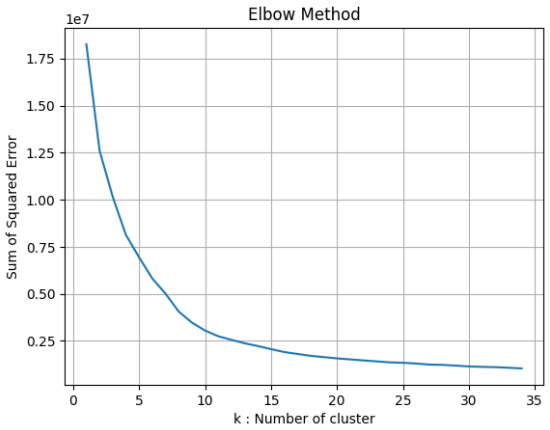

Figure 2: Elbow method to compute the value of k in K-means algorithm

2.5. Algorithm used

Renowned for its high accuracy, robustness to outliers, scalability to large datasets via parallelization, and interpretability features, XGBoost is a popular decision tree-based ensemble learning algorithm that sequentially builds a stronger model by iteratively adding weak learners designed to correct the errors of their predecessors, achieving exceptional performance across regression, classification, and ranking tasks [16].

2.6. Train/Test Data Split

The overall data of 1,382,658 rows were divided into two parts: training data set and test data set in the ratio of 80-20 percent.

3. Results and Discussion

This study used the CICIoT2023 dataset for anomaly detection. The dataset was processed and most relevant feature were selected based on the hybrid model of SelectKBest algorithm and Recursive Feature Elimination method. Then the value of K was computed to apply K-means clustering algorithm. Figure 2 shows the SSE vs. number of clusters to identify the elbow point. The optimal k was found to be 8, and clustering was then applied. The performance of the k-means clustering algorithm was measured computing the silhouette coefficient. The silhouette score obtained was 0.7048. The XGBOOST algorithm was trained and tested, the overall runtime of the algorithm was 85.52 seconds with overall accuracy of 0.9991. Table 4 shows the overall classification report obtained after testing the well-trained XGBOOST algorithm with unknown data.

Table 4: Classification Report of XGBOOST algorithm

Label	Precision	Recall	F1-Score	Support
0	0.99	1.00	0.99	10760
1	1.00	1.00	1.00	125410
2	1.00	1.00	1.00	17702
3	1.00	1.00	1.00	54591
4	1.00	1.00	1.00	15123
5	0.97	0.97	0.97	387
6	1.00	1.00	1.00	19818
7	0.00	0.00	0.00	15

There were total of 243,806 test datasets distributed among the 8 different clusters. The classification reports shows that the Precision, Recall and F1-score value for attack type 1,2,3, 4 and 6 are perfect 1.0 which means our model identified the unknown attack types perfectly for these five types. Whereas for the attack type 0, the precision score is 0.99, Recall score is 1.0 and F1-score is also perfect 1.0. Likewise, for attack type 5 the model performance was good with score of 0.97 across all the performance metrics. The model performs exceptionally well on most labels, with precision, recall, and F1-scores close to 1.00, indicating high accuracy. The only exception is label 7, which has a very low support of 15, resulting in poor performance (0.00 for precision, recall, and F1-score). The classification report indicates that the model demonstrated strong performance when applied to the dataset partitioned into multiple cluster labels.

Table 5: Performance Metrics-Based Result Comparison

Ref.	Algorithm	Multiclass
[1]	Ensemble (NB, DT, LR)	88.96
[2]	SSC-OCSVM	89
[4]	CNN-LSTM	82
[8]	K-means	81.61 (11 clusters) 65.40 (22 clusters) 61.30 (44 clusters) 55.43 (88 clusters)
[7]	KNN, SVM, DT, RF, XGBOOST, MLP, LSTM	77.9, 73.9, 80.1, 82.8, 82.4, 81.1, 81.6
[17]	ADASYN, RENN	99.65
[18]	Recursive Feature Elimination	90.79
[19]	XGBoost-DNN	96
[20]	ANN, RF	96.4
[21]	XGBoost	97
[22]	CRCF	99
[23]	One Class SVM	94
Proposed Research	K-means, XGBOOST	99.91 (8 clusters)

Table 5 shows the comparison of the result obtained by proposed study and the previous research. The table compares various algorithms and their performance in terms of multiclass accuracy. The proposed research utilizing K-means and XGBOOST stands out with the highest multiclass accuracy (99.91%) and provides a Silhouette Score (0.7043), indicating well-defined clusters. This suggests the proposed method is highly effective for the given task compared to the other referenced methods. This result shows that the model performance

using hybrid feature selection and combination of K-means and XGBoost algorithm has performed with high accuracy than the ensemble method used in the research [1]. Figure 3 shows a steady log loss decrease, from 1.026 to 0.0003 (training) and 1.026 to 0.0021 (validation), indicating improved model learning. Figure 4 shows classification error dropping to 0.00001 (training) and 0.00080 (validation), indicating strong predictive performance and good generalization.The overall methodology of the research model included multiple algorithms with respective computation time. The research was performed to design an automated aggregation of the attack types in the network based on the clustering of best feature attributes. The integration of a hybrid feature selection technique with the SelectKBest algorithm and Recursive Feature Elimination algorithm along with the clustering technique with a machine learning model, XGBOOST has given good results for multi-class classification as compared to previous research [1], [2], [4], [8] and [7]. The cons of k-means clustering that the values of k should be defined primarily was tackled using the elbow method. The optimal value of k was computed.

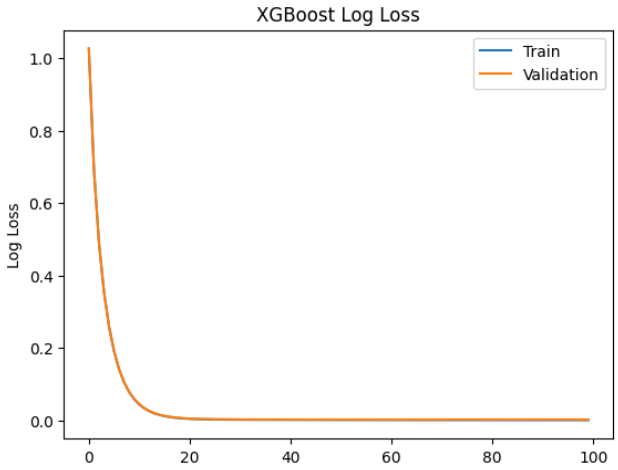

Figure 3: XGBoost Log Loss

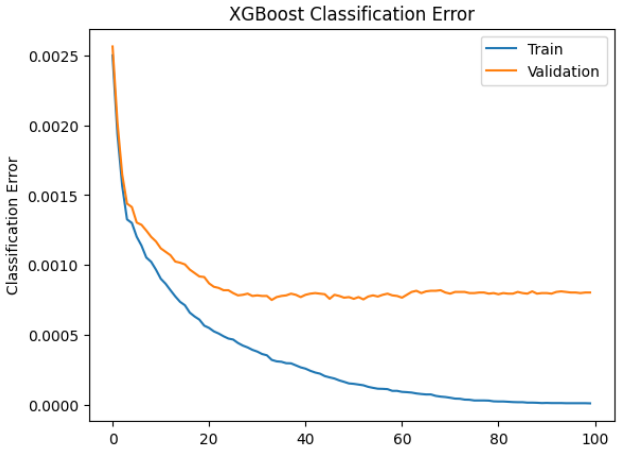

Figure 4: XGBoost Classification Error

The proposed model can be implemented in home IoT appliances, medical appliances or industry based IoT applications to defend against zero-day attacks. The rapid response cyber security team can tackle with immediate response for any unknown attacks and mitigate the risk accordingly. Moreover, the research helps on classifying zero-day attacks to the most

likely one and will make easier to act accordingly. This has unlimited potentials not bound to home and corporate IoT but in Global network as well.

4. Conclusions

The proposed algorithm significantly advances anomaly detection by enhancing clustering methods for real-time identification of new attacks. Misclassifying novel threats remains a concern, highlighting the need for refined clustering validation. Extensive research confirmed the effectiveness of combining hybrid feature selection with integrated clustering and classification models, achieving 99.91% accuracy. This approach enables automated aggregation and detection of diverse network attacks, marking a major step forward in proactive cyber defense and threat mitigation.

Conflicts of Interest Statement

The authors declare no conflicts of interest for this study.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] A. Abbas, M. A. Khan, S. Latif, M. Ajaz, A. A. Shah and J. Ahmad, "A New Ensemble-Based Intrusion Detection System for Internet of Things," *Arabian Journal for Science and Engineering,* 2022.

[2] G. Pu, L. Wang, J. Shen and F. Dong, "A Hybrid Unsupervised Clustering-Based Anomaly Detection Method," *Tsinghua Science and Technology,* vol. 26, no. 2, 2021.

[3] I. Sharafaldin, A. H. Lashkari, S. Hakak and A. A. Ghorbani, "Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy," Canadian Institute for Cyber Security.

[4] A. Halbouni, T. S. Gunawan, M. H. Habaebi, M. Halbouni, M. Kartiwi and R. Ahmad, "CNN-LSTM: Hybrid Deep Neural Network for Network Intrusion Detection System," *IEEE Access,* vol. 10, 2022.

[5] R. Abdulhammed, H. Musafer, A. Alessa, M. Faezipour and A. Abuzneid, "Features Dimensionality Reduction Approaches for Machine Learning Based Network Intrusion Detection," *Electronics MDPI,* vol. 8, 2019.

[6] M. A. Talukder, K. F. Hasan, M. M. Islam, M. A. Uddin, A. Akhter, M. A. Yousuf, F. Alharbi and M. A. Moni, "A Dependable Hybrid Machine Learning Model for Network Intrusion Detection," *Journal of Information Security and Applications,* 2023.

[7] X. Larriva-Novo, C. Sanchez-Zas, V. A. Villagra, M. Vega-Barbas and D. Rivera, "An Approach for the Application of a Dynamic Multi-Class Classifier for Network Intrusion Detection Systems," *electronics MDPI,* vol. 9, no. 1759, 2020.

[8] S. Duque and M. N. b. O. Dr., "Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)," *Procedia Computer Science,* vol. 61, pp. 46-51, 2015.

[9] S. Sandosh, V. Govindasamy and G. Akila, "Enhanced Intrusion Detection System through Agent Clustering," *Peer-to-Peer Networking and Applications,* vol. 13, pp. 1038-1045, 2020.

[10] M. Amru, R. J. Kannan, E. N. Ganesh, S. Muthumarilakshmi, K. Padmanaban, J.

Jeyapriya and S. Murugan, "Network intrusion detection system by applying ensemble model for smart home," *International Journal of Electrical and Computer Engineering,* vol. 14, no. 3, pp. 3485-3494, 2024.

[11] H. Nandanwar and R. Katarya, "Deep learning enabled intrusion detection system for Industrial IOT environment," *Expert Systems with Applications,* vol. 249, 2024.

[12] F. Ullah, A. Turab, S. Ullah, D. Cacciagrano and Y. Zhao, "Enhanced Network Intrusion Detection System for Internet of Things Security Using Multimodal Big Data Representation with Transfer Learning and Game Theory," *sensors,* 2024.

[13] H. Ghasemi and S. Babaie, "A new intrusion detection system based on SVM–GWO algorithms for Internet of Things," *Wireless Networks,* vol. 30, pp. 2173-2185, 2024.

[14] H. Jeon and S. Oh, "Hybrid-Recursive Feature Elimination for Efficient Feature Selection," *applied sciences MDPI,* vol. 10, no. 3211, 2020.

[15] D. M. Saputra, D. Saputra and L. D. Oswari, "Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method," in *Sriwijaya International Conference on Information Technology*, 2019.

[16] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," arxiv, Washigton, 2016.

[17] B. Cao, C. Li, Y. Song, Y. Qin and C. Chen, "Network Intrusion Detection Model Based on CNN and GRU," *MDPI,* 2022.

[18] B. Venkatesh and J. Anuradha, "A Hybrid Feature Selection Approach for Handling a High-Dimensional Data," *Innovations in Computer Science,* 2019.

[19] P. Devan and N. Khare, "An efficient XGBoost–DNN-based classification model for network," *Neural Computing and Applications,* 2020.

[20] E. E. Abdallah, Wafa' Eleisah and A. F. Otoom, "Intrusion Detection Systems using Supervised Machine Learning Techniques: A survey," in *The 13th International Conference on Ambient Systems, Networks and Technologies (ANT)*, Porto, Portugal, 2022.

[21] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *MDPI,* 2021.

[22] K. Pramilarani and P. V. Kumari, "Cost based Random Forest Classifier for Intrusion Detection System in Internet of Things," *Applied Soft Computing,* vol. 151, 2024.

[23] A. Kaushik and H. Al-Raweshidy, "A novel intrusion detection system for internet of things devices and data," *Wireless Networks,* vol. 30, pp. 285-294, 2023.

[24] A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman and A. Alazab, "A Novel Ensemble of Hybrid Intrusion Detection System for Detecting Internet of Things Attacks," *electronics,* vol. 8, no. 1210, 2019.