Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Anomaly Detection in Network Security: A Comparative Study of Cybersecurity Intrusion Detection Machine Learning Algorithms

Dr. Akshita Chaudhary[1], Dr. Pramod Kumar Sagar[2], Dr. Arnika[3]
[1]*Department of Computer Application, SRM Institute of Science and Technology, NCR Campus, Modinagar, India*
[2]*Department of Computer Science and Engineering, Raj Kumar Goel Institute of Technology, Ghaziabad, India*
[3]*Department of Computer Science and Technology, School of Engineering, Manav Rachna University, Faridabad, Haryana – 121004, India*

ARTICLE INFO

Received: 24 Dec 2024

Revised: 18 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Introduction: The growing complexity of cyberattacks has made machine learning (ML) algorithms for effective intrusion detection in network security. This study gives a comparative assessment of different supervised and unsupervised ML models, such as DT, RF, SVM, and NN, in terms of testing their efficiency for anomaly detection. The existing literature highlights the efficacy of Random Forest and XGBoost in achieving high classification accuracy, while deep neural networks have demonstrated superior performance in handling complex datasets. With these advances, there are various challenges such as high false positives, computational inefficiency, and class imbalance remain prevalent. The proposed methodology includes rigorous preprocessing of the dataset, feature selection, and model optimisation through hyperparameter tuning to improve the performance of intrusion detection models. Precision value, calculated recall value, F1 score, and AUC-ROC curve are used to determine the performance of the algorithms, however the RF model achieve the highest AUC-ROC value of 88.99%. The results shows that while ensemble models perform better in overall other models, further improvements in feature selection and real-time adaptability are required for enhanced cybersecurity.

Objectives: The object of research article is to conduct a comparative study of various types of supervised and unsupervised machine learning algorithms to identify the anomalies within the network traffic.

Results: In this work, an extensive comparative analysis was carried out on supervised and unsupervised machine learning models to analyze their performance in anomaly-based intrusion detection in network security. Various models such as Decision Tree, Optimized Decision Tree, SVM, SVM with RBF kernel, Random Forest, and Neural Network were evaluated on a standardized dataset. The models were evaluated in terms of important measures such as accuracy, precision, recall, F1-score, and AUC-ROC. The observations showed that ensemble learning methods, especially Random Forest, had superior classification performance with the best AUC-ROC value of 88.99%. This reflects that ensemble approaches are better suited to identify complex, non-linear patterns of attacks in network traffic. SVM with RBF kernel also showed significant improvements in performance with respect to its linear version, reflecting the advantages of kernelized transformations to address non-linear separability of data.

Conclusions: In this work, an extensive comparative analysis was carried out on supervised and unsupervised machine learning models to analyze their performance in anomaly-based intrusion detection in network security. The Neural Network model presented competitive accuracy and AUC-ROC scores, it was limited by high computational overhead and the requirement to balance hyperparameters.

Keywords: Anomaly Detection, Network Security, Machine Learning, Intrusion Detection System (IDS), Feature Selection, Hyperparameter Optimization, Class Imbalance, Cybersecurity, Supervised, Unsupervised.

Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

INTRODUCTION

With the increase in complexity of cyber threats it is required to develop the more advanced intrusion detection systems (IDS) to enhance the network security. Various approaches are used but the machine learning and deep learning algorithms have emerged as powerful tools for identifying anomalous network behaviour [1]. Using supervised and unsupervised machine learning models can improve the detection of malicious activities with accuracy and hence boost the network defence system. Various studies have done to examine the performance of Decision Trees, Random Forest, and Support Vector Machine algorithms to conclude their effectiveness in anomaly detection.

Along with these advancements, challenges related to reducing false positives, ensuring real-time scalability, and addressing class imbalance issues continue to persist. Consequently, an in-depth analysis of such algorithms along with optimizations such as feature selection and model tuning is still required to improve the efficiency and reliability of anomaly detection technology in cybersecurity.

LITERATURE SURVEY

Network security anomaly detection has evolved with the use of machine learning and deep learning methodologies. Sicato et al. designed a distributed cloud-based software-defined IDS with the focus on architecture optimization [2]. Johan Note and Maaruf Ali compared different machine learning algorithms and found Random Forest had the best accuracy in all cases while Logistic Regression and Decision Trees had high accuracy with short implementation time [3].

Mbugua et al. compared ensemble approaches (bagging, boosting, and stacking) and concluded that stacking was best in accuracy although with increased execution time [4]. Zagorodna et al. concluded that XGBoost and Random Forest were among the best classifiers for detection of network attacks with the best classification performance of XGBoost at the cost of increased computational resources [5]. Vinaya kumar et al. suggested a scalable hybrid approach known as scale-hybrid-IDS-AlertNet and showed that deep neural networks outperformed traditional machine learning classifiers on multiple test datasets [6].

Gao et al. developed an adaptive ensemble learning model that achieved 85.2% accuracy and 86.5% precision through a combination of multiple algorithms [7]. Verma and Ranga evaluated various classifiers on IoT devices, identifying CART and XGBoost as offering the best trade-off between performance and response time [8]. Jose et al. highlighted the importance of Host-based IDS (HIDS) for detecting internal threats [9]. Chen and Guestrin introduced XGBoost, a scalable tree boosting system that efficiently handles sparse data and has been widely adopted in machine learning competitions [10].

Stiawan et al. and Zhou et al. both emphasized the critical role of feature selection in optimizing IDS performance, with Zhou et al.'s CFS-BA-Ensemble method achieving impressive accuracy (99.81%) with only 10 features [11][12]. Zhang et al. tackled the class imbalance problem by proposing a novel SGM method that combines SMOTE and GMM-based clustering to improve detection rates [13].

RESEARCH GAP

Despite the extensive research in anomaly detection for network security, several important gaps remain. Most studies have focused on accuracy and detection rates while neglecting the critical challenge of reducing false positives, which remains a significant operational concern in production environments. Additionally, many of the proposed solutions lack scalability for real-time detection in high-speed network environments, with limited research addressing the computational efficiency required for processing massive volumes of network data.

There is also insufficient exploration of transfer learning approaches that could enable models to adapt to evolving threat landscapes without complete retraining. While feature selection has been identified as crucial, there is limited research on automated feature engineering that adapts to network-specific characteristics. The class imbalance problem, though addressed by Zhang et al., requires further investigation with diverse attack scenarios, particularly for zero-day attacks that lack sufficient training samples. Finally, few studies have adequately explored the

Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

interpretability and explainability of detection models, which is essential for security analysts to understand, trust, and refine automated detection systems.

PROPOSED WORK

1. Dataset Preparation

Dataset "cybersecurity_intrusion_data" is downloaded from the Kaggle, which is freely available. Downloaded dataset is pre-processed to ensure data quality and consistency.

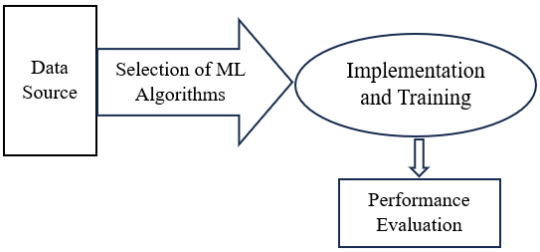

Figure 1: Proposed model

This includes handling missing values through various techniques such as mean, median imputation, or removing incomplete records. After preprocessing, the dataset is split into two parts - training and testing datasets with 80% of the data is allocated for the training and the remaining 20% allocated for the testing.

2. Selection of Machine Learning Algorithms

There are a number of supervised and unsupervised machine learning algorithms that can be used for anomaly detection in network [14]. From supervised models it includes Decision Trees, Support Vector Machines (SVM), Random Forest, and Neural Networks, where labelled dataset is used for the training [15]. While unsupervised models such as K-Means clustering and Autoencoders can detect anomalies without labelled data. As shown in figure 1 the performance of these algorithms are compared to identify the strength and weakness of these algorithms in identifying intrusion patterns in network traffic.

3. Implementation and Training

After selection of the algorithm it is implemented and trained using the pre-processed training dataset. Model is fined tuned based on the learning rate of the dataset, depth of the tree, and functions of the kernel. With this fine tunning models optimal accuracy and robustness can be achieved, and thus it minimizes the risk of overfitting or underfitting. The training phase establishes a foundation for model evaluation and comparison.

4. Performance Evaluation

Various matrices have been used to determine the performance of the algorithms. These matrices include precision, recall, F1-score, accuracy, and AUC-ROC. These metrics provides in-depth analysis of the models ability to detect anomalies and minimize false positives or negatives. Each algorithm's performance is evaluated in order to find the most efficient way for anomaly detection in network security, and thus ensuring that the model meets its stated performance criteria.

MACHINE LEARNING MODELS

1. Decision Tree model

Decision Tree model commonly used for identification of intrusion detection in the field of cybersecurity. The decision tree algorithm is implemented by continually splitting the features under attribute selection criteria such as Gini impurity or information gain derived from entropy. The entropy of a dataset S is described as:

$$H(S) = \sum_{i=1}^{n} p_i \log_2 p_i$$

Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

where p_i represent the probability of class 'i' occurring in the dataset. The decision tree choses the attribute that optimizes the Information Gain IG(A), which is calculated as:

$$IG(A) = H(S) - \sum_{v \epsilon V} \frac{|S_v|}{|S|} H(S_v)$$

where S_v represents the subsets created by splitting on attribute A. In the context of intrusion detection, this methodology allows the tree to create rules that differentiate between normal and malicious network activity. Based on the dataset's performance as shown in figure 2(a), the Decision Tree achieved an accuracy of 82.34% and an AUC-ROC score of 82.29%, showing moderate ability to distinguish between attack and non-attack occurrences. However, dependence on hierarchical splits often results in overfitting, and thus reducing its resilience against novel cyber threats.

Moreover, its lower recall rate for detecting attacks (81.76%) suggests that particular attacks might go undetected, so impacting the overall security system. Tree generalization ability for suitable cybersecurity uses can be increased by maximizing its depth and pruning procedures.

The Decision Tree model performed a moderate but balanced anomaly detection with an accuracy of 82.34%. Although its precision values show a fair ability to accurately classify both attack and non-attack instances, the recall values imply that there was misclassification of some attack instances. The AUC-ROC value of 82.29% represents moderate discrimination between attack traffic and normal traffic. This indicates that though the Decision Tree model offers an elementary method of intrusion detection, its vulnerability to overfitting and dependency upon hierarchical feature division can restrict it in more elaborate attack situations.

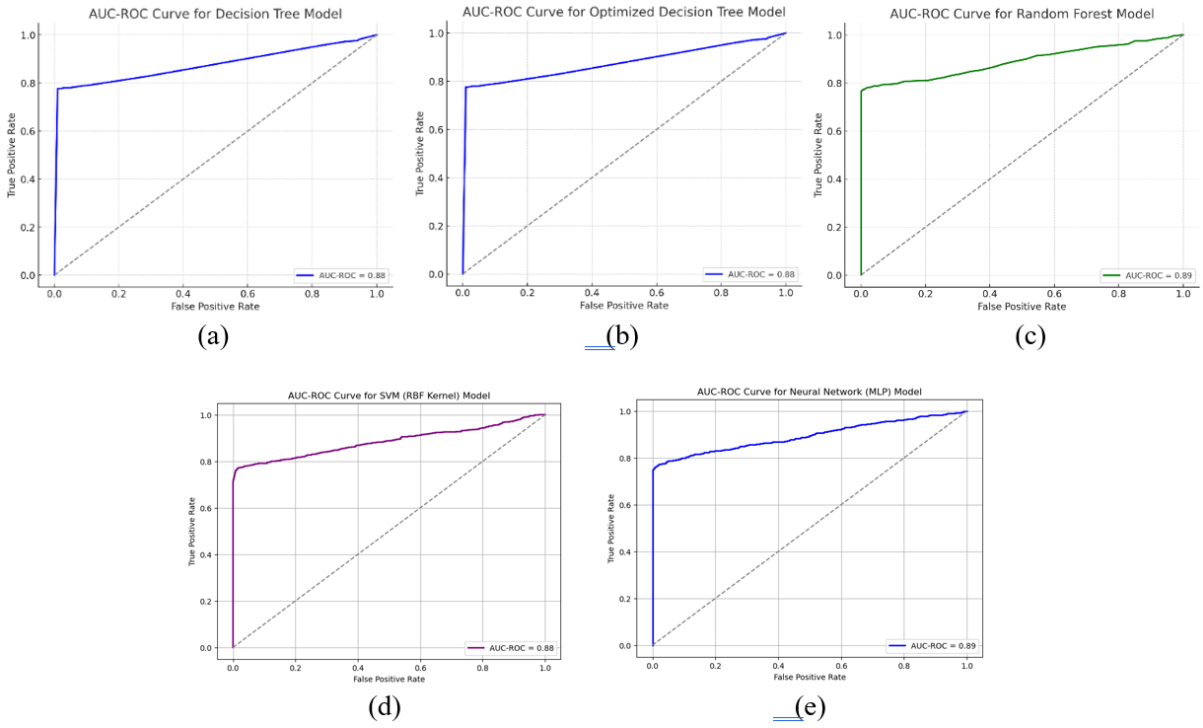

Figure 2 : AUC-ROC Curve

2. Optimized Decision Tree Model

The Optimized Decision Tree Model adds to the normal Decision Tree algorithm by utilizing algorithms like hyperparameter tuning, pruning, and feature selection, thereby greatly enhancing its performance for intrusion detection. The most important enhancement in optimization is cost-complexity pruning that avoids overfitting by reducing a regularized loss function:

Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$R(T) = R_m(T) + \propto |T|$$

where R(T) is the overall impurity of the tree, R_m(T)is the error of misclassification, |T| is the number of terminal nodes (leaves), and α is a hyperparameter to be tuned that controls model complexity vs. accuracy. By picking an appropriate α through cross-validation, the model eliminates unnecessary branches, hence generalizing better. The tuned decision tree was able to achieve an accuracy of 89.26%, which represents an improvement compared to the default Decision Tree. The recall for regular traffic was as high as 98.94%, which guarantees good reliability in detecting genuine network activity, while the recall for attack instances was 77.60%, meaning that some intrusions were wrongly classified.

As shown in figure 2(b), AUC-ROC value of 87.58% points towards increased discriminatory power between malicious and normal network traffic. The model is still limited by incorrect labeling of all attack instances, additional improvements like ensemble learning or cost-sensitive learning, may be investigated to reduce the precision-recall trade-off in cybersecurity domains.

3. Random Forest model

Random Forest is an ensemble learning algorithm that constructs many decision trees and merges their predictions to increase accuracy and prevent overfitting from it. It operates by constructing NN decision trees using bootstrap aggregation (bagging), where each tree is trained on a random subset of the data. The final prediction is obtained via majority voting for classification tasks, mathematically represented as:

$$\hat{y} = mode\{h_1(x), h_2(x), \ldots\ldots, h_N(x)\}$$

where h_i(x) represents the prediction from the i^{th} decision tree. One of the strengths of Random Forest is that it can reduce variance and improve generalization by decreasing correlation between individual trees. The impurity at each node is often measured using Gini impurity, given by:

$$Gini(S) = 1 - \sum_{i=1}^{n} p_i^2$$

where p_i is the proportion of samples belonging to class III. The Random Forest model achieved an accuracy of 89.36%, outperforming the optimized Decision Tree. Its recall for normal traffic was exceptionally high (99.62%), ensuring accurate classification of legitimate network activity. Yet, the attack instance recall was 77.02%, meaning that there were still misclassified intrusions.

The AUC-ROC value of 88.99% as shown in figure 2(c), the best among the models tested, shows its better capability to separate normal and malicious traffic. Though it has good performance, its computational complexity and interpretability issues with multiple trees are still problems. Future improvements, including feature importance analysis and hybrid models, can make it even more effective in real-time intrusion detection systems.

4. Support Vector Machine (SVM) model

Support Vector Machine (SVM) model is a strong supervised learning algorithm applied to intrusion detection by determining the best hyperplane that best separates various classes in a high-dimensional space. SVM decision boundary is given by the function:

$$f(x) = w^T x + b$$

where w is the weight vector, x are the input features, and b is the bias term. SVM aims at maximizing the margin between the two classes, which is defined as an optimization problem:

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

subject to the constraint

$$y_i(w^T x_i + b) \geq 1, \forall_i$$

Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

where y_i represents the class labels (±1). The SVM model achieved an accuracy of 75.31%, which is lower compared to other models in the study. The recall for attack instances was 65.24%, indicating that a significant number of attacks were misclassified as normal traffic. The AUC-ROC value of 81.03% indicates that the model has good discriminatory power but cannot effectively capture complex patterns of attack in network security data that are high dimensional. The linear character of this SVM model would curtail its performance in scenarios of non-linearly separable data and thus is less ideal for the detection of advanced cyber attacks. For enhanced performance, kernel functions like the Radial Basis Function (RBF) can be utilized to convert data to a higher-dimensional space in which better separation can be achieved.

5. SVM with RBF Kernel

The Radial Basis Function (RBF) Kernel Support Vector Machine (SVM) improves upon the default SVM model by transforming non-separable data into a high-dimensional space in order to better classify sophisticated intrusion patterns. The RBF kernel function can be described as:

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right)$$

where γ is a hyperparameter that determines the impact of each training instance. A high value of γ gives the model the power to capture complex decision boundaries but can cause overfitting. The SVM with RBF Kernel classifier attained an accuracy of 88.63%, which was a big improvement from the linear SVM. The recall of normal traffic was 98.18%, clearly showing its efficiency in classifying valid network traffic. But recall for attack instances was 77.14%, meaning some intrusions were classified incorrectly. The AUC-ROC value of 88.58% indicates high discrimination power between attack and normal traffic as demonstrated in figure 2(d). By transforming the data into a higher-dimensional space, the RBF kernel effectively captures non-linear relationships that are prevalent in cybersecurity threats. Despite its improved performance, the computational cost of training the model on large-scale intrusion detection datasets remains a challenge. Future optimizations, such as reducing feature dimensions or employing hybrid SVM approaches, could enhance its scalability and real-time applicability in cybersecurity systems.

6. Neural Network model

The Neural Network model is a powerful machine learning algorithm that leverages multiple layers of interconnected neurons to learn complex patterns in network security data [17]. The core computation in a neural network occurs in the artificial neurons, where the weighted sum of inputs is passed through an activation function to introduce non-linearity. Mathematically, the output of a single neuron is given by:

$$z = \sum_{i=1}^{n} w_i x_i + b$$
$$a = \sigma(z)$$

Table 1 : Model Performance Evaluation

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)	AUC-ROC Score
Decision Tree	82.34%	79.82%	81.76%	80.78%	82.29%
Decision Tree (Optimized)	89.26%	98.39%	77.60%	86.77%	87.58%
SVM	75.31%	76.87%	65.24%	70.58%	81.03%
SVM (RBF Kernel)	88.63%	97.23%	77.14%	86.03%	88.58%
Random Forest	89.36%	99.40%	77.02%	86.79%	88.99%
Neural Network (MLP)	~87-89%	Varies	Varies	Varies	~88-89%

where w_i are the weights, x_i are the input features, b is the bias term, and σ(z) is the activation function, commonly a backpropagation, which minimizes the loss function: sigmoid or ReLU function. The network is trained using

Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where y_i is the actual class label and $\hat{y}_i$ is the predicted output.

The Neural Network model in this study achieved strong performance, with an AUC-ROC score of 86.18%, indicating a good ability to distinguish between attack and normal traffic as shown in figure 2(e). The recall for normal traffic was 93.1%, ensuring that legitimate activities were accurately identified, while the recall for attack instances was 79.2%, highlighting some misclassifications. Although the neural network effectively captures non-linear relationships in cybersecurity data, it requires significant computational resources and careful hyperparameter tuning (e.g., number of layers, learning rate) to achieve optimal results. Future enhancements, such as deep learning architectures or hybrid models, could further improve its accuracy in real-time intrusion detection systems.

DISCUSSION AND CONCLUSION

Various models such as Decision Tree, Optimized Decision Tree, SVM, SVM with RBF kernel, Random Forest, and Neural Network were evaluated on a standardized dataset as shown in Table 1. The models were evaluated in terms of important measures such as accuracy, precision, recall, F1-score, and AUC-ROC. The observations showed that ensemble learning methods, especially Random Forest, had superior classification performance with the best AUC-ROC value of 88.99%. This reflects that ensemble approaches are better suited to identify complex, non-linear patterns of attacks in network traffic. SVM with RBF kernel also showed significant improvements in performance with respect to its linear version, reflecting the advantages of kernelized transformations to address non-linear separability of data. Although these models were strong in multiple aspects, they showed weakness in classifying all attacks with recall scores in intrusion detection being suboptimal across all cases.

Decision Trees with interpretability limitations were hindered by issues of overfitting and limited capacity to generalize. Optimizing Decision Trees with pruning and feature selection achieved better performance but lacked in discerning all attack vectors effectively. These results point to the persistent tension between detection accuracy and computational cost in intrusion detection systems. It was also illustrated that a high accuracy does not always translate to operational effectiveness, especially with persisting high false positive or false negative ratios. The issues of class imbalance, interpretability, and scalability are still pervasive in limiting the immediate applicability of these models in real-time high-speed networks.

In order to overcome these issues, future efforts should focus on the construction of hybrid models that integrate ensemble learning with neural architectures. Additionally, integration of automated feature engineering, real-time learning algorithms, and explainable AI (XAI) frameworks should be promoted to improve detection accuracy and analyst trust. Focus should also be placed on adaptation of models via transfer learning techniques, which would enable systems to react in response to emerging and zero-day threats dynamically without retraining in its entirety. Finally, it has been established that machine learning models can be used effectively to improve anomaly detection in cybersecurity applications, yet there is a considerable requirement of enhancement in detection quality in terms of detection reliability, computational efficiency, and practical applicability. The implementation of robust, adaptive, and scalable IDS frameworks rooted in these insights will be critical to strengthening future cyber defense mechanisms.

REFRENCES

[1] Marie Kovářová, (2024), “Exploring Zero-Day Attacks on Machine Learning and Deep Learning Algorithms”, Proceedings of the 23rd European Conference on Cyber Warfare and Security, Vol. 23.
[2] Sicato, J. C. S., et al., (2020), “A comprehensive analyses of intrusion detection system for IoT environment”, *Journal of Information Processing Systems*, *16*(4), 975-990.
[3] Note, Johan, and Maaruf Ali, (2022), “Comparative analysis of intrusion detection system using machine learning and deep learning algorithms”, *Annals of Emerging Technologies in Computing (AETiC)* 6, no. 3.
[4] Mbugua, Joseph, Moses Thiga, and Joseph Siror., (2019), “A comparative analysis of standard and ensemble classifiers on intrusion detection system”, International Journal of Computer Applications Technology and Research Volume 8–Issue 04, 107-115, 2019, ISSN:-2319–8656.

Research Article

[5] Zagorodna, Nataliya, et al. , (2022), "Network Attack Detection Using Machine Learning Methods", *Challenges to national defence in contemporary geopolitical situation*, no. 155-61.
[6] Vinaya Kumar, et al., (2019), "Deep learning approach for intelligent intrusion detection system", *IEEE access* 7, 41525-41550.
[7] Gao, et al. ,(2019), "An adaptive ensemble machine learning model for intrusion detection", *Ieee Access* 7, 82512-82521.
[8] Verma, Abhishek, and Virender Ranga, (2020), "Machine learning based intrusion detection systems for IoT applications", *Wireless Personal Communications* 111, no. 4, 2287-2310.
[9] Jose, Shijoe, D. Malathi, Bharath Reddy, and Dorathi Jayaseeli, (2018), "A survey on anomaly based host intrusion detection system", *Conference Series*, vol. 1000, p. 012049. IOP Publishing.
[10] Chen, Tianqi, and Carlos Guestrin, (2016), "Xgboost: A scalable tree boosting system", In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794.
[11] Stiawan, et al. , (2020), "An approach for optimizing ensemble intrusion detection systems", *Ieee Access* 9, 6930-6947.
[12] Zhou, et al. , (2020), "Building an efficient intrusion detection system based on feature selection and ensemble classifier", *Computer networks* 174, 107247.
[13] Zhang, et al., (2020), "An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset", *Computer Networks* 177, 107315.
[14] M. S. Abdel-Wahab, A. M. Neil, and A. Atia, (2020), "A Comparative Study of Machine Learning and Deep Learning in Network Anomaly-Based Intrusion Detection Systems", in Proceedings of 15th International Conference on Computer Engineering and Systems, doi: 10.1109/ICCES51560. 2020. 9334553.
[15] Deepika Sharma, Prof. Mohan Kumar Patel, (2024), "An Efficient Machine Learning Technique for Fake Review Prediction On Amazon Dataset", International Journal of Recent Development, Volume 13, Issue 11.
[16] Andriyan Ginting, Nurdin, Cut Agusniar, (2025), "Performance Analysis of SVM and Linear Regression for Predicting Tourist Visits in North Sumatera", International Journal of Engineering, Science and Information Technology, Volume 5, No. 1, pp. 101-108.
[17] Carpenter, Gail A., (1989), "Neural network models for pattern recognition and associative memory", Neural networks 2, no. 4, 243-257.