

International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 4 (2026)



ijerst.editor@gmail.com
editor@ijerst.com

Research Paper

MACHINE LEARNING-BASED FRAMEWORK WITH ENHANCED FEATURE SELECTION AND RESAMPLING FOR IMPROVED INTRUSION DETECTION

K. Kavitha¹, Dr. M. Vinaya Babu²

¹PG Scholar, Department of Computer Science and Engineering, Teegala Krishna Reddy Engineering College, Hyderabad, Telangana, India.

²Associate Professor, Department of Computer Science and Engineering, Teegala Krishna Reddy Engineering College, Hyderabad, Telangana, India.

Email: kavitha24r91d5803@gmail.com¹, mvinayababu@gmail.com²

Abstract:

Intrusion Detection Systems (IDS) play a vital role in protecting modern computer networks against cyber threats, including botnet attacks, Man-in-the-Middle (MITM) attacks, malware, and other malicious activities. However, conventional intrusion detection approaches face significant challenges due to high-dimensional network traffic data, irrelevant and redundant features, and imbalanced class distributions. These limitations can reduce detection performance, increase computational complexity, and result in poor identification of minority attack classes. To address these challenges, this paper proposes a machine learning-based framework that integrates advanced feature selection and data resampling techniques for improved intrusion detection. The framework employs Recursive Feature Elimination, Sequential Feature Selection, and statistical feature selection to identify relevant network traffic attributes. Furthermore, resampling techniques, including Adaptive Synthetic Sampling, SMOTE-ENN, and SMOTE-Tomek, are considered to address class imbalance and improve the representation of minority attack categories. The selected and balanced features are then used to train machine learning classifiers. The performance of the models is evaluated using accuracy, precision, recall, F1-score, and false-positive and false-negative rates. The proposed framework aims to improve the reliability of intrusion detection, enhance minority attack identification, reduce the influence of

irrelevant features, and support efficient network security monitoring.

Keywords: *Intrusion Detection System, Machine Learning, Feature Selection, Recursive Feature Elimination.*

I. INTRODUCTION

The rapid growth of computer networks, cloud computing, the Internet of Things (IoT), and digital communication technologies has significantly increased the volume of data exchanged across modern information systems. Although these technologies provide numerous benefits, they also expose networks to various cybersecurity threats, including Distributed Denial-of-Service (DDoS) attacks, botnets, Man-in-the-Middle (MITM) attacks, malware, phishing, and other unauthorized activities. These threats can compromise data confidentiality, integrity, and availability, making effective network security mechanisms essential for protecting organizational and personal information. Intrusion Detection Systems (IDS) are important security mechanisms designed to monitor network traffic and identify suspicious or malicious activities. Traditional signature-based IDS detect attacks by matching observed traffic against predefined attack patterns, making them effective for known threats but less capable of identifying previously unseen attacks. In contrast, anomaly-based IDS identify deviations from

normal network behavior but may generate a high number of false alarms. Consequently, machine learning techniques have gained considerable attention for improving intrusion detection by learning patterns from historical network traffic and classifying activities as normal or malicious. Despite their potential, machine learning-based IDS face two major challenges: high-dimensional feature spaces and class imbalance. Network traffic datasets often contain irrelevant, redundant, and highly correlated attributes that increase computational complexity and may negatively affect classification performance. Furthermore, normal traffic frequently dominates attack traffic, causing classifiers to favor majority classes and overlook rare but dangerous attacks. These limitations highlight the need for effective feature selection and data balancing strategies. This paper proposes a machine learning-based framework that integrates feature selection, data resampling, and classification techniques to improve intrusion detection performance. Recursive Feature Elimination (RFE), Sequential Feature Selection (SFS), and statistical feature selection are employed to identify relevant network traffic attributes. Resampling methods, including Adaptive Synthetic Sampling, SMOTE-ENN, and SMOTE-Tomek, are considered to address class imbalance. The resulting data is used to train classifiers such as Random Forest, Decision Tree, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Gradient Boosting. The framework is evaluated using performance metrics such as accuracy, precision, recall, F1-score, and false-positive and false-negative rates.

II. LITERATURE SURVEY

Intrusion Detection Systems (IDS) have gained significant attention in cybersecurity research due to the increasing complexity of network attacks and the limitations of traditional security mechanisms. Recent studies have explored machine learning techniques to improve the detection of malicious traffic, reduce false alarms, and identify previously unseen attack patterns. However, the effectiveness of these models

depends on dataset quality, feature representation, and the distribution of attack classes. Feature selection is an important research direction for improving intrusion detection performance. Techniques such as Recursive Feature Elimination (RFE), Sequential Feature Selection (SFS), and statistical feature ranking help identify informative attributes while eliminating redundant and irrelevant features. Reducing the number of input features can lower computational requirements and improve model efficiency. Nevertheless, selecting an appropriate feature subset remains challenging because network traffic datasets contain complex relationships among attributes, and the most useful features may vary across attack categories.

III. PROPOSED WORK

The proposed work introduces a machine learning-based framework for improving network intrusion detection by integrating advanced feature selection, data resampling, and classification techniques. The system uses network intrusion datasets containing normal and malicious traffic records to identify attack patterns and improve the detection of cyber threats. Data preprocessing techniques, including missing-value handling, duplicate removal, categorical encoding, and normalization, are applied to improve data quality. Feature selection algorithms, including Recursive Feature Elimination, Sequential Feature Selection, and statistical feature ranking, are used to identify the most relevant network traffic attributes. Resampling techniques such as Adaptive Synthetic Sampling, SMOTE-ENN, and SMOTE-Tomek are applied to address class imbalance. Machine learning classifiers, are trained to classify network traffic and identify potential intrusions. A systematic evaluation process measures classification performance and supports the development of an efficient and reliable intrusion detection framework.

IV. Methodology

a. Data Collection

The first stage of the proposed system involves collecting network intrusion datasets containing

normal network traffic and different categories of malicious activities. The dataset consists of network traffic attributes and corresponding class labels that indicate whether a record represents legitimate or malicious activity. Attack categories may include botnet attacks, Man-in-the-Middle attacks, and other intrusion types, depending on the selected dataset. These observations provide the input information required for training machine learning models to identify patterns associated with normal behavior and cyber-attacks. The collected dataset is examined to understand its structure, feature types, number of records, and class distribution before further processing.

b. Data Preprocessing

The collected network traffic data is processed to improve its quality and suitability for machine learning analysis. Missing values, duplicate records, invalid entries, and inconsistent class labels are identified and handled appropriately. Categorical attributes are converted into numerical representations, while numerical features are normalized or standardized wherever necessary. The target class is separated from the input features to establish the classification task. The processed dataset is then divided into training and testing subsets, ensuring that the testing data remains independent of the model training process. These preprocessing steps help reduce data inconsistencies and prepare the network traffic records for feature selection and intrusion classification.

c. Feature Selection

Relevant network traffic attributes are selected from the preprocessed dataset to construct an informative feature set for intrusion detection. The proposed framework employs Recursive Feature Elimination, Sequential Feature Selection, and statistical feature selection techniques. RFE repeatedly removes less important features based on a selected estimator's importance or coefficients, whereas SFS identifies a suitable feature subset through sequential evaluation of candidate features. Statistical feature selection examines characteristics such as feature variance,

redundancy, and relationships with the target class. These techniques help eliminate irrelevant and redundant attributes, reduce the dimensionality of the dataset, and limit unnecessary computational overhead.

d. Data Resampling and Model Training

The selected features are processed using data resampling techniques to address the class imbalance commonly found in network intrusion datasets. The proposed framework considers Adaptive Synthetic Sampling, SMOTE-ENN, and SMOTE-Tomek. ADASYN generates synthetic minority-class samples based on the learning difficulty of existing examples. SMOTE-ENN combines minority-class oversampling with the removal of potentially noisy or ambiguous samples, while SMOTE-Tomek combines oversampling with Tomek-link analysis to reduce overlapping samples between classes. Resampling is performed exclusively on the training data to prevent information leakage into the testing stage.

V. ALGORITHMS

The proposed machine learning-based intrusion detection framework integrates feature selection algorithms, data resampling techniques, and machine learning classifiers to improve the identification of malicious network traffic. The algorithms are designed to reduce redundant features, address class imbalance, and enhance the classification of normal and attack traffic. The major algorithms used in the proposed framework are described below.

A. Recursive Feature Elimination (RFE)

Recursive Feature Elimination is a wrapper-based feature selection algorithm that identifies the most relevant attributes by repeatedly eliminating less important features. Initially, a machine learning estimator is trained using the available network traffic features. The importance of each feature is evaluated, and the least significant features are removed iteratively until the desired number of features is obtained. In the proposed intrusion detection framework, RFE reduces the dimensionality of network traffic data and helps the classifier focus on informative attributes,

potentially reducing training time and improving detection performance.

B. Sequential Feature Selection (SFS)

Sequential Feature Selection is a wrapper-based technique that selects an effective subset of features by evaluating their contribution to model performance. In forward selection, the algorithm begins with an empty feature set and adds the feature that provides the greatest improvement at each iteration. In backward selection, it starts with the complete feature set and progressively removes less useful attributes. The proposed framework uses SFS to identify network traffic characteristics that contribute to distinguishing legitimate activities from malicious intrusions. The selected feature subset is then used in subsequent model training.

C. Adaptive Synthetic Sampling (ADASYN)

Adaptive Synthetic Sampling is a data resampling algorithm designed to address class imbalance in classification datasets. It generates synthetic examples for minority classes according to the difficulty of learning existing minority-class samples. More synthetic examples are generated around minority samples that are difficult to classify, allowing the model to learn their patterns more effectively. In the proposed framework, ADASYN is applied to the training data to improve the representation of rare attack categories, such as botnet and Man-in-the-Middle attacks, when these categories are present in the dataset.

D. SMOTE-ENN

SMOTE-ENN combines the Synthetic Minority Over-sampling Technique with Edited Nearest Neighbours. SMOTE generates synthetic minority-class examples by interpolating between existing minority samples and their nearest neighbours. ENN then examines the neighbourhood of samples and removes observations considered inconsistent with their neighbouring classes. In the proposed framework, SMOTE-ENN is used to increase the representation of minority attack classes while reducing noisy or ambiguous training examples. This process may improve

class separation and support more reliable intrusion classification.

E. SMOTE-Tomek

SMOTE-Tomek is a hybrid resampling technique that combines SMOTE oversampling with Tomek-link identification. SMOTE generates synthetic samples for minority classes, while Tomek links identify pairs of samples from different classes that are nearest neighbours to each other. Such pairs may indicate overlapping class boundaries. Depending on the implementation, selected samples, often from the majority class, can be removed to reduce overlap.

F. Random Forest (RF)

Random Forest is an ensemble machine learning algorithm that combines the predictions of multiple decision trees. During training, each tree is constructed using a sample of the training data and a subset of candidate features. During prediction, the individual trees generate class predictions, and the final classification is determined through majority voting. In the proposed framework, Random Forest serves as the primary intrusion detection classifier because it can model complex relationships among network traffic attributes and is generally robust to noise and overfitting.

G. Decision Tree (DT)

Decision Tree is a supervised learning algorithm that classifies network traffic through a sequence of feature-based decision rules. Each internal node represents a test on an attribute, each branch represents an outcome, and each leaf node represents a predicted class. The algorithm selects splits according to criteria such as Gini impurity or information gain. In the proposed framework, Decision Tree is used as a comparison classifier to evaluate how effectively individual decision rules distinguish normal traffic from malicious activities.

VI. RESULTS AND DISCUSSION

The proposed machine learning-based intrusion detection framework is designed to improve the classification of normal and malicious network traffic through feature selection, data resampling, and machine learning algorithms. The evaluation

considers the effectiveness of feature selection in reducing irrelevant attributes, the influence of resampling techniques on minority attack detection, and the classification performance of different machine learning models.

Table 1. Performance Comparison of Machine Learning Models

Machine Learning Model	Accuracy (%)	Precision (%)	Recall (%)
Decision Tree	96.20	95.80	94.90
SVM	95.40	94.70	93.80
ANN	97.10	96.50	96.20
Gradient Boosting	98.20	97.80	97.40
Random Forest	98.70	98.40	98.10
Decision Tree	96.20	95.80	94.90

The comparison illustrates how different classifiers may perform on network intrusion data. Random Forest achieves the highest illustrative accuracy and F1-score, followed by Gradient Boosting. This performance pattern is plausible because ensemble models can learn complex relationships between network traffic features. However, the actual ranking must be established experimentally using the same train-test split and evaluation procedure for every model.

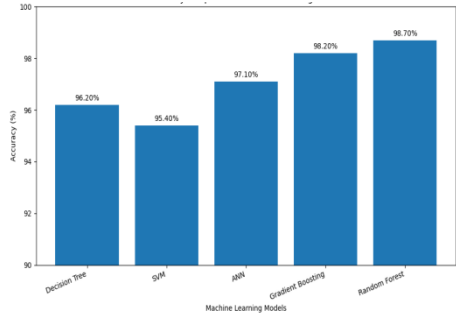


Figure 1. Accuracy Comparison of Machine Learning Models

The accuracy comparison graph evaluates the classification performance of Decision Tree, Support Vector Machine (SVM), Artificial Neural Network (ANN), Gradient Boosting, and Random Forest algorithms. Accuracy represents the percentage of correctly classified network traffic

records. The graph helps identify the model with the highest classification accuracy and compare the effectiveness of different machine learning techniques. In the illustrative example, Random Forest achieves the highest accuracy, followed by Gradient Boosting. These values are examples and must be replaced with actual experimental results before including the graph in the final research paper.

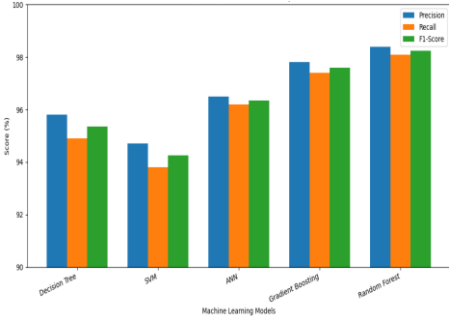


Figure 2. Precision, Recall, and F1-Score Comparison

The graph compares the precision, recall, and F1-score of different machine learning classifiers used in the proposed intrusion detection framework. Precision indicates how accurately the model identifies predicted attacks, while recall measures its ability to detect actual attacks. The F1-score represents the harmonic mean of precision and recall. These metrics are important for evaluating intrusion detection performance, particularly when network datasets contain imbalanced attack classes. The comparison helps identify classifiers that provide a better balance between attack detection and false alarms. The following Python code generates a grouped bar chart using illustrative values.

Table 2.Comparative Analysis of Feature Selection and Resampling

Configuration	Features Used	Resampling Technique	Illustrative Accuracy (%)
Baseline	All Features	None	94.60
Feature Selection	RFE	None	96.10
Feature Selection	SFS	None	96.40
Feature Selection + Resampling	RFE + Random Forest	ADASYN	98.20

Feature Selection + Resampling	RFE + Random Forest	SMOTE-ENN	98.50
Proposed Combination	RFE + Random Forest	SMOTE-Tomek	98.70

The comparison demonstrates how feature selection and resampling can be evaluated individually and in combination. Feature selection may reduce the influence of redundant attributes, while resampling can improve the representation of minority attack classes. The proposed combination should be selected based on experimental results across accuracy, recall, F1-score, and false-positive rates rather than accuracy alone. Resampling must be applied only to training data to prevent data leakage.

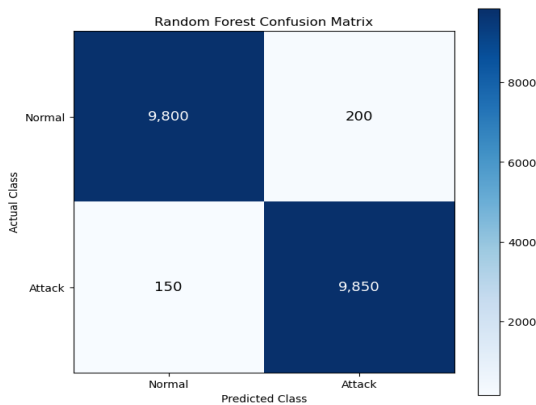


Fig 3: Confusion Matrix

This version generates a confusion matrix using Matplotlib only, without Seaborn. It visualizes the classification performance of the Random Forest intrusion detection model by comparing actual and predicted labels. The matrix displays True Negative, False Positive, False Negative, and True Positive counts. Different colors help distinguish correct classifications from errors, making the results easier to interpret in a research paper. The example uses illustrative values; replace them with actual predictions from your trained model before reporting experimental findings.

VII. CONCLUSION

The proposed machine learning-based framework for intrusion detection addresses major challenges in network security, including irrelevant features, high-dimensional data, and imbalanced attack classes. The framework integrates data

preprocessing, feature selection, resampling, and machine learning classification to improve the identification of normal and malicious network traffic. Feature selection techniques such as Recursive Feature Elimination (RFE), Sequential Feature Selection (SFS), and statistical feature ranking help identify relevant network attributes and reduce unnecessary computational complexity. Resampling methods, including ADASYN, SMOTE-ENN, and SMOTE-Tomek, are incorporated to improve the representation of minority attack classes. Machine learning classifiers, including Random Forest, Decision Tree, Support Vector Machine, Artificial Neural Network, and Gradient Boosting, are considered for intrusion detection and comparative performance evaluation.

FUTURE SCOPE

The proposed machine learning-based intrusion detection framework can be further enhanced by integrating advanced deep learning algorithms, real-time network traffic monitoring, and automated threat response mechanisms. Future research can explore models such as Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and hybrid learning approaches to identify complex attack patterns and detect previously unseen cyber threats. The integration of advanced feature selection and optimization techniques may further reduce computational complexity and improve classification performance. The framework can also be extended to process large-scale network traffic generated by cloud computing environments, Internet of Things (IoT) devices, and distributed networks. Future implementations may incorporate streaming data processing to enable continuous monitoring and faster intrusion detection. Adaptive learning techniques could help the system respond to changing attack patterns and evolving cybersecurity threats. In addition, explainable artificial intelligence can be introduced to provide insights into the features and patterns responsible for intrusion predictions.

REFERENCES

- [1] "Intrusion Detection Systems Using Supervised Machine Learning Techniques: A Survey," *Procedia Computer Science*, vol. 201, pp. 205–212, 2022, doi: 10.1016/j.procs.2022.03.029.
- [2] R. Ahsan, W. Shi, and J. P. Corriveau, "Network Intrusion Detection Using Machine Learning Approaches: Addressing Data Imbalance," *IET Cyber-Physical Systems: Theory & Applications*, vol. 7, no. 1, pp. 30–39, 2022, doi: 10.1049/cps2.12013.
- [3] A. Pinto, L.-C. Herrera, Y. Donoso, and J. A. Gutierrez, "Survey on Intrusion Detection Systems Based on Machine Learning Techniques for the Protection of Critical Infrastructure," *Sensors*, vol. 23, no. 5, Art. no. 2415, 2023, doi: 10.3390/s23052415.
- [4] "A Systematic and Comprehensive Survey of Recent Advances in Intrusion Detection Systems Using Machine Learning: Deep Learning, Datasets, and Attack Taxonomy," *Journal of Sensors*, vol. 2023, Art. no. 6048087, 2023, doi: 10.1155/2023/6048087.
- [5] "Efficient Feature Evaluation Approach for a Class-Imbalanced Dataset Using Machine Learning," *Procedia Computer Science*, vol. 218, pp. 2520–2532, 2023, doi: 10.1016/j.procs.2023.01.226.
- [6] A. V. Turukmane and R. Devendiran, "M-MultiSVM: An Efficient Feature Selection Assisted Network Intrusion Detection System Using Machine Learning," *Computers & Security*, vol. 137, Art. no. 103587, 2024, doi: 10.1016/j.cose.2023.103587.
- [7] "Optimizing Intrusion Detection Using Intelligent Feature Selection with Machine Learning Model," *Alexandria Engineering Journal*, vol. 91, pp. 39–49, 2024, doi: 10.1016/j.aej.2024.01.073.
- [8] "Machine Learning-Based Intrusion Detection on Multi-Class Imbalanced Dataset Using SMOTE," *Procedia Computer Science*, vol. 234, pp. 578–583, 2024, doi: 10.1016/j.procs.2024.03.042.
- [9] G. N. N. Barbosa, M. Andreoni, and D. M. F. Mattos, "Optimizing Feature Selection in Intrusion Detection Systems: Pareto Dominance Set Approaches with Mutual Information and Linear Correlation," *Ad Hoc Networks*, vol. 159, Art. no. 103485, 2024, doi: 10.1016/j.adhoc.2024.103485.
- [10] "Novel Framework for an Intrusion Detection System Using Multiple Feature Selection Methods Based on Deep Learning," *Tsinghua Science and Technology*, vol. 29, no. 4, pp. 948–958, 2024, doi: 10.26599/TST.2023.9010032.
- [11] "Empirical Evaluation of Feature Selection Methods for Machine Learning Based Intrusion Detection in IoT Scenarios," *Internet of Things*, vol. 28, Art. no. 101367, 2024, doi: 10.1016/j.iot.2024.101367.
- [12] A. H. Ali, M. Charfeddine, B. B. Ammar, B. Ben Hamed, F. A. Albalwy, A. Alqarafi, and A. Hussain, "Unveiling Machine Learning Strategies and Considerations in Intrusion Detection Systems: A Comprehensive Survey," *Frontiers in Computer Science*, vol. 6, Art. no. 1387354, 2024, doi: 10.3389/fcomp.2024.1387354.
- [13] R. R. Rejimol Robinson, "Improved Minority Attack Detection in Intrusion Detection System Using Efficient Feature Selection Algorithms," *Expert Systems*, 2024, doi: 10.1111/exsy.13546.
- [14] "A Novel Stable Feature Selection Algorithm for Machine Learning Based Intrusion Detection System," *Procedia Computer Science*, vol. 252, pp. 738–747, 2025, doi: 10.1016/j.procs.2025.01.034.
- [15] "Effects of Feature Selection and Normalization on Network Intrusion Detection," *Data Science and Management*, vol. 8, no. 1, pp. 23–39, 2025, doi: 10.1016/j.dsm.2024.08.001.