National College of Ireland

Novel Approach in Intrusion Detection Systems Using Mutual Information-based Gradient Boosting Machine

Academic Internship
MSc Cyber Security

Benhur Kachhap
Student ID: 22168494

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Benhur Kachhap
Student ID:	22168494
Programme:	MSc Internship **Year:** 2023
Module:	Academic Internship
Supervisor:	Vikas Sahni
Submission Due Date:	14/12/2023
Project Title:	Novel Approach in Intrusion Detection Systems Using Mutual Information-based Gradient Boosting Machine
Word Count:	5932 **Page Count** 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Benhur Kachhap

Date: 14/12/2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Novel Approach in Intrusion Detection Systems Using Mutual Information-based Gradient Boosting Machine

Benhur Kachhap
22168494

Abstract

This study aims to improve the performance of intrusion detection systems (IDS) by implementing the Mutual Information-based Gradient Boosting Machine (MIGBM) feature selection approach. The significance of this study arises from the increasing sophistication of cyberattacks, highlighting the urgent need to innovate and strengthen IDS capabilities. Although numerous scholars have put forth a plethora of approaches to enhance the identification of unauthorized access attempts, this paper introduces a conceptual technique that leverages the utilization of mutual information (MI) feature selection. MIGBM was rigorously tested as a unique feature selection technique to enhance detection accuracy while simultaneously lowering computing time. The objective is to compare the top-performing techniques across multiple performance metrics—recall, precision, classification accuracy, and F1 score with MIGBM with and without MI feature selection. Each method is also critically evaluated based on its limitations. The evaluation involves generating confusion matrices to assess the system's performance, utilizing an updated and pertinent dataset. The leading approach demonstrated MIGBM with an impressive 95% accuracy, just 2% lower than the baseline approach, showcasing remarkable efficiency and precision with a notable timeframe reduction of 40%.

1 Introduction

1.1 Background

In today's world, the widespread use of technology and reliance on interconnected systems have changed how things work. This shift, known as digital transformation, has had a big impact because people depend a lot on information technology and networks. This revolution has initiated several challenges in terms of protecting sensitive networks and data security. Intrusion Detection Systems (IDS) are an integral part of cyber security mechanisms, enabling the safeguarding of computer systems and networks against malicious activities and unauthorized access (Markevych and Dawson, 2023).

IDS can be categorized into signature-based and anomaly-based systems. A signature-based system depends on a predefined database of the identified attack pattern, whereas an anomaly-based system intends to identify deviation from normal network behavior. To improve the efficiency of the IDS system based on signature and anomaly the Mutual

Information-based Gradient Boosting Machine (MIGBM) would be helpful. MIGBM integrates mutual information and the power of gradient boosting (Dash *et al.*, 2022). Mutual information acts as a measure of the statistical dependency among the existing variables whereas gradient boosting allows the use of machine learning techniques to manage non-linear and high dimensional data effectively. MIGBM is primarily used in data analysis and machine learning to improve data representation and model training effectiveness and efficiency. This also relates to dealing with data sets, particularly found in IDS, making it an appropriate technique in intrusion detection.

1.2 Importance

The research is crucial in the context of data protection and cyber security, critical to the increasing challenges of intrusion detection and response. A better understanding of the performance evaluation criteria of IDS using the MIGBM feature selection technique allows improved accuracy with reduced training time in protecting networks, critical infrastructure, and sensitive data. It addresses the gap required in the evolving digital landscape, which is pertinent to several security challenges (Al-Sarem *et al.*, 2021). The study determines how ideas can become more effective in detecting anomalies by improving the accuracy of identifying previously unknown and novel threats. The research plays a crucial role in customizing specific environments, as fine-tuning intrusion detection involves selecting the most relevant features through MIGBM. This allows customization according to organizational requirements. The research is also important in addressing the gap of the need to select accurate detection methods by applying machine learning techniques in IDS. As the study evaluates these models' functions, selecting appropriate features relevant to MIGBM would be more relevant. Comprehensively, the challenges in selecting features for IDS would be better.

1.3 Research question

Does the Mutual Information-based Gradient Boosting Machine (MIGBM) feature selection approach improve the detection accuracy, reduce computational time, and increase the robustness of intrusion detection systems (IDS), and by how much?

1.4 Research aims and objectives.

The research aims to use the feature selection approach of a Mutual Information-based Gradient Boosting Machine (MIGBM) to increase accuracy in detecting intrusions, reducing computational time, and enhancing the robustness of Intrusion Detection Systems (IDS).
The following are the objectives of the research:

- To evaluate various processes of selecting the most relevant features for IDS with and without mutual information.
- To synthesize gradient boosting machines with relevance to other techniques that can contribute to increased intrusion detection accuracy with low computation time.

1.5 Contribution to the Scientific Literature

The study assesses the efficacy of an IDS employing MIGBM, offering advancements in research concerning enhanced feature selection, machine learning, and intrusion detection.

This paves the way for further exploration in the domain by introducing innovative approaches to IDS and MIGBM, fostering the creation of more proficient and effective IDS system techniques.

1.6 Structure of the report

The report consists of several key sections: an introductory section that sets the context and outlines research objectives, a critical review of existing literature highlighting the gaps, a detailed methodology section describing research procedures and setups, specifications for implementing an IDS using MIGBM, the practical implementation of the proposed solution, a thorough evaluation of results, and a conclusive summary presenting findings and implications while also suggesting future research directions.

2 Related Work

The section explains the need for a better machine learning model for bigger dimensions of data sets that require better training and testing to monitor data transmissions and network traffic suggesting a need to use adequate methods for feature selection for IDS.

2.1 Wrapper feature selection techniques

(Almasoudy *et al.*, 2020) identify the method of dealing with the challenges of high-dimensional data containing redundant and irrelevant features while using IDS. It applies dimensionality reduction as a wrapper feature selection model established on the technique of differential evolution. It is based on such an IDS system to reduce the number of features by identifying minimum features without impacting system performance. Differential evolution helps in selecting some features and evaluating them using extreme machine learning techniques to identify the best feature selection process for IDS.

(Almaghthawi *et al.*, 2022) also discuss wrapper feature selection techniques to overcome the challenges of the negative impact of classification accuracy. It selects certain techniques such as sequential forward selection, genetic algorithm, and sequential backward selection to analyze, mitigate, and compare their performance. Further two classification methods are selected, which include multi perceptron and support vector machine. These are considered across different datasets to determine the effectiveness and accurate outcomes of detecting intrusions. Therefore, the wrapper features election technique facilitates better access to selecting features for reducing computational complexity and detection time.

2.2 Machine learning

(Hossain and Islam, 2023) establishes the use of detection systems and machine learning for intrusions. It uses an ensemble-based machine learning approach, which focuses on various models and different times over the year. They used a novel approach for creating decision tree classifiers across genetic algorithms that provides detection criteria for abuse system. This provides an average accuracy rate of 89% and identification accuracy of 97%.

Alternatively, (Tripathy and Behera, 2023) suggest the use of a machine learning algorithm for IDS and evaluate its performance. It determines IDS with machine learning to

improve the accuracy of detecting security attacks. Such a method would be to resize and be effective at detecting network assaults while dealing with large dimensional spaces and data. It would execute a feasible feature removal technique to get rid of features that can impact the classification process.

(Awotunde *et al.*, 2021) discuss a rule-based feature selection method for intrusion detection of IoT networks based on a deep learning model. The use of a rule-based model and genetic search tool provides hybrid feature selection. It enables the highest correlation between the class relationship and attribute. The merits of each attribute are evaluated by selecting functions as per genetic surge methods producing attributes of the greatest values.

Moreover, (Jaw and Wang, 2021) propose a hybrid feature selection with an ensemble classifier that allows consistent classification of attacks and relevant feature selection of this method indicates excellent accuracy of 99.99%. The benefit of this model is that it enables combining learning techniques to enhance output accuracy for feature selection.

In addition, (Qadir Mohammed and A. Hussein, 2022) identify the use of machine learning models for dealing with inclusion detection and performance analysis of different machine learning models based on supervised machine learning algorithms. Confusion matrix metric conducted a performance-wise analysis to facilitate comparison between classifiers. Pearson, F test, and information gain facilitated feature selection techniques to determine the results between all features. However, the random forest classifier provided the best performance with 99.96% accuracy, which superseded other classifiers.

(Upadhyay *et al.*, 2021) also use gradient boosting for feature selection with classifiers of machine learning for detecting Intrusions on the power grid. Its approach is to use an integrated IDS system, combining feature engineering-based preprocessing with classifiers of machine learning. This approach helps in selecting the most relevant features of the data set and allows better detection rate and execution speed. Therefore, decision-tree-based machine learning techniques help in selecting the most prominent features and execution time.

Further (Nimbalkar and Kshirsagar, 2021) discuss IDS as an ensemble classifier for feature selection. The IDS system uses a correlation coefficient and ensemble classifier for detecting intrusions. Classifiers such as decision trees, Naive Bayes, and Artificial Neural networks provided 98.54% accuracy in detecting denial of service attacks. This indicates achieving a higher accuracy of 89.76% with top-ranking feature selection. This discloses the use of information gains in the feature selection process for IDS.

(Souhail Et. Al., 2019) discuss using recursive feature elimination and random forest, along with other techniques for selecting the most promising features of the data set for machine learning purposes. Further binary classification is performed for detecting intrusive traffic through data mining techniques such as the Gradient Boost Mechanism, Logistic Regression, and Support Vector Machine. This indicates support vector machine has the highest accuracy of 82.11%.

(Karthigha and L, 2022) also supports multi-level modified gated recurrent unit as a means of better feature selection and classification. The model includes classification, accuracy, and reduced call alarm rate. Therefore, IDS become convenient using machine learning aspects for detecting intrusions.

(Kasongo and Sun, 2020) opine the use of IDS based on machine learning as an accurate and effective method for detecting network intrusions. XGBoost Algorithm helps in

overcoming the existing challenges of low detection accuracy by machine learning approaches such as K nearest neighbor, logistic regression, decision tree, support vector machines, and artificial neural network. Multiclass and binary classification configuration results in a feature selection method with an accuracy of 88.13 to 90.85% using a decision tree.

Moreover, (Seth *et al.*, 2021) propose the use of a smart inclusion detection system through a hybrid feature selection approach. This method uses a light gradient boosting machine as the gradient boosting model for analyzing datasets. The hybrid feature selection and light gradient boosting mechanism as the proposed model provides a 99.3% precision rate, 96% sensitivity, and 97.73% accuracy.

(Disha and Waheed, 2022) also support the appropriate performance of the decision tree in selecting features as compared to other models. This was based on the analysis of several techniques, such as decision tree, multilayer perceptron gradient boosting tree, AdaBoost, and gated recurrent unit.

(Le *et al.*, 2022) propose the use of the extremely gradient ghosting model for detecting intrusions and feature selection in industrial IoT. This extremely gradient-boosting model accomplishes significant detection in attacks with 99.9% accuracy in the data set. Thus, it comprehensively indicates how the decision tree and light gradient boosting mechanism offer a precise feature selection for IDS.

2.3 IDS Feature Selection Technique in IoT Environment

(Li *et al.*, 2021) propose a linear nearest neighbor lasso step (LNNLS-KH) for feature selection of network intrusion detection. This solves the problem of high false positives and low efficiency in intrusion detection. The number of classification accuracy and selected features are initiated within the fitness evaluation function. The linear nearest neighbor performs lasso step optimization to drive global optimal solutions. LNNLS-KH algorithm retains seven features and effectively eliminates redundant features with better accuracy. This proposed method reduces intrusion detection time by 14.41% and 4.03% on average.

In addition, (Albulayhi *et al.*, 2022) discuss IoT intrusion detection with machine learning. It uses features and attributes for IDS, the approach initiates through two entropy-based approaches for selecting and extracting required features in different ratios. This comparison with other state-of-the-art approaches reveals that 11 and 28 relevant features use the union and intersection. However, the two-entropy approach is competent in superior providing 99.98% accuracy in classification in the IoT ecosystem.

(Verma *et al.*, 2021) also propose a machine learning ensemble for intrusion detection. It uses a gradient boosting machine to improve precision by 96.40% and accuracy by 98.27%. Comprehensively, it indicates the effectiveness against cyber threats and performing successfully in IoT environments.

2.4 Particle Swarm Optimization

(Louk and Tama, 2022) initiate a unique particle swarm optimization (PSO) driven feature selection approach. This approach derives final features from various IDS datasets. They are effectively trained for hybrid ensembles comprising ensemble learners. This proposed scheme

leads to crucial refinement of existing baselines, providing majority voting and other ensemble-based IDS. This model has surfaced with data sets with wide accuracy for feature selection.

In addition, (Almomani, 2020) uses PSO with a genetic algorithm, firefly optimization, and grey wolf optimizer. These proposed models are used to derive features based on support vector machines. It indicates that future selection using a rule-based pattern provides a better scope of symmetrical recognition for intrusions.

2.5 Artificial Intelligence

(Agyapong *et al.*, 2023) discuss employing a soft voting-based ensemble learner for intrusion detection system networks to classify networks between malicious and normal data. It uses LoGD-ai which uses logistic regression, decision tree, and gradient boosting for detecting intrusions. LoGD-ai performs its classification, which is compared to other gradient boosting machines, AdaBoost, and random forest. This comparison indicates that LoGD-ai offers better accuracy by 0.52% as compared to other approaches.

In addition, (Saha *et al.*, 2022) demonstrate the use of AI methods, namely machine learning, deep learning, and unsupervised learning for feature selection and performance evaluation. This means that the ensemble feature selection technique is the most relevant feature selection process, providing universal prominent features for all AI models.

Further (Farhan and Jasim, 2022) discusses the use of machine learning techniques in detecting attacks and preventing them, particularly with the use of deep learning. It offers the ability to extract features with high accuracy and self-learning to use the learning for analyzing real data set network traffic. This helps in analyzing attacks and normal behavior while evaluating deep model long, short-term memory. The analysis provides 99% accuracy in detecting intrusions.

(Alqahtani *et al.*, 2019) encourage the wireless sensor network detection system to outperform other state-of-the-art approaches with a 98.2% high detection rate for flooding. Simultaneously, 92.9%, 98.9%. 99.5% for scheduling, grey hole, and black hole attacks.

2.6 Summary

Table 1: Summary of the related works

Paper	***Findings***	***Gap***
Almasoudy et al., 2020	Wrapper feature selection based on differential evolution to reduce redundant and irrelevant features in IDS; Evaluation using extreme machine learning techniques.	The gap in accessing the system with certain privileges to authorized users or unauthorized users.
Almasoudy et al., (2020)	Wrapper feature selection methods (e.g., sequential forward selection, genetic algorithm) compared for IDS using multi perceptron and support vector machine.	High dimensionality and feature redundancy.

Louk & Tama (2022)	Particle swarm optimization (PSO)-driven feature selection; Hybrid ensemble models with improved accuracy.	Underexplored hybrid ensemble feature selection technique.
Hossain & Islam (2023)	Ensemble-based ML approach for IDS using decision tree classifiers across genetic algorithms; Average accuracy rate of 89%.	Limited datasets to select adequate features for intrusion detection.
Tripathy & Behera (2023)	Application of ML algorithm to resize and effectively detect network assaults; Feature removal to impact classification process.	Biased conventional categorization indicators for selecting adequate features.
Nimbalkar & Kshirsagar (2021)	Ensemble classifier approach for IDS in IoT with decision trees, Naive Bayes, and ANN achieving 98.54% accuracy; Use of information gain in feature selection.	Increasing attacks in IOT, make it difficult for machine learning models to detect intrusions.
Disha & Waheed (2022)	Analysing machine learning performance for IDS through weighted random forest feature selection technique.	Overcomes the gaps of machine learning model, particularly gradient boosting tree.
Verma et al., (2021)	Used ensemble machine learning technique for detecting novel intrusions in IOT environment.	The method has a significant retrieval time as compared to the learning time for IDS.
Agyapong et al., (2023)	Enabled detecting intrusions in the network using soft voting-based ensemble learner.	Limitations in the data set applicable to Adaboost, Random Forest, gradient boost machine, and LoGD-ai.
Alqahtani et al. (2019)	Using genetic-based extreme gradient boosting model for IDS in wireless sensor networks.	The proposed model has gaps in terms of false positives and false negative alarms.

The research project focuses on Mutual Information-based feature selection coupled with Gradient Boosting Machine due to its promising potential in addressing the gaps observed in existing studies. Specifically, the niche lies in the lack of exploration and optimization of feature selection techniques that leverage Mutual Information to enhance the accuracy and efficiency of IDS using Gradient Boosting Machine

3 Research Methodology

The research methodology has been thoughtfully designed to guarantee thoroughness, reproducibility, and clarity. This section outlines the systematic process followed, beginning

from the initial data collection phase, and concluding with the comprehensive assessment and evaluation.

3.1 Data Sourcing and Preprocessing

The dataset for the IDS evaluation was sourced from the author of the University of Nevada - Reno Intrusion Detection Dataset (UNR-IDD) that provides researchers with a wider range of intrusion samples and scenarios. The authenticity and relevance of the dataset to multiple modern intrusion patterns were key considerations, ensuring the results apply to current cybersecurity challenge (Das, 2023).

The investigation begins with a thorough data preprocessing stage, encompassing data cleansing to eliminate discrepancies and outliers, data splitting in an 80:20 ratio for train and test data, feature value normalization to establish a consistent scale, categorical variable encoding as required, and variance threshold for noise and redundancy reduction. By performing this procedure, the data's quality and compatibility are verified in preparation for the feature selection and model training that follows.

3.2 Feature Selection Technique (MIGBM)

The MIGBM methodology is the focus of this study. To determine if a network's activity is benign or malicious, it uses mutual information to quantify the relationship between each feature and the target variables. Using this method, it may isolate the most informative and non-redundant characteristics that improve the model's prediction ability based on their mutual information scores.

3.3 Default vs. Feature Subset Training

Gradient Boosting Machine, Random Forest Classifier, KNN Neighbours, and Gaussian Naive Bayes are a few of the machine learning algorithms that are utilized in the comparison of the technique implemented. To establish a performance baseline, each model is initially trained and then tuned using GridSearch hyperparameters to the default feature set. To assess the effect of the feature selection on the performance of the model, the models are subsequently retrained and retuned utilizing the feature subset that was chosen by the MIGBM technique.

3.4 Model Evaluation Framework

Performance Metrics: The models were graded on their accuracy, precision, recall, and f1 score. The algorithms' ability to accurately classify various intrusion types was prioritized in selecting these measures.

Each model's efficacy is measured both before and after the MIGBM feature selection is used, and the results are then compared using the evaluation framework. Measures of effectiveness include accuracy, precision, recall, and f1-score. These metrics are computed for every dataset class and combined to show the overall model performance. To show how MIGBM feature selection has enhanced IDS performance, a comparison study is carried out.

Statistical paired t-tests: This provides an interpretation of the paired t-test results and helps in understanding the significance of the difference in accuracy scores before and after feature selection.

3.5 Results Analysis

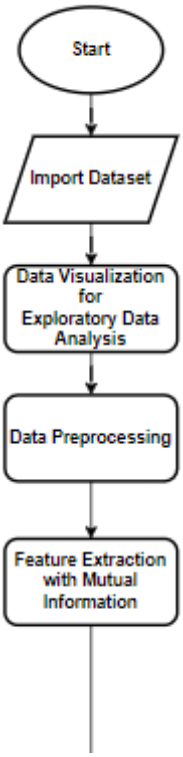
By analyzing the outcomes of the model evaluations, conclusions regarding the efficacy of the MIGBM feature selection method can be drawn. A comprehensive analysis of the classification reports and confusion matrices is required for this. The objective of the analysis is to determine whether the MIGBM method yields models with enhanced precision and recall in detecting diverse categories of intrusions.

Rationale for Methodology

The methodology was shaped by a comprehensive literature study that revealed weaknesses in standard IDS evaluation methods and showed the promise of mutual information in feature selection. By merging Mutual Information with Gradient Boosting Machine, the research presents a novel technique for feature selection in IDS.

4 Design Specification

The design specification for the proposed research incorporates a distinctive amalgamation of machine learning modeling and feature selection, with a particular focus on its application in intrusion detection systems (IDS). It delineates a comprehensive roadmap for the research methodology employed in the study. It incorporates modules dedicated to data preprocessing, feature selection, model training and evaluation, and model analysis and comparison.
The implementation framework for the MIGBM technique is structured in the given flow chart:



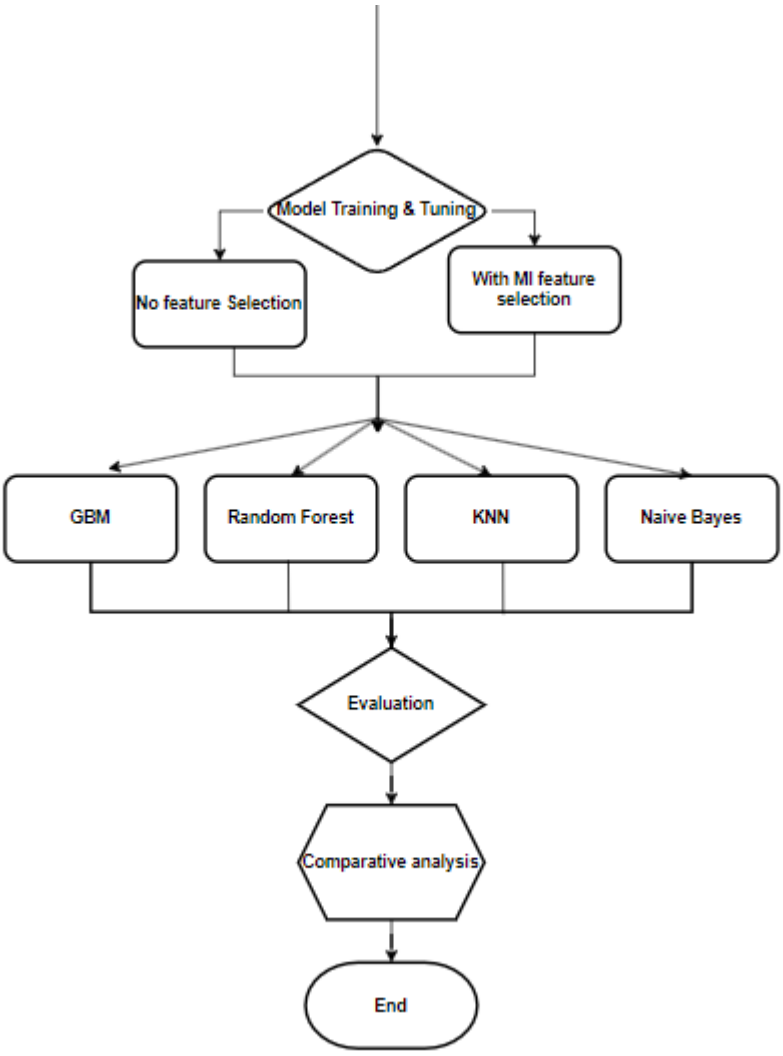


Figure 1: Workflow diagram

New Algorithm/Model Functionality

Based on the idea that not all features equally contribute to an IDS model's accuracy, the MIGBM feature selection method was developed. It uses machine learning to determine which features are most important for increasing detection rates. The model is then able to zero in on these characteristics, resulting in a more refined and potent IDS. This method's flexibility and usefulness in a variety of IDS implementations stem from the fact that it is model agnostic, meaning it may be used with multiple machine learning techniques.

5 Implementation

Model Implementation

For implementation, Python was used with Scikit-learn, Pandas, NumPy, Seaborn, Matplotlib, and SciPy was used.

5.1 Data Loading and Preprocessing

The process starts by requesting the dataset path from the user. Using Pandas, the script loads the dataset and ensures it's handled for any potential issues. It further explores the dataset using data.info() to reveal essential details like entry count, column information, and data types, also, displaying the initial dataset rows for a quick overview. Following this exploration, the script verifies the dataset for duplicates and missing values, confirming the absence of both issues in this instance.

5.2 Exploratory Data Analysis (EDA)

The project utilizes a diverse range of graphical representations to visually analyze the dataset. These visualization techniques play a pivotal role in uncovering key insights and understanding the dataset's underlying patterns. Histograms, box plots, pair plots, scatter plots, count plots, and bar plots, reveal diverse aspects of the dataset's distributions, relationships, outliers, and class composition, aided in comprehensive analysis and interpretation.

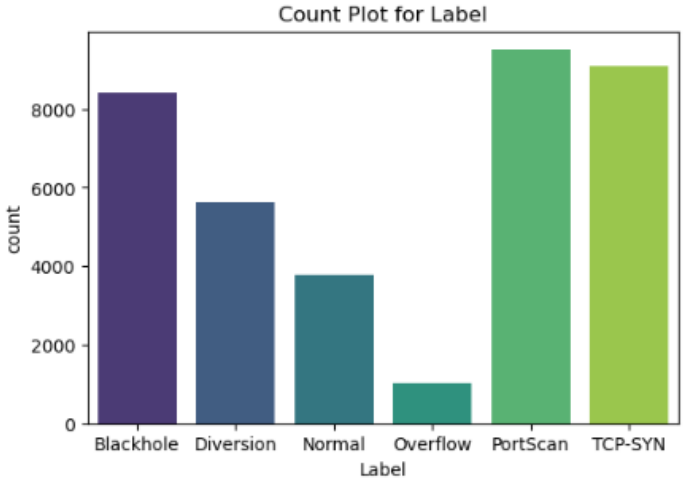


Figure 2: Number of instances per attack type

5.3 Model Training and Evaluation

Data Preparation: The data preparation involves separating features from the target variable. It encodes categorical variables into numerical format using Label Encoding if present. The dataset is then split into training and testing sets (80% and 20% respectively) while maintaining class distribution. Features are scaled using StandardScaler for accurate algorithm performance. Finally, VarianceThreshold is applied to eliminate low-variance features, reducing redundancy and noise in the dataset.

Feature Importance/Selection: The significance of various features is assessed using mutual information, which reveals which features are most informative about the classification task, and features are subsequently ranked based on their MI scores, facilitating the selection of pertinent attributes.

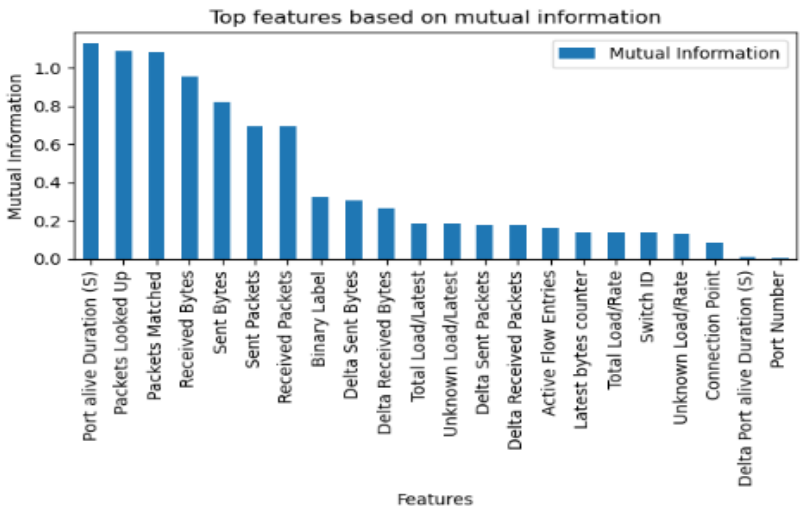


Figure 3: Top features based on Mutual Information

Model Selection and Training

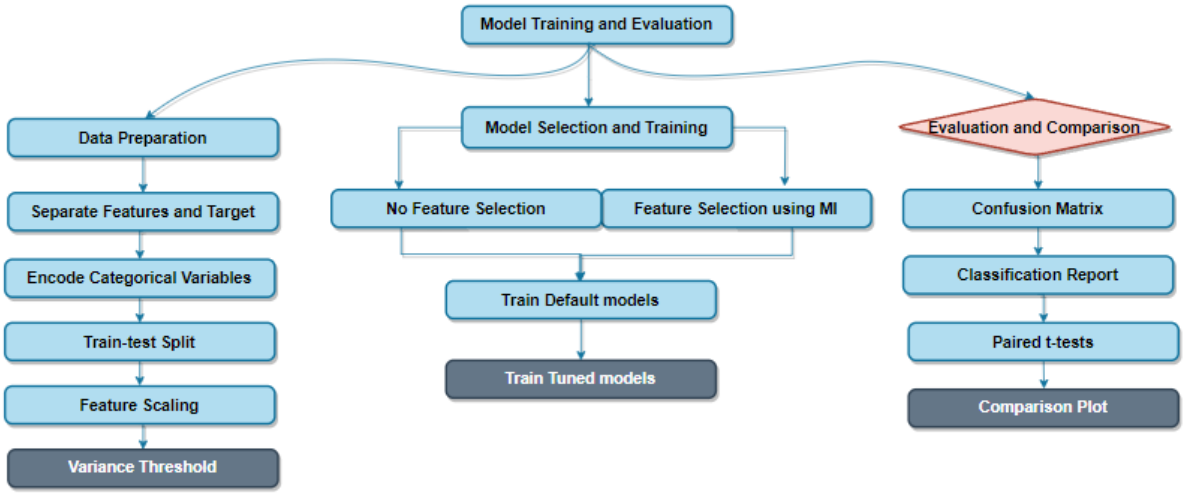


Figure 4: Model Training and Evaluation Chart

The code iterates the classifier over a selection of machine learning models which consists of the following: Gradient Boosting Classifier, Random Forest Classifier, KNN Neighbours, and Gaussian Naive Bayes. For comparison, both the default and tuned iterations of the model are trained using GridSearchCV. Once the evaluation is generated, the same iteration is run over the models with a new subset of MI features.

5.4 Evaluation, Model Comparison, and Analysis:

The script evaluates default and tuned models, presenting classification reports and confusion matrices with detailed metrics like precision, recall, and f1-score for each class. Furthermore, it conducts a comparative analysis between IDS models which aims to reveal the effectiveness of the MIGBM technique in enhancing overall model performance.

6 Evaluation

The evaluation of multiple machine learning models for intrusion detection was conducted using two different feature selection techniques: one without Mutual Information (MI) feature selection and another with MI feature selection. Four diverse models—Gradient Boosting Classifier, Random Forest Classifier, KNN Neighbours, and Gaussian Naive Bayes—were trained and evaluated in both scenarios.

6.1 Without Mutual Information Feature Selection:

The models were initially trained and tested on the dataset without using MI feature selection. The evaluation metrics, including precision, recall, F1-score, support, and accuracy, were recorded for both default and tuned models.

Table 2: Model Performance Observation & Comparison (without MI)

Model	Accuracy (Default)	Precision (Default)	Recall (Default)	F1-score (Default)	Accuracy (Tuned)	Precision (Tuned)	Recall (Tuned)	F1-score (Tuned)
Gradient Boosting Classifier	92.98%	0.931952	0.929832	0.930075	97.42%	0.976855	0.973202	0.974364
Random Forest Classifier	94.91%	0.949952	0.949078	0.949105	94.69%	0.940835	0.937686	0.937950
KNN Neighbours	86.66%	0.867762	0.866613	0.866381	88.31%	0.883052	0.883053	0.881314
Gaussian Naive Bayes	67.70%	0.733904	0.676958	0.673253	67.70%	0.733904	0.676958	0.673253

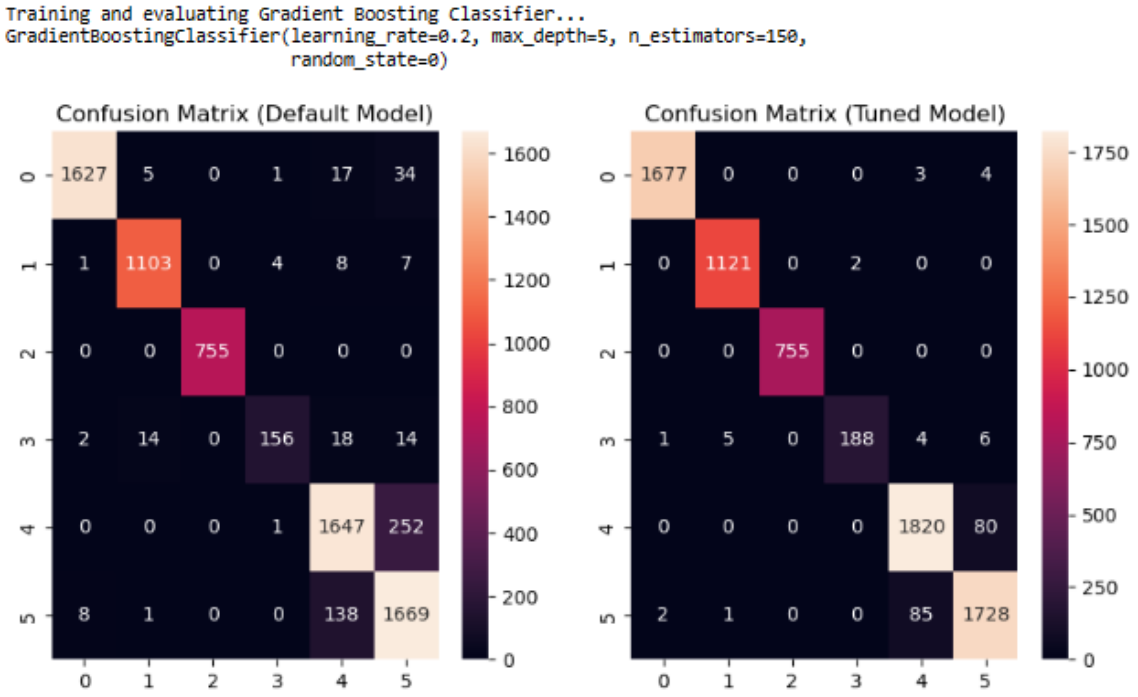


Figure 5: Confusion Matrix of GBM (without MI)

6.2 With Mutual Information (MI) Feature Selection:

The same set of models underwent evaluation after applying Mutual Information feature selection. The evaluation metrics were calculated for both default and tuned models.

Model Performance Observation & Comparison:

Table 3: Model Performance Observation & Comparison (with MI)

Model	Accuracy (Default)	Precision (Default)	Recall (Default)	F1-score (Default)	Accuracy (Tuned)	Precision (Tuned)	Recall (Tuned)	F1-score (Tuned)
Gradient Boosting Classifier	87.93%	0.884740	0.879310	0.878280	94.64%	0.974391	0.968381	0.969363
Random Forest Classifier	93.24%	0.933262	0.932371	0.931999	93.21%	0.932162	0.931215	0.930780
KNN Neighbours	82.40%	0.821020	0.823978	0.821564	86.65%	0.866512	0.866791	0.864707
Gaussian Naive Bayes	62.15%	0.689965	0.621492	0.618380	62.15%	0.689965	0.621492	0.618380

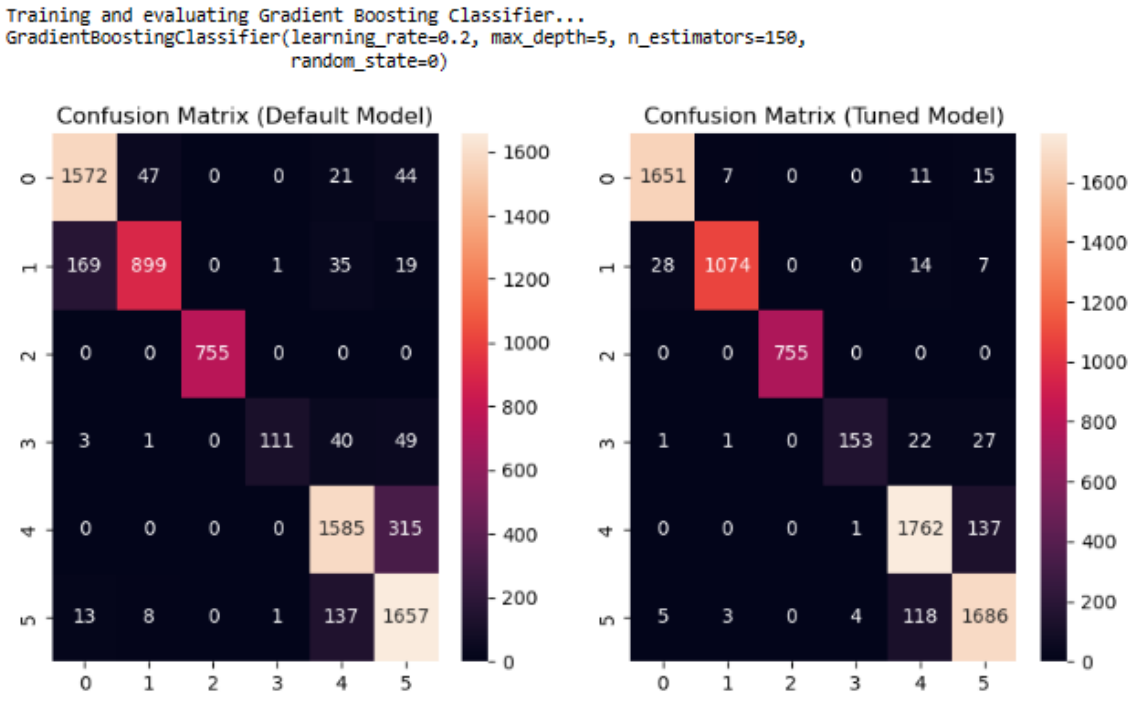


Figure 6: Confusion Matrix of GBM (with MI)

The performance between models with and without Mutual Information (MI) feature selection highlights varying outcomes.

- **Gradient Boosting Classifier** showcased significant improvements in accuracy in both cases, emphasizing its adaptability to tuning and feature selection. Without MI, the accuracy and precision were notably higher in the baseline model (97%) than with MI (95%).

- **Random Forest Classifier** exhibited consistent performance, maintaining accuracy levels despite changes, indicating robustness in different scenarios.
- **KNN Neighbours** demonstrated higher accuracy and precision in both default and tuned settings with MI (87%) versus without MI (88%) highlighting the sensitivity of this model to optimization techniques.
- **Gaussian Naive Bayes** demonstrated limited adaptability to tuning and feature selection, remaining consistent in accuracy scores.

Overall, all the models showed decrease in just 2% in accuracy scores after MI while Naïve Bayes showed 5% decrease.

6.3 Observations on Training Time Changes:

Table 4: Training time Observation & Comparison

Classifier	Training Time (Without MI)	Training Time (With MI)	Change in Training Time	% Change in Training Time
Gradient Boosting	34.98 seconds	20.84 seconds	-14.14 seconds	-40.41%
Random Forest	3.56 seconds	3.83 seconds	+0.27 seconds	+7.58%
KNN Neighbours	4.02 seconds	0.17 seconds	-3.85 seconds	-95.77%
Gaussian Naive Bayes	0.03 seconds	0.03 seconds	No change	0.00%

- **Gradient Boosting Machine:** With MI it experienced a significant reduction in training time, down to around 20.84 seconds, showcasing a notable decrease of about 14.14 seconds after incorporating MI feature selection. The training time was reduced by approximately 40.41% when MI feature selection was applied.
- **Random Forest Classifier:** With MI it maintained a similar training time, approximately 3.83 seconds, reflecting minimal change even after MI feature selection. The training time increased by approximately 7.58% when MI feature selection was applied.
- **KNN Neighbours:** Without MI it recorded a training time of approximately 4.02 seconds whereas with MI it displayed a marginal reduction in training time to about 0.17 seconds, indicating a substantial reduction of approximately 3.85 seconds with MI feature selection which is approximately 95.77% reduction.
- **Gaussian Naive Bayes:** With MI it remained consistent with an extremely low training time of about 0.03 seconds, demonstrating no significant change even after MI feature selection.

6.4 Paired T-Test Report for Accuracy Scores Before and After Feature Selection:

Below is the result of the T-Statistic and P-value obtained from the code:

T-Statistic: 5.305877239085448
P-Value: 0.0011162269423481833
Cohen's d: 0.27246029501026686

With a calculated t-statistic of approximately 5.3058 and a corresponding p-value of approximately 0.0111, we derive that there is sufficient statistical evidence to conclude that there is a significant difference in accuracy scores before and after feature selection at a significance level of 0.05, while Cohen's d of 0.27 signifies a moderate effect size, suggesting a small-to-medium practical difference between the models.

6.5 Discussion

Overall, all the models showed decrease in just 2% in accuracy scores after MI feature selection was applied to the training data while Naïve Bayes showed 5% reduction.

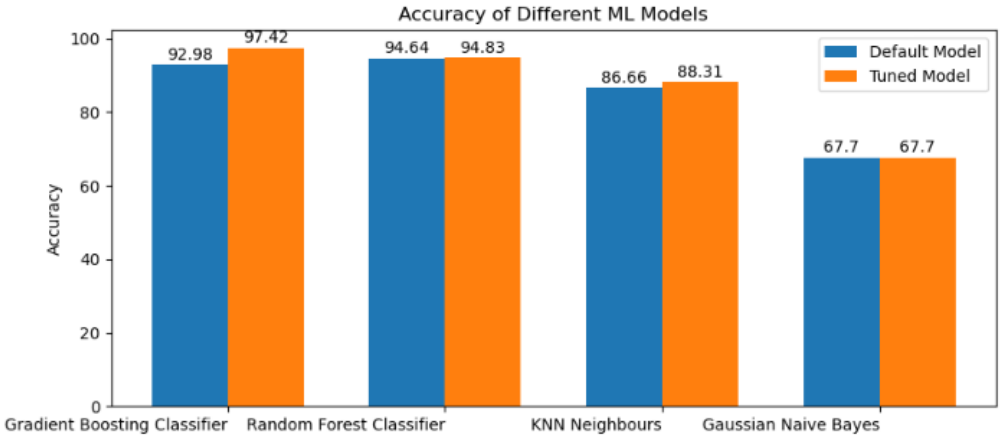


Figure 7: Comparison of Accuracy of all ML model (without MI)

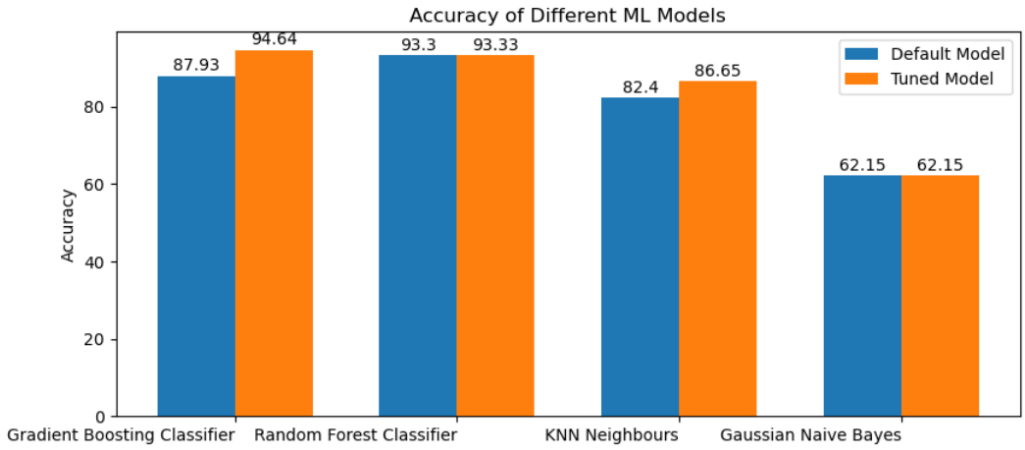


Figure 8: Comparison of Accuracy of all ML models (with MI)

Insights on Training Time Changes:
The most noticeable training time reductions were observed in the Gradient Boosting Classifier and KNN Neighbours models after employing the Mutual Information (MI) feature selection.

Gradient Boosting Classifier exhibited a substantial decrease in training time by approximately 40%, emphasizing the efficiency gained through feature selection. KNN Neighbours, also, demonstrated a significant reduction of about 95% when considering MI feature selection. While Random Forest saw a slight increase in training time of 7.58%, Gaussian Naïve showed no significant change.

This exemplifies how optimizing feature selection techniques not only maintains model accuracy but also contributes to substantial reductions in training time, thereby enhancing model efficiency without compromising performance. Gradient Boosting Classifier combined with Mutual Information illustrates the significance of feature selection in improving both accuracy and training efficiency for intrusion detection models.

7 Conclusion and Future Work

The Gradient Boosting Classifier with Mutual Information (MI) feature selection significantly reduced its training time from approximately 34.98 seconds without MI to around 20.84 seconds with MI. This reduction in training time by almost half highlights the efficiency gained through feature selection. Considering that the reduced dataset after MI just used 6 features compared to the baseline model training of 33 features, the model maintained good accuracy scores with only a 2% downfall, showcasing the dual advantage of faster training and enhanced performance.
The theory of feature selection in machine learning substantiates this variance, highlighting how the choice of features profoundly influences model outcomes. MI feature selection aims to extract the most relevant features, enhancing model robustness and reducing overfitting. However, the observed slight reduction in performance metrics with MI suggests a potential trade-off between simplicity (with fewer features) and model accuracy but achieving a significant advantage in computational time thus reducing computation cost as well.

7.1 Limitations:

Because of limitations in time and resources, creating a unique intrusion dataset through simulation or other means wasn't feasible. Hence, the research project relied on the widely used UNR-IDD dataset. Additionally, due to resource limitations, a larger dataset wasn't utilized, which might have demonstrated improved results. Factors such as quality constraints, and the inherent complexities of the domain might have influenced the study's outcomes.

7.2 Future Work:

The research aims to extend by implementing real-time strategies for dynamic threat detection and mitigation. Also, performing the test upon additional larger dataset from diverse sources or creating a unique dataset through simulation becomes imperative to train and test the model across various intrusion types. Additionally, the plan includes transforming this system into an API, enabling integration with multiple languages and platforms for versatile use.

8 References

Agyapong, I. *et al.* (2023) *LoGD-Ai: An Efficient Network Intrusion Detection System Using a Soft Voting-Based Ensemble Learner*. In Review DOI: 10.21203/rs.3.rs-3329365/v1.

Albulayhi, K. *et al.* (2022) 'IoT Intrusion Detection Using Machine Learning with a Novel High Performing Feature Selection Method'. *Applied Sciences*, 12(10), p. 5015. DOI: 10.3390/app12105015.

Almaghthawi, Y., Ahmad, I. and Alsaadi, F.E. (2022) 'Performance Analysis of Feature Subset Selection Techniques for Intrusion Detection'. *Mathematics*, 10(24), p. 4745. DOI: 10.3390/math10244745.

Almasoudy, F.H., Al-Yaseen, W.L. and Idrees, A.K. (2020) 'Differential Evolution Wrapper Feature Selection for Intrusion Detection System'. *Procedia Computer Science*, 167, pp. 1230–1239. DOI: 10.1016/j.procs.2020.03.438.

Almomani, O. (2020) 'A Feature Selection Model for Network Intrusion Detection System Based on PSO, GWO, FFA and GA Algorithms'. *Symmetry*, 12(6), p. 1046. DOI: 10.3390/sym12061046.

Alqahtani *et al.* (2019) 'A Genetic-Based Extreme Gradient Boosting Model for Detecting Intrusions in Wireless Sensor Networks'. *Sensors*, 19(20), p. 4383. DOI: 10.3390/s19204383.

Al-Sarem, M. *et al.* (2021) 'An Aggregated Mutual Information Based Feature Selection with Machine Learning Methods for Enhancing IoT Botnet Attack Detection'. *Sensors*, 22(1), p. 185. DOI: 10.3390/s22010185.

Awotunde, J.B., Chakraborty, C. and Adeniyi, A.E. (2021) 'Intrusion Detection in Industrial Internet of Things Network-Based on Deep Learning Model with Rule-Based Feature Selection' Jolfaei, A. (ed.). *Wireless Communications and Mobile Computing*, 2021, pp. 1–17. DOI: 10.1155/2021/7154587.

Das, T. (2023) 'UNR-IDD Dataset'. Available at: https://www.tapadhirdas.com/unr-idd-dataset.

Dash, S. *et al.* (2022) 'Multiscale Domain Gradient Boosting Models for the Automated Recognition of Imagined Vowels Using Multichannel EEG Signals'. *IEEE Sensors Letters*, 6(11), pp. 1–4. DOI: 10.1109/LSENS.2022.3218312.

Disha, R.A. and Waheed, S. (2022) 'Performance Analysis of Machine Learning Models for Intrusion Detection System Using Gini Impurity-Based Weighted Random Forest (GIWRF) Feature Selection Technique'. *Cybersecurity*, 5(1), p. 1. DOI: 10.1186/s42400-021-00103-8.

Farhan, B.I. and Jasim, A.D. (2022) 'Performance Analysis of Intrusion Detection for Deep Learning Model Based on CSE-CIC-IDS2018 Dataset'. *Indonesian Journal of Electrical Engineering and Computer Science*, 26(2), p. 1165. DOI: 10.11591/ijeecs.v26.i2.pp1165-1172.

Hossain, Md.A. and Islam, Md.S. (2023) ‘Ensuring Network Security with a Robust Intrusion Detection System Using Ensemble-Based Machine Learning’. *Array*, 19, p. 100306. DOI: 10.1016/j.array.2023.100306.

Jaw, E. and Wang, X. (2021) ‘Feature Selection and Ensemble-Based Intrusion Detection System: An Efficient and Comprehensive Approach’. *Symmetry*, 13(10), p. 1764. DOI: 10.3390/sym13101764.

Karthigha, M. and L, L. (2022) *Clustered Ensemble Feature Selection with M-GRU Classification for Efficient Intrusion Detection System of Industrial Systems*. In Review DOI: 10.21203/rs.3.rs-1571372/v1.

Kasongo, S.M. and Sun, Y. (2020) ‘Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset’. *Journal of Big Data*, 7(1), p. 105. DOI: 10.1186/s40537-020-00379-6.

Le, T.-T.-H., Oktian, Y.E. and Kim, H. (2022) ‘XGBoost for Imbalanced Multiclass Classification-Based Industrial Internet of Things Intrusion Detection Systems’. *Sustainability*, 14(14), p. 8707. DOI: 10.3390/su14148707.

Li, X. *et al.* (2021) ‘LNNLS-KH: A Feature Selection Method for Network Intrusion Detection’ Díaz-Verdejo, J. (ed.). *Security and Communication Networks*, 2021, pp. 1–22. DOI: 10.1155/2021/8830431.

Louk, M.H.L. and Tama, B.A. (2022) ‘PSO-Driven Feature Selection and Hybrid Ensemble for Network Anomaly Detection’. *Big Data and Cognitive Computing*, 6(4), p. 137. DOI: 10.3390/bdcc6040137.

Markevych, M. and Dawson, M. (2023) ‘A Review of Enhancing Intrusion Detection Systems for Cybersecurity Using Artificial Intelligence (AI)’. *International Conference KNOWLEDGE-BASED ORGANIZATION*, 29(3), pp. 30–37. DOI: 10.2478/kbo-2023-0072.

Nimbalkar, P. and Kshirsagar, D. (2021) ‘Feature Selection for Intrusion Detection System in Internet-of-Things (IoT)’. *ICT Express*, 7(2), pp. 177–181. DOI: 10.1016/j.icte.2021.04.012.

Qadir Mohammed, S. and A. Hussein, M. (2022) ‘Performance Analysis of Different Machine Learning Models for Intrusion Detection Systems’. *Journal of Engineering*, 28(5), pp. 61–91. DOI: 10.31026/j.eng.2022.05.05.

Saha, S. *et al.* (2022) ‘Towards an Optimized Ensemble Feature Selection for DDoS Detection Using Both Supervised and Unsupervised Method’. *Sensors*, 22(23), p. 9144. DOI: 10.3390/s22239144.

Seth, S., Singh, G. and Kaur Chahal, K. (2021) ‘A Novel Time Efficient Learning-Based Approach for Smart Intrusion Detection System’. *Journal of Big Data*, 8(1), p. 111. DOI: 10.1186/s40537-021-00498-8.

Tripathy, S.S. and Behera, B. (2023) ‘PERFORMANCE EVALUATION OF MACHINE LEARNING ALGORITHMS FOR INTRUSION DETECTION SYSTEM’. *Journal of Biomechanical Science and Engineering*.

Upadhyay, D. *et al.* (2021) ‘Gradient Boosting Feature Selection With Machine Learning Classifiers for Intrusion Detection on Power Grids’. *IEEE Transactions on Network and Service Management*, 18(1), pp. 1104–1116. DOI: 10.1109/TNSM.2020.3032618.

Verma, P. *et al.* (2021) ‘A Novel Intrusion Detection Approach Using Machine Learning Ensemble for IoT Environments’. *Applied Sciences*, 11(21), p. 10268. DOI: 10.3390/app112110268.