Fortifying network security: machine learning-powered intrusion detection systems and classifier performance analysis

Arar Al Tawil[1], Lara Al-Shboul[2], Laiali Almazaydeh[3], Mohammad Alshinwan[1,4]
[1]Faculty of Information Technology, Applied Science Private University, Amman, Jordan
[2]King Abdullah II School of Information Technology, The University of Jordan, Amman, Jordan
[3]College of Information Technology, Al-Hussein Bin Talal University, Ma'an, Jordan
[4]MEU Research Unit, Middle East University, Amman, Jordan

Article Info

Article history:

Received Feb 21, 2024
Revised Jun 18, 2024
Accepted Jul 1, 2024

Keywords:

Class imbalance handling
Classification
Feature selection
Intrusion detection systems
Preprocessing

ABSTRACT

Intrusion detection systems (IDS) protect networks from threats; they actively monitor network activity to identify and prevent malicious actions. This study investigates the application of machine learning methods to strengthen IDS, explicitly emphasizing the comprehensive CICIDS 2017 dataset. The dataset was refined by implementing stringent preprocessing methods such as feature normalization, class imbalance management, feature reduction, and feature selection to ensure its quality and lay the foundation for developing robust models. The performance evaluation of three classifiers-support vector machine (SVM), extreme gradient boosting (XGBoost), and naive Bayes was highly impressive. Vital accuracy, precision, recall, and F1-score values of 0.984389, 0.984479, 0.984375, and 0.984304, respectively, were achieved by SVM. Notably, XGBoost demonstrated exceptional performance across all metrics, attaining flawless scores of 1.0. naive Bayes demonstrated noteworthy accuracy, precision, recall, and F1-score performance, which were recorded as 0.877392, 0.907171, 0.877007, and 0.876986, respectively. The results of this study emphasize the critical importance of preparation methods in improving the effectiveness of IDS via machine learning. This further demonstrates the potential of particular classifiers to detect and prevent network intrusions efficiently, thereby substantially contributing to cybersecurity measures.

Corresponding Author:

Arar Al Tawil
Faculty of Information Technology, Applied Science Private University
Amman, Jordan
Email: ar_altawil@asu.edu.jo

1. INTRODUCTION

An intrusion detection system (IDS) is a device or application that monitors a network or systems to detect and prevent malicious activity or policy violations are known as an IDS. It is possible to find IDS variants customized to accommodate diverse levels of security, spanning from individual computer systems to vast networks. The primary classifications are network intrusion detection systems (NIDS) and host-based intrusion detection systems (HIDS). Frequently, intrusion detection systems are divided into two categories based on the detection method [1]. An IDS may be deployed as a physical device or a software application to oversee system operations or network activity. Its principal objective is identifying and responding to malevolent activities or violations of predetermined regulations. IDS varieties demonstrate various applications, encompassing server rooms and enterprise networks. HIDS and NIDS are the primary standard categories. Frequently, classifications of intrusion detection systems are based on their detection methodologies [2], [3]. The fundamental classification is predicated on signature detection, which compares distinctive patterns in network traffic (e.g., byte sequences) with an established repository of recognized attack signatures. In contrast, the anomaly-based detection approach assesses the current state of a network about a predetermined reference point. This empowers it to identify and discern both established and novel perils. Furthermore, it is critical to note that dimensional reduction is a prevalent technique utilized in machine learning, mainly when dealing with feature spaces comprising numerous dimensions.

Learning by machines is analogous to instructing computers to improve their task performance without being explicitly instructed on each step. It is everything about developing programs that can use data to enhance intelligence. Observing the data and learning from it to identify patterns and generate more precise predictions is the initial step in the learning process. The primary objective is for computers to acquire knowledge autonomously and adjust their behavior without requiring continuous human supervision [4]. Preprocessing can significantly impact the overall predictive performance of a supervised machine learning algorithm in the context of generating hypotheses using novel data. One of the most formidable challenges encountered in inductive machine learning pertains to detecting and eliminating chaotic instances. These cases commonly demonstrate substantial departures from the standard, frequently distinguished by many absent or inconsequential attribute values. Often, these exceptionally aberrant characteristics are denoted as outliers. In addition, in situations where working with huge datasets is impractical, it is typical to select a representative sample from the massive set while also addressing the problem of missing data [5]. Our study employed A comprehensive preprocessing strategy to improve data quality and maximize the efficiency of our machine-learning models. The approach utilized various methods, including data normalization for consistent scaling. Data normalization entails reducing the magnitude of numerical characteristics in a dataset to a standard range, typically from 0 to 1. This mechanism prevents any one feature from exerting an excessive influence on machine learning models by ensuring that all features have an equal impact [6]. Feature selection by correlation entails identifying and retaining the most pertinent characteristics present in a given dataset. The primary objective is to decrease the dimensionality of the data without altering the attributes that maintain the most robust associations with the target variable. This streamlines the process of modeling [7]. Managing missing data techniques entails the implementation of approaches to address data instances or attributes that contain null or incomplete values. Conventional approaches to managing missing values encompass imputation and exclusion. Imputation entails employing statistical techniques to compensate for missing values, while exclusion entails excluding instances containing missing data from the analysis [5]. Class imbalance strategies aim to alleviate the problem when one class is significantly underrepresented relative to the others in a given dataset. These methods aim to restore equilibrium to the class distribution so that machine-learning models can generate accurate predictions for all classes, including those with fewer instances and do not favor the majority class. Methods include oversampling, undersampling, and applying suitable evaluation metrics [8]. Implementing these preprocessing procedures was critical in empowering our machine learning models to generate precise and resilient forecasts, even when confronted with intricate and practical datasets.

This paper tackles the critical issue of improving IDS to ensure that they can accurately identify and mitigate network intrusions, which is essential for maintaining a robust cybersecurity system. The proposed solution entails using machine learning techniques, with a particular emphasis on preprocessing methods such as feature normalization, class imbalance management, feature reduction, and feature selection, to enhance the quality of the data and construct robust models. The study assesses the efficacy of three classifiers: Naive Bayes, extreme gradient boosting (XGBoost), and support vector machine (SVM). The results suggest that SVM obtained high accuracy, precision, recall, and F1-score, whereas XGBoost exhibited extraordinary performance with flawless scores across all metrics. Although Naive Bayes was less effective than the other two, it still demonstrated significant precision and accuracy. This research expands upon previous research by utilizing rigorous preprocessing techniques and assessing the efficacy of various classifiers on the CICIDS 2017 dataset. The results emphasize the superior performance of XGBoost and the critical role of data preparation in enhancing the effectiveness of IDS.

Following this, the remaining sections are structured as follows: an examination of the literature about intrusion detection systems and machine learning algorithms is presented in section 2. The methodology utilized in this study is delineated in section 3, encompassing the selection of datasets, preprocessing procedures, and experimental configuration. The evaluation and implementation of multiple machine learning classifiers for intrusion detection are described in section 4. The results and analysis of the experiments are detailed in section 5, emphasizing performance metrics, including accuracy, false positives, and detection rate (DR). In conclusion, the paper is summarized in section 6, which also analyzes the main findings' implications and proposes potential directions for future research.

2. LITERATURE REVIEW

The study [9] utilizes a variety of established machine learning classification algorithms, including the Bayesian network, naive Bayes classifier, decision tree (DT), random decision forest, random tree, decision table, and artificial neural network (ANN). The objective is to identify intrusions and improve cyber-security services. The researchers do tests using the KDD'99 cup dataset, which encompasses a wide range of cyber-attack categories. The assessment of these algorithms includes performance indicators such as precision, recall, F1-score, and accuracy. The random forest (RF) classifier stands out as the best performer, with an excellent accuracy of 0.94. This highlights its effectiveness in the field of cyber-security intrusion detection.

Regarding the research referenced as [10], the authors have presented a feature selection model known as ID3-BA. This model is intricately crafted to maximize the selection of a subset of attributes within the area of IDS. The technique integrates the ID3 classifier algorithm with the bees algorithm, with the bees algorithm being crucial in creating the necessary subset of features, while the ID3 algorithm is used to create the classifier. The study used the KDD Cup99 dataset, a well-recognized dataset in the field of knowledge discovery and data mining. This dataset consists of 41 characteristics that are used for both training and testing. The performance assessment criteria consist of three primary metrics: false alarm rate (FAR), detection rate (DR), and accuracy (AR). The empirical data obtained from this study activity provide persuasive outcomes. The ID3-BA model regularly achieves a high detection rate of 91.02% and an exceptional accuracy rate of 92.002%. It also maintains a low FAR of 3.917%. The research's main conclusion emphasizes that carefully choosing a subset of characteristics, rather than employing all of them, greatly improves the effectiveness of IDS in terms of detection rate, accuracy, and a decrease in false alarm rate.

As described in reference [2], the authors have developed a methodical approach for selecting features in the field of IDS. This strategy entails the collaborative employment of a clustering algorithm performed via filter and wrapper approaches. The wrapper approach utilizes the linear correlation coefficient algorithm (FGLCC), while the filter method utilizes the cuttlefish algorithm (CFA). The suggested technique also integrates a decision tree for constructing the classifier, and its performance evaluation is based on the well-established KDD Cup 99 dataset. Throughout the experimental phase, performance assessment involves crucial variables such as accuracy, detection rate, false positives, and a fitness function. The assessment findings are rigorously compared to those acquired by the 10-fold cross-validation approach and other techniques based on features. The result of this meticulous testing is convincing. The FGLCC-CFA algorithm regularly surpasses other approaches, with a remarkable detection rate of 95.23%, an accuracy rate of 95.03%, and an incredibly low false positive rate of 1.65%. The results highlight the effectiveness of the suggested technique in improving the performance of IDS and its significant benefits compared to alternative feature selection algorithms.

The main aim of research [11] is to improve the effectiveness of IDS by combining rule-based approaches with learning-based algorithms for the purpose of detecting and categorizing intrusions. The study utilizes neural networks (NN), RF, and SVM techniques to accomplish this objective. The study used conventional datasets, such as KDD Cup 99, as input in their system. Before doing analysis, the KDD 99 dataset undergoes a preprocessing stage to remove data noise and assure data consistency. Consequently, the researchers get pristine and uniform input data. The processed data is then inputted into machine learning algorithms such as SVM, NN, and RF to carry out classification tasks. The categorization results are then used as training data for prediction tasks. Practically, when fresh data about infiltration attempts is included into the framework, the system utilizes the acquired patterns from the training data to forecast if the new data is typical or atypical. The remarkable accomplishment of this study is the SVM algorithm achieving the best accuracy score of 0.94, highlighting its usefulness in detecting and classifying intrusions in this specific context.

Chung *et al.* [12] have created simplified swarm optimization (SSO), a new and efficient variation of particle swarm optimization (PSO) designed explicitly for feature selection. This approach integrates a localized search technique to accelerate the selection process of features by discovering the most optimum surrounding solution. The suggested SSO method has a crucial capability to significantly decrease the number of characteristics needed to capture the behavioral patterns in network traffic data accurately. More precisely, it reduces the original collection of 41 factors from the KDD Cup 99 dataset to just six elements while also attaining higher accuracy than the typical PSO technique. The SSO technique achieves a notable accuracy rate of 93.3%, highlighting its usefulness in enhancing feature selection for enhanced performance in network traffic analysis and intrusion detection.

Almseidin *et al.* [13] provided valuable insights by examining specific classifiers. The decision table classifier demonstrated superior performance by achieving the lowest false negative rate. On the other hand, the RF classifier excelled with an accuracy rate of 93.77%, backed by the least root mean square error

(RMSE) and minimum false positives. The random tree classifier had the lowest mean accuracy rate yet with the smallest receiver operating characteristic (ROC) value. Meanwhile, the multi-layer perceptron (MLP) and naive Bayes classifiers showed similar average accuracy rates. The Bayes network algorithm has demonstrated exceptional performance in accurately recognizing regular packets. On the other hand, while the Decision Table algorithm did not achieve the maximum level of accuracy, it exhibited the lowest rate of false negatives and efficient model construction. Ultimately, rule-based classifiers such as the decision table provide a favorable balance by achieving satisfactory accuracy and instilling a greater sense of certainty, mainly because they have the lowest rates of false negatives when used for intrusion detection.

Agarwal *et al.* [14] featured a thorough examination that used three different machine learning classification algorithms: naïve Bayes (NB), SVM, and k-nearest neighbor (KNN). The main objective was to determine their efficacy in improving accuracy and reducing processing time using the UNSW-NB15 dataset. The primary goal was to identify the most appropriate algorithm for acquiring knowledge about the complexities of suspicious network activity. The selection of the most suitable algorithm for training the IDS was facilitated by conducting a comparative study of feature sets. The selected algorithm was then used to forecast and analyze future incursion behavior. During the testing phase of the model, performance measures such as accuracy, recall, and F1-score were systematically produced. Additionally, confusion matrices were created and compared to determine the best validation and support status achieved. The derived results show that the SVM outperformed the other algorithms, achieving an impressive accuracy rate of 0.977. This highlights the outstanding appropriateness of SVM in the study model, showcasing its capacity to handle the dataset successfully and improve intrusion detection skills.

Emanet *et al.* [15] focuses on developing a sophisticated IDS that prioritizes enhanced accuracy using strategic feature selection and ensemble learning techniques. Using the CIC-CSE-IDS2018 dataset, the study progresses through two crucial phases, substantially contributing to its overall effect. The first refining of the dataset entails carefully selecting features and using ensemble learning methods to enhance the performance of IDS by combining the capabilities of several classifiers. Implementing ensemble learning afterward results in a resilient model, improving attack detection and substantially decreasing detection time. The suggested ensemble model achieves an impressive accuracy rate of 98.82% by using under-sampling and feature selection techniques. This results in a significant decrease of 73% in intrusion detection time and a modest improvement of 3% in accuracy. Spearman's correlation analysis, recursive feature elimination (RFE), and chi-square test procedures are used to determine the essential elements that enhance the efficiency of IDS. A comparative comparison of classifiers, such as additional trees, decision trees, and logistic regression, demonstrates reasonable accuracy rates while considering actual implementation time. The significance of this research is its contribution to the advancement of IDS capabilities through the proposal of an ensemble learning model that surpasses individual classifiers. This affirms the model's potential impact on future intrusion detection systems and strengthens computer security across various domains. Additionally, it paves the way for innovative approaches in the field.

Fitni and Ramli [16] addresses the growing concerns about data security in organizational information systems. It emphasizes the necessity for more robust defensive mechanisms to counter sophisticated assaults that may bypass standard security technologies such as firewalls and antivirus software. This study aims to overcome the constraints of existing IDSs by using an ensemble learning technique. The approach combines logistical regression, decision trees, and gradient boosting as effective classifiers. Using the CSE-CIC-IDS2018 dataset and employing Spearman's rank correlation coefficient, the research improves the model by carefully choosing 23 essential characteristics from a pool of 80, considerably boosting its concentration. The experimental results illustrate the strength of the ensemble model, displaying exceptional performance metrics: a final accuracy of 98.8%, precision, and recall rates of 98.8% and 97.1%, respectively, resulting in an excellent F1-score of 97.9%. These results highlight the effectiveness of ensemble learning in strengthening IDS capabilities, making significant progress in tackling current difficulties and enhancing network security.

Al Tawil and Sabri [17] introduces a novel feature selection algorithm for IDS that employs the moth flame optimization (MFO) algorithm. The objective of the proposed algorithm is to reduce the time required for training and improve the precision of the model by selecting pertinent features. The algorithm was evaluated on the CIC-2017 dataset, resulting in a reduction of the number of features from 78 to 4. It obtained a high detection rate (100%) and accuracy (99.9%) with a lower false alarm rate.

Table 1 provides a comprehensive summary of the machine learning algorithms applied to intrusion detection systems, as documented in the relevant literature. Every cell in the table represents a distinct study, providing comprehensive information regarding the algorithms utilized, datasets incorporated, performance metrics assessed, and significant discoveries attained. This comparative analysis illuminates the efficacy of various methodologies in detecting and classifying intrusions, providing essential perspectives for improving cybersecurity protocols.

Table 1. Machine learning approaches in intrusion detection

Ref.	Algorithms used	Dataset	Performance metrics	Key findings
9	Bayesian network, NB, DT, random decision forest, random tree, decision table, ANN	KDD'99 cup	Precision, Recall, F1-score, Accuracy	RF classifier achieved the highest accuracy of 0.94.
10	ID3-BA (ID3 classifier + bees algorithm)	KDD Cup99	FAR, DR, AR	ID3-BA model achieved a DR of 91.02%, AR of 92.002%, and FAR of 3.917%.
2	FGLCC-CFA (Filter: FGLCC, Wrapper: CFA)	KDD CUP99	Accuracy, DR, False Positives, Fitness Function	FGLCC-CFA algorithm achieved a DR of 95.23%, AR of 95.03%, and false positives rate of 1.65%.
11	NN, RF, SVM	KDD Cup 99	Accuracy	SVM algorithm achieved the highest accuracy score of 0.94.
12	SSO	KDDCUP 99	Accuracy	SSO achieved an accuracy rate of 93.3% and reduced the number of features from 41 to 6.
13	Decision table, RF, random tree, MLP, NB, Bayes network	-	False negative rate, Accuracy, RMSE, False positives	Decision table showed the lowest false negative rate, RF achieved the highest accuracy of 93.77%.
14	NB, SVM, KNN	UNSW-NB15	Accuracy, Recall, F1-score	SVM outperformed others with an accuracy rate of 0.977.
15	Ensemble learning techniques with feature selection	CIC-CSE-IDS2018	Accuracy	Ensemble model achieved an accuracy rate of 98.82% with a 73% decrease in detection time.
16	Ensemble learning (logistic regression, decision trees, gradient boosting)	CSE-CIC-IDS2018	Accuracy, Precision, Recall, F1-score	Ensemble model achieved an accuracy of 98.8%, precision of 98.8%, recall of 97.1%, and F1-score of 97.9%.
17	MFO	CIC-IDS 2017	Accuracy, F-score, Sensitivity, Time	Reduced features from 78 to 4. Achieved accuracy of 99.9%, high detection rate of 100%, and lower false alarm rate.

3. METHOD

This section provides an overview of the structure and techniques used to carry out the study. The procedure includes defining the dataset, thoroughly analyzing preprocessing approaches used to improve the data quality, and choosing appropriate classifiers. This section thoroughly explains the methodical technique used in this study, guaranteeing transparency and the ability to reproduce the research process. Figure 1 depicts the approach used in this study report.

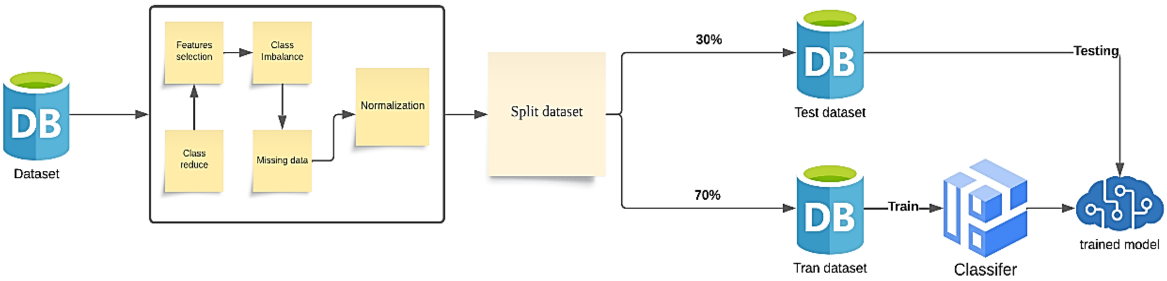

Figure 1. Research paper methodology

3.1. Dataset

The CICIDS2017 dataset [18], developed by the Canadian Institute for Cybersecurity, is an extensive compilation of network traffic data specifically tailored for study and assessment in the realm of cybersecurity and intrusion detection. This dataset offers a wide variety of network traffic situations, including both harmless and harmful actions. As a result, it is an invaluable asset for the development and evaluation of intrusion detection systems. The dataset contains labeled data for various network traffic features, enabling the training and evaluation of intrusion detection models in a controlled and realistic environment.

The dataset consists of unique labels, each linked to a certain quantity of cases as shown in Table 2. These categories contain a wide range of network activity, including both normal traffic and different types of attacks and intrusion attempts. The CICIDS2017 dataset may be used by cybersecurity professionals and academics to improve and evaluate intrusion detection algorithms, therefore strengthening network security and reducing risks. Regrettably, the offered information does not provide the specific count of occurrences for the "Heartbleed" category.

Table 2. Dataset

Label	Number of instances
Benign	227,3097
DoS Hulk	23,1073
PortScan	15,8930
DDoS	12,8027
DoS GoldenEye	10,293
FTP-Patator	7,938
SSH-Patator	5,897
DoS slowloris	5,796
DoS Slowhttptest	5,499
Bot	1,966
Web Attack Brute Force	1,507
Web Attack XSS	652
Infiltration	36
Web Attack SQL Injection	21
Heartbleed	11

XSS: Cross-site scripting

3.2. Preprocessing methods

In the context of this research, several preprocessing techniques were applied to enhance the quality of the dataset and prepare it for further analysis and modeling. These methods aimed to address data imbalances, handle missing values and standardize the feature set for a more robust and accurate analysis. Using these preprocessing techniques, the dataset was converted into a more appropriate format for analysis and modeling, effectively tackling problems such as class imbalance, missing data, and feature relevance. These processes provide the groundwork for more precise and dependable outcomes in the subsequent phases of the study.

3.2.1. Class reduction based

To simplify the dataset and improve computational efficiency, classes with more than 10,000 instances were retained while others were reduced or excluded. This reduction process ensures that the dataset remains manageable and that computational resources are used effectively. By focusing on labels with a substantial number of instances, the analysis can prioritize the most prevalent and significant classes for a more efficient and targeted study.

3.2.2. Feature selection using correlation

Feature selection based on correlation [19] was employed to identify and retain the most relevant features while eliminating redundant or highly correlated ones. A correlation threshold of 0.6 was applied, selecting the best 40 features from the dataset. This step optimizes the feature set by focusing on those attributes that have the most significant impact on the analysis, while also ensuring that the selected features are not highly correlated with each other. As we show in (1) present the formula of Correlation.

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \qquad (1)$$

3.2.3. Class imbalance

To address class imbalance [20] issues, the RandomUnderSampler [21] technique was applied. This method reduces the number of instances in the overrepresented classes, effectively balancing the class distribution and preventing biased model training. By ensuring a more equitable representation of each class, the RandomUnderSampler helps improve the performance and reliability of the machine learning models. This approach ensures that the models can accurately predict outcomes for all classes, including those with fewer instances.

3.2.4. Missing data

The missing of data may lead to the introduction of disturbances and errors in the analysis. The dataset was imputed using the SimpleImputer [22] using the 'mean' technique to replace missing values. This approach involves substituting missing values with the average value of the related characteristic, hence maintaining the integrity and significance of the data.

3.2.5. Normalization

To ensure that all characteristics are measured on the same scale, the StandardScaler technique [23] was used. This approach normalizes the elements, providing a mean of 0 and a standard deviation of 1.

Standardization facilitates the attainment of a consistent and comprehensible dataset, which is especially crucial for specific machine-learning algorithms.

4. CLASSIFIER

The dataset was converted into a more appropriate format for analysis and modeling using these preprocessing techniques, effectively tackling problems such as class imbalance, missing data, and feature relevance. These processes provide the groundwork for more precise and dependable outcomes in the subsequent phases of the study.

4.1. Support vector machine

The support vector machine (SVM) is a widely used and reliable classification technology known for its effectiveness in handling complex and high-dimensional data. We used SVM to capture intricate, non-linear relationships within the dataset and assess its suitability. The key parameters used for the support vector machine model were:

- Kernel function: radial basis function (RBF)
- Regularization parameter (C): 1.0

4.2. XGBoost

XGBoost is a widely used ensemble learning technique known for its resilience and exceptional forecast precision. We used the gradient boosting methodology to construct a collection of decision trees, allowing the model to capture complex patterns within the data effectively. The primary parameters used for the XGBoost model included:

- Learning rate: 0.1
- Max Depth: 6
- Subsample: 0.8
- Colsample_bytree: 0.8
- Number of estimators: 100

4.3. Naive Bayes

Naive Bayes is a probabilistic technique that is based on Bayes' theorem. It is particularly advantageous for tasks like text categorization when the feature independence assumption holds. We evaluated the appropriateness of naive Bayes in our dataset to estimate its efficacy in the classification task. The key parameters for the naive Bayes model were – Distribution: Gaussian naive Bayes.

5. EXPERIMENTAL SETUP AND RESULTS

This section explains the sequential procedure for performing experiments, which encompasses data preparation, dataset separation, and the application of the chosen classifiers. The experimental configuration was devised to guarantee a meticulous and uniform assessment of the model's efficacy. Subsequently, we report the outcomes of these trials, highlighting crucial assessment criteria.

5.1. Dataset splitting

The dataset was divided into training (80%) and testing (20%) subsets to ensure a robust evaluation of the classifiers. This split ratio was chosen to provide a balance between model training and independent model evaluation while preventing overfitting. By maintaining this proportion, the model benefits from sufficient data to learn effectively while having enough separate data to evaluate its performance reliably.

5.2. Classifier selection and model training

Three classifiers were chosen for the analysis: the SVM, XGBoost, and naive Bayes classifiers. These classifiers were trained using the training dataset and specific parameter configurations to optimize their performance. Afterward, the models were assessed for their performance using the testing dataset, allowing for a thorough evaluation of their predictive accuracy and effectiveness.

5.3. Metrics

5.3.1. Model accuracy

Almazaydeh *et al.* [24] quantifies the extent to which the model's predictions align with the actual outcomes, measuring the model's overall reliability. The metric measures the ratio of accurately identified examples to the total number of occurrences in the dataset. A greater level of accuracy signifies a more significant proportion of accurate forecasts. Equation (2) represents the formal to calculate the accuracy.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \tag{2}$$

5.3.2. Precision

Almazaydeh *et al.* [24] is a quantitative measure that assesses the model's capacity to accurately forecast favorable outcomes. The calculation involves dividing the number of correct optimistic predictions by the total number of positive predictions generated by the model. Greater accuracy indicates that the model has a higher probability of correctly predicting a joyous event. Equation (3) represents the formal to calculate the precision.

$$Precision = \frac{TP}{(TP+FP)} \tag{3}$$

5.3.3. Recall

Almazaydeh *et al.* [24] sometimes called sensitivity or actual positive rate, quantifies the model's capacity to detect all positive cases accurately. It computes the proportion of accurate optimistic predictions relative to the dataset's overall number of positive cases. A higher recall signifies that the model can capture more good cases. Equation (4) represents the formal to calculate the recall.

$$Recall = \frac{TP}{(TP+FN)} \tag{4}$$

5.3.4. F1-score

Almazaydeh *et al.* [25] is calculated as the reciprocal of the arithmetic mean of the reciprocals of accuracy and recall. It compromises accuracy and recall by including erroneous positives and false negatives into a unified score. The F1-score is precious when there is a disparity between the number of positive and negative classifications in the dataset. Equation (5) represents the formal to calculate the F1-score**.**

$$F1 - Score = 2 * \frac{(Recall*Precision)}{(Recall+Precision)} \tag{5}$$

5.4. Results

The classifiers' performance was evaluated using several measures, such as model accuracy, model precision, model recall, and model F1-score. These metrics provide a comprehensive assessment of how well each classifier performs in terms of both prediction accuracy and the balance between precision and recall. The findings, as shown in Table 3, highlight the comparative effectiveness of the SVM, XGBoost, and naive Bayes classifiers.

Table 3. Result

Model	Model accuracy	Model precision	Model recall	Model F1-score
SVM	0.984389	0.984479	0.984375	0.984304
XGBoost	1	1	1	1
Naive Bayes	0.877392	0.907171	0.877007	0.876986

In Figure 2 illustrates the model accuracy scores for SVM, XGBoost, and naive Bayes. It provides a comparative view of the accuracy achieved by each model, showing the performance in correctly predicting instances across the dataset. In Figure 3 displaying precision scores for SVM, XGBoost, and Naive Bayes, this figure highlights the precision achieved by each model. Precision is a metric that quantifies the proportion of accurate optimistic predictions produced by the model out of all the positive predictions it made.

Figure 4 displays the recall scores for SVM, XGBoost, and naive Bayes, illustrating the models' capacity to identify all positive cases accurately. The term "true positive rate" refers to the proportion of correctly predicted positive cases in the dataset relative to the total number of positive instances. Figure 5 displays the F1-scores for SVM, XGBoost, and naive Bayes. This figure represents a composite measure that considers both precision and recall. The F1-score is precious when there is a disparity between the number of positive and negative classifications in the dataset.

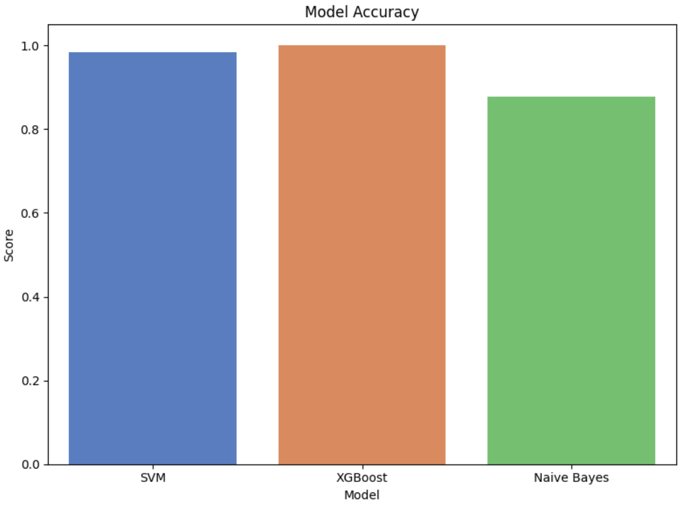

Figure 2. Model accuracy across different models

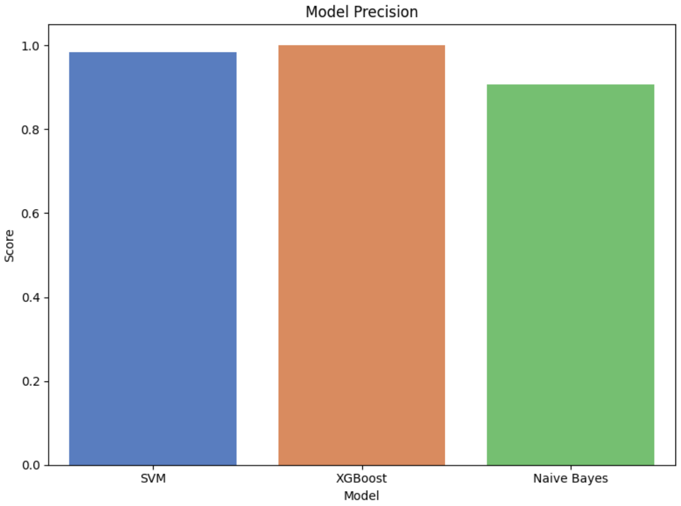

Figure 3. Model precision across different models

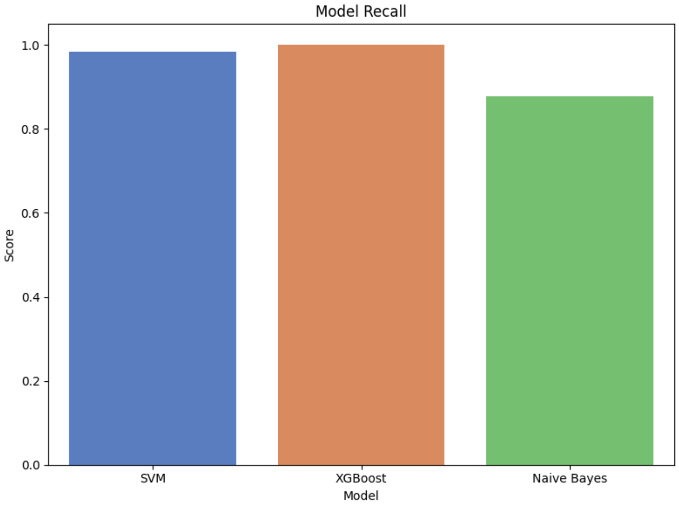

Figure 4. Model recall across different models

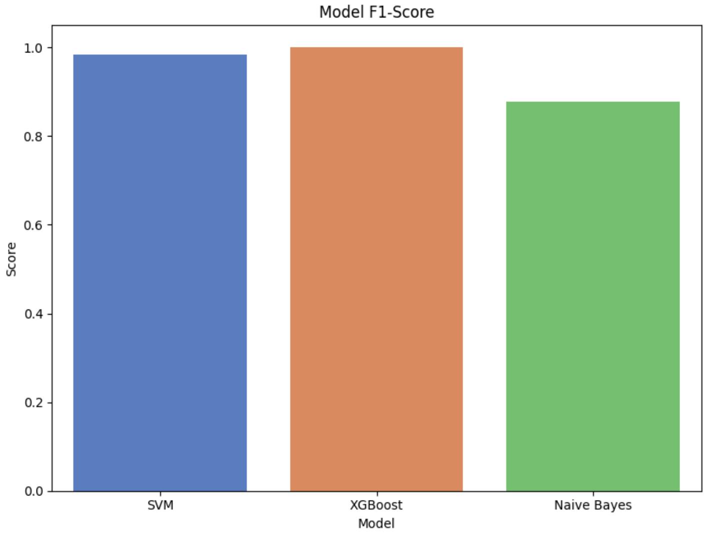

Figure 5. Model F1-score across different models

6. CONCLUSION

In conclusion, this investigation tackles the critical issue of improving IDS to accurately identify and mitigate network intrusions, a prerequisite for a robust cybersecurity system. In order to enhance the quality of the data, the study implements sophisticated machine learning techniques and rigorous preprocessing strategies, such as class reduction, feature selection, class imbalance management, missing data treatment, and feature normalization, by utilizing the CICIDS 2017 dataset. The results of the evaluations of the SVM, XGBoost, and naive Bayes classifiers were compelling. XGBoost exhibited extraordinary performance, attaining near-perfect scores in the F1-score, precision, recall, and accuracy metrics. SVM also demonstrated commendable performance, consistently achieving high ratings across various metrics. Despite its comparatively inferior performance compared to SVM and XGBoost, naive Bayes still achieved significant results. These results emphasize the critical role of preprocessing techniques in improving the efficiency of IDS through machine learning. The potential of specific classifiers, notably XGBoost, to accurately identify and mitigate network intrusions is underscored by their exceptional performance, significantly enhancing cybersecurity measures. This research enhances the existing body of work by employing rigorous preprocessing techniques and evaluating various classifiers on the CICIDS 2017 dataset. In addition to demonstrating promising results, additional research utilizing a broader range of classifiers and diverse datasets could provide a more profound understanding of the potential of machine learning to improve IDS capabilities and guarantee robust network security.

REFERENCES

[1] H.-J. Liao, C.-H. Richard Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: a comprehensive review," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 16–24, Jan. 2013, doi: 10.1016/j.jnca.2012.09.004.

[2] S. Mohammadi, H. Mirvaziri, M. Ghazizadeh-Ahsaee, and H. Karimipour, "Cyber intrusion detection by combined feature selection algorithm," *Journal of Information Security and Applications*, vol. 44, pp. 80–88, Feb. 2019, doi: 10.1016/j.jisa.2018.11.007.

[3] K. Rajasekaran and K. Nirmala, "Classification and importance of intrusion detection system," *International Journal of Computer Science and Information Security*, vol. 10, no. 8, pp. 44–47, 2020.

[4] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, Jan. 2021, doi: 10.1613/jair.1.12228.

[5] D. Kotsiantis, S. B. Kanellopoulos and P. E. Pintelas, "Data preprocessing for supervised leaning," *International journal of computer science*, vol. 1, no. 2, pp. 111–117, 2006.

[6] P. J. M. Ali, R. H. Faraj, E. Koya, P. J. M. Ali, and R. H. Faraj, "Data normalization and standardization: a technical report," *Mach Learn Tech Rep*, vol. 1, no. 1, pp. 1–6, 2014.

[7] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: a review," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, Apr. 2022, doi: 10.1016/j.jksuci.2019.06.012.

[8] H. Ali, M. N. Mohd Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, "Imbalance class problems in data mining: a review," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3, pp. 1552–1563, Jun. 2019, doi: 10.11591/ijeecs.v14.i3.pp1552-1563.

[9] H. Alqahtani, I. H. Sarker, A. Kalim, S. M. Minhaz Hossain, S. Ikhlaq, and S. Hossain, "Cyber intrusion detection using machine learning classification techniques," in *Communications in Computer and Information Science*, vol. 1235, Springer Singapore, 2020, pp. 121–131.

[10] A. S. Eesa, Z. Orman, and A. M. A. Brifcani, "A new feature selection model based on ID3 and bees algorithm for intrusion detection system," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 23, pp. 615–622, 2015, doi: 10.3906/elk-1302-53.

[11] G. S. Sajja, M. Mustafa, R. Ponnusamy, and S. Abdufattokhov, "Machine learning algorithms in intrusion detection and classification," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 6, pp. 12211–12219, 2021.

[12] Y. Y. Chung and N. Wahid, "A hybrid network intrusion detection system using simplified swarm optimization (SSO)," *Applied Soft Computing*, vol. 12, no. 9, pp. 3014–3022, Sep. 2012, doi: 10.1016/j.asoc.2012.04.020.

[13] M. Almseidin, M. Alzubi, S. Kovacs, and M. Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system," Sep. 2017, doi: 10.1109/sisy.2017.8080566.

[14] A. Agarwal, P. Sharma, M. Alshehri, A. A. Mohamed, and O. Alfarraj, "Classification model for accuracy and intrusion detection using machine learning approach," *PeerJ Computer Science*, vol. 7, Apr. 2021, doi: 10.7717/peerj-cs.437.

[15] S. Emanet, G. Karatas Baydogmus, and O. Demir, "An ensemble learning based IDS using voting rule: VEL-IDS," *PeerJ Computer Science*, vol. 9, Sep. 2023, doi: 10.7717/peerj-cs.1553.

[16] Q. R. S. Fitni and K. Ramli, "Implementation of ensemble learning and feature selection for performance improvements in anomaly-based intrusion detection systems," in *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, Jul. 2020, pp. 118–124, doi: 10.1109/IAICT50021.2020.9172014.

[17] A. Al Tawil and K. E. Sabri, "A feature selection algorithm for intrusion detection system based on moth flame optimization," in *2021 International Conference on Information Technology (ICIT)*, Jul. 2021, pp. 377–381, doi: 10.1109/ICIT52682.2021.9491690.

[18] UNB, "Intrusion detection evaluation dataset (CIC-IDS2017)," University of New Brunswick, 2023, https://www.unb.ca/cic/datasets/ids-2017.html (accessed Feb. 21, 2024).

[19] R. Duangsoithong and T. Windeatt, "Correlation-Based and Causal Feature Selection Analysis for Ensemble Classifiers," in *Artificial Neural Networks in Pattern Recognition (ANNPR); Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2010, pp. 25–36.

[20] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, Oct. 2018, doi: 10.1016/j.neunet.2018.07.011.

[21] G. LemaÃŽtre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of machine learning research*, vol. 18, no. 17, pp. 1–5, 2017.

[22] M. Kuhn and K. Johnson, *Feature engineering and selection: a practical approach for predictive models*. Chapman and Hall/CRC, 2019.

[23] H. T. Nguyen, A. H. Cao, and P. H. D. Bui, "Electrocardiogram-based heart disease classification with machine learning techniques," 2023, pp. 689–701.

[24] L. Almazaydeh, R. Alsalameen, and K. Elleithy, "Herbal leaf recognition using mask-region convolutional neural network (MASK R-CNN)," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 11, pp. 3664–3671, 2022.

[25] L. Almazaydeh, M. Abuhelaleh, A. Al Tawil, and K. Elleithy, "Clinical text classification with word representation features and machine learning algorithms," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 04, pp. 65–76, Apr. 2023, doi: 10.3991/ijoe.v19i04.36099.

BIOGRAPHIES OF AUTHORS

Arar Al Tawil he earned his BSc. in computer science from Al-Hussein Bin Talal University, Jordan, in 2018, followed by an MSc. from Jordan University in 2021. He currently serves as a lecturer and developer specializing in virtual reality and game design. He holds a prominent position as a lecturer in the esteemed Faculty of Information Technology at Applied Science Private University, Amman. His professional pursuits are deeply rooted in virtual reality, augmented reality environments, and the intricate intersection of machine learning and data analysis. His dedication to these fields is reflected in his ongoing research endeavors, where he continually explores new dimensions of technology. Moreover, he remains at the forefront of innovation and is deeply interested in cutting-edge domains such as deep learning and natural language processing (NLP). This commitment to staying abreast of the latest advancements underscores his dedication to pushing the boundaries of technology and contributing significantly to its ever-evolving landscape. He can be contacted at email: ar_altawil@asu.edu.jo.

Lara Al-Shboul a dedicated researcher and educator, holds an MSc from the University of Jordan in addition to her current pursuit of a Ph.D. in computer science at the same institution. Driven by a fervent passion for advancing knowledge within her field, Lara focuses her research endeavors on artificial intelligence, striving to construct more precise models. Alongside her doctoral studies, Lara excels as an instructor at the University, imparting her expertise with excellence, particularly in the realm of biology. She can be contacted at email: lar9220473@ju.edu.jo.

Laiali Almazaydeh received her doctorate degree in computer science and engineering from University of Bridgeport in USA in 2013, specializing in human computer interaction. She is currently a full professor and the dean of College of Computer Information Technology, The American University in the Emirates, UAE. Laiali has published more than seventy research papers in various international journals and conferences proceedings, her research interests include human computer interaction, pattern recognition, and computer security. She received best paper awards in 3 conferences, ASEE 2012, ASEE 2013 and ICUMT 2016. Recently she has been awarded two postdoc scholarships from European Union Commission and Jordanian-American Fulbright Commission. She can be contacted at emails: laiali.almazaydeh@ahu.edu.jo.

Mohammad Alshinwan received the Ph.D. degree from the School of Computer Engineering, Inje University, Gimhae, Republic of Korea, in 2017. He was an assistant professor with the Department of Computer and Information Sciences, Amman Arab University, Jordan. He is currently an associate professor with applied science private University, Jordan. His research interests include computer networks, mobile networks, information security, artificial intelligent, and optimization methods. He can be contacted at email: m_shinwan@asu.edu.jo.