International Journal of Communication Networks and Information Security
ISSN: 2073-607X, 2076-0930
Volume 15 Issue 02 Year 2023

Hybrid Intrusion Detection Model for Enhancing the Security and Reducing the Computational Cost

Hutaf Alqwifli
Master, Department of Information Technology, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia
421200326@qu.edu.sa
Afef Selmi*
Assistant Professor, Department of Information Technology, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia
a.selmi@qu.edu.sa

Article History

Received: 1 March 2023
Revised: 18 April 2023
Accepted: 16 May 2023

Abstract

Artificial Intelligence (AI) is becoming essential technology in Cybersecurity. It represents a revolution in detecting and analyzing intrusions based on predictive models and classification methods. Various recent studies discussed the applications of artificial intelligence in Intrusion Detection Systems to improve the accuracy of the classifiers in detecting cyber-attacks but ignored the computational cost of running the algorithm which is considered a crucial factor of the model evaluation. The aim of this paper is to solve this security issue by using dimensionality reduction techniques and machine learning algorithms. To raise their effectiveness and thus enhance network security, a hybrid classifier with high accuracy and low computational cost is proposed. It combines Decision Tree (DT) and Linear Regression (LR) techniques with AdaBoost technique to build a powerful model for detecting cyber-attacks. The hybrid model included 5 stages, (i) selecting and analyzing the dataset, (ii) pre-processing it, (iii) reducing the dimensions using the Principal Component Analysis (PCA), (iv) classifying stage and (v) evaluating the model using the dataset UNSW-NB15. The model has been compared with several state-of-the-art algorithms. The results have shown that the proposed hybrid model achieved a high accuracy (99%) and the runtime was significantly reduced by half using PCA principle.

Keywords: Cybersecurity, Intrusion Detection System, Decision Tree, Linear Regression, Feature Reduction, AdaBoost Technique, Computational Cost

1. Introduction

Recently, the need to enhance cyber security has increased with the increase of cyber-attacks over Internet (2.200 attacks per day) [1]. This is due to the increasing tendency to use technology and network in various fields such as medicine, industry, marketing, and others. Cyber-attacks have increased in line with the increase in Internet users and the dependence of institutions and governments on technology. According to a report of Cybersecurity Ventures, cyber-attacks have been ranked as the fifth most significant risk in 2020 that can affect the public and private sectors [2]. They can affect the organization in different manners from simple operational disruptions to major

financial losses which is very costly for the organization in terms of effort and cost. With the rapid increase in attacks and their negative effects on organization, the detection rate of intrusions is very limited, around 0.05% in United States based on the report of the World Economic Forum's 2020 Global Risks[3]. Cyber-attacks are becoming more sophisticated than before and traditional methods are not enough to prevent them. Therefore, with these security issues, the need to provide a secure environment for technology and network users has increased. Cyber Security is responsible for protecting data, systems, and networks from cyber-attacks. It is based on three concepts including confidentiality, integrity, and availability [4].

Therefore, researchers turned to Artificial Intelligence to improve defence methods and thus enhance cybersecurity. Artificial intelligence provides smart and effective solutions to protect against cyber-attacks. It has an effective and positive impact on cyber security [5] . Some of the cybersecurity tasks that an AI can be adapted to perform are Classification, Clustering and Predictive Analysis [6] . Based on these tasks, AI can be applied in cyber security fields, such as network protection, endpoint protection, application security, suspect user behaviour, and others [6].

Intrusion detection is a cybersecurity task. Since intrusion detection is a classification issue, artificial intelligence can be used specifically machine learning to improve its effectiveness [7]. IDS represent a security tool deployed in network to detect and analyze intrusions. With the sophisticated cyber-attacks, intrusion detection process must be fast and more accurate and the traditional methods are not enough to prevent them. Therefore, machine learning helps build powerful classifiers that can detect attacks through network traffic [8]. The performance of the classifier depends on several concepts such as the algorithms used, the pre-processing methods and the type of dataset. Various metrics can be used to measure the classifiers' performance such as accuracy, precision, F-measure, false positive rate (FPR), specificity, detection rate, etc. Computational cost must be considered and kept as low as possible. This can be achieved by using dimensionality reduction techniques and machine learning classifiers. The research study focused on the use of AI approaches for detecting intrusions, determining their effectiveness and evaluating the accuracy of these approaches.

Intrusion Detection System (IDS) aims to raise the level of security in the system and to keep the network from penetration. Intrusion Detection System used AI and specifically machine learning to effectively detect cyber-attacks. It provides effective classifiers to increase the classification accuracy. Various recent studies discussed the applications of artificial intelligence in cybersecurity. The study [9] presents a model for building a powerful classifier based on 5 machine learning algorithms to solve the intrusion detection problem. The research study[10] makes recommendations for proper use of Artificial Intelligence approaches to identify cyber-attacks. Most of the existing studies provided solutions to enhance the effectiveness of the classifiers in detecting cyber-attacks but ignored the computational cost of running the algorithm which is considered a crucial factor the evaluate the model.

Therefore, this paper proposes a solution to increase the accuracy and reduce the computational cost. It starts with an empirical analysis of the proposed approaches for detecting cyber-attacks and then selects the appropriate machine learning classifiers for building a powerful model to enhance the IDS' performance in term of accuracy. Finally, it selects the appropriate feature reduction techniques to reduce the computational cost of the algorithm.

The objectives of this paper are:

- Provide a comprehensive analysis on the use of AI on intrusion detection and discussed their strengths and weaknesses.
- Select the most recent and appropriate network traffic datasets, UNSW-NB15, to evaluate the proposed model for binary classification.
- Select feature reduction techniques to reduce the dimensions which help reducing the cost.
- Propose a hybrid classifier for intrusion detection by combining Decision Tree (DT) and Linear Regression (LR) classifiers with AdaBoost technique as meta-classifier to build a powerful model for detecting cyber-attacks.
- Evaluate the model and shows its effectiveness in terms of accuracy and computational cost compared to others classifiers.

2. Related Work

Various research studies have explored Artificial Intelligence approaches in the area of Intrusion Detection Systems (IDS) to detect and efficiently classify attacks, improve defense methods and thus enhance cybersecurity. This section focusses on the proposed Machine Learning based approaches for IDS, and thus, discusses the related studies and shows their strengths and weaknesses.

In the work in [11] , authors provided an Artificial Intelligence based-typology that helps organizations that helps organizations especially managers to understand the effect of AI on their industries. In [5] , authors studied the impact of different Artificial Intelligence technologies on enhancing cyber security and its main role to improve cyber security by facing cyber-attacks, specifically in Iraq. The study is limited to a relatively small geographical area and a small sample size as well.

A solution for cybersecurity problems such as the problem of intrusion detection has been proposed in the study [6]. Authors used Naive Bayes classifier to solve the problems. Authors have also discussed some cybersecurity problems studied how artificial intelligence approaches and data mining techniques can solve them.

The study [12] discusses the application of artificial intelligence and data mining techniques in cybersecurity to improve cyber security in three major fields including intrusions detection, malware examination and spam detection. The weakness of this study is that it does not depend on experience or accurate analysis. Instead, it depends entirely on the results of previous studies.

Table 1 summarizes the related studies that have investigated the use of Artificial Intelligence techniques for Intrusion Detection. The studies are presented based on chronological order.

Table 1. Comparative Study of the Presented Studies

Study	Year	Domain	Description	Result	Limitation
[9]	2021	Artificial Intelligence and Cyber Security	Describe AI Applications and Techniques in cybersecurity	Machine learning, deep learning, and data mining are used to improve cyber security in three fields: intrusions detection, malware examination, and spam detection	The study's outcome isn't quantified in any way.
[2]	2021	Artificial Intelligence and Cyber Security	study the effect of AI techniques on cyber security and its main role to improve cyber security by facing cyber-attacks, specifically in Iraq	AI has a significant and positive impact on cybersecurity to enhance it. Except for the expert system, it had no effect	small Geographical area, limited variables, small sample size.
[8]	2020	Artificial intelligence and industry	Provide a typology of AI-enabled innovations that helps managers to understand the effect of AI on their industries	determining the impact of AI on organizations using the proposed topology	-

[3]	2019	Machine Learning algorithms and Cyber Security	Study How can use machine learning algorithms in cyber security to enhance it	Naive Bayes and multilayer Perceptron (MLP) give good results for intrusion detection. For IP classification, intrusion detection system and anomaly detection, Bayes Net is used. Clustering techniques are used to detect the malware and detect attacks and signatures in real-time	The study's outcome isn't quantified in any way.

Another study [11] has proposed an intrusion detection model based on the principle of Adaptive Boosting. Authors have used 5 weak classifiers, namely KNN, C4.5, MLP, SVM and LDA. The dataset UNSW-NB15 was used to evaluate the model. The results showed a good accuracy but the computational cost is completely neglected.

Authors in [13] proposed a cloud intrusion detection system based on Deep Neural Networks (DNNs) to enhance the accuracy. To measure the effectiveness of the classifier for both binary and multiclass classification, the dataset CSE-CIC-IDS2018 was selected. The study results have shown that the models achieved good accuracy, it was 98.97% for binary classification and 98.41% for multi-class.

In [14], authors have proposed a hybrid classifier model instead of an individual classifier to enhance the classification process using deep learning algorithms. They relied on the principle of stacked generalization to build the hybrid classifier and used the DNN and LSTM models as basic classifiers and the LR model as a meta-classifier. The proposed model appears to be effective, but in fact, the researchers ignored the computational cost. Deep learning algorithms are more complex than machine learning algorithms, so they consume more computational costs. One of the advantages of the paper is that they evaluated the proposed model on 3 datasets, which are IoT-23, LITNET-2020 and NetML-2020.

Authors in [15] have used the layering principle to build an effective intrusion detection model. They proposed a model consisting of two layers. The first layer classifies the data set into two classes (cyber-attack / normal). The second layer classifies the cyber-attack class into several sub-classes as well. The study results have shown that the proposed model reached 95% of accuracy.

In contrast to previous studies, the study [16] proposes a simple model. Despite the simplicity of the model, its performance is good. The model was based on the decision tree algorithm, in addition to selecting the characteristics and arranging them according to the most important. The study results have shown that the proposed model reached 96.7% of accuracy.

In [10], the researchers have evaluated the most popular and used algorithms in recent studies to build intrusion detection models, which are 12 algorithms using the same environment and the same datasets (CICIDS-2017, UNSW-NB15, ICS). The study results have shown that machine learning algorithms performed better results than deep learning algorithms because deep learning algorithms need huge data sets to train. Also, authors have found that Decision Tree (DT) and Random Forest (RF) gave the best performance results compared to Naive Bayes.

Authors of the study [17] have used dimension reduction techniques in building intrusion detection models and compares the two most popular techniques: PCA and SVD. The study applied

dimensionality reduction techniques with 7 different algorithms to demonstrate their effectiveness in increasing classification accuracy and reducing the computational cost. It was found that the use of decision tree with SVD technique for dimension reduction saved computational cost a lot while maintaining classification accuracy.

When building an intrusion detection model, the study claims that the stage of the pre-processing dataset, regardless of the algorithm used for classification, has a significant impact on the result [18] . In this study, authors have conducted two experiments: the first one without pre-processing the dataset. The second experiment was with care in the pre-processing stage in addition to reducing the dimensions. In both experiments, 3 different classification algorithms were used. They found that classification accuracy increased in each of the three classifiers after pre-processing and dimensionality reduction.

Table 2 discusses the studies related to intrusion detection system. As shown in the table, studies are compared in terms of dataset used, pre-processing techniques, basic algorithms used for classification and performance accuracy.

Table 2. A Summary of Studies that Suggested A Classification Approach

Study	**Dataset**	**Pre-processing Technique**	**Machine Learning classifier**	**Computational cost (Yes/No)**	**Accuracy**
[10]	CSE-CIC-IDS-2018	One Hot Encoder Standard Scaler	MLP-BP MLP-PSO	No	98.97% (binary classification) 98.41% (multi-class)
[6]	UNB-CIC	CFS	Adaboost.M1	No	99.98
[11]	IoT-23	DSAE	LR	No	99.7
	LITNET-2020				100
	NetML-2020				100
[12]	Power System Datasets	MDI	RF	No	95.44
[13]	UNSW-NB 15	Gini index	DT	No	96.7
[7]	CICIDS-2017	_	Set of Machine Learning techniques	No	ML performed better results than Deep Learning
	UNSW-NB15		Deep Learning		
	ICS				
[14]	NSL-KDD	PCA	LR	Yes	95
			KNN		99
			SVM		95
			NB		90
			DT		99
			AdaBoost		97
			RF		99
[15]	KDD Cup99	CFS with PSO	KNN	No	99.8
			SVM		99.9
			NB		91.4
	Kyoto 20062		KNN		99.7
			SVM		99.7
			NB		99.1
	UNSWNB15		KNN		92.8
			SVM		92.2
			NB		84.7

Based on the comparative analysis shown in Table 2, we observed that the most used datasets for evaluating models is UNSWNB15 dataset, it gives the highest performance for a model with

general network traffic. In the case of a hybrid classifier, using the AdaBoost method gives a high accuracy result.

The majority of studies have ignored the computational cost. Only the study [17] has interested to enhance the classifier's accuracy and minimize the computational cost by combining the dimension reduction technique with different machine learning classifiers.

Therefore, this study proposes a hybrid classifier for intrusion detection system. It combines Decision Tree (DT) and Linear Regression (LR) techniques with AdaBoost to build a powerful model for detecting cyber-attacks. PCA method was used to reduce the dimensions, and thus, to minimize the computational cost, while maintaining the high performance of the model.

3. Methodology

This section discusses the followed methodology for Intrusion detection systems (see Figure 1). It includes five main phases: dataset selection, data preprocessing, dimension reduction, classification and evaluation.

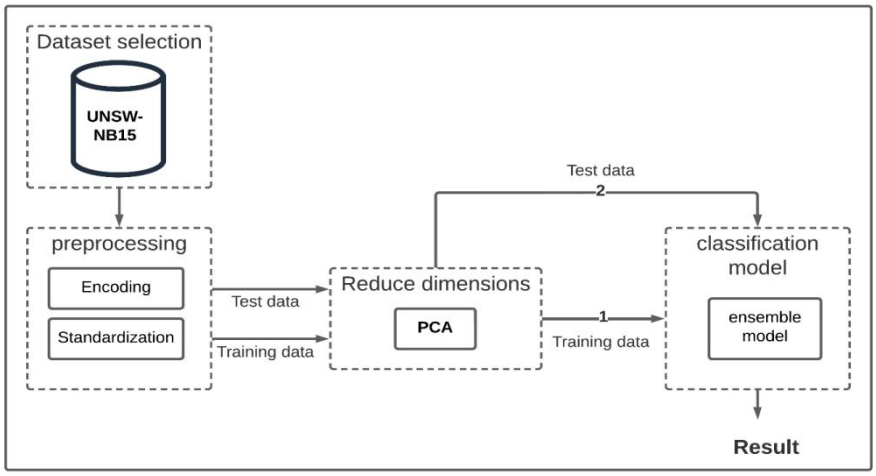

Figure 1. The Overall Methodology of Proposed Model

The first phase is the dataset selection. For this study, the dataset UNSW-NB15was selected datasets to evaluate the proposed model. The second phase is data pre-processing, which includes Feature Encoding and Feature Standardization. Before moving to the third phase, the dataset was divided into two sets: training data and test data in a ratio of 80:20. The third phase is dimensions reduction using Principal Component Analysis. The fourth phase is the classification model. The last phase is to evaluate the classifier in terms of accuracy. Also, the runtime is used to evaluate the computational cost. The following subsections describe the methodology phases in detail.

3.1 Dataset Selection

Various datasets are available to the public such as KDD CUP 99, NSL-KDD, power system ICS cyber-attack dataset and UNSW-NB15 [19]. This study selects the most used dataset to evaluate the performance of the classifier. Looking at KDD CUP 99, it is considered one of the most used in this filed, but the study indicates that it is very old as it was generated in 1998, outdated, does not represent the new network structures and can only perceive a limited number of attacks. For the same reasons explained previously in KDD CUP 99, NSL-KDD was not selected. Compared to UNSW-NB15, power system ICS cyber-attack dataset has a higher attack rate than normal behavior instances. The study [10] shows that the effectiveness of machine learning techniques on the general-purpose network dataset such as UNSW-NB15 is better than its performance on the ICS cyber-attack dataset. Therefore, the dataset UNSW-NB15 was selected.

The dataset UNSW-NB15 is a publicly available dataset which includes modern real activities and compound attack activities. It was generated in 2015 with a duration of 31 hours. UNSW-NB15 contains 49 features, which include 2 labelled features [20].

3.2 Data Preprocessing

It consists in preparing the data before its use. Pre-processing is the process of converting data into more organized and understandable data. This process increases the classification accuracy and effectiveness. Two main stages are needed for this process: Feature Encoding and Feature Standardization.

Feature Encoding: machine learning algorithm only deals with numbers, so some features that contain letters or symbols must be converted into numbers to deal with the classifiers easily. The features needed to be encrypted are srcip, dstip, proto, state, attack_cat and service.

Feature Standardization: it places all variables in a specific range to facilitate comparison and classification. This is done in one of two ways: Standardization or Normalization. Normalization means shifting and re-scaling values so that they are in the range between 0 and 1. This study adopts the standardization technique for feature scaling. According to the study [15] , the use of the Standardization method gives higher accuracy of the model. Equation (1) shows the method for calculating Feature Standardization:

$$X' = x - \text{mean}(x)/\sigma \quad (1)$$

Where;

- $\boldsymbol{x}$: is the feature values in the dataset.
- $\boldsymbol{mean(x)}$: is the mean of the feature values.
- $\boldsymbol{\sigma}$: is the standard deviation.

3.3 Dimension Reduction

This study aims to build a classifier with high accuracy and low computational cost. So, it is crucial to reduce the dimensions in order to improve the computational cost [21]. PCA was used to reduce the size of the dataset. The statistical method was used for dimensionality reduction in high dimensional data without losing any important information [22]. It converts data from n-dimensions to K-dimensions (n > k).

Principal Components are the dimensions in which there is the most variance and the data is the most scattered. PCA is a linear transformation that places the dataset in new coordinates by finding the most significant variance in the first coordinate. Then the coordinates are formed perpendicular to the last and have less variance. Using Principal Components, the complexity of the original dataset can be approximated. All Principal Components are called Eigenvectors and have Eigenvalues, which mean the amount of variance in the data in that vector. Therefore, the Eigenvectors with the highest Eigenvalues are the first Principal Component [23].

3.4 Classification

To build a classifier from several classifiers (machine learning algorithms), we need a way to combine these different classifiers. Several methods exist to ensemble different classifiers [24]. Two main methods are used; averaging and boosting methods [25].

Averaging methods depend on the principle of training each classifier independently and then averaging their predictions.

Boosting methods are based on the principle of training all classifiers in a sequential manner and reducing bias in the final classification. It aims to build a powerful model from several weak models.

In this study, the AdaBoost method was adapted and it belongs to the boosting methods. Two machine learning algorithms have been presented for use as weak classifiers in the proposed model: Linear Regression (LR) and Decision Tree (DT). They were selected based on two scales, namely accuracy and runtime, according to the study of [10] and [17]. The AdaBoost method is based on the principle of constructing a strong classifier from several weak classifiers [26] . It works in a sequential manner. The first classifier is trained, then the second. Then it combines the predictions of all the classifiers by voting to get the final prediction.

All samples in the original dataset (training data) are given weights. The Initial weight of each example is 1/n. After training the first algorithm on the original dataset, all the weights of the examples are individually modified so that the weights of the incorrectly classified examples

increase. To make it easier for the following classifier to focus on this high-weight (difficult to classify) examples. We train the second algorithm on the updated weights dataset. The sequential training process continues until it reaches the second and final classifier in the proposed model. Then, we combine the classifier' predictions with the voting method to produce the final prediction. The proposal model is shown in Figure 2.

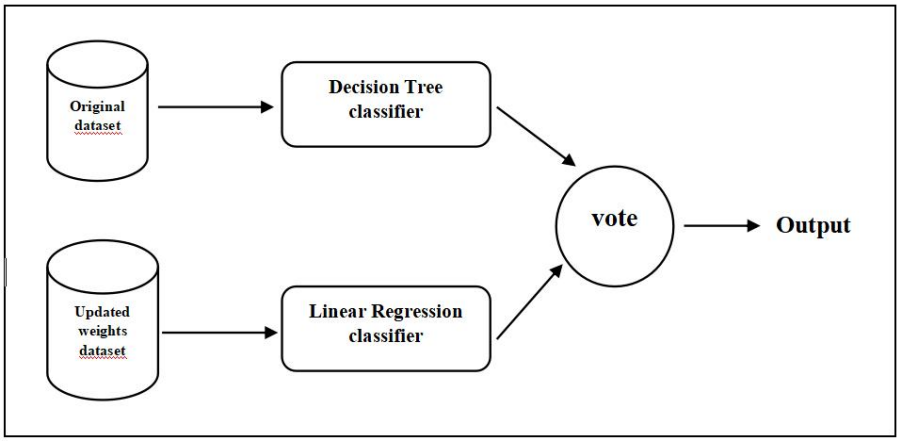

Figure 2. The Proposed Model

3.5 Model Evaluation

In this phase, we evaluate the performance of the model using different measures: Confusion Matrix, Accuracy, Precision, Recall, and F1 Score [27]. To measure the computational cost of the classifier, the run time was calculated.

3.5.1 Confusion Matrix

A confusion Matrix is not a measure in itself, but all performance measures depend on it and its components. The Confusion Matrix is simply a table with two dimensions. The columns in the table represent the actual classifications, and the rows in the table represent predictions. The table includes Four classes in both dimensions. These four terms represent the components of confusion matrices on which all performance measures depend (see Table 3).

Table 3.Components of Confusion Matrices

		Actual	
		Positive	**Negative**
Predicted	**Positive**	TP	FP
	Negative	FN	TN

3.5.2 Accuracy Evaluation

This section describes the different measures used to evaluate the classifier. The measures include: accuracy, precision, recall and F1-score which represent the classifiers output with desirable representation for purposes.

The accuracy is the number of adjust predictions for the model partitioned by all predictions. It is calculated by the following formula (see Equation (2)):

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{Total of instances}} \qquad (2)$$

The precision is the number of correct positive expectations of the model partitioned by the positive expectations. It is defined as follows (see Equation (3)):

$$P = \frac{TP}{TP + FP} \quad (3)$$

The recall is defined by the number of adjust positive expectations of the model partitioned by all positive examples in the dataset. It is defined as follows (see Equation (4)):

$$R = \frac{TP}{TP + FP} \quad (4)$$

The F-score calculates the adjust between Precision and Recall. It considers false positives and false negatives. It is defined as follows (see Equation (5)):

$$F1 = 2 * \frac{P * R}{P + R} \quad (5)$$

4. Results and Discussion

In this section, we review the implementation details of the model. The implementation includes 5 stages which are: (i) reviewing and analyzing the dataset, (ii) pre-processing the dataset, (iii) applying PCA to reduce dimensions, (iv) building the model and (v) testing and evaluating the classifier.

4.1 Implementation Details

This study was implemented using the Jupyter Notebook platform, which is a platform that provides a working environment for the implementation and development of artificial intelligence applications. The programming language used to write code is Python, and using some machine learning libraries such as Scikit-learn and some other libraries to modify and manipulate raw data such as Pandas and NumPy libraries. The working environment used to implement the model is a Windows 10 PC with a Core i7 processor at 2.39 GHz and 8 GB RAM.

4.2 Implementation Steps

4.2.1 Dataset Exploration

This study uses the dataset UNSW-NB15. The dataset records were divided into four CSV files named UNSW-NB15_1, UNSW-NB15_2, UNSW-NB15_3 and UNSW-NB15_4. The size of the dataset is 2.540.044 records. The dataset contains 49 features, including the attack_cat feature to specify class label, it has binary values 0 for normal and 1 for attack.

The research study aims to minimize the features to 35. It deletes some unnecessary features; features that have a high correlation and the attack_cat feature, because the study is based on binary classification. It is, also, important to deal with null values before the pre-processing stage, as they affect the performance of the classifier and give unrealistic results. Several methods exist to deal with null values, the most popular methods are: (i) delete instances of null values, (ii) replace instances with the average value, (iii) replace instances with the majority value and (iv) replace instances with zero. According to the study [15], replacing the null values with zero is the best method which gives a higher accuracy for the performance of the classifier [15]. Table 4 shows a description of the used dataset UNSW-NB15.

Table 4. Dataset Description

Dataset	Size	# of features before preprocessing	# of features after preprocessing
UNSW-NB15	2.540,044	49	35

4.2.2 Data Preprocessing

It consists of making data tidy, clear, understandable and usable. The field of machine learning needs to change raw data into data that the machine can understand and deal with it.

For this study, the preprocessing included two stages: (i) feature encoding, (ii) feature balancing.

Feature Encoding means converting feature values that contain words into numbers to make it easier for machine learning algorithms to handle. While Feature Balancing means making all feature values in the same range.

The two most popular Feature Encoding methods are One Hot Encoding and Label Encoding [12].

- One Hot Encoding assigns to each category a vector of size n, when n defines the number of groups in the feature, containing 1 to denote the presence of the feature and the rest 0.
- Label Encoding assigns a numerical label to each category. If the number of categories in the feature is n, then the numerical label will start from 0 to N-1.

Figure 3 shows the difference between One Hot Encoding and Label Encoding methods, in an example of state feature encoding.

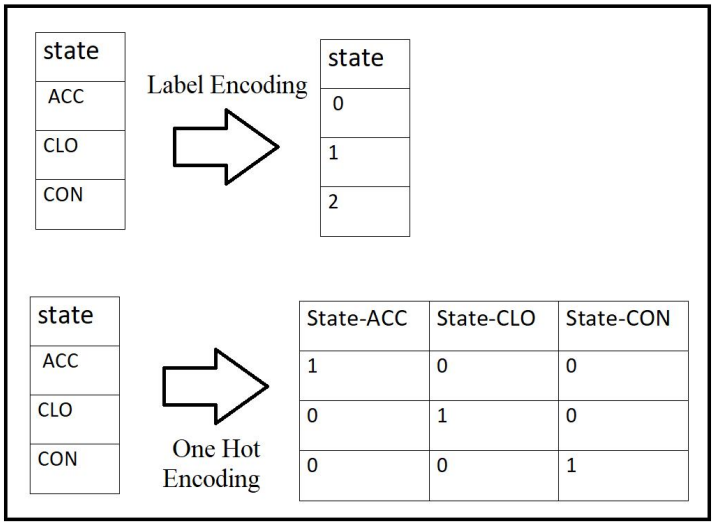

Figure 3. Difference between Label Encoding and One Hot Encoding

As it becomes clear, encoding using One Hot Encoding will increase the features, thus slow down the classification process and raise the computational cost of the proposed model. Therefore, Label Encoding was used in our study. In the dataset, there are 3 features that need to be encoded: proto, state and service. The encoding process was carried out using the Label Encoder from the scikit-learn library.

The second stage is feature scaling, which places features in the same distribution or same range. All features must be in the same range to give accurate results for the model.

The two most popular methods to feature scaling are normalization and standardization [15]:

- Standardization means rescaling all features so that it has a mean of 0 and a standard deviation of 1, as defined in equation (1).
- Normalization means rescaling all features to be in a specified range, usually between 0 and 1, as defined in equation (1).

In this study, the features were scaled using the standardization approach because it gives higher accurate results in binary classification. The StandardScaler class is used to rescale the features in the dataset.

After the pre-processing stage, we divide the dataset into a training set and a testing set in a ratio of 20:80. The dataset was divided before applying PCA to give more accurate and realistic results for the performance of the proposed classifier.

4.2.3 Principal Component Analysis (PCA)

It analyzes data that contains descriptions of objects through several dependent and quantitatively correlated variables. As mention in [23] , this method aims to: (i) extract the significant data, (ii) minimize the dimensions, (iii) simplify the data in a descriptive way and (iv) describe the structure of the objects and attributes of the dataset.

This study uses PCA method to minimize the size of the dataset and thus minimize the cost of the classifier. It divides the dataset into a training set and a testing set, before reducing the dimensions, to avoid getting false and inaccurate results.

4.2.4 Model Building

To build the intrusion detection's classifier with high performance and low computational cost, we are based on the meta_classifier AdaBoosting, the Decision Tree (DT) and Linear Regression (LR) as basic classifiers. The concept of AdaBoosting helps us to build an effective model from several others. It adopts a successive training method to take the advantage of each weak classifier. The basic classifiers of the proposed model are DT and LR, which are simple algorithms, therefore, they do not need a high computational cost, but they have a relatively good performance at the same time, so they were used as basic classifiers. Figure 4 shows the model's framework.

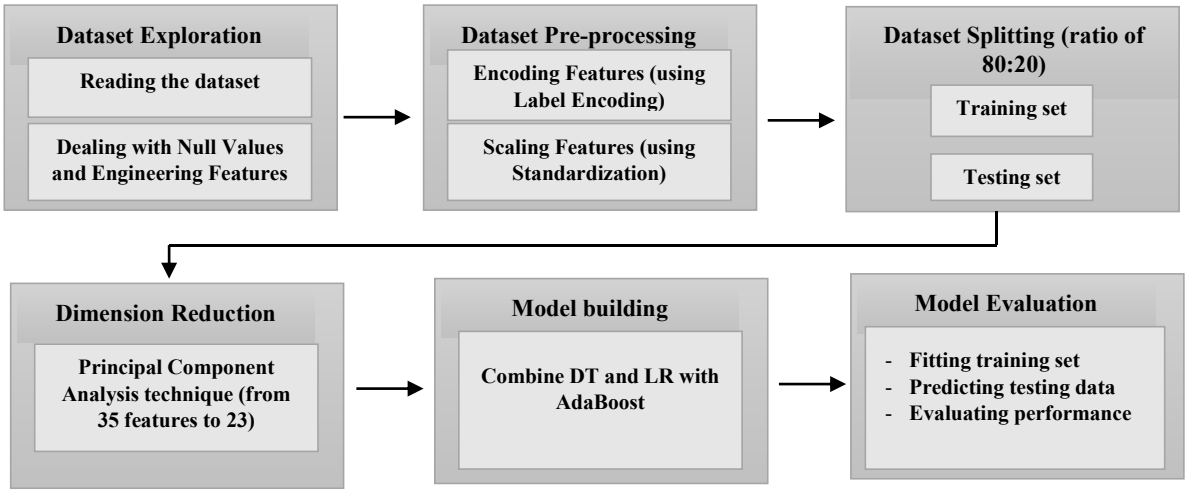

Figure 4. The Model's Framework

4.2.5 Results

We evaluate the effectiveness of the model compared to some of the discussed models in the related works. It was compared individually with three classifiers such as Decision Tree (DT), Support Vector Machine (SVM) and Logistic Regression (LR).

This study opts a binary classification 0 for normal and 1 for attack. The experiment included several stages, namely exploring the dataset to clean it, engineering the features and deleting the high correlation and unnecessary features of the classification process, pre-processing the dataset (Encoding and Scaling the features), and finally building the classifier. Decision Tree (DT) and Linear Regression (LR) were the basic classifiers and AdaBoosting was a meta-classifier.

To evaluate the model for binary classification, the study uses six metrics which are confusion matrix, accuracy, precision, recall, F1-score and run time.

Confusion matrix: the output is presented as true positives, true negatives, false positives and false negatives. Figure 5 shows the results of confusion matrix using PCA method to reduce the dimensions.

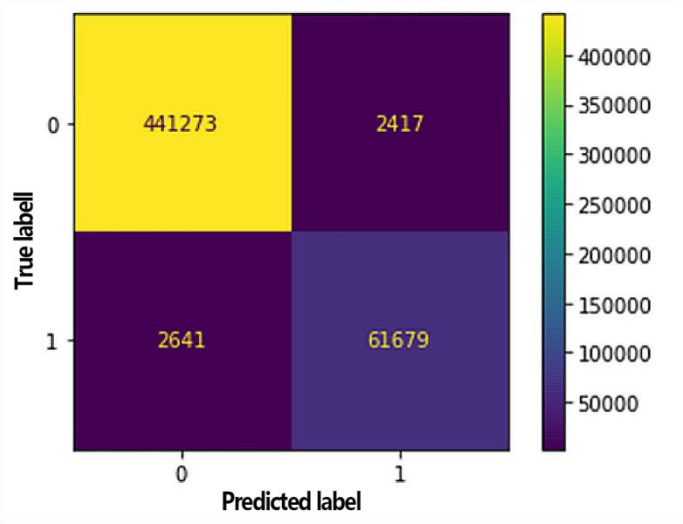

Figure 5. Confusion Matrix with PCA Method

Accuracy evaluation: various measures were used to evaluate the model. These parameters represent the classifiers output with desirable representation for different purposes. Table 5 shows the results of the binary classification.

Table 5. Classification Report

Class	Precision	Recall	F1-score	Support
Normal	0.99	0.99	0.99	443690
Attack	0.96	0.96	0.96	64320
macro avg	0.98	0.98	0.98	508010
weighted avg	0.99	0.99	0.99	508010
Accuracy: 0.99004				

To make sure that the PCA method achieves the desired goal, the same experiment was applied, but without applying the PCA principle. Figure 6 shows the results of confusion matrix and Table 6 shows the obtained results without applying the PCA principle to the dataset.

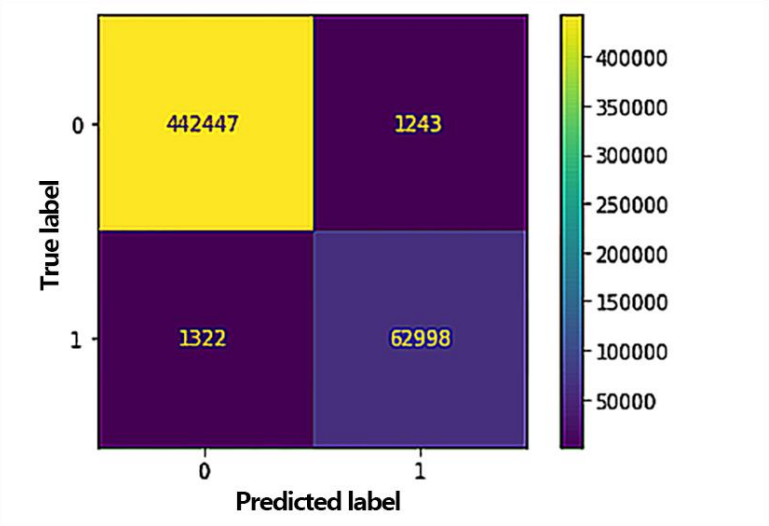

Figure 6. Confusion matrix without PCA method

Table 6. Classification Report of the Model

Model	Accuracy	Precision	Recall	F1-score	Run time (ms)
with PCA	99%	98%	98%	98%	353

without PCA	99.4%	99%	99%	99%	686

Since the classifier is based on Decision Tree and Linear Regression, two additional experiments were made: First, each classifier is evaluated separately. Second, the model is compared with one of the most used algorithms in this field, which is the Support Vector Machine (SVM) [17], [18]. Table 7 shows the results using the same processed dataset, the same preprocessing steps and the PCA method to reduce the dimensions.

Table 7. Performance scores for Each Classifier

Classifier	accuracy	precision	recall	f1-score
LR	98.2%	95%	97%	96%
DT	98.6%	96%	99%	97%
SVM	98.3%	95%	98%	96%

Figure 7 shows the performance of each classifier compared to the others classifiers (DT, LR and SVM).

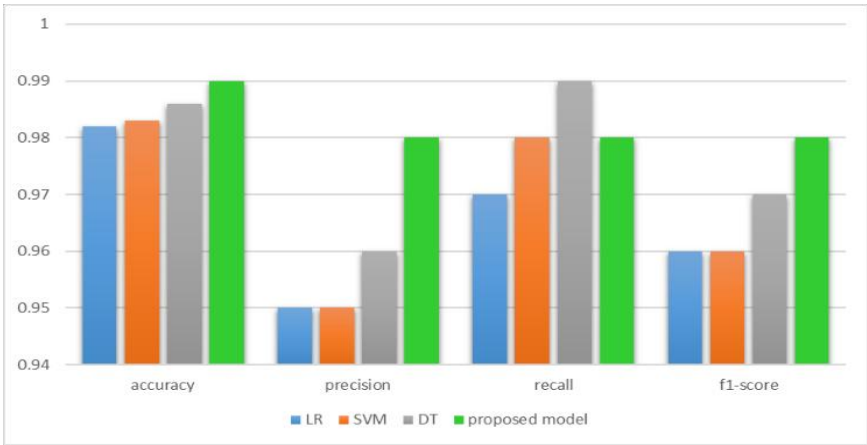

Figure 7. Comparison of the Performance Scores for the Classifiers

According to the results above, we observed that the result of the experiment was very interested, as the accuracy reached 99%.

In the proposed model, the PCA method was used to minimize the dimensions in order to minimize cost, as it reduced the number of features from 35 features to 23 features. The model succeeded in reducing the computational cost while maintaining the high performance of the model, as the runtime was reduced by half after applying PCA, while the accuracy maintained the same percentage, which is 99% (see Table 8).

The comparison was carried out in the same environment, the same selected dataset and with the same pre-processing steps. All the classifiers in general gave relatively good results and this is due to the preprocessing steps followed in the proposed model. Compared to DT, LR and SVM classifiers, the performance of our model is the best. Removal of unnecessary features and high correlation features in the feature engineering phase had an effective effect on the classification process as it reduced complexity. Also, encoding the features using the label encoding method preserved the features' number and did not affect the complexity. Choosing a standardization approach in the feature balancing process improved the model's performance.

In [9], the model showed an accuracy of 96.7% and used the same selected dataset, while our model achieved good accuracy of 99%. The hybrid model has a very high performance and low computational cost, unlike models that have a very high complexity, where a large number of algorithms are used as base classifiers and completely ignore the computational cost.

The techniques used for feature engineering, encoding and scaling features made a big difference in simplifying the classification process, and thus the model's performance results.

Therefore, this must be considered and the best techniques should be chosen. Therefore, to obtain the best performance of the model, the dataset must be processed well.

5. Conclusion and Future Works

This paper studied the integration of artificial intelligence (AI) into the cybersecurity infrastructure. AI was used for many tasks in cyber security such as; Network protection, Endpoint protection, Application security and Suspect user behaviour. In addition, many applications based on AI exist in cybersecurity, the most important are: network intrusion detection and prevention, fraud detection, etc.

This study proposed a hybrid model to enhance the accuracy and minimize the cost for detecting cyber-attacks. It combines Decision Tree and Linear Regression classifiers using AdaBoost as a meta-classifier. For model' evaluation, we used the dataset UNSW-NB15, the latest and most realistic datasets for network traffic. To reduce the cost, the method PCA is applied and succeeded in reducing the run time by half while maintaining the accuracy of the model. The model was evaluated and compared with the classifiers DT, LR and SVM using the metrics confusion matrix, accuracy, precision, recall, F-score, and runtime. The study results have shown that the performance of the proposed model is better. In addition, the quality of the pre-processing and feature engineering stages using optimum technologies in order to simplify the dataset helps in improving the classification process, thus improving the model effectiveness.

In the future, we aim to evaluate the performance of the model with other datasets such as KDD CUP 99, NSL-KDD and power system ICS cyber-attack. To more reduce the computational cost of the model, AdaBoost can be replaced by voting as a combination method between DT and LR. The phases of pre-processing, feature engineering and dimensionality reduction achieved the desired results. Therefore, it can be applied with other hybrid models and also with simple classifiers to increase accuracy and reduce computational cost.

6. Acknowledgement

The authors gratefully acknowledge Qassim University, represented by the Deanship of "Scientific Research, on the financial support for this research under the number (COC-2022-1-2-J-30588) during the academic year 1444 AH / 2022 AD".

References

[1] Packetlabs. "239 Cybersecurity statistics (2023)" packetlabs.net. https://www.packetlabs.net/posts/239-cybersecurity-statistics-2023/ (accessed Feb. 15, 2023).

[2] Internet Crime Complaint Center. "Federal bureau of innvestigation: Internet crime report," 2022. [Online]. Available: https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf

[3] M. McLean. "2023 Must-Know cyber attack statistics and trends" embroker.com. https://www.embroker.com/blog/cyber-attack-statistics/ (accessed Feb. 15, 2023).

[4] D. Coss and S. Samonas, "The CIA Strikes Back: Redefining Confidentiality, Integrity and Availability in Security," *Journal of Information System Security*, vol. 10, no. 3, pp. 21-45, 2014.

[5] B. Alhayani, H. Jasim Mohammed, I. Zeghaiton Chaloob, and J. Saleh Ahmed, " WITHDRAWN: Effectiveness of artificial intelligence techniques against cyber security risks apply of IT industry," *Mater Today Proc*,Mar. 2021, doi: 10.1016/j.matpr.2021.02.531.

[6] S. Soni and B. Bhushan, " Use of Machine Learning algorithms for designing efficient cyber security solutions," *in 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies*, (ICICICT), Jul.2019, doi: 10.1109/ICICICT46008.2019.8993253.

[7] S. Zeadally, E. Adi, Z. Baig, and I. A. Khan, "Harnessing artificial intelligence capabilities to improve cybersecurity," *IEEE Access*, vol. 8, pp.1-1, 2020.

[8] Y. Xin et al., "Machine Learning and Deep Learning Methods for Cybersecurity," *IEEE Access*, vol. 6, 2018.

[9] P. Sornsuwit and S. Jaiyen, "A New Hybrid Machine Learning for Cybersecurity Threat Detection Based on Adaptive Boosting," *Applied Artificial Intelligence*, vol. 33, no. 5, pp. 462-482, Mar. 2019.

[10] N. Elmrabit, F. Zhou, F. Li, and H. Zhou, "Evaluation of Machine Learning Algorithms for Anomaly Detection," *IEEE Xplore*, 2020.

[11] U. Paschen, C. Pitt and J. Kietzmann, " Artificial intelligence: Building blocks and an innovation typology," *Bus Horiz*, vol. 63, no. 2, 2020.

[12] K. Ramasubramanian and S. Yerram, "Applications and Techniques of Artificial Intelligence in Cyber Security," 2021.

[13] S. Alzughaibi and S. El Khediri, "A Cloud Intrusion Detection Systems Based on DNN Using Backpropagation and PSO on the CSE-CIC-IDS2018 Dataset," *Applied Sciences* , vol. 13, no. 4, Jan 2023.

[14] V. Dutta, M. Choraś, M. Pawlicki, and R. Kozik, "A deep learning ensemble for network anomaly and cyber-attack detection," *Sensors* , vol. 20, no. 16, pp.4583, Aug 2020.

[15] Y. A. Farrukh, Z. Ahmad, I. Khan, and R. M. Elavarasan, "A Sequential Supervised Machine Learning Approach for Cyber Attack Detection in a Smart Grid System," *IEEE Xplore,* Nov. 01, 20210.

[16] M. Al-Omari, M. Rawashdeh, F. Qutaishat, M. Alshira'H, and N. Ababneh, "An Intelligent Tree-Based Intrusion Detection Model for Cyber Security," *Journal of Network and Systems Management*, vol. 29, no. 2, 2021.

[17] H. Zhou, G. Yang, Y. Xu, and W. Wang, "Effective matrix factorization for recommendation with local differential privacy, " *Lecture Notes in Computer Science*, pp. 235-249, Jan. 2019.

[18] T. Ahmad and M. N. Aziz, "Data preprocessing and feature selection for machine learning intrusion detection systems," *ICIC Express Letters,* vol. 13, no. 2, 2019.

[19] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," *Computers and Security*, vol. 86. 2019.

[20] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," *IEEE Xplore*, Nov. 01, 2015.

[21] L. J. P. Van Der Maaten, E. O. Postma, and H. J. Van Den Herik, "Dimensionality Reduction: A Comparative Review," *Journal of Machine Learning Research*, vol. 10, 2009.

[22] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, "Dimensionality Reduction with Principal Component Analysis," *in Mathematics for Machine Learning*, 2020.

[23] F. L. Gewers et al., "Principal component analysis: A natural approach to data exploration," *ACM Comput Surv*, vol. 54, no. 4, 2021.

[24] H. Rajadurai and U. D. Gandhi, "A stacked ensemble learning model for intrusion detection in wireless network," *Neural Comput Appl*, vol. 34, no. 18, 2022.

[25] M. Mohandes, M. Deriche, and S. O. Aliyu, " Classifiers Combination Techniques: A Comprehensive Review," *IEEE Access*, vol. 6. 2018.

[26] Y. CAO, Q.-G. MIAO, J.-C. LIU, and L. GAO, " Advance and Prospects of AdaBoost Algorithm," *Acta Automatica Sinica*, vol. 39, no. 6, 2013.

[27] H. M and S. M. N, "A Review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process,* vol. 5, no. 2, 2015.