© Universiti Tun Hussein Onn Malaysia Publisher's Office

JSCDM

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/jscdm

Journal of Soft Computing and Data Mining

Evaluation of Classification Algorithms for Intrusion Detection System: A Review

Azar Abid Salih[1]*, Adnan Mohsin Abdulazeez[2]

[1]Department Information Technology Management,
Duhok Polytechnic University, Duhok, Kurdistan Region, IRAQ

[2]Duhok Polytechnic University,
Duhok, Kurdistan Region, IRAQ

*Corresponding Author

DOI: https://doi.org/10.30880/jscdm.2021.02.01.004
Received 19 January 2021; Accepted 28 March 2021; Available online 15 April 2021

Abstract: Intrusion detection is one of the most critical network security problems in the technology world. Machine learning techniques are being implemented to improve the Intrusion Detection System (IDS). In order to enhance the performance of IDS, different classification algorithms are applied to detect various types of attacks. Choosing a suitable classification algorithm for building IDS is not an easy task. The best method is to test the performance of the different classification algorithms. This paper aims to present the result of evaluating different classification algorithms to build an IDS model in terms of confusion matrix, accuracy, recall, precision, f-score, specificity and sensitivity. Nevertheless, most researchers have focused on the confusion matrix and accuracy metric as measurements of classification performance. It also provides a detailed comparison with the dataset, data preprocessing, number of features selected, feature selection technique, classification algorithms, and evaluation performance of algorithms described in the intrusion detection system.

Keywords: Classification algorithm, confusion matrix, intrusion detection, feature selection, dimension reduction, data preprocessing

1. Introduction

Nowadays, intrusion detection system has gained an essential role in computer and network security. IDS monitoring and analyzing network traffic is used to classify different types of attacks [1]. The network traffic action consists of many features collected in the form of a dataset to detect different types of attacks [2]. The increase of the massive amount of data being generated daily via the internet has caused the world of technology to face a big challenge [3]. Datasets represent instances that consist of several features and are related to the intrusion detection system [4]. So, it is essential to realize the type of data containing different types of attacks and features [5]. The most popular data set that is being used for the intrusion detection system is a KDD'99 cup to develop predictive models for distinguishing the relationship between intrusions or several attacks [6]. The intrusion detection system builds the model based on security data sets such as KDD99 and NSL-KDD [7]. It contains different types of features similar to a predictor to distinguish the normal attacks from the abnormal ones as a features target [8]. The classification model splits the data set into stage training and testing [9]. The massive number of features with high dimensions leads to complexity in the training phase and wastes time. Therefore, it needs to select some useful and relevant features from the whole range of features to improve the performance of the model in the testing phase [10]. The critical stage to improve a classification model's quality is data preprocessing machine learning algorithms [11]. It is such a crucial step to solving numerous types of big data sets [12].

*Corresponding author: azar.abid@dpu.edu.krd

Machine Learning (ML) techniques widely used in computer security data sets have recently become a trend in security technology [13]. It contributes to analyses and handling the massive amount of data and extracts the essential features that are used in various techniques for feature selection [14]. IDS is a commonly used machine learning classifier to distinguish between various attacks as a class [15]. Many supervised classification algorithms are applied to IDS, such as Decision Trees, Naïve Bayes, K-Nearest Neighbor, Tree C4.5, Random Forest, Support Vector Machine, and Logistic Regression [16]. Evaluation of classification algorithms depends on various statistical metrics, especially confusion matrix results, to classify and predict different types of threats [17].

The rest of the paper is organized as follows: Section 2 classification model, Section 3 effective of dimensionality reduction for feature selection, Section 4 performance evaluation appropriate metrics, Other, sections are related works reviewed and compared with discussion finally, and conclusion.

2. Classification Model

Classification is one of the machine learning tasks. It is a supervised learning model. It is used for intrusion detection systems based on binary or multi classes [18]. In supervised learning, data is always labeled, which takes each record in a dataset assigned to a particular class [19]. A classification model-based IDS classifies all the network traffic into either normal or abnormal classifier algorithms. The obstacle to building the model is the massive amount of data [20]. Classification algorithms, facing many problems in building a model, need data preprocessing stage, especially in high data dimensionality [21]. Choosing the best classification algorithm depends on the performance evaluation metrics in terms of confusion matrix and accuracy [22].

The data classification process in the dataset includes the two stages of training and testing [23]. During the training and learning stage, a classifier is learned as a target, while during the second stage, the testing phase, the built model is used to predict the class labels for a given data [24]. It is essential to analyze each classifier's required time for both stages of the training and testing. Before applying the classifiers, preprocessing of the data helps the classification model decrease time and complexity by removing irrelevant data to improve the classifier algorithms efficiency [25]. The whole dataset's cross-validation process is divided equally into two groups for network traffic dataset classification; one group for testing, and the rest will be used for the training model [26], [27]. Few algorithms are capable of distinguishing among the different attacks and normal ones with sufficient results. The most popular classifiers are used Decision Tree (DT), Random Forest, SVM (Support Machine Learning), KNN (K-Nearest Neighbor), Naïve Bayes, and Logistic Regression [28], [29].

3. Effectiveness of Dimensionality Reduction for Feature Selection

The feature selection process requires dimensionality reduction for the lessening of redundant and irrelevant data. Moreover, the removal of useless features enhances the accuracy of the model. Simultaneously, it speeds up the training and testing time [30]. Dealing with big data sets is a difficult and time-consuming task, especially with different categorical data types. Reducing the high dimensionality of data improves the process of feature selection. In general, many data sets are used in the IDS. Each dataset covers various kinds of features to detect and prevent different malicious attacks [31]. Hence, increasing the space of data, the computations need more complex calculations. Handling cases of the high number of features by reducing useless features by using dimension reduction techniques [32]. Feature selection and feature extraction are two main techniques to overcome high dimensionality. The feature selection requires finding a subset of relevant features of the original data set. The feature extraction reduces the data in the original high-dimensional data set space to a lower dimension space [33].

There are many techniques used for dimension reduction, such as Principal Component Analysis (PCA) as a linear method, Linear Discriminant Analysis (LDA), Generalized Discriminant Analysis (GDA), and Support Vector Machine (SVM). The advantage of dimension reduction is to save storage by decreasing data, reducing computation time, removing irrelevant features, and eliminating redundant features. Also, it helps with data visualization. However, some drawbacks could lead to the loss of some features in the data set, which causes the "curse of dimensionality" [34], [35]. Generally, the IDS feature space faces the curse of dimensionality on a large scale. The curse dimension happens when big data set contains extra dimension space that does not occur in low dimensions [36].

The quality of the building model in the classification task depends on the features selected in the data. The most crucial point in the process of feature selection is meant to overcome the curse of high dimensionality [37], [38]. This operation removes unwanted features based on the feature importance top score and uses the feature ranking, leading to increased learning algorithm performance [39], [40]. Also, this process provides the model with the removal of the redundant information and improvement in the generalization [41]. Many techniques are used for feature selection, such as Gain Ratio (GR), Symmetrical uncertainty, Chi-Square analysis, Information Gain (IG), and Practical Swarm Optimization (PSO) [42], [43].

4. Evaluation Performance Appropriate Metrics

Evaluation metrics describe the performance of the classification model. The critical point behind the classification is an evaluation metric used to understand the performance and efficiency of an algorithm [44]. Building an intelligent detection system capable of detecting various types of network intrusions, one must evaluate the performance of the model via using different evaluation metrics, then compare the results to find the best fit model [45], [46]. During the classification process, the Cross-validation method as a statistical approach is used to enhance the accuracy. Cross-validation is a process that splits data into two parts of train and test sets [47]. The number of k-folds in cross-validation is separated into k equal-sized folds [48]. After applying a training classification algorithm, implementing a model, and getting the output of classification, the last step is a validation to find how effective the model is based on several different metrics in the phase of the testing dataset [49]. Various performance metrics are used to evaluate different supervised learning algorithms, as shown in Table 1 [50]. Choosing the right metric is essential during the models' evaluation because different metrics are proposed to evaluate different problems and application models [51]. Several measurements are appropriate for a classification model, but the most commonly applied one is the confusion matrix [52], [53]. A confusion matrix is a statistical measurement used in machine learning classification algorithms performance for finding the accuracy of the model. The confusion matrix includes four measures: True Positive (TP), False Positive, True Negative (TN), and False Negative (FN). A good model result would be the one that contains zero false positives and negative. The impact of splitting data set ratio into training and testing phase affects the result of a confusion matrix [54].

Table 1 - Metrics for classification algorithms.

Formula	Description Evaluation Metrics
Accuracy = (TP+TN) /(TP+TN+FP+FN)	total correct classified over the total number of records.
Error rate =1-Accuracy	misclassification error ratio of incorrect predictions
Precision = TP / (TP+FP)	True positive that are correctly predicted from the total predicted patterns in a positive class.
Recall = TP / (TP+FN)	Positive patterns that are correctly attack classified.
Specificity= TN / (TN+FP)	Negative patterns that are correctly classified.
F-Measure=2*recall*precision ⁄ recall+ precision	This metric represents relation between recall and precision values
Sensitivity or True Positive Rate (TPR) = TP / (TP+FN)	Sensitivity: correctly classified over the total amount of abnormal network. True Positive Rate (TPR): Attacks correctly classified as predicted attacks It called (detection rate)
False Positive Rate (FPR) =FP / (FP+TN)	False positive (FP): Incorrectly classified normal as predicted attacks.
True Negative Rate (TNR) =TN/(TN+FP)	True negative (TN): Normal correctly classified as normal (false alarm)
False Negative Rate (FNR)=FN/ (FN+TP)	False negative (FN): Incorrectly classified attacks as a normal
AUC = $\int_0^1 (\frac{TP}{TP+FN} + \frac{FP}{TN+FP})$	Area Under Curve measures the performance of a binary classification. It is the area under the ROC curve.
$MAE = \frac{1}{N}\sum_{I=1}^{N} \lvert Y^{\sim} - Y \rvert$	Mean Absolute Error predictions to the actual outcome and is the average of the absolute errors
$RMSE = \sqrt{\frac{1}{N}\sum_{I=1}^{N} \lvert Y^{\sim} - Y \rVert}$	Root mean squared error measure to calculate the values predicted by a model when compared to the actual observed values.
(G-Mean) = Specificity * Sensitivity	Geometric Mean balance between normal and attacks classification accuracy
Log Loss= $\frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(p_{ij})$	Probability method used for multi-classification
Gini coefficient = 2*AUC-1	derived from the AUC ROC number.

5. Review of Classification Algorithms for IDS

During the last decade, many works have been presented to improve the IDS to detect and prevent different malicious attacks from accessing computer information. This section discusses some techniques and algorithms of machine learning in the classification process that are used for intrusion detection, including data preprocessing, feature selection techniques, number of features selected, classification algorithms, and metrics evaluation algorithms.

In 2018, Abdulhammed et al. [55] proposed a wireless network intrusion detection system based on machine learning techniques that included classification algorithms and feature selection methods. Before the training phase, the preprocessing stage is applied, such as dataset value conversion to the integer, scaled the range of big data, and normalizing them into smaller ranges. This work utilized different classifiers such as AdaBoost, Random Forest, Random Tree, J48, Logit Boost, Multi-Layer Perceptron, and ZeroR. The core of this work focuses on the effectiveness of feature reduction of the classification algorithms, leading to a better result in terms of detection and speed accuracy. They selected four useful feature sets 32, 10, 7, and 5, to be applied to the training model. The experimental results show that the best performance is associated with the random forest classifier with selected 32 features. The performance evaluation of classification algorithms is 99.64% for accuracy, precision 0.995, and recall 0.966. The proposed system was applied on a wireless AWID dataset. Moreover, in order to validate the results, a comparison is made between the proposed system and other classification algorithms.

In 2018, Belouch et al. [56] paper evaluated the performance of four classification algorithms, namely SVM, Naïve Bayes, Decision Tree, and Random Forest. The approach applies Apache Spark tools to classify intrusion detection on network traffic. The public dataset for network intrusion detection UNSW-NB15 is applied with 42 features to build the model. The experimental results demonstrate a random forest classifier to be the best among other classifiers with the accuracy of 97.49% sensitivity 93.53%, and specificity of 97.75%.

In 2018, Bhosale and Nenova [57] proposed a filter-based Hybrid Feature Selection Algorithm (HFSA) for a suitable process of selection features. HFSA optimized a subset of the most relevant and top-rank features used to build classifiers for respective multi classes. This model is working on the real-time packets, which are captured using the Jpcap library. Naïve Bayes classification algorithm is used to classify normal attacks from malicious ones. The preprocessing phase includes two stages. Firstly, data transformation converts symbolic data into a numerical value. Secondly, during the data normalization phase, features are scaled from the biggest range to the lowest between (0,1), and every record is standardized. Then, applying Naïve Bayes, feature selection is carried out to detect six standard classes: normal, R2L, U2R, DoS, Probe, and Brute force attacks. HFSA is applied for upgrading purposes to enhance the classification system. Overall, the model obtained a total accuracy ratio of 92%, 95% of precision, and 90% of recall.

Gulla et al. [58] presented an intrusion detection system based on Naïve Bayes and Support Vector Machine (SVM) as a classification algorithm. The correlation subset type of feature selection was used for selecting only 24 features out of 42 of the NSL-KDD data set. Moreover, in data preprocessing, the attributes are converted to binary values, and data normalization is applied. The experimental results, depending on accuracy detection, demonstrate the SVM as the best classifier with an overall accuracy result of 93.95% compared to the Naïve Bayes classifier.

In 2019, Kazi et al. [59] presented a novel supervised method to classify and analyze network traffic for detecting malicious attacks. This study utilized Artificial Neural Network (ANN) and Support Vector Machine (SVM) algorithms for classification purposes. Both types of feature selection have used filter method-based Chi-Square and wrapper method-based Correlation for feature selection. The NSL-KDD data set with 25,191 records used as the training model. The approach adopts the Correlation-based wrapper method with selected 17 features more relevant out of 41 features. On the other hand, applying a chi-square-based filter, 35 features are chosen that are more informative and relevant for the training model stage. The experimental results demonstrate that the performance of ANN with wrapper method selecting 17 features gets the highest accuracy of 94.02% compared to all other techniques.

In 2020, Zina et al. [60], proposed a novel method for classification and feature selection applying Regression Trees (CART) combining with Random Forest. This system is called the Hybrid Anomaly-based Intrusion Detection System (HAIDS). The hybrid approach is used to improve the efficiency of the model rather than in a single algorithm. Moreover, the process of removing irrelevant features is applied to overcome the case of high dimensionality. The proposed model was applied to the UNSW-NB15 dataset and selected the highest-ranked thirteen features. The hybrid method achieved the highest performance and accuracy in terms of false alert rate with 11.86% and accuracy rate 87.74%.

In 2020, Iman and Ahmad [61] presented the random forest as a classification algorithm with feature selection Boruta algorithm used to build IDS applied on NSL-KDD dataset. Moreover, the method is provided to find entropy and Gini index as z-score for the number of tree depth values. The satisfied number of optimal features is 34 out of 41. The proposed model results were evaluated in terms of accuracy, sensitivity, and specificity, which is 0.99.

In 2020, Latah and Toker [62] presented an effective hybrid intrusion detection Software-Defined Networking (SDN). The hybrid system was a combination of K Nearest Neighbor (KNN), Extreme Learning Machine (ELM), and Hierarchical Extreme Learning machine (HELM) algorithms. The proposed system results, applying on the KDD Cup 99 dataset, illustrated an outstanding accuracy accessing 84.29%. Moreover, the method provides the detection rate of new attacks rating 77.18%.

In 2020, Jie Gua and Shan Luc. [63] proposed a novel method, an embedding system for intrusion detection system based on Support Vector Machine (SVM) with Naive Bayes feature embedding. The naïve Bayes is used for feature transformation to convert data state. The SVM algorithm is implemented as a classifier. The embedding model was applied to multiple data sets to detect different types of attacks such as UNSW-NB15, NSL-KDD, CICIDS2017, and Kyoto 2006+ using different features for each data dataset. The proposed method, embedding system result compared with a single SVM algorithm, concluded that detection's highest accuracy gets with embedding Naive Bays with SVM. The experiment demonstrated NSL-KDD as the best data set with the highest accuracy of 99.36%. DR 99.25, FAR0.54.

In 2020, Pokharel and P. et al. [64] presented IDS depended on a hybrid classification algorithm and profile improvement to detect anomalous user behavior. The hybrid approaches contain Naïve Bayes and Support Vector Machine (SVM) algorithms for classification. Moreover, it provides data preprocessing. The excellent effect on model accuracy such as data normalization scaled features between (0,1) and selecting the right features on the real-time data set. In this hybrid approach, classifiers get a total accuracy of 0.931 and a precision of 0.958. Also, it provides the accuracy for Classifier Enhancement (CE) 0.953 and precision 0.958.

In 2020, Kumari and Mehta [65], suggested a hybrid classification method for IDS. The hybrid is a combination between Decision Tree J48 and Support Vector Machine (SVM). The SVM has the ability to overcome the problem of high dimensionality. Moreover, for feature extraction, the approach used Particle Swarm Optimization (PSO), selecting nine relevant features out of 42. The paper applies the KDD99 dataset for both stages of training and test. The data set is split into different ratios. The results showed that using 70% for testing and 30% for the training data set is optimal because it increases the accuracy and decreases the false alarm rate. Generally, the hybrid model achieves a total ratio of 99.1 % in accuracy, detection rate 99.6 %, and FAR 0.9 %.

In 2020, Shahmiri et al. [66] introduced an intrusion detection system on the Internet of Things (IoT) technology and upgrading the power grid to a Smart Grid (SG) to detect normal malicious attacks. In this work, they suggested the Hybrid of three Decision Trees (HDT) to detect different types of attacks. Additionally, the performance of the hybrid proposed method was compared with the K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree (DT) methods. The experiments results showed that the (HDT) proposed approach more efficient with evaluation measurement in terms of accuracy 83.1485%, precision 97.2193%, recall 72.4694%, F-score 83.0394% applying NSLKDD.

In 2020, Kachavimath et al. [67] proposed a Distributed Denial of Service (DDoS) detection model to improve network security cases by using machine learning techniques. The K- Nearest Neighbour and Naïve Bayes algorithms were used for classification, and for feature extraction, the correlation was utilized. The proposed model was compared with the conventional learning models being applied on NSL-KDD and KDD Cup 99 datasets. The experimental performance showed that the KNN algorithm with eight features obtained the best results than Naïve Bayes. The different parameters are used for measurement of classification algorithm performance accuracy 98.51%, precision 98.9%, recall 97.8% f-measure 1.005%, sensitivity 97.8%, specificity 99.12%, efficiency 98.48%, error rate 1.50% and ROC 0.99%.

In 2020, Bhosale and Nenova [68] proposed a new method for attack classification, which is Modified Naïve Bayes Algorithm (MNBIDS) with hybrid feature selection to improve the system accuracy of detecting attacks. The hybrid feature selection method that ranks features according to the value of G_corrof for each selected feature. Additionally, compared the CNN, ANN, KNN, SVM algorithms and proposed (MNBIDS). The performance of MNBIDS is measured with the highest accuracy of 97%, precision 98%, and recall of 99%. The IDS performs data preprocessing, data normalization, and features extraction implemented in real-time data to KDD cup 99.

In 2020, Sah and Banerjee [69] described the purpose of feature reduction on the classification model. In this work, they proposed intelligent IDS, using different machine learning algorithms, such as Naive Bayes, K-Nearest Neighbours, Random Forest, and SVM. The Recursive feature Elimination (RFE) and Principal Component Analysis (PCA) methods are used for feature reduction. This work tested the accuracy of classification algorithms with all 41 features compared with feature reduction in different sets 11,12,13,15 of feature selected. The experimental illustrated that accuracy be improved with feature reduction. The random forest classification algorithm gets the best results with the DoS class in terms of accuracy 99.63%, precision 99.53%, recall 99.6%, and f-score 99.58%.

In 2020 Waskle et al. [70] introduced an approach to developing an intrusion detection system using the random forest as a classifier and Principal Component Analysis (PCA) as a dimension reduction technique. The proposed method is compared with other classifiers like decision tree, naïve Bayes, and SVM. The experimental result showed that the proposed method gets the highest performance in terms of accuracy 96.78%, an error rate of 0.21%, and it takes less time for building model 3.42.

In 2020, Fitni and Ramli [71] presented a method for anomaly intrusion detection systems based on Decision Tree, Logistics regression, and gradient boosting as a classifiers ensemble. The model applied to the CSE-CIC-IDS2018 dataset contains 80 features divided by 80% for the training model and 20% for validation. For feature selection, the present method of ensembles feature selection includes Chi-square to calculate the score of high features with the rank correlation of features. After applying the hybrid method, only selected 23 features out of 80. The results of the proposed model compared with seven single classifications showed that the outperform of ensemble three classifiers in terms of accuracy 98.8%, recall 97.1%, precision 98.8%, and F1 97.9%.

6. Comparison and Discussion

The implementation of classification algorithms for IDS to classify different types of attack are presented in Table 2. Machine learning techniques have been applied to the field of network security to improve intrusion detection systems. Previous sections reviewed some researches about classification algorithms applied to build the IDS model and evaluated the performance by different metrics in terms of accuracy, recall, precision, f-score, specificity, sensitivity, error rate, and dependable tool confusion matrix. The dimension reduction and feature selection had a good effect on the classification model performance because it reduces training and testing time via removing the irrelevant features, making the classification process more accurate and less complicated.

A combination of multi-classification algorithms and called hybrid classification could be the optimal solution to classify attacks type. The different data types of attacks may deal with different types of classification algorithms. Most studies now have focused on the hybrid classification algorithm rather than a single classification because it provides very satisfying results in different performances measurement.

The best results for most reviewed studies showed that the Random Forest algorithm achieved the best accuracy of classification because it combines many decision trees that then decide the type of attack, leading to the decrease of the risk of overfitting. The random forest can deal with various big types of features that do not require data scaling. Moreover, the Practical Swarm Optimization (PSO) gets the best result for feature selection.

In this paper, the comparison is performed in terms of data set; data-preprocessing techniques, a number of features selected, feature selection techniques, classification algorithms, and evaluation metrics. This study aims to show different classification algorithms' performance by using different measurements to select a suitable classifier best model to gain speed and accuracy.

Table 2 - Comparison of evaluation of different classification algorithms performance.

Ref	Data set	Data preprocessing Techniques	Number of features selected	Feature Selection Techniques	Classification Algorithm	Evaluation Metrics
[55] 2018	AWID	Transformation values into integer Normalization scale	32 set,10set 7 set,5 set	ZeroR	AdaBoost, Random Forest, Random Tree, J48, logit Boost, MLP	best performance Random Forest with 32 features accuracy 99.64%, precision 0.995, recall 0.966
[56] 2018	UNSW-NB15	Apache Spark processing tools	42 features out of 49	-	SVM, Naïve Bayes, Decision Tree and Random Forest	best results Random Forest accuracy 97.49, Sensitivity 93.53, specificity 97.75
[57] 2018	KDD Cup 99	Data transformation Data normalization Data standardization	rank all 41 features	HFSA	Naïve Bayes	multi classes accuracy 92% precision 95%, recall 90%
[58] 2019	NSL–KDD	convert nominal attribute to binary attribute non-numeric, dimension reduction, Normalization	24	CfsSubsetEval	SVM Naïve Bayes	SVM best accuracy of 93.95
[59] 2019	NSL-KDD	Reduce features	17 35	Correlation Chi-Square	ANN SVM	Highest ANN with Wrapper(correlation) 17 features, accuracy 94.02%,
[60] 2020	UNSW-NB15	categorical features remove redundant and irrelevant features	top rank 13	Random Forest	Classification and Regression Trees (CART)	accuracy 87.74
[61] 2020	NSL-KDD	outlier detection	when 34 Accepted features	Boruta Algorithm	Random Forest	accuracy 0.99892798 Sensitivity 0.99852158 Specificity 0.99939955
[62] 2020	NSL-KDD	without need preprocessing	features F1, F2, F5, F6, F23, F24	Software SDN	KNN, ELM, H-ELM	accuracy 84.29, False alarm rate 6.3
[63] 2020	UNSWNB , CICIDS2017 NSL-KDD Kyoto 2006+	Data normalization Data transformation	Different number for each data set	Naïve Bayes embedding feature	Embedding SVM Naive Bayes	the highest score on NSL-KDD data set with accuracy 99.36%., DR 99.25%, FAR 0.54%
[64] 2020	real-world log	Normalization min-max	event logs generated on	-	Naïve Bayes	for CE accuracy 0.953,

	dataset.	Selecting right features	user actions		SVM	precision 0.958
[65] 2020	KDD'99	Dimension reduction	9	PSO	decision trees J48, SVM	99.1 %, detection rate 99.6 %, FAR 0.9 %
[66] 2020	NSLKDD	Data Normalization	-	CART tree	Hybrid three decision tree	accuracy 83.1485, Precision 97.2193, recall 72.4694, F-score 83.0394
[67] 2020	NSL-KDD KDD Cup 99	Data cleaning selected the removal of the essential feature of non-values	8	feature extraction using the correlation	KNN Naïve Bayes	Best result KNN accuracy 98.51, Precision 98.9% Recall 97.8%, F-measure 1.005%, Sensitivity 97.8% Specificity 99.12%, efficiency 98.48%, error rate 1.50%, BCR 98.5%, ROC 0.99%
[68] 2020	Real Time Data KDD Cup 99 dataset	data normalization and feature extraction	Sort 41 feature according to the value of G_corrof	Hybrid feature selection	MNBIDS	accuracy 97%, Precision 98%, recall 99%
[69] 2020	NSL KDD	removing duplicate records non-numerical objects to numerical feature scaling feature reduction	Selected different set 11,12,13,15	feature reduction PCA - RFE	KNN, SVM, Random Forest Naive Bayes	Best result Random Forest with DoS class, accuracy 99.63%, precision 99.53, recall 99.6%, F-score 99.58%
[70] 2020	KDD	reduction high dimension using Python	-	Feature reduction PCA	Random forest decision tree, naïve bayes and SVM	Best result Random Forest accuracy 96.78% and error rate 0.21%.
[71] 2020	SE-CIC-IDS2018	missing values small sample of data	23	Chi-square Correlation	Decision Tree, Logistics regression, , and gradient boosting ensembled	accuracy 98.8%, recall 97.1%, precision 98.8%, F1 97.9%.

7. Conclusion

IDS improvement performance depends on different machine learning techniques. Classification algorithms have a significant role in helping IDS to distinguish different types of attacks. This paper aims to test different classifier algorithms and find the evaluation performance by using different metrics. The study, applying various metric measurements to evaluate classifiers' performance, noticed that the random forest algorithm achieved sufficient results and the highest accuracy to classify different types of attacks. Obtaining high performance of the model, most researchers used the hybrid classification algorithm for building intrusion detection systems rather than individual classification. The effectiveness of dimension reduction to reduce big data sets' complexity leads to select optimal features to obtain better performance in classification in terms of accuracy and speed.

Acknowledgement

The authors would like to thank the Department Information Technology Management, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.

References

[1] Dang, Q. V. (2020, October). Active learning for intrusion detection systems. In 2020 RIVF International Conference on Computing and Communication Technologies (RIVF) (pp. 1-3). IEEE

[2] Singh, R., Kalra, M., & Solanki, S. (2020). A hybrid approach for intrusion detection based on machine learning. International Journal of Security and Networks, 15(4), 233-242

[3] Lee, J., Kim, J., Kim, I., & Han, K. (2019). Cyber threat detection based on artificial neural networks using event profiles. IEEE Access, 7, 165607-165626

[4] Abdulazeez, A., Salim, B., Zeebaree, D., & Doghramachi, D. (2020). Comparison of VPN Protocols at Network Layer Focusing on Wire Guard Protocol

[5] Ugochukwu, C. J., Bennett, E. O., & Harcourt, P. (2019). An intrusion detection system using machine learning algorithm. LAP LAMBERT Academic Publishing

[6] Salih, A. A., & Abdulrazaq, M. B. (2019, April). Combining best features selection using three classifiers in intrusion detection system. In 2019 International Conference on Advanced Science and Engineering (ICOASE) (pp. 94-99). IEEE

[7] Ghanem, W. A. H., Jantan, A., Ghaleb, S. A. A., & Nasser, A. B. (2020). "An Efficient Intrusion Detection Model Based on Hybridization of Artificial Bee Colony and Dragonfly Algorithms for Training Multilayer Perceptrons". IEEE Access, 8, 130452-130475

[8] Alamiedy, T. A., Anbar, M., Alqattan, Z. N., & Alzubi, Q. M. (2019). "Anomaly-based intrusion detection system using multi-objective grey wolf optimisation algorithm". Journal of Ambient Intelligence and Humanized Computing, 1-22

[9] Bhumgara, A., & Pitale, A. (2019, July). Detection of Network Intrusions using Hybrid Intelligent Systems. In 2019 1st International Conference on Advances in Information Technology (ICAIT) (pp. 500-506). IEEE

[10] Rai, A. (2020, June). Optimizing a New Intrusion Detection System Using Ensemble Methods and Deep Neural Network. In 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184) (pp. 527-532). IEEE

[11] Mirza, A. H. (2018, May). Computer network intrusion detection using various classifiers and ensemble learning. In 2018 26th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE

[12] Ahmad, S., Arif, F., Zabeehullah, Z., & Iltaf, N. (2020, June). Novel Approach Using Deep Learning for Intrusion Detection and Classification of the Network Traffic. In 2020 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA) (pp. 1-6). IEEE

[13] Hasan, D. A., & Abdulazeez, A. M. (2020). A Modified Convolutional Neural Networks Model for Medical Image Segmentation. learning, 20, 22

[14] Phadke, A., Kulkarni, M., Bhawalkar, P., & Bhattad, R. (2019, March). A Review of Machine Learning Methodologies for Network Intrusion Detection. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 272-275). IEEE

[15] Golrang, A., Golrang, A. M., Yayilgan, S. Y., & Elezaj, O. (2020). A Novel Hybrid IDS Based on Modified NSGAII-ANN and Random Forest. Electronics, 9(4), 577

[16] Shashank, K., & Balachandra, M. (2018, August). Review on Network Intrusion Detection Techniques using Machine Learning. In 2018 IEEE Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER) (pp. 104-109). IEEE

[17] Yihunie, F., Abdelfattah, E., & Regmi, A. (2019, May). Applying machine learning to anomaly-based intrusion detection systems. In 2019 IEEE Long Island Systems, Applications and Technology Conference (LISAT) (pp. 1-5). IEEE

[18] Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. (2019, April). Machine learning and region growing for breast cancer segmentation. In 2019 International Conference on Advanced Science and Engineering (ICOASE) (pp. 88-93). IEEE

[19] Talingdan, J. A. (2019, May). Performance comparison of different classification algorithms for household poverty classification. In 2019 4th International Conference on Information Systems Engineering (ICISE) (pp. 11-15). IEEE

[20] Jahwar, A. F., & Abdulazeez, A. M. (2020). Meta-Heuristic Algorithms For K-Means Clustering: A Review. PalArch's Journal of Archaeology of Egypt/Egyptology, 17(7), 12002-12020

[21] Ahmed, M. R. A. G., & Ali, F. M. A. (2019, September). Enhancing Hybrid Intrusion Detection and Prevention System for Flooding Attacks Using Decision Tree. In 2019 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE) (pp. 1-4). IEEE

[22] Ahmad, U., Asim, H., Hassan, M. T., & Naseer, S. (2019, November). Analysis of Classification Techniques for Intrusion Detection. In 2019 International Conference on Innovative Computing (ICIC) (pp. 1-6). IEEE

[23] Zeebaree, D. Q., Haron, H., & Abdulazeez, A. M. (2018, October). Gene selection and classification of microarray data using convolutional neural network. In 2018 International Conference on Advanced Science and Engineering (ICOASE) (pp. 145-150). IEEE

[24] Zebari, D. A., Zeebaree, D. Q., Abdulazeez, A. M., Haron, H., & Hamed, H. N. A. (2020). Improved Threshold Based and Trainable Fully Automated Segmentation for Breast Cancer Boundary and Pectoral Muscle in Mammogram Images. IEEE Access, 8, 203097-203116

[25] Abdulqader, D. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). Machine Learning Supervised Algorithms of Gene Selection: A Review. Machine Learning, 62(03)

[26] Abhale, A. B., & Manivannan, S. S. (2020). Supervised Machine Learning Classification Algorithmic Approach for Finding Anomaly Type of Intrusion Detection in Wireless Sensor Network. Optical Memory and Neural Networks, 29(3), 244-256

[27] Shenfield, A., Day, D., & Ayesh, A. (2018). Intelligent intrusion detection systems using artificial neural networks. ICT Express, 4(2), 95-99

[28] Omar, N., Abdulazeez, A. M., Sengur, A., & Al-Ali, S. G. S. (2020). Fused faster RCNNs for efficient detection of the license plates. Indonesian Journal of Electrical Engineering and Computer Science, 19(2), 974-982

[29] Aboueata, N., Alrasbi, S., Erbad, A., Kassler, A., & Bhamare, D. (2019, July). Supervised machine learning techniques for efficient network intrusion detection. In 2019 28th International Conference on Computer Communication and Networks (ICCCN) (pp. 1-8). IEEE

[30] Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. (2019, April). Trainable Model Based on New Uniform LBP Feature to Identify the Risk of the Breast Cancer. In 2019 International Conference on Advanced Science and Engineering (ICOASE) (pp. 106-111). IEEE

[31] Punitha, A., Vinodha, S., Karthika, R., & Deepika, R. (2019, March). A Feature Reduction Intrusion Detection System using Genetic Algorithm. In 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-7). IEEE

[32] Zeebaree, D. Q., Abdulazeez, A. M., Zebari, D. A., Haron, H., & Hamed, H. N. A. Multi-Level Fusion in Ultrasound for Cancer Detection Based on Uniform LBP Features

[33] Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. Journal of Applied Science and Technology Trends, 1(2), 56-70

[34] Terol, R. M., Reina, A. R., Ziaei, S., & Gil, D. (2020). A Machine Learning Approach to Reduce Dimensional Space in Large Datasets. IEEE Access, 8, 148181-148192

[35] Jiang, K., Wang, W., Wang, A., & Wu, H. (2020). Network intrusion detection combined hybrid sampling with deep hierarchical network. IEEE Access, 8, 32464-32476

[36] Marir, N., Wang, H., Feng, G., Li, B., & Jia, M. (2018). Distributed abnormal behavior detection approach based on deep belief network and ensemble svm using spark. IEEE Access, 6, 59657-59671

[37] Wei, W., Chen, S., Lin, Q., Ji, J., & Chen, J. (2020). "A multi-objective immune algorithm for intrusion feature selection. Applied Soft Computing", 95, 106522

[38] Adhao, R. B., & Pachghare, V. K. (2019, December). Performance-Based Feature Selection Using Decision Tree. In 2019 International Conference on Innovative Trends and Advances in Engineering and Technology (ICITAET) (pp. 135-138). IEEE

[39] Devan, P., & Khare, N. (2020). An efficient XGBoost–DNN-based classification model for network intrusion detection system. Neural Computing and Applications, 1-16

[40] Uysal, E. İ., Demircioğlu, G., Kale, G., Bostanci, E., Güzel, M. S., & Mohammed, S. N. (2019, October). Network Anomaly Detection System using Genetic Algorithm, Feature Selection and Classification. In 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-5). IEEE

[41] Davis, A., Gill, S., Wong, R., & Tayeb, S. (2020, September). Feature Selection for Deep Neural Networks in Cyber Security Applications. In 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-7). IEEE

[42] Selvamani, D., & Selvi, V. (2019). A comparative study on the feature selection techniques for intrusion detection system. Asian J. Comput. Sci. Technol, 8(1), 42-47

[43] Sarvari, S., Sani, N. F. M., Hanapi, Z. M., & Abdullah, M. T. (2020). An efficient anomaly intrusion detection method with feature selection and evolutionary neural network. IEEE Access, 8, 70651-70663

[44] Saranya, T., Sridevi, S., Deisy, C., Chung, T. D., & Khan, M. A. (2020). Performance analysis of machine learning algorithms in intrusion detection system: A review. Procedia Computer Science, 171, 1251-1260

[45] Rashid, A., Siddique, M. J., & Ahmed, S. M. (2020, February). Machine and deep learning based comparative analysis using hybrid approaches for intrusion detection system. In 2020 3rd International Conference on Advancements in Computational Sciences (ICACS) (pp. 1-9). IEEE

[46] Bargarai, F., Abdulazeez, A., Tiryaki, V., & Zeebaree, D. (2020). Management of Wireless Communication Systems Using Artificial Intelligence-Based Software Defined Radio

[47] Danjuma, K. J. (2015). Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients. arXiv preprint arXiv:1504.04646

[48] Sujatha, J., & Rajagopalan, S. P. (2017). Performance evaluation of machine learning algorithms in the classification of Parkinson disease using voice attributes. International Journal of Applied Engineering Research, 12(21), 10669-10675

[49] Chkirbene, Z., Erbad, A., Hamila, R., Mohamed, A., Guizani, M., & Hamdi, M. (2020). TIDCS: A dynamic intrusion detection and classification system based feature selection. IEEE Access, 8, 95864-95877

[50] Dahiya, P., & Srivastava, D. K. (2018). Network intrusion detection in big dataset using spark. Procedia computer science, 132, 253-262

[51] bhai Gupta, A. R., & Agrawal, J. (2020, April). A Comprehensive Survey on Various Machine Learning Methods used for Intrusion Detection System. In 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 282-289). IEEE

[52] Haggag, M., Tantawy, M. M., & El-Soudani, M. M. (2020). Implementing a Deep Learning Model for Intrusion Detection on Apache Spark Platform. IEEE Access, 8, 163660-163672

[53] Hussien, Z. (2020). Anomaly Detection Approach Based on Deep Neural Network and Dropout. Baghdad Science Journal, 17(2 (SI)), 0701-0701

[54] Panda, M., Abraham, A., Das, S., & Patra, M. R. (2011). Network intrusion detection system: A machine learning approach. Intelligent Decision Technologies, 5(4), 347-356

[55] Abdulhammed, R., Faezipour, M., Abuzneid, A., & Alessa, A. (2018, June). Effective features selection and machine learning classifiers for improved wireless intrusion detection. In 2018 International Symposium on Networks, Computers and Communications (ISNCC) (pp. 1-6). IEEE

[56] Belouch, M., El Hadaj, S., & Idhammad, M. (2018). Performance evaluation of intrusion detection based on machine learning using Apache Spark. Procedia Computer Science, 127, 1-6

[57] Bhosale, K. S., Nenova, M., & Iliev, G. (2018, December). Data Mining Based Advanced Algorithm for Intrusion Detections in Communication Networks. In 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS) (pp. 297-300). IEEE

[58] Gulla, K. K., Viswanath, P., Veluru, S. B., & Kumar, R. R. (2020). Machine learning based intrusion detection techniques. In Handbook of computer networks and cyber security (pp. 873-888). Springer, Cham

[59] Taher, K. A., Jisan, B. M. Y., & Rahman, M. M. (2019, January). Network intrusion detection using supervised machine learning technique with feature selection. In 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) (pp. 643-646). IEEE

[60] Chkirbene, Z., Eltanbouly, S., Bashendy, M., AlNaimi, N., & Erbad, A. (2020, February). Hybrid machine learning for network anomaly intrusion detection. In 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT) (pp. 163-170). IEEE

[61] Iman, A. N., & Ahmad, T. (2020, February). Improving Intrusion Detection System by Estimating Parameters of Random Forest in Boruta. In 2020 International Conference on Smart Technology and Applications (ICoSTA) (pp. 1-6). IEEE

[62] Latah, M., & Toker, L. (2020). An efficient flow-based multi-level hybrid intrusion detection system for software-defined networks. CCF Transactions on Networking, 3(3), 261-271

[63] Gu, J., & Lu, S. (2021). An effective intrusion detection approach using SVM with naïve Bayes feature embedding. Computers & Security, 103, 102158

[64] Pokharel, P., Pokhrel, R., & Sigdel, S. (2020, October). Intrusion Detection System based on Hybrid Classifier and User Profile Enhancement Techniques. In 2020 International Workshop on Big Data and Information Security (IWBIS) (pp. 137-144). IEEE

[65] Kumari, A., & Mehta, A. K. (2020, October). A Hybrid Intrusion Detection System Based on Decision Tree and Support Vector Machine. In 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA) (pp. 396-400). IEEE

[66] Taghavinejad, S. M., Taghavinejad, M., Shahmiri, L., Zavvar, M., & Zavvar, M. H. (2020, April). Intrusion Detection in IoT-Based Smart Grid Using Hybrid Decision Tree. In 2020 6th International Conference on Web Research (ICWR) (pp. 152-156). IEEE

[67] Kachavimath, A. V., Nazare, S. V., & Akki, S. S. (2020, March). Distributed Denial of Service Attack Detection using Naïve Bayes and K-Nearest Neighbor for Network Forensics. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 711-717). IEEE

[68] Bhosale, K. S., Nenova, M., & Iliev, G. (2018, December). Modified Naive Bayes Intrusion Detection System (MNBIDS). In 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS) (pp. 291-296). IEEE

[69] Sah, G., & Banerjee, S. (2020, July). Feature Reduction and Classifications Techniques for Intrusion Detection System. In 2020 International Conference on Communication and Signal Processing (ICCSP) (pp. 1543-1547). IEEE

[70] Waskle, S., Parashar, L., & Singh, U. (2020, July). Intrusion Detection System Using PCA with Random Forest Approach. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 803-808). IEEE

[71] Fitni, Q. R. S., & Ramli, K. (2020, July). 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) (pp. 118-124). IEEE