# A comparative study of Machine learning Algorithms on the UNSW-NB 15 Dataset

*Rachid* Tahri[1*] *Abdessamad* Jarrar[2]*, Abdellatif* Lasbahani[3], and *Youssef* Balouki[1]

[1]Faculty of Sciences and Technics, Hassan First, Settat, Morocco
[2]Faculty of Sciences, Mohammed First University, Oujda, Morocco[1]
[3]National School of Applied Sciences, Sultan Moulay Slimane University, Bni Mellale, Morocco

**Abstract.**

The world has experienced a radical change due to the internet. Internet became a crucial element in our daily life, therefore, the security of our DATA could be threatened at any time. This safety is handled using systems to detect network intrusion called Intrusion Detection Systems (IDS). Machine learning techniques are being implemented to improve these systems. In order to enhance the performance of IDS, different classification algorithms are applied to detect various types of attacks. Choosing a good one for building IDS is not an easy task. The best method is to test the performance of the different classification algorithms. Nevertheless, most researchers have focused on the confusion matrix as measurements of classification performance. Therefore, many papers use this matrix to present a detailed comparison with the dataset, data preprocessing, feature selection technique, algorithms classification and performance evaluation. The goal of this paper is to present a comparison of application of different Machine Learning algorithms used to build and improve intrusion detection systems in terms of confusion matrix, accuracy, recall, precision, FAR, specificity and sensitivity using the UNSW-NB15 Dataset. Furthermore, we introduce some lesson learnt to shoot more researchers in their future works.

## 1 Introduction

Many researches especially concerning NIDS (Network Intrusion Detection Systems), have been conducted under specific conditions in terms of systems, algorithms, and data sets used, limiting the extrapolation of results to a larger target of systems. These researches are based on statistical learning methods allowing to design algorithms never implemented in a real system. As a result, new attacks are detected and many problems due to the constant delay of the security tools on the threats are frequently faced (data theft, espionage...).these informatics attacks are grouped together as a dataset.

Indeed, datasets represent instances that consist of several features and are related to the intrusion detection system. So, it is essential to realize the type of data containing different types of attacks and features. The most popular data set that is being used for the intrusion detection system is UNSW-NB 15.After classification; the dataset will be treated using machine learning.

Machine Learning (ML) techniques widely used in computer security data sets have recently become a trend in security technology. 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. IDS is a commonly used machine learning classifier to distinguish between various attacks as a class. Many supervised classification algorithms are applied to IDS, such as Decision Trees, Naïve Bayes, K-Nearest Neighbor, Random Forest, Support Vector Machine, and Logistic Regression. Evaluation of classification algorithms depends on various statistical metrics, especially confusion matrix results, to classify and predict different types of threats.

## 2 Classification method

Classification is one of the tasks of machine learning. It is a supervised learning model. It is used for intrusion detection systems based on binary or multiple classes [1]. In supervised learning, data is always labeled.

Each record in a data set is assigned to a particular class. An IDS based on a classification model classifies all network traffic into normal or anomalous classification algorithms. The obstacle to building the model is the massive amount of data. Classification algorithms, facing many problems when building a model, need a data-preprocessing step, especially when the dimension of the data is high.

The choice of the best classification algorithm depends on the performance evaluation in terms of confusion matrix and accuracy.

[*] Corresponding author: rachid.tahritr@gmail.com

The process of classifying the data in the dataset includes the two stages of training and testing. During the training and learning phase, a classifier is learned as a target, while during the second stage, the testing phase, the constructed model is used to predict class labels.

It is essential to analyze the time required for each classifier for both training and testing stages. Before applying the classifiers, data preprocessing helps the classification model to reduce time and complexity by eliminating irrelevant data.

Before applying the classifiers, data preprocessing helps the classification model reduce time and complexity by eliminating irrelevant data to improve the efficiency of the classification algorithms.

The cross-validation process of the dataset is divided into two equal groups for the classification of the network traffic dataset.

One group for testing, and the rest will be used for the training model. Few algorithms are able to distinguish different attacks from normal attacks with sufficient results.

The most popular classifiers are Decision Tree (DT), Random Forest, Support Machine Learning (SVM), K-Nearest Neighbor (KNN), Naïve, Decision Tree and Logistic Regression Supervised Algorithms

## 3 Supervised Algorithms

Machine learning (ML) and data mining have proven many effective applications in gene expression analysis. ML is used in every area of computational work where algorithms are designed and performance is increased. The main kinds of machine learning are supervised learning (SL) and unsupervised learning (USL). The classification pattern is used in discrete value problems, while the regression pattern is used to make decisions in persistent value problems.

From the above mentioned machine learning models supervised learning used in this paper. Supervised learning is learning that involves an expert well-versed in the environment and the expected response from the learning system [2]. The modelling of some maps between the input and output vectors is a problem of learning under the supervision of some real phenomena. Supervised learning techniques include classification, regression, and band techniques where the target variable is categorical in classification and continues to decline

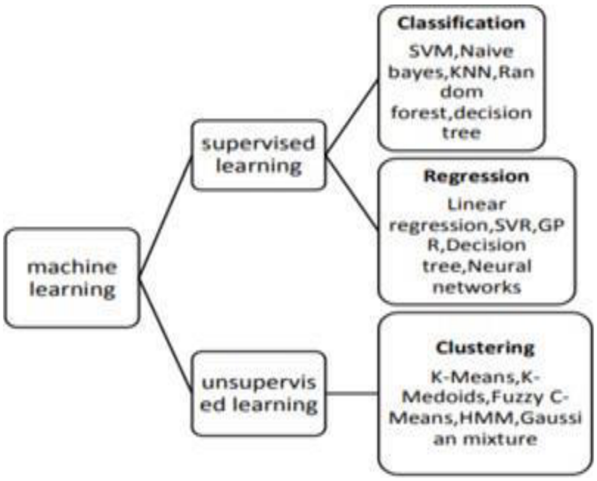
machine learning
supervised learning
unsupervised learning
Classification
SVM,Naive bayes,KNN,Random forest,decision tree
Regression
Linear regression,SVR,GPR,Decision tree,Neural networks
Clustering
K-Means,K-Medoids,Fuzzy C-Means,HMM,Gaussian mixture

**Fig. 1.** Machine Learning Algorithms

## 4 Evaluation Performance

The performance of the intrusion detection models was achieved by evaluating the measurements of the values of the coincidence matrix also known as the confusion matrix. The confusion matrix shows the distribution of instances that are either correctly classified or misclassified by the models, the confusion matrix is an N X N matrix, where N is any integer greater than 1, The diagonal elements represent the number of points for which the predicted label is equal to the true label, while the off-diagonal elements are those that are mislabelled by the classifier. The higher the diagonal values of the confusion matrix, indicating a large number of correct predictions

It produced four results, which are:

- True positive (TP) means correct detection of the intrusion.
- False positive (FP) means normal traffic is considered as a cyber-attack.
- True negative (TN) refers to normal traffic correctly labelled as normal.
- False negative (FN) means that the intrusion disclosure fails.

Authors should use the forms shown in Table 1 in the final reference list.

**Table 1.** Metrics for classification algorithms.

| **Element Formula** | **Description Evaluation Metrics** |
|---|---|
| Accuracy=(TP+TN)/( TP+TN+FP+FN) | Total correct classified over the total number of records |
| Precision = TP / (TP+FP) | True positive that are correctly predicted from the total predicted patterns in a positive class |
| Book Recall = TP / (TP+FN) | Positive patterns that are correctly attack classified. |
| Specificity= TN / (TN+FP) | Negative patterns that are correctly classified. |
| 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 |
| Page True Negative Rate (TNR) =TN/(TN+FP) | True negative (TN): Normal correctly classified as normal (false alarm) |
| Year False Negative Rate (FNR)=FN/(FN+TP) | False negative (FN): Incorrectly classified attacks as a normal |

## 5 Review of Classification Algorithms for IDS

In the last decade, many works have been presented to improve IDSs to detect and prevent various malicious attacks of accessing computer information. This section discusses some of the machine learning techniques and algorithms in the classification process used for intrusion detection, including data preprocessing, feature selection techniques, classification algorithms, and metric evaluation algorithms

- In 2018, Belouch et al. [4] 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 2020, Zina et al. [5] , 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, Jie Gua and Shan Luc. [6] 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. [7] 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.

## 6 Comparison and Discussion

The implementation of classification algorithms for IDS to classify different types of attack are presented in the table [1] below.to improve intrusion detection systems, Machine learning techniques have been applied to the field of network security. Previous sections reviewed some researches about classification algorithms used 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.

Indeed, 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. So 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, 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 pre-processing 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 in order 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 |
|---|---|---|---|---|---|---|
| [4] 2018 | UNSW<br>NB15 | Apache Spark processing tools | 42 features out of 49 | - | SVM, Naïve Bayes,<br>Decision Tree and Random Forest | best results Random Forest accuracy 97.49, Sensitivity 93.53, specificity 97.75 |
| [5] 2020 | UNSW<br>NB15 | categorical features<br>remove redundant and<br>irrelevant features | top rank 13 | Random Forest | Classification and Regression Trees (CART) | accuracy 87.74 |
| [6] 2020 | UNSWNB<br>,<br>CICIDS2017<br>NSL-KDD<br>Kyoto 2006+ | Data normalization<br>Data transformation | Different number for each data set | Naïve Bayes embedding feature | Embedding SVM<br>Naive Bayes | the highest score on NSL<br>KDD data set with<br>accuracy 99.36%., DR<br>99.25%, FAR 0.54% |

## 7 Conclusion

To conclude, IDS improvement performance depends on different machine learning techniques. Classification algorithms have a significant role in helping IDS to distinguish different types of attacks. The present document aims to test different classifier algorithms and find the evaluation performance by using different metrics, by 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.

## References

1. B. Charbuty, and A. Adnan. "*Classification based on decision tree algorithm for machine learning*." Journal of Applied Science and Technology Trends 2.01, pp. 20-28, (2021).
2. Krishnan, Deepa. "*Analysis of Accuracy of Supervised Machine Learning Algorithms in Detecting Denial of Service Attacks*." Advances in Signal and Data Processing. Springer, Singapore, pp. 519-529, (2021).
3. A. Agarwal, et al. "*Classification model for accuracy and intrusion detection using machine learning approach.*" PeerJ Computer Science **7** (2021): e437
4. M. Belouch, S. El Hadaj, & M. Idhammad, *Performance evaluation of intrusion detection based on machine learning using Apache Spark*. Procedia Computer Science, **127**, pp. 1-6, (2018).
5. Z. Chkirbene, S. Eltanbouly, M. Bashendy. N. AlNaimi, & A. Erbad. *Hybrid machine learning for network anomaly intrusion detection*. IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), pp. 163-170. IEEE, (2020).
6. J. Gu, & S. Lu. *An effective intrusion detection approach using SVM with naïve Bayes feature embedding*. Computers & Security, **103**, 102158, (2021).