# Intrusion Detection System Using machine learning Algorithms

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

[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. As a matter of fact, it assists people in maintaining their social networks and links them to other members of their social networks when they require assistance. In effect sharing professional and personal data comes with several risks to individuals and organizations. Internet became a crucial element in our daily life, therefore, the security of our DATA could be threatened at any time. For this reason, IDS plays a major role in protecting internet users against any malicious network attacks. (IDS) Intrusion Detection System is a system that monitors network traffic for suspicious activity and issues alerts when such activity is discovered. In this paper, the focus will be on three different classifications; starting by machine learning, algorithms NB, SVM and KNN. These algorithms will be used to define the best accuracy by means of the USNW NB 15 DATASET in the first stage. Based on the result of the first stage, the second one is used to process our database with the most efficient algorithm. Two different datasets will be operated in our experiments to evaluate the model performance. NSL-KDD and UNSW-NB15 datasets are used to measure the performance of the proposed approach in order to guarantee its efficiency.

## 1 Introduction

The number of computing devices has grown at a rapid rate. Laptops and desktop computers, as well as smartphones and tablets, have become nearly vital tools in everyday life, and many people use them on a regular basis. The main issue comes here the data which we get through internet has to be secured; This security of data over network is done by Intrusion Detection System (IDS).

An intrusion detection system (IDS) is a software application or device that monitors system or network activity for policy violations or malicious behavior, and generates reports for the management system. The need for an intrusion detection system is undeniable; thus, an accurate model must be developed. In this field, machine learning has proven to be an effective investigation device that can detect any irregular event taking place in any system's traffic. To build a good IDS it well be able to detect malicious traffics with a high efficacy; the accuracy of algorithms of classification well decide that efficacy.

In this work, we propose an IDS approach for detecting malicious network traffic with more efficiency and higher accuracy at the first a presentation of our DATASET that will be trained by 3 different algorithms of classification; the next section represents ours second DATASET but this time it well be trained by the higher accuracy of the tree algorithms bellows; the last section is a conclusion as well as some issues which have been highlighted for future research.

## 2 Dataset Descriptions

Many datasets are publicly available online for research purposes. According to an examination of the literature, some of them were created decades ago and may not be very useful in detecting recent threats. Some examples of such datasets include KDD98 and KDD'99.

UNSW-NB15 was created in 2015 in the cyber range lab of the Australian Centre for Cybersecurity, according to (Slay N. M., 2016). CSV files are one of the dataset's formats. We are not using the original CSV files because they include over 2.5 million records split into four files.

We are using the polished CSV files in our research since they have 175,341 transactions and 82,332 entries in the training and testing sets, respectively. There are 47 features in the dataset, including numeric, nominal, and categorical data types. It is a binary and multi-class labelled dataset. The distribution of each assault in training and testing sets is shown in Table 1.

**Table 1.** UNSW-NB15 Datasets.

| Dataset | Class | Train-set | Test-set |
|---|---|---|---|
| UNSW-NB15 | normal | 56 000 | 37 000 |
| | generic | 40 000 | 18 871 |
| | exploits | 33 393 | 11 132 |
| | fuzzers | 18 184 | 6 062 |

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

| | | | |
|---|---|---|---|
| | dos | 12 264 | 4 089 |
| | reconnaissance | 10 491 | 3 496 |
| | analysis | 2 000 | 677 |
| | backdoor | 1 746 | 583 |
| | shellcode | 1 133 | 378 |
| | worms | 130 | 44 |
| | Total | 175 341 | 82 332 |

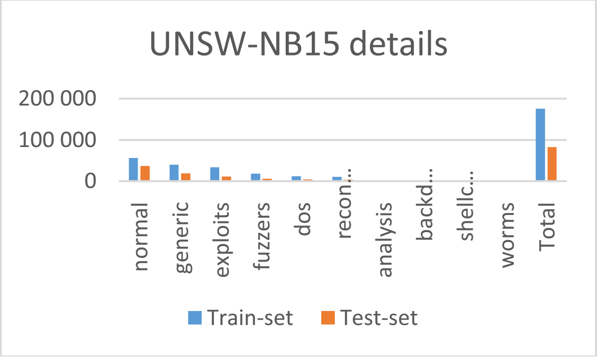

**Fig. 1.** UNSW-NB15 details

### 2.1 NSL-KDD

KDD'99 is outdated and contains redundant records, resulting in network intrusion detection inaccuracy. The problem is solved in NSL-KDD, which is a developed version of KDD'99. The training set of NSL-KDD has 125973 data points, whereas the testing set contains 22544 data points. It features 41 variables with numeric, binary, and nominal data types, as well as a label. Dos, probe, r2l, u2r, and regular class are the four major groups of attack types in the dataset. The distribution of each assault in training and testing sets is shown in Table 2.

**Table 2.** NSL-KDD Datasets.

| Dataset | Class | Train-set | Test-set |
|---|---|---|---|
| NSL-KDD | normal | 67 343 | 9,711 |
| | dos | 45 927 | 7 458 |
| | probe | 11 656 | 2 421 |
| | r2l | 995 | 2 754 |
| | u2r | 52 | 200 |
| | Total | 125 973 | 22 544 |

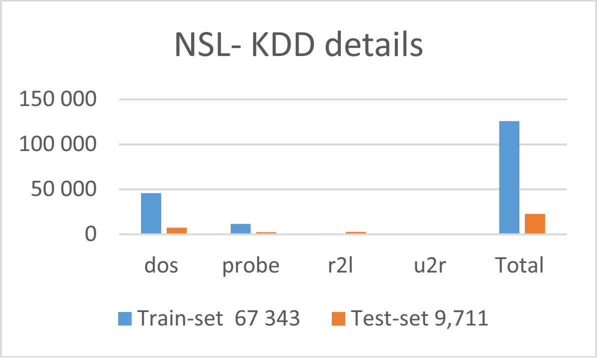

**Fig. 2.** NSL-KDD details

## 3 Related works

Abhishek Divekar et al (A. Divekar, 2018) used classification algorithms such as Naıve. Bayes, K-means, neural network, RF, SVM, and DT and compared performances for alternatives KDD'99. They found that UNSW-NB15 is a better and modern alternative for the KDD'99. The result of the study showed that classifiers trained in terms of f1-score were much better than those trained with KDD'99 and NSL-KDD.

The authors of (Srivastava, 2018) have attempted to assess the performance and effectiveness of NIDS. They have used two characteristic reduction methods, LDA and CCA. Seven classifiers were applied with different measurement parameters and metrics such as FPR, training time, accuracy, the ROC zone. The algorithms used are the random tree, the naive bayes, the rep tree, the RF, random committee, randomizable bagging, and filtered. The result with LDA and random tree on UNSW-NB15 was declared best.k2

In (M. Belouch, Performance evalution of intrusion detection based on machine learning using apache spark, 2018), the authors conducted experimental studies and evaluated the performance of some most commonly used classification ML algorithms such as NB, SVM, DT and RF on apache spark big data environment. They measured time of detection, time of building and the time of prediction for network intrusion detection systems. They used UNSW-NB15 data-set for the purpose of performance evaluation and claimed that RF technique was outperformer with respect to specificity, accuracy, sensitivity, and also execution time among all the four other tested algorithms

In (Slay N.M., 2015), the authors proposed a hybrid feature reduction approach based on attribute values' CP followed by an ARM. First, the dataset was splitted into partitions equally, the reason behind was to reduce the processing-time, then the output of CP technique was given as input to ARM to reduce number of feature. In the decision-engine expectation maximization clustering, logistic regression and naive bayes algorithms were employed for network intrusion detection to compare and evaluate the results. They claim that the model was able to boost accuracy, reduce false alarm rate and shorten processing time. Datasets were NSL-KDD and UNSW-NB15.

## 4 Classification Algorithms

Market analysis, science exploration, production control, and other applications can all benefit from the retrieved data. One of the key principles in the machine learning method is classification algorithms. They're used to sort unlabeled data into different categories. The following are the algorithms that were employed in the work:

Support Vector Machine (SVM): When compared to other algorithms, SVM is one of the most reliable classification algorithms in machine learning since it offers a rapid and easy prediction process. It creates a hyperplane that separates the class labels into their associated classes by classifying data points based on support vectors in a data source.

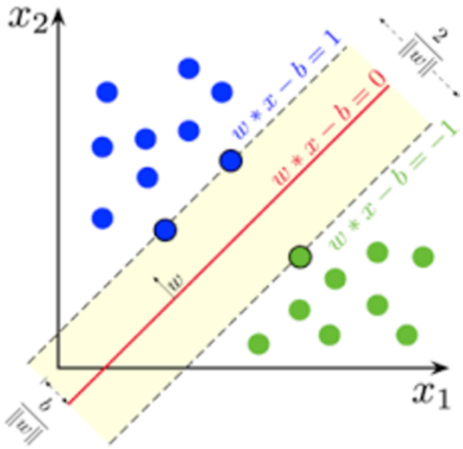

**Fig. 3.** SVM

K-Nearest Neighbor (KNN): is another reliable classification algorithm used for classifying data classes. One of its promising features is that it can be used for both classification regression purposes.

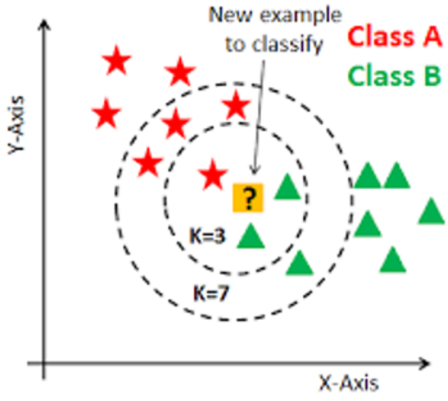

**Fig. 4.** NSL-KDD details

Naïve Bayes (NB): They are capable to forecast the probability that whether the given model fits to a particular class. It is based on Bayes' theorem. It constructed on the hypothesis that for instance, for a given class, the attribute value is independent to the values of the attributes. This theory is called Class Conditional Independence.

$$P(H/X) = P(X/H).P(H)/P(X) \quad (1)$$

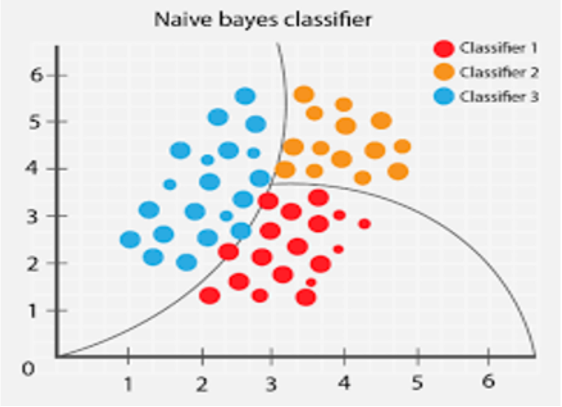

**Fig. 5.** Naïve Bayes

## 5 Methodology

Comparative analysis done between SVM KNN and Naïve Bayes for classification of dataset, to analyze their accuracy. At first raw dataset is taken and the class attribute contains 19 different types of attack which get labeled under 5 categories. They are normal, Dos, Probe, r2l and u2r. Figures

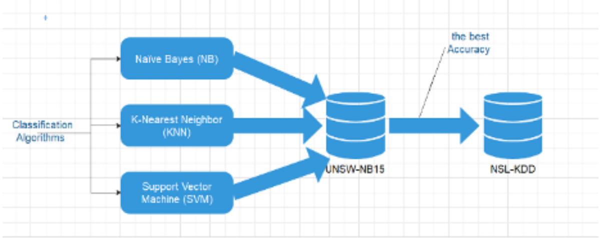

**Fig. 6.** Processes of testing

**Table 3.** Accuracy rate of attacks (UNSW-NB15)

| | Accuracy |
|---|---|
| KNN (k=3) | 93.3333% |
| NB | 95.55555% |
| SVM | 97.77777% |

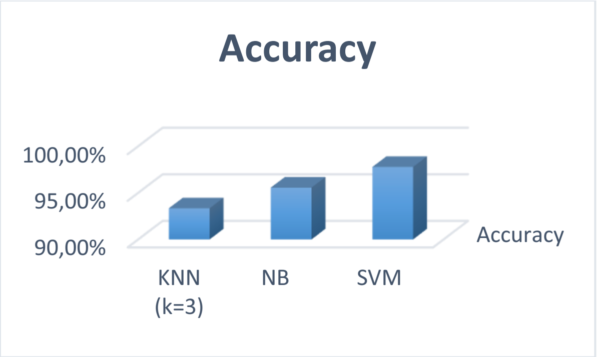

**Fig. 7.** Comparison of classifiers' performance on UNSW-NB15

**Table 4.** Accuracy rate of attacks (NSL-KDD).

| | Accuracy<br>Rate on NSL-KDD |
|---|---|
| SVM | 97,29 |
| Naïve Bayes | 67,26 |

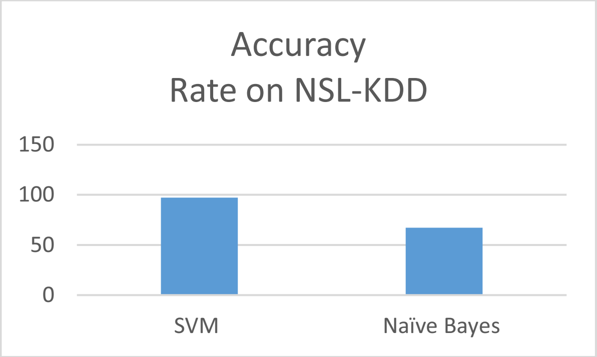

**Fig. 8.** Comparison of classifiers' performance on NSL-KDD

The SVM algorithms showed a better accuracy again with a similar type of DATASET that they have deferent attacks and size

## 6 Conclusion

In this search the first database is treated by the three algorithms SVM NB and KNN with neighbourhood of 3. The elimination of KNN is made following the weak result obtained then the second database is treated by the two other algorithms. SVM has shown a good performance whatever the size of the database or the type of attacks it contains this model will be optimized in future works in terms of processing time and also we will work on its implementation in a firewall and test it in real time

## References

1. M. Belouch, S. El Hadaj, and M. Idhammad. *A two-stage classifier approach using reptree algorithm for network intrusion detection*. International Journal of Advanced Computer Science and Applications, **8**(6), pp.389-394 (2017)
2. M. Belouch, S. El Hadaj, & M. Idhammad. Performance evaluation of intrusion detection based on machine learning using Apache Spark. *Procedia Computer Science, 127,* 1-6,(2018).
3. N. Moustafa, N. (2017). Designing an online and reliable statistical anomaly detection framework for dealing with large high-speed network traffic (Doctoral dissertation, University of New South Wales, Canberra, Australia). (2017)
4. W. Richert, L. P. Coelho, “Building Machine Learning Systems with Python”, Packt Publishing Ltd., ISBN 978-1-78216-140-0
5. M. Bkassiny, Y. Li, and S. K. Jayaweera, “A survey on machine learning techniques in cognitive radios,” IEEE Communications Surveys & Tutorials, **vol. 15**, no. 3, pp. 1136–1159, 2012.
6. A. Iftikhar, M. Basheri, M. Javed Iqbal, A. Raheem; “Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection”, IEEE ACCESS, Survivability Strategies for Emerging Wireless Networks, **6** ,pp.33789-33795, (2018).