Intrusion detection system based M-best features selection in MANET

Rachid KHALLADI, Mohammed REBBAH, Omar SMAIL
dept of Computer Science
University oh Mustapha Stambouli – Mascara Algeria
r.khalladi@univ-mascara.dz , rebbahmed@univ-mascara.dz, o.smail@univ-mascara.dz

***Abstract*— MANETs networks are mobile units interconnected between them without infrastructures which make them vulnerable to different types of attacks. Although several techniques have been proposed as an endeavour to remedy this issue, they are still insufficient. In this work, a technique based on machine learning, more precisely on the random forest algorithm with the selection of the best features, is proposed. The latter is tested on the NSL-KDD dataset. The results found were very satisfying in terms of Accuracy 99,625%, Precision 99,85 %, Recall 99,83 % and F1-Score 99,84%. Thus, the results have improved when compared with those of other techniques.**

***Keywords*—** *Machine learning, Intrusion detection, NSL-KDD, Random Forest, Most Best Features Selection*.

I. Introduction

The growing evolution of computing, particularly in the field of networks and communication has given rise to several types of networks. The ad-hoc mobile networks (MANET) are the networks of today, given their simplicity, especially with the absence of a fixed infrastructure. Security is a serious issue in this type of network since attacks can come from inside or outside and influence traffic control or traffic data [1]. However, several research studies have been under taken for the sake of creating a robust and efficient Intrusion Detection System (IDS). Among the modern techniques used for the creation of these, the adoption of the classification of data is very interesting, particularly the use of Machine Learning.

This technique requires two things; first, a Dataset (for training and testing) and a classifier to make proper sense of the huge amount of data available in the dataset.
Supervised and unsupervised methodologies are used to make a classification of data. In addition, in the supervised method, known as predictive, all the possible classes are known in advance unlike the unsupervised method that is known as descriptive or indirect [2].

In this research, we apply of the technique of random forests, with a selection of the most important features, on incoming MANET packets to classify the data in an efficient way in order to properly predict whether it is a packet coming from an attack or not.
The major contribution in this work is to select the most appropriate features (M-best features) by calculating the influence of each feature in the total accuracy. The features that have a great influence will be selected as the best features and will be used in the learning phase.

This paper is organized into six sections. The second section provides a complete background on the algorithm and dataset used and presents related works in this area. The proposed approach is discussed in details in section three. Section four is dedicated for experimentation while section five is devoted to the analysis of the results and compared with others proposed solutions. Section six concludes the results of this paper.

II. Related works

MANET networks are exposed to several types of attacks due to their vulnerability caused by the absence of a fixed infrastructure. On the other hand, several researches have been made to remedy this problem. In what follows a set of this research is presented.

[3] proposes a CFS with Ensemble Classifiers (Bagging and Adaboost) which has high accuracy, high packet detection rate, and low false alarm rate, to improve the Intrusion Detection System (IDS).
The authors built machine Learning Ensemble Models with base classifiers (Random Forest, Reptree and J48). Multiclass and binary classification was done for KDD99 and NSLKDD datasets. All the attacks were deemed an anomaly and normal traffic.
Five major attacks are labled , namely Denial of Service (DoS), Probe, User-to-Root (U2R), Root to Local attacks (R2L), and Normal class attacks.

[4] offers a method for classifying incoming data into MANETs and reducing the data set using the RF / ET technique (Random Forest / Ensemble Tree).

This approach facilitates a specific classification of the large mass of data (unmanageable data).

The evaluation was applied on NSL-KDD dataset and the results indicated that the proposed model achieved a level of accuracy of 86% in the management of data packets in MANETs.

[5] propose an approach for the interception and detection of blackhole attacks in the DYMO protocol. Their method is summarized in three phases: Planting, Detection, and Interception. Several classifiers have been used noting Decision tree, Support Vector Machine, K-Nearest Neighbor, and neural network. The work is simulated in MATLAB and the results revealed that the SVM classifiers gave more precision than the others did.

[6] have proposed a DSR protocol alarm to defend themselves against black hole attacks in MANET networks. It is called AIS-DSR (Artificial Immune System DSR) and employs AIS (Artificial Immune System) which is inspired by the human's immune system mechanism. The objective is to be able to identify the isolated nodes and remove them from the routing in accordance with the behavior of the nodes in the system. In-depth simulations on the ns-2 environment have been developed to evaluate this algorithm. AIS-DSR results are excellent by around 20% compared to DSR in terms of PDR, Throughput and End to end Delay.

[7] proposed two protocols, namely BDD-AODV and Hybrid that were constructed by modifying the original AODV. In the BDD-AODV protocol, each MANET node has three tables: a trusted table including the trusted nodes of the networks, a black table including the malicious nodes and a counting table, which contains all the statistics concerning the RREP. This protocol uses the BDD dataset [8].

The hybrid protocol is a combination of the BDD-AODV protocol and the MI-AODV [9]protocol. The MI-AODV protocol has a confidence field indicating whether the response node is reliable or not. The hybrid protocol has two tables: a trust table and a black table. Moreover, each source node in the network has a counting table.

DPAA-AODV is a technique proposed by [10] for detection and prevention against active attacks such as Black hole and Gray hole. This method has two phases. The first phase (offline phase), contains the following modules: Data selection , Features selection, and detection. In the Features selection module, the Relief F classification algorithm is used on the BDD dataset (choose the most relevant features and eliminate redundancies to increase the detection rate). In the second phase (Online phase), If the previous features are frequently detected for network nodes and exceeded a predefined threshold, this node will be perceived as a Black-Hole node and will be excluded and avoided from the routes.

The classical methods of cryptography have become limited for the detection of intrusion into MANETs. To reach this, several recent searches are based on deep learning to guarantee security and eliminate intruders. To facilitate the choice of a DL algorithm, [11] made a comparison between several of them noting, including CNN, Inception-CNN, Bi-LSTM and GRU. These algorithms were applied on NSL-KDD dataset.

[12] proposed a method based on clustering for the detection and prevention of black hole attacks in the AODV routing protocol in MANET networks.
This model suggests that the physical characteristics are similar for all nodes. The default-trusted nodes are the destination node and the predecessor nodes. Any node dropping half of the total number of packets is considered a black hole node. The cluster head is chosen as a node located in the centre of the cluster.

To detect the exclusive difference between the number of data packets received and transmitted by the nodes, each member of the unit sends a ping once to the head of the cluster. If the defect is perceived, all nodes will mask the contagious nodes of the network. The experiments are done on ns2 by comparing the PDR, Energy, Throughput and End to End Delay.

[2] collected his own dataset by doing simulations on GloMoSim2 for three attacks including Blackhole, Flood, and Packet Loss. Then, he applied on several machine-learning algorithms. Analysis of the results showed that Multilayer perceptrons (MLP), logistic regression (LR), and support vector machines (SVM) have a higher level of detection.

[13] propose a random forest classifier optimized for intrusion detection. Optimization consists in the selection of features. They opted for the use of genetic algorithms for the selection of features. The proposed model is tested on two different Datasets; NSL-KDD and UNSW-NB15. The results obtained in terms of accuracy, F1-Score, Recall and precision are respectively 96.12%, 88.25%, 86.35%, 90.23% for NSL-KDD and 92.06%, 94.27 %, 96.26%, 92.36% for UNSW-NB15. In this work, an optimized random forest classifier is also proposed. This model is based on the selection of the most important features by measuring their impact. The results obtained will be compared in the results and discussions section.

[14] offer an Intrusion Detection System (IDS) using Deep Neural Network (DNN) and Recurrent Neural Network (RNN) for classification with recursive feature elimination to select features. The proposed technique was tested using the NSL-KDD dataset, which provided a high accuracy rate of up to 94%.

Two approaches have been implemented in [15]. The first is to use CNN to reduce the number of NSL-KDD packages then reduce the number of attributes by a selection of relevant characteristics. The selected attributes are 22 of 42.

The second approach is called an adaptive RBF neural network. It is used for the selection of attributes via its objective function.

III. APPROACH

A. *Random Forest Classifier*

It is a technique covering a large part of Machine Learning problems given its simplicity of interpretation and its stability. It generally has good accuracies and can be used for regression or classification tasks.

In Random Forest, there is first the word "Forest" which implies that this algorithm will be based on trees referred to as a decision tree.

A decision tree helps making a decision thanks to a series of questions (also called tests) whose answer (yes / no) will lead to the final decision. On the tree, each question corresponds to a node, that is, a place where a branch splits into two branches. At each node, the algorithm enquires which question to ask.

Random Forests can be made up of several tens or even hundreds of trees. The number of trees is a parameter that is generally adjusted by cross-validation (is a technique for evaluating a Machine Learning algorithm consisting in training and test the model on pieces of the starting dataset).

Each tree is trained on a subset of the dataset and gives a result. The results of all the decision trees contribute to a final answer. Each tree "votes" (yes or no) and the final answer determines which one had the majority vote. (This is called a bagging method) In this way, we build a robust model from several models that are not necessarily so robust.

To conclude, this algorithm is very popular for its ability to combine the results of its trees to obtain a more reliable final result. Its efficiency has enabled it to be used in many areas, including intrusion detection.

The proposed solution can be summed up in two phases before learning: a phase of data normalization and a phase of selection of the best features.

B. *Flow diagram of proposed model*

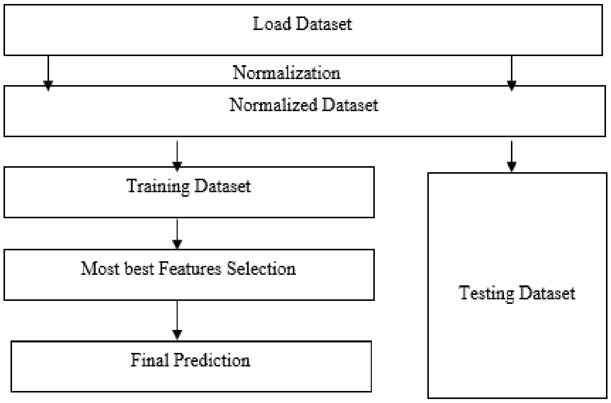

Fig1. Proposed model process.

C. *Features Selection*

There are several techniques for selecting important features. In this study, a popular feature selection method is used. It consists in directly measuring the influence of each characteristic on the accuracy of the model. The general principle is that the value of each feature is permuted, then measured to what extent this permutation decreases the accuracy of the model.

Obviously, for the important variables, the permutation considerably decreases the accuracy of the model. On the other hand, the permutation of the unimportant variables will have no effect or a minimal effect.

The authors propose keeping only the best features (M-best features), and removing the other features which have an impact less than or equal an impact threshold (Delta) on total accuracy.

The delta threshold is used to select the best features by comparing it with the influence of each feature on the total accuracy value. The authors opted for a delta value = 10%.

Algorithm

```
Input : Original set of features   F={ $F_1$ , $F_2$ , ....... $F_n$ }
        Delta  // Threshold
Output : Most-Best Features
Start
  Calculate $\text{Accuracy}_{\text{Total}}$   //  Accuracy Total of dataset
   For i = 1 to n Do
     Permute Value ( $F_i$ )  //  Permute the value of feature i
     Calculate Accuracy ( Permuted value ( $F_i$ ))
     If Accuracy(Permuted value( $F_i$ ))/$\text{Accuracy}_{\text{Total}}$ $\leq$ Delta
          F ← F - { $F_i$ } // Remove the feature that has
                               minimal or no impact
       Else  F ← F  // No feature removed
End.
```

IV. EXPERIMENT

A. *Dataset*

Since 1999, KDD Cup 99 [16] has been used as an example dataset in intrusion detection systems. Each packet (instance) consists of 41 fields and is labeled as normal packet or abnormal packet with attack types, noting 37 are numeric type fields and 4 are non-numeric type fields. KDD99 contains 37 attack types divided into four major classes: DOS, U2R, R2L and Probes .

DOS (Denial of service attaks): are attacks that aim to undermine the availability of services by saturating the resources of the target machine, server or network. These successful attacks in networks have the immediate consequence blocking of network traffic.

Probes: aims to gather information on the susceptible target to help the attacker initiate an attack.

R2L (Remote To Local): these attacks bypass or spoof the authentication settings of a target in order to execute commands. Most of these attacks have come from social engineering.

U2R (User To Root): This type of attack comes from within. The attacker spoofs the password of the super administrator and consequently of other users. Most of these attacks result from the saturation of the buffer caused by the errors of programming.

KDD99 data is full of redundant packets in both training and test data. Redundant data is able to give one type of attack more importance than it deserves. NSL-KDD is an excellent dataset for comparing network IDSs. Our experimentation is carried out with NSL-KDD [17].

B. Normalization

There are several methods of normalization, in this research. The min-max normalization is used. It is one of the most common ways to normalize data. Its principle is quite simple. The minimum value of each feature is transformed into 0 and its maximum value is transformed into 1. Therefore, all other values are transformed into a decimal number between 0 and 1 with the following formula:

$$\frac{Value - Min}{Max - Min} \qquad (1)$$

As an example, if the minimum value of a feature was 10 and the maximum value was 50, then 10 would be transformed into 0, 50 would be transformed into 1 and 30 would become 0.5 (this is halfway between 10 and 50).

After applying this technique, all the features are in the same scope and have the same weighting.

C. Evaluation Metrics

TP (True Positives): cases where the prediction is positive and the real value is indeed positive.

TN (True Negatives): cases where the prediction is negative and the real value is indeed negative.

FP (False Positive): cases where the prediction is positive, but the real value is negative.

FN (False Negative): cases where the prediction is negative, but the real value is positive.

$$\text{Confusion Matrix} = \begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} \qquad (2)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (5)$$

$$\text{F1-Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (6)$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN} \qquad (7)$$

V. RESULATS AND DISCUSSION

A. Impact of data normalization and Features Selection

To test the efficiency of the proposed model, three (3) scenarios were launched on Python using the same dataset, and based on the same algorithm (Random Forest). We applied the random forest classifier directly to the dataset without pre-processing (Normalization). The second scenario consists in normalizing the dataset before the application of Random Forest. Finally, we applied random forest with selection of important features on the normalized dataset. The results obtained are demonstrated in the table 1.

The normalization of the dataset brought about an increase in accuracy of about 7.7%. Thus, the selection of features on the standardized dataset gave an accuracy of 99.66% that is almost perfect in terms of efficient detection. Table 1 indicates that the other metrics noting Precision, Recall and F1-Score have also been improved.

TABLE 1. IMPACT OF NORMALIZATION & MOST BEST FEATURES SELECTION ON RANDOM FOREST CLASSIFIER.

Method	**RF with based Dataset**	**RF with normalized dataset**	**RF with normalized dataset & M-best features selection**
Accuracy (%)	85,92	93,63	99,66
Precision (%)	95	99	99,85
F1-Score (%)	85	95	99,84
Recall (%)	77	91	99,83

B. Comparison

Table 2 reveals a comparison between the proposed model with other models discussed in [10]. It is clear that RF / M-Best Features provide the best results in terms of accuracy, precision, recall, and f1-Score

TABLE 2. COMPARISON OF PROPOSED MODEL(RF/M-BEST) WITH OTHERS.

Method	**RF/M-Best**	**CNN**	**Inc-CNN**	**Bi-LSTM**	**GRU**
Accuracy(%)	99,66	85,99	89,03	84,33	78,98
Precision(%)	99,85	90,90	85,08	93,98	81,08
F1-Score(%)	99,84	85,76	85,33	89,82	84,20
Recall (%)	99,83	81,17	85,58	86,01	87,56

Table 3 and 4 display a performance comparison between the proposed model and other methods based on features selection discussed in [12], [13], [14]. All models are tested on the NSL-KDD dataset. RF / M-Best gives more efficiency compared to others with an accuracy of 99.66%.

TABLE 3. COMPARISON OF PROPOSED MODEL WITH OTHERS TECHNIQUES BASED FEATURES SELECTION.

Method	**RF/M-Best**	**GA-RF**	**SVM/ Fselect**	**DNN(4 LAyer)**
Accuracy(%)	99,66	96.12	98.27	94
Precision(%)	99,85	90.23	97.80	91
F1-Score(%)	99,84	88.25	97.64	92
Recall (%)	99,83	86.35	97.49	77
FPR (%)	1,08	2,91	1,7	6

TABLE 4. COMPARISON OF ACCURACY, FPR AND TIME PROCESSING FOR THE SELECTION OF IMPORTANT ATTRIBUTES.

Method	**RF/M-Best**	**KNN**	**CNN**	**C4.5**	**IBK**	**RBF**
Accuracy(%)	99,66	87,45	95,54	87,95	88,03	94,28
FPR (%)	1,08	3,07	4,64	2,75	3,25	4,64
Processing Time (s)	5,07	165	2,5	12	124	-

VI. CONCLUSION

The use of machine learning in the field of intrusion detection in MANET networks has become a topical issue. However, several techniques have been discussed in this article.

The authors proposed the use of the random Forest Classifier with a data normalization and an automatic selection of the best features. These last ones were carefully selected according to their impact on the total accuracy. Only features that have a strong impact will be selected. The experiments are done on python using the NSL-KDD dataset that gave 99.625%, 99.85%, 99.93% and 99.84% for the Accuracy, Precision, Recall and F1-Score respectively.

In a future work, several perspectives are possible noting the application of RF / M-Best on other datasets, the implementation of the suggested solution in an NS2 / NS3 environment.

REFERENCES

[1] Khalladi, R., Rebbah, M., & Smail, O. (2021). A new efficient approach for detecting single and multiple black hole attacks. *Journal of Supercomputing, 77*(7), 7718–7736. https://doi.org/10.1007/s11227-020-03596-1

[2] Y. Xu, “Research on Intrusion Detection Method of Industrial Internet Based on Machine Learning,” *Journal of Physics: Conference Series*, vol. 1802, no. 4, p. 042029, Mar. 2021, doi: 10.1088/1742-6596/1802/4/042029.

[3] C. Iwendi, S. Khan, J. H. Anajemba, M. Mittal, M. Alenezi, and M. Alazab, “The Use of Ensemble Models for Multiple Class and Binary Class Classification for Improving Intrusion Detection Systems,” *Sensors*, vol. 20, no. 9, p. 2559, Apr. 2020, doi: 10.3390/s20092559.

[4] A. Nayyar and B. Mahapatra, “Effective Classification and Handling of Incoming Data Packets in Mobile Ad Hoc Networks (MANETs) Using Random Forest Ensemble Technique (RF/ET),” *(eds) Data Management, Analytics and Innovation. Advances in Intelligent Systems and Computing*, vol. 1016, pp. 431–444, 2020, doi: 10.1007/978-981-13-9364-8_31.

[5] S. H. Mahin, F. Taranum, L. N. Fatima, and K. U. R. Khan, “Detection and interception of black hole attack with justification using anomaly based intrusion detection system in MANETs,” *International Journal of Recent Technology and Engineering*, vol. 8, no. 2 Special Issue 11, pp. 2392–2398, 2019, doi: 10.35940/ijrte.B1274.0982S1119.

[6] S. Behzad, “An Artificial Immune Based Approach for Detection and Isolation Misbehavior Attacks in Wireless Networks,” *Journal of Computers*, vol. 13, no. 6, pp. 705–720, 2018, doi: 10.17706/jcp.13.6.705-720.

[7] Y. Khamayseh, M. B. Yassein, and M. Abu-Jazoh, “Intelligent black hole detection in mobile AdHoc networks,” *International Journal of Electrical and Computer Engineering*, vol. 9, no. 3, pp. 1968–1977, 2019, doi: 10.11591/ijece.v9i3.pp1968-1977.

[8] M. Yassein, Y. Khamayseh, and M. Abujazoh, “Feature selection for black hole attacks,” *Journal of Universal Computer Science*, vol. 22, no. 4, pp. 521–536, 2016.

[9] Y. khamayseh, A. Bader, W. Mardini, and M. BaniYasein, “A new protocol for detecting black hole nodes in Ad Hoc Networks,” *International Journal of Communication Networks and Information Security*, vol. 3, no. 1, pp. 36–47, 2011.

[10] F. Albalas, M. B. Yaseen, and A. Nassar, “Detecting Black Hole Attacks in MANET using Relieff classification algorithm,” *ACM International Conference Proceeding Series*, pp. 0–5, 2019, doi: 10.1145/3330431.3330454.

[11] S. Laqtib, K. El Yassini, and M. L. Hasnaoui, “A deep learning methods for intrusion detection systems based machine learning in MANET,” *ACM International Conference Proceeding Series*, no. October, 2019, doi: 10.1145/3368756.3369021.

[12] V. K. Saurabh, R. Sharma, R. Itare, and U. Singh, “Cluster-based technique for detection and prevention of black-hole attack in MANETs,” *Proceedings of the International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017*, vol. 2017-Janua, pp. 489–494, 2017, doi: 10.1109/ICECA.2017.8212712.

[13] Z. Liu and Y. Shi, “A Hybrid IDS Using GA-Based Feature Selection Method and the Random Forest,” *International Journal of Machine Learning and …*, vol. 12, no. 2, pp. 1–14, 2019, doi: 10.18178/ijmlc.2022.12.2.1077.

[14] B. Mohammed and E. K. Gbashi, “Intrusion Detection System for NSL-KDD Dataset Based on Deep Learning and Recursive Feature Elimination,” *Engineering and Technology Journal*, vol. 39, no. 07, pp. 1069–1079, 2021, doi: 10.30684/etj.v39i7.1695.

[15] F. Z Belgrana, N. Benamrane, M. A. Hamaida, A. M. Chaabani, and A. Taleb-ahmed, “Network Intrusion Detection System Using Neural Network and Condensed Nearest Neighbors with Selection of NSL-KDD Influencing Features,” *2020 IEEE International Conference on Internet*

of Things and Intelligence System (IoTaIS), pp. 23–29, 2021, doi: 0.1109/IoTaIS50849.2021.9359689.

[16] https://archive.ics.uci.edu/ml/datasets/kdd+cup+1999+data

[17] https://web.archive.org/web/20150205070216/http://nsl.cs.unb.ca/NSL-KDD/