Comparative Analysis of Machine Learning Algorithms Based on the Outcome of Proactive Intrusion Detection System

1 Sivaprasad Abirami, 2 S.Palanikumar
1 Information Technology Department, Noorul Islam Centre for Higher Education, Thuckalay, Kumaracoil-629180, Tamilnadu, India.
   abirami.sivaprasad@sakec.ac.in
2 Department of Information Technology, Noorul Islam Centre for Higher Education, Thuckalay, Kumaracoil-629180, Tamilnadu, India.
   palanikumar@niuniv.com

Received: 22nd August 2020, Accepted: 18th September 2020, Published: 31st October 2020

Abstract
Machine learning plays a vital role in the current world. In the case of an Intrusion detection system to classify the normal and the malicious packet, the machine learning classification algorithms are used. When the live packets are captured and classified without allowing it to enter the network or the host represents a proactive intrusion detection System. Every machine learning algorithm has its pros and cons. This article works on the KDD dataset to classify the captured packets using the machine learning-based classification algorithms of Decision tree and Random forest. The implementation work with Denial of Service attack, Man in the middle attack, and buffer overflow attack. The attack packets are captured and classified using Random forest and decision tree and based on that, the classification algorithms are also compared with Accuracy, precision, and recall parameters.

Keywords
Machine Learning, Packet Classification, Packet Analysis, Data Mining, Network Attacks, Security.

Introduction
The intrusion detection system will prevent unwanted packets to enter the system. The system can be divided into two categories of reactive and proactive IDS. The reactive IDS react after the malicious packet enters the network and damages the network. In the case of reactive IDS, the system classifies the packets and identifies the malicious packets before entering the network [1]. The system explained in this paper is about the proactive IDS and classify the packets using various classification algorithms.

The dataset is one of the famous dataset used in the world of the intrusion detection system. The dataset carries 42 attributes and the attacks are classified into 5 major categories of DoS, Probe, R2L, U2R, Normal. The features, attributes, the attacks are all explained in [2]. The features are selected and based on the attack it’s used in this implementation.

Machine learning is a part of artificial intelligence that makes the system learn and take decisions. For this implementation, a machine learning-based classification algorithm is used which is based on supervised learning. Machine learning can be done in three ways: Supervised learning, Unsupervised Learning, and Reinforcement Learning. Supervised learning works based on training happens with the guidance of a labeled dataset. Unsupervised learning happens through self-learning through previous experiences. Reinforcement learning is the combination of both means for input if a labeled dataset is available then it will act as a supervised learning else as unsupervised learning.

Survey on Classification Algorithms
A survey on the KDD dataset is done and the needed and the unwanted attributes are identified for the research. The survey to identify the best classification algorithm is done with the help of implementing the various classification algorithms using python. In this survey, the unwanted attributes which do not contribute to the attacks are dropped and the remaining attributes are considered for the implementation. For this comparative analysis, the algorithm considered is Random Forest, Decision Tree, Naïve Bayes. To select the classification algorithm, 10% of the KDD dataset is used with 70% of the same as the training dataset and 30% as the testing dataset and the duplication entries in the dataset are also removed. The parameters used for this comparative analysis of the classification algorithms are Accuracy, Precision, Recall, and F-Score. The algorithms are chosen based on the survey carried out in [3][7][8][9].

Proposed System & Implementation
The proposed system collects the incoming packet using wireshark [4] and identifies the feature needed for the proposed implementation and give as the input to the classifier. The classifiers used in this system are based on
Decision tree and random forest. The implantation identifies the following attacks DoS, Buffer overflow, the man in the middle attack. [10]

Steps of the proposed system:
Step 1: Capture the incoming packets
Step 2: Feature selection of packet parameter
Step 3: Classify the packet with the machine learning algorithms.
Step 4: Update the dataset for future decisions.
Step 5: Compare classification algorithms

Random Forest
This classifier has a collection of a Decision tree which improves the accuracy in the prediction with less classification error [5]. Sometimes the accuracy gets affected by providing false decisions for some new inputs, this happens when the random forest creates a noisy tree.

Decision Tree
This classification method is one of the popular machines learning technique for various applications which involve large data and need good decision making. These use a divide and conquer method and each branch of the decision tree represents an output decision [6].

For the implementation to be tested, a dummy website is created to launch the attack which is shown in Fig.1. This dummy website is designed to show the attack and which helps the system to generate attack packets and the packets are collected to classify and to update the dataset for the upcoming prediction.

The GUI console which is shown in Fig.2 is used by the user to mention the URL to launch the attack; in this case, it’s the URL of the dummy website address. This GUI guides the user to start & stop the attack, to capture the packets for the classification, to detect the type of attack, and also to show the graph which compares the classification algorithms.

Once the dummy website URL link is inserted and clicked on “Start Attack” as shown in Fig.3. A response code is received which shows that the DoS attack has been launched.
To test whether the DoS attack has been launched, the Dummy website is reloaded which provides an error as “this site can’t be reached” as shown in Fig. 4 which means it’s unable to load the Site. The error message represents that the attack is launched successfully.

In Fig. 5 the captured packet details after the attack launch are shown. The packets are captured and the features are selected based on the dataset to update the dataset for the future classification.

Similarly, the buffer overflow is launched and the data packets are captured and maintained in the dataset as shown in Fig. 6 and Fig. 7.
Now comes the Man in the middle attack which is shown in Fig.8 and Fig.9.

Fig 6: Buffer Overflow Attack

Fig 7: Buffer Overflow Packets

Fig 8: Man in the Middle Attack Panel
As per the attack launched and the packet captured the classification algorithms detect the attacks which are shown in Fig.10.

The comparative analysis of the classification algorithm based on Precision, recall, and accuracy are shown in Fig.11.
Conclusion
The system is tested under three attacks under two classification algorithms of Random forest and Decision tree. As per the analysis both the algorithms are providing almost equally good prediction. Random Forest is slightly better than the Decision Tree. The attacks involved in this study are Denial of Service, Man in the Middle, Buffer Overflow attacks. As the future scope, more attacks can be tested with a more updated feature of the dataset and classified with more algorithms to get a better outcome.

References
[1] Abirami Sivaprasad  Department of Information Technology, Shah and Anchor Kutchhi Engineering College “Secured Proactive Network Forensic Framework” IEEE CTCEEC 2017.
[2] P. Gifty Jeya, Department of Computer Science and Engineering Sri Venkateswara College of Engineering, Chennai, “Efficient Classifier for R2L and U2R Attacks”, International Journal of Computer Applications (0975 – 8887), Volume 45– No.21, May 2012.
[3] Abirami Sivaprasad Department of Information Technology, Shah and Anchor Kutchhi Engineering College “ Packet analysis based on Machine learning-based Classification Algorithms” International Conference On Networks, Image, And Security, April 2019.
[4] Abirami Sivaprasad Department of Information Technology, Shah and Anchor Kutchhi Engineering College “Machine learning-based traffic classification using statistical analysis” International Journal on Recent and Innovation Trends in Computing and Communication March 2018.
[5] Nabila Farnaaz and M. A. Jabbar, MJCET Hyderabad, India” Random Forest Modeling for Network Intrusion Detection System” Twelfth International Multi-Conference on Information Processing-2016 IMCIP-2016.
[6] Himani Sharma and Sunil Kumar “A Survey on Decision Tree Algorithms of Classification in Data Mining”, International Journal of Science and Research (IJSR) 5(4), April 2016.
[7] Tavallaee, Mahbod & Bagheri, Ebrahim & Lu, Wei & Ghorbani, Ali. (2009). A detailed analysis of the KDD CUP 99 data set. IEEE Symposium. Computational Intelligence for Security and Defense Applications, CISDA. 2. 10.1109/CISDA.2009.5356528.
[8] S. Zargari and D. Voorhis, "Feature Selection in the Corrected KDD-dataset," 2012 Third International Conference on Emerging Intelligent Data and Web Technologies, Bucharest, 2012, pp. 174-180, doi: 10.1109/EIDWT.2012.10.
[9] Staudemeyer, Ralf & Omlin, Christian. (2014). Extracting salient features for network intrusion detection using machine learning methods. South African Computer Journal. 10.18489/sacj.v52i0.200.
[10] FarouqAliyu, TarekSheltami, Elhadi M. Shakshuki,” A Detection and Prevention Technique for Man in the Middle Attack in Fog Computing”, The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2018), https://doi.org/10.1016/j.procs.2018.10.125