An Intrusion Detection System Using Machine Learning Algorithm

Chibuzor John Ugochukwu, & E. O Bennett
Department of Computer Science,
Rivers State University,
Port Harcourt.
Nigeria.
ugochukwuchibuzor@gmail.com, bennett.okoni@ust.edu.ng

Abstract
Security of data in a network based computer system has become a major challenge in the world today. With the high increase of network traffic, hackers and malicious users are devising new ways of network intrusion. In other to address this problem, an intrusion detection system (IDS) is developed which will detect attacks in a computer network. In this research, the KDDCup99 Test datasets is analyzed using certain machine learning algorithms (Bayes Net, J48, Random Forest, and Random Tree) to determine the accuracy of these algorithms by classifying these attacks into their various classes. A constructive research methodology is adopted throughout this research. The experimental results show that the Random Forest and Random Tree algorithms appear to be the most efficient in performing the classification technique on the Test dataset. The experimental tool used is WEKA which is used to perform a correlation based feature selection on the dataset with a Best First search method, and the parameters used for the computation are Precision, Recall and F-measure.

Keywords: *Intrusion detection system; KDDCup99; Machine learning; WEKA*

1. Introduction

Network intrusion have predominantly increased following the rapid growth of network or internet technologies in different areas of social networking, e-learning, e-business etc. this has made the security of data from malicious Hackers more challenging. An Intrusion Detection System is an application used for monitoring the network and protecting it from the intruder. With the rapid progress in the internet based technology new application areas for computer network have emerged (Kabiri & Ghorbani, 2005). An intrusion detection system (IDS) can be classified into Network-bases IDS, Host-based IDS and Application-Based IDS.

Network-based IDS: Network intrusion detection system gathers information directly from a network and performs auditing on the attacks in the network as packets travels in the network. This type of IDS grants users the privilege to specify its signature.

Host-based IDS: Host based IDS views the sign of intrusion in the local system. For analysis they use host system's logging and other information. Host based handler is referred as sensor (Bace, 1998).

Application-based IDS: will check the effective behavior and event of the protocol. The system or agent is placed between a process and group of servers that monitors and analyzes the application protocol between devices (Karthikeyan & Indra, 2010).

2. Review of Related Work

The authors of this research (Perez et al., 2017) demonstrated the use of a hybridized machine learning models with the expectation of showing the capability to the job of intrusion detection in a computer network. The research used the training and testing versions of the NSL-KDD datasets in other to illustrate the effectiveness of the model against known and unknown entries in the model. This work made use of Neural Network (NN) and Support Vector Machine (SVM) algorithms for the supervised learning, K-Means algorithm for the unsupervised learning and PCA and GFR for feature selection on the datasets.

The authors of this paper (Noureldien & Yousif, 2016) used seven machine learning algorithms to perform a supervised technique on the NSL-KDD dataset using WEKA as their desired data mining tool. The algorithms used to carry out this experimental work are: PART, Bayes Net, IBK, Logistic, J48, Random Committee and Input Mapped.
This paper authored by (Vijayarani & Sylviaa, 2015) was centered on the overview and the major importance as it relates to intrusion detection system (IDS). The study gave a general insight on the major types of intrusion detection system, the attack types, diverse domains, attack tools and IDS lifecycle.

The authors in this paper (Nalavadel & Meshram, 2014) used the Apriori algorithm alongside the association rules to solve the intrusion detection issues. This research applied an evasion technique in other to detect new attacks using information gotten from the set of known attacks in the datasets. This framework was actually done using the KDDCup'99 datasets, a widely used and known signature Apriori algorithm was applied to these datasets to detect intrusion.

3. Methodology and Design

Research methodology is the study of how a specific research project is been carried out using some laid down techniques or approach. It can also be seen as the scientific study of how a research problem is solved.

3.1 Research Methodology

This research adopted a Constructive research methodology. The Constructive research method is mostly used in software engineering and computer science research by constructing diagrams, models, plans.

3.2 Design Methodology

The design methodology used in this research is a hybridized design method which combines the Top Down design approach and the Object Oriented design approach. The purpose for the top down design approach is to follow the TCP/IP 5 layer architecture in other to analyze the requirements of the new system. The top down approach gives an opportunity to troubleshoot a system when a layer in the TCP/IP suite is having a problem. However, the object-oriented technique is used to specify the classes and objects of a system and the relationship between them.

3.3 System Architecture

System architecture or systems architecture is the conceptual model that defines the structure, behavior, and more views of a system (Jaakkola & Thalheim, 2011).

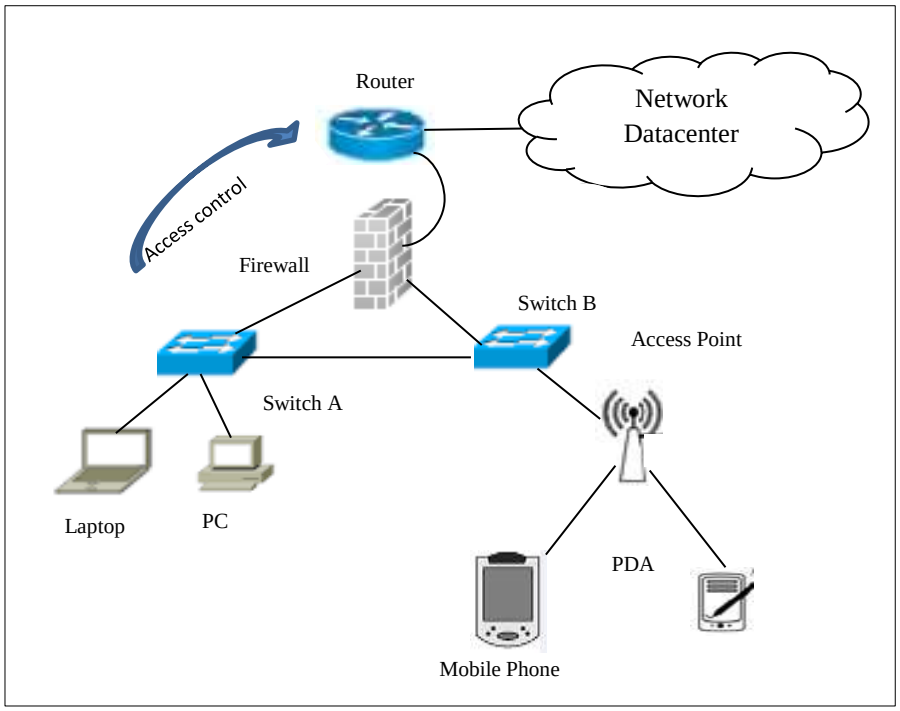

Figure 1: System Security Architecture

In figure (1), the end nodes (PC, laptop, mobile phones and PDA) which are the major entry point of intrusion are connected to the switches either using a wired or wireless technique. Both switches are connected to the firewall which create a block or unblock mechanism in other to protect the system from unauthorized access; the firewall is then connected to the router via a gateway.

3.4 Machine Learning Techniques

Machine learning is programming computers to optimize a performance criterion using example data or past experience (Alpaydın, 2010). There are several machine learning techniques adopted to predict the attacks in the Test datasets which was used to train the system. These algorithms were used to classify the attacks in other to ascertain an efficient technique in predicting and classifying attacks.

- **i.** Bayes Net
- **ii.** J48
- **iii.** Random Forest
- **iv.** Random Tree

Bayes Net: Bayes Net learns Bayesian networks under the assumptions like nominal attributes and no missing values. These two are completely dissimilar elements for estimating the conditional probability tables of the network (Modi & Jain, 2016).

$P(X_i \dots X_n) = \prod_{i=1}^{n} P(X_i | Parents\ (X_i))$ - - - - - - - - - - - - - - - (1)

In modeling a probability distribution using Bayesian Network, each variable in (1) must be conditionally independent of all its non-descendants in the graph given the value of all its parents.

The J48 (Decision tree) algorithm is WEKA's implementation of the C4.5 decision tree learner. The algorithm uses a greedy technique to induce decision trees for classification and uses reduced-error pruning (Quinlan, 2003). In the process of building a tree in J48 (Decision tree), there is a distinctive attributes which is the internal node of the tree and the branches

between the internal node holds the details of the values that these attributes can assume. The final value which is the classification of the dependent variable emerged from the terminal node. This algorithm has the capability to model or classify both discrete and continuous attributes; and can ignore missing attribute values in a dataset.

Random Tree is an algorithm for constructing a tree that considers *K* random features at each node. This algorithm performs no pruning (Witten & Frank, 2005). In Random Tree, every tree stands a chance of being sampled due to the uniform distribution of the trees. This trees can be in labeled, unlabeled and dendrogram (rooted by definition) forms.

Random forest is a machine learning classifier which consist of a collection of tree structured classifiers

$\{h\,(x, \Theta k), k = 1, \ldots\}$ - - - - - - - - - - - - (2)

From (2) $\{\Theta k\}$ represents random vectors distributed independently identical and each tree has a vote for the most famous class at input *x*. The nature and dimensionality of Θ depends on its use in tree construction (Breiman, 2001).

4. Experimental Results

The experimental analysis of this research was performed using WEKA 3.8 (Waikato Environment for Knowledge Analysis). WEKA is an open source machine learning scripting software which was developed in Java by the Waikato University, New Zealand (Hall et al., 2009). Table (1) shows the distribution of records in different classes for testing dataset used in the experiments.

Table 1: Distribution for Test Dataset

Attack category	Number of Samples
Dos	65776
R2L	490
U2R	35
Probe	1042
Normal	23872
Total	91059

The experiment was conducted using the correlation based feature selection (CFS) with Best First search method in other to remove the irrelevant features from the datasets. The original datasets has 42 attributes including the class label, by performing a feature selection on the attributes using the CFS, the attributes on the dataset is now reduced to 9 as seen in table (2).

Table 2: Feature Selection of Attributes

Correlation based feature selection (using Best First)
service
dst_bytes
num_failed_logins
logged_in
lnum_file_creations
Count
srv_diff_host_rate
dst_host_srv_diff_host_rate
Label

4.1 Performance Evaluation

An intrusion detection system (IDS) is evaluated by the measure of accuracy, detection rate and F-measure. An intrusion detection system should have a very low false alarm.

Precision is the percentage of the total number of attacks that are properly detected. It is measure with the equation below

$$\text{Accuracy(Precision)} = \frac{TP}{TP+FP} \quad \text{---------} \quad (3)$$

Detection Rate or Recall is described as the number of attacks detected by the proposed technique to the total number of attacks truly there (Modi & Jain, 2016).

$$\text{DetectionRate (Recall)} = \frac{TP}{TP+FN} \quad \text{---------} \quad (4)$$

$$F - \text{Measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad \text{---------} \quad (5)$$

True Positive (TP): it is the number of connections that were correctly classified as an intrusion

False Positive (FP): this is the number of intrusion connections that were incorrectly classified as normal

False Negative (FN): this is the number of normal connections that were incorrectly classified as intrusion

Table 3: Precision of Classifiers

Bayes Net	J48	Random Forest	Random Tree	Class
0.994	0.999	0.999	0.999	**DoS**
0.606	0.989	0.974	0.966	**Probe**
0.675	0.882	0.970	0.971	**U2R**
0.979	0.958	0.958	0.951	**R2L**
0.979	0.983	0.984	0.984	**Normal**
0.847	**0.962**	**0.977**	**0.974**	

From table (3), the four machine learning algorithm performed a classification technique against the classes of attacks and it shows that the Random Forest algorithm has the highest precision in classifying the attacks in the class label.

Table 4: Detection Rate (Recall) of Classifiers

Bayes Net	J48	Random Forest	Random Tree	Class
0.993	0.999	0.999	0.999	**DoS**
0.844	0.629	0.645	0.650	**Probe**
0.771	0.857	0.914	0.971	**U2R**
0.961	0.823	0.820	0.820	**R2L**
0.961	0.999	0.999	0.999	**Normal**
0.906	**0.861**	**0.875**	**0.888**	

Table (4) shows that Bayes Net has the highest detection rate or recall followed by the Random Forest algorithm amongst other classifiers in the experiment.

Table 5: F-Measure of Classifiers

Bayes Net	J48	Random Forest	Random Tree	Class
0.993	0.999	0.999	0.999	**DoS**
0.705	0.769	0.776	0.770	**Probe**
0.720	0.870	0.941	0.971	**U2R**
0.970	0.886	0.884	0.881	**R2L**
0.970	0.991	0.992	0.992	**Normal**
0.872	**0.903**	**0.918**	**0.923**	

The result in table (5) shows that the Random Tree algorithm outperforms the other classifiers in carrying out the F-Measure experiment. The Random Tree is followed by Random Forest algorithm in classifying the attacks in the class label.

5. Discussion of Results

This experiment was done using a laptop running Windows 8.1 operating system with 1.6GHz Dual core processor and 2GB of RAM memory. The analysis was performed using four (4) algorithms (Bayes Net, J48, Random Forest, and Random Tree) to carrying out classification technique on the dataset. Feature engineering was performed on the Test dataset using the Correlation-based Feature Selection (CFS) algorithm for attribute selection with Best First search method. A 10 fold cross-validation was performed on the Test dataset using the four aforementioned algorithms.

Table 6: Percentages of Weighted Average of the Four Classifiers

	Bayes Net (%)	J48 (%)	Random Forest (%)	Random Tree (%)
Precision	86.1	96.2	97.7	97.4
Recall	90.6	86.1	87.5	88.8
F-Measure	87.2	90.3	91.8	92.3

Table (6) describes the percentages of the weighted average of the machine learning classifiers that were used to perform the experiment.

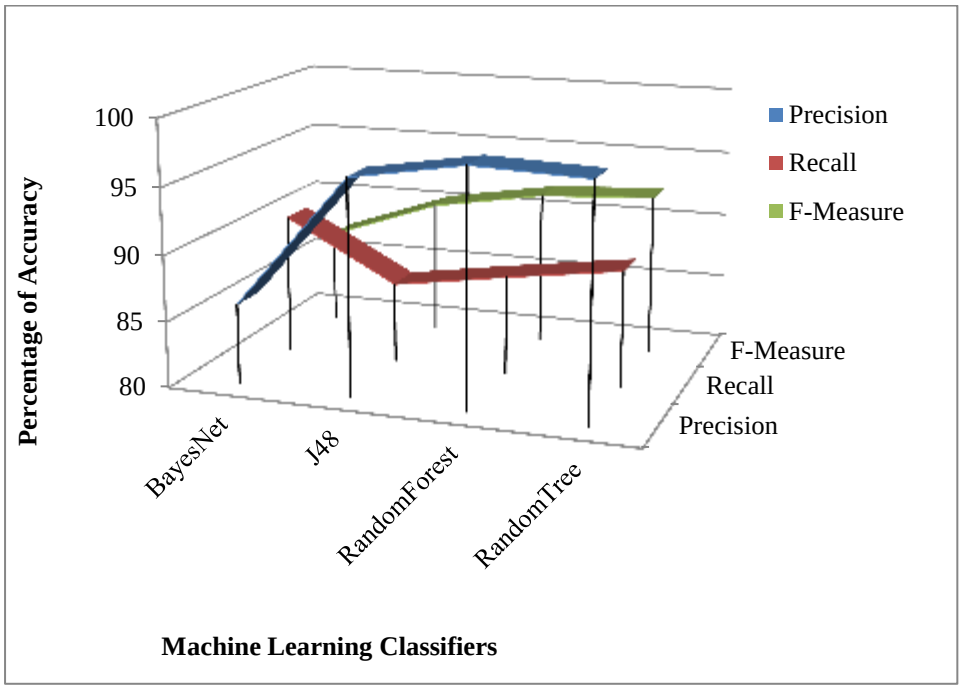

Fig 2: Accuracy of Classification of Four (4) Machine Learning Algorithm

The graph in figure (2) is generated from Table (6); the Y-axis denotes the percentage of accuracy while the X-axis represents the Machine Learning Classifiers. The graph was plotted in other to obtain the percentage of accuracy in the four (4) classifiers. The comparison shows that Random Forest and Random Tree algorithms outperform the other algorithms in their level of precision and F-measure as they are above 97%, while the Bayes Net outperforms the others by its detection rate. However, the Random Forest and Random Tree algorithms are more efficient in performing classification exercise on the Test datasets.

6. Future Works

Future research should consider other machine learning algorithms to ascertain more efficient ways to perform the classification technique on the datasets. It is recommended that further research should be carry out on other parameters that can improve the accuracy of detection

References

Alpaydın, E. (2010). An Introduction to Machine Learning. 2nd Ed.The MIT Press Cambridge, Massachusetts London, England.

Bace, R. (1998). An Introduction to Intrusion Detection & Assessment. Infidel, Inc. for ICSA, Inc

Breiman, L. (2001). Random Forests. University of California Berkeley, CA 94720

Hall, M. Frank, E. Holmes, G. Pfahringer, B. Reutemann, P. & Witten, I. H. (2009). The WEKA Data Mining Software: an update. ACM SIGKDD explorations newsletter 11, no. 1: 10-18

Jaakkola, H. & Thalheim, B. (2011). Architecture-driven modeling methodologies. conference on Information Modeling and Knowledge Bases XXII. IOS Press. p. 98

Kabiri, P. & Ghorbani, A. A. (2005). Research on Intrusion Detection and Response Survey. International Journal of Network Security, Vol.1, No.2, PP.84–102

Karthikeyan, K. R. & Indra, A. (2010). Intrusion Detection Tools and Techniques a Survey. International Journal of Computer Theory and Engineering, Vol.2, No.6, 1793-8201

Modi, U. & Jain, A. (2016). An improved method to detect intrusion using machine learning algorithms. Informatics Engineering, an International Journal (IEIJ), Vol.4, No.2,

Nalavadel, K. & Meshram, B. B. (2014). Mining Association Rules to Evade Network Intrusion in Network Audit Data. International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)Volume-4 Number-2 Issue-15

Noureldien, N. A. & Yousif, I. M. (2016). Accuracy of Machine Learning Algorithms in Detecting DoS Attacks Types. Science and Technology,6(4):89-92DOI: 10.5923/j.scit.20160604.01

Perez, D. Astor, M. A. Abreu, D. P & Scalise, E. (2017). Intrusion Detection in Computer Networks Using Hybrid Machine Learning Techniques. Central University of Venezuela, Caracas, Venezuela

Quinlan, J. R. (2003). C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA

Vijayarani, S. & Sylviaa, S. M. (2015). Intrusion Detection System – a study. International Journal of Security, Privacy and Trust Management (IJSPTM) Vol 4, No 1

Witten, I. H & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition