International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

An Efficient NIDS Using Ensemble Techniques for Multinominal Classification

M. Sreya[1], J. Ganga Bhavani[2], M. Navya Sri[3], M. Bhumika[4], S.R. Shailaja[5]

[1,2,3,4]*B.Tech. Student, Department of Computer Science and Engineering,*

mudumbasreya4@gmail.com, jorrigala.gangabhavani@gmail.com,

navyasrimunukuntla@gmail.com, bhumikamachhaka@gmail.com,

shailaja3793@gmail.com .

[5]*Assistant Professor, Department of Computer Science and Engineering,*

Nalla Malla Reddy Engineering College, Hyderabad, India.

Abstract: **Intrusion detection is one of the major concerns in network administration and security. An intrusion refers to an attack on a system through the network. There is a need to safeguard networks from known vulnerabilities and detect new and unseen, but possible, system abuses. This is done by developing more reliable and efficient intrusion detection systems. The system must detect attacks accurately with the minimum number of false alarms (wrong detections) to be reliable and more efficient. Therefore, a NIDS is developed to improve the accuracy of intrusion detection. Despite high network traffic, this network intrusion detection system efficiently identifies threats. By using various Ensemble techniques and Data mining techniques of classification, the NIDS is developed. In addition to identifying the intrusion, the system also specifies the type of attack. DOS, PROBE, U2R, and R2L attacks can be detected by the developed system. The dataset used is the benchmark NSL-KDD dataset. In the dataset examined, the full NSL-KDD set including attack-type labels and difficulty level is described. Using various attributes of the dataset, by training the model, intrusion detection can be done and the attack can be identified**.

Keywords: Intrusion, NIDS, Host-based NIDS Signature-based NIDS, Ensemble, Probe, Dos R2L, U2R, Multinominal, parameter tuning.

1. Introduction

Network security is currently very important. Thus, intrusion detection is one of the major concerns in network security. There's a need to guard networks against known vulnerabilities. An intrusion detection system (IDS) inspects all inbound and outbound network communications and identifies suspicious patterns that may indicate a network or system attack from someone trying to break into or compromise a system's security. An IDS can also be used to detect indecorous use and policy violations, like a user downloading large quantities of confidential data. The system then cautions administrators about a possible security breach so that they can take action to stop it. There are two types of intrusion detection systems: host-based (HIDS) and network-based (NIDS). HIDS is software installed on an end device that analyses system activities, logs, and events to identify suspicious gestures on that device. NIDS is a security tool that monitors network operations and detects intrusions. It protects the entire network structure. The NIDS can be further classified as(i) Signature-based NIDS: In this, known attack patterns are identified. When network traffic matches the signature patterns, it is considered an intrusion. (ii) Anomaly-based NIDS: In Anomaly-based NIDS, normal behaviour is established, and if there's a deviation from it, it's considered an intrusion. (iii) Hybrid NIDS: A combination of Signature-based NIDS and Anomaly-based NIDS is Hybrid NIDS. Strengths of both networks can be combined to provide a more effective

approach. Various classifier algorithms and ensemble ways can be used for effective discovery. The intrusion detection system can be deployed over a network. It examines network packets and identifies attacks. The four major attacks that cause intrusions are: PROBE, DoS, R2L, and U2R.

2. Literature Survey

Research says that intrusion detection causes a huge loss of money. Hence many efforts were made to detect intrusions. The intrusion detection system is built using Machine Learning techniques. Unsupervised approach was used in intrusion detection. K-means method is used in the unsupervised approach [4]. Improved Genetic K-Means Algorithm is also an unsupervised technique which is proved to be better than K-means [1]. Modified K-means algorithm builds a high-quality training dataset that contributes significantly to improving the performance of classifiers [15]. The dataset used for intrusion detection is NSL-KDD. It is one of the benchmark datasets. The dataset used previously was KDDCUP'99. Statistical analysis on the KDDCUP'99 data set was conducted and resulted in poor anomaly detection evaluation [2]. Intrusion detection can also be done by tree-based classifiers. We can build an effective intrusion detection system using tree-based classification techniques like BF Tree, FT, NB Tree, Random Tree, Random Forest [3,8]. In this paper Towards Near-Real-Time Intrusion Detection for IoT Devices using Supervised Learning and Apache Spark [12], the performances of several machine learning algorithms in identifying cyber-attacks (namely SYN-DOS attacks) to IoT systems are compared. Supervised machine learning algorithms are used that are included in the ML library of Apache Spark, a fast and general engine for big data processing. NIDS was also deployed over the SDN (Software Defined Networks) controller. As NIDS listens to the network and actively compares all traffic against predefined attack signatures, it detects the attacker's scanning attempts [6]. Reinforcement Learning Approach for Anomaly Network Intrusion Detection System has the ability of self-updating to reflect new types of network traffic behavior [13]. Evaluation of machine learning techniques for network intrusion detection [14]. As described in this paper, early research work in this area and commercially available Intrusion Detection Systems (IDS) are mostly signature-based. In this, seven different machine learning techniques were applied with information entropy calculation to Kyoto 2006+ data set and evaluation of performances of these techniques was done. The recent trend is developing the IDS using Machine Learning and Deep Learning [10]. A NIDS developed using ML and DL methods usually involves following three major steps they are: Data pre-processing phase, Training phase, and Testing phase. ML algorithms used for IDS are Decision Tree, K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Mean Clustering, Fast Learning Network, and Ensemble Methods [7,8]. The tuning of the ML model's parameters is a critical topic since it can improve detection quality. The procedure is called Hyper Parameter Optimization [5]. Deep Learning algorithms include recurrent neural networks, auto encoder, deep belief network and convolutional neural networks [7]. DL methods use deep confidence neural network to extract features of network monitoring data, and uses BP neural network as top-level classifier to classify intrusion types [13]. Deep Neural Network Based Real-Time Instruction Detection System, identifies intrusions by analyzing the inbound and outbound network data in real –time. It consists of a deep neural network (DNN) [9].

3. Existing System

In intrusion detection, two types of techniques are employed: anomaly detection and misuse detection, also known as signature detection. While anomaly detection explains the abnormal behaviour pattern, misuse detection focuses on the use of known patterns of unauthorized behaviour. Misuse detection learns the attack patterns in order to detect the type of attack. Generally, the major attacks include Probe, DoS, R2L, U2R. Growing complexity and the number of attacks, machine learning is left out as the only option for building and maintaining an intrusion detection system with the least human intervention. Various unsupervised machine learning techniques like K-means were used. KDDCUP'99 dataset was used which had redundant data. NSL-KDD dataset has overcome all the disadvantages of KDDCUP'99. Anomaly detection is the trend, it studies the normal behaviour of the system and considers any deviation from normal behaviour as an intrusion. Therefore, the probability of false positives is higher when it comes to anomaly detection. Supervised Anomaly detection provides a better detection rate as compared to the unsupervised method. The supervised learning methods that were used comprise of Fuzzy logic, neural network, support vector machine, decision tree, Bayesian network, etc.

4. Proposed System

The proposed system aims to develop a Network Intrusion Detection System that detects abnormal network behaviour. By applying a number of machine learning algorithms, a system can be developed. The model is trained and evaluated for better performance. Firstly, the data is loaded and preprocessing is done to remove outliers and noisy data. The dataset used is NSL-KDD dataset which overcomes the drawbacks of KDD CUP'99 dataset. There will be no redundant data in NSL-KDD like in KDD CUP'99. Exploratory Data Analysis is performed on the data to identify relevant features. Various classifier algorithms are employed, such as decision trees and logistic regression. Using these algorithms, a model can be built. Various ensemble techniques such as boosting and bagging are used. Ensemble techniques refer to combining diverse algorithms to improve system accuracy. The system, along with the detection of attacks, will specify the type of attack. The system can detect Probe, DoS, R2L and U2R attacks.

5. NSL-KDD Dataset

As an improvement over KDDCUP'99, NSL-KDD is a new dataset. With the NSL-KDD dataset, KDDCUP'99's drawbacks have been overcome, making it more efficient than before. Kaggle offers the NSL-KDD dataset. In the NSL-KDD dataset, the selected data file contains attack-type labels and difficulty levels. The NSL-KDD data is advantageous over the original KDDCUP'99 data set for the following reasons: In the train set, redundant records are not included. Record selection from each difficulty-level group is inversely proportional to record selection from the original KDD data set. There are a reasonable number of records in the train and test sets. This will ensure consistency and comparability among research results.

6. Methodology

6.1 Working

i. **Data Preprocessing and exploratory data analysis:** The preprocessing of the data is done and as part of it, outliers are handled. In exploratory data analysis, the data is analysed considering various features. After analysing and observing the data, KBest selects some features for final interpretation.

ii. **Building model:** Various machine learning algorithms are used and models are built. Cross validation is done on the models to find algorithms with greater accuracy. Parameter tuning is done improve the performance of the task. Parameter tuning refers to the process of selecting the optimal values for hyperparameters of a model Description of some of the algorithms is as follows.

a) **Logistic regression:** The logistic regression method is a widely used technique for predicting binary outcomes. Using this method, input variables are mapped to probability values between 0 and 1. Based on network traffic values which are predictor variables, the model predicts the intrusion.

b) **K Nearest Neighbours:** The k-nearest neighbours (k-NN) algorithm is a classification algorithm that assigns a label to a new data point based on the majority label of its k nearest neighbours in the training dataset. The value of k is a hyperparameter that can be tuned to optimize performance. KNN can be effective in detecting anomalies in network traffic by identifying instances that are significantly different from their neighbours.

c) **Discriminant Analysis:** Discriminant analysis is a classification method in machine learning that seeks to find a linear combination of features that maximizes separation between classes. Discriminant analysis can be used in intrusion detection by finding a linear combination of features that maximally separates normal and intrusive instances in a training set.

d) **Decision Tree:** Decision trees are a type of algorithm used for classification and regression problems in machine learning. They work by recursively partitioning the data into subsets based on the values of input features. This model can hence detect the intrusion by the recursive partitioning.

e) **Neural Network Model:** A neural network model is a machine learning model inspired by the human brain structure and function. Neural networks can be used in intrusion detection by learning a non-linear mapping between input features (e.g., network traffic features) and output labels (e.g., normal or intrusive).

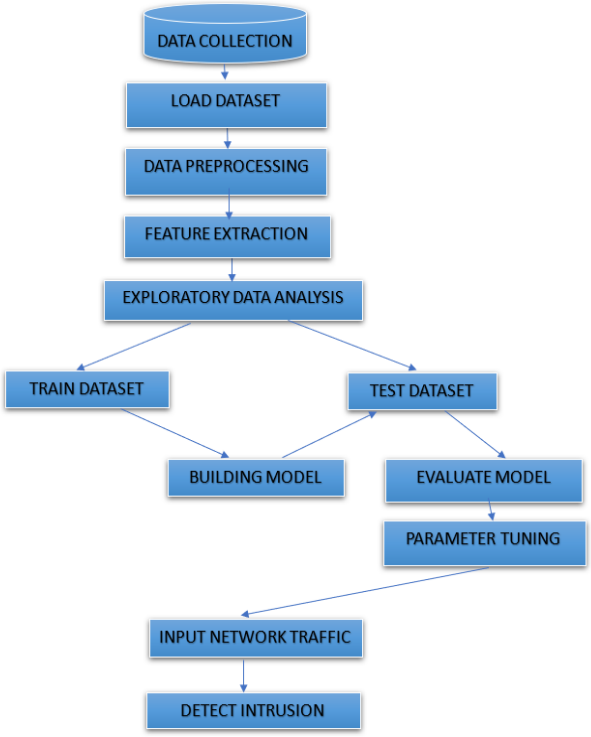
f) **AdaBoost:** Adaptive Boost is a boosting algorithm and a type of ensemble technique. Ensemble refers to combining two diverse algorithms, in order to improve prediction accuracy. AdaBoost combines multiple weak classifiers to create a strong classifier. Therefore, the model will have improved performance.

iii. **Parameter Tuning:** For the improvement of efficiency, the model is tuned. The parameters are changed for a better performance of the model.

iv. **Detection of Intrusion:** The intrusion is detected based on various features. When the network traffic is given as input to the system, it detects abnormality in the system.

6.2 Architecture

Initially, data has been collected from data set. Then, data is preprocessed by removing missing values, handling outliers. Followed by feature extraction, specific features are selected based on variable reduction where K-Best technique is used. Then data set has two categories: Train dataset and Test dataset. By using Train dataset, a Machine Learning Model is using various techniques like – Logistic Regression, K Neighbours Classifier, Decision Tree, Naïve Bayes and Ensemble predictions like boosting algorithms. Then the model is evaluated for test dataset. Parameter tuning is done in order to improve the efficiency of the model. Finally, intrusions are detected, based on the provided input.



7. Software and Hardware used

Software Used:

- Windows OS
- Python
- Jupyter Notebook
- Anaconda
- Flask

Hardware Used:

- Hard Disk – 1 TB
- Memory – 4 GB RAM

8. Conclusion and Future Work

As part of the proposed NIDS, a dataset is taken, preprocessed, and analysed. ML models are built using different algorithms based on the training data. Classifier algorithms such as Decision Trees, Logistic Regression are used and ensemble techniques such as AdaBoost Voting Classifier are employed. The model is designed to classify whether there is an attack in the network. Additionally, it specifies the type of attack among Probe, DoS, R2L, and U2R attacks. In order to achieve optimal accuracy, learning models were trained and parameter-tuned according to network traffic details and configuration parameters. Some

models have achieved a higher level of accuracy than others. The model is limited to intrusion detection. Further, the model can be developed and employed for other websites where networking is crucial. It can be made to notify users directly while communication is going on. In that case, not only detection, but prevention can be made so that the data does not lose its confidentiality, integrity and availability.

References

[1] Network Intrusion Detection Using Improved Genetic k-means Algorithm. S. McElwee, "Active learning intrusion detection using k-means clustering selection", *Conf. Proc.* - IEEE SOUTHEASTCON, 2017

[2] Intrusion Detection Using Tree-Based Classifiers. Ahmim, M. Derdour, and M. A. Ferrag. An intrusion detection system based on combining probability predictions of a tree of classifiers, International Journal of Communication System, vol. 31, pp.1–14, 2018.

[3] A Survey of Intrusion Detection Models based on NSL-KDD Data Set. M. R. Parsaei, S. M. Rostami, and R. Javidan, "A Hybrid Data Mining Approach for Intrusion Detection on Imbalanced NSL-KDD Dataset," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 6, pp. 20–25, 2016.

[4] Intrusion Detection Using Unsupervised Approach. Mirsky Y, Doitshman T, Elovici Y, Shabtai A (2018) Kitsune: an ensemble of autoencoders for online network intrusion detection. arXiv Preprint.

[5] Arindam Sarkar, Hanjabam Saratchandra Sharma & Moirangthem Marjit Singh. A supervised machine learning-based solution for efficient network intrusion detection using ensemble learning based on hyperparameter optimization International Journal of Information Technology volume 15, pages423–434 (2023)

[6] Abdulsalam O. Alzahrani and Mohammed J. F. Alenazi. Designing a Network Intrusion Detection System Based on Machine Learning for Software Defined Networks. Future Internet 2021, 13,111.

[7] Zeeshan Ahmad, Adnan Shahid Khan, CheahWai Shiang, Johari Abdullah, Farhan Ahmad. Network intrusion detection system: A systematic study of machine learning and deep learning approaches.

[8] J. Olamantanmi Mebawondu, OlufunsoD. Alowolodu , JacobO. Mebawondu , Adebayo O. Adetunmbi. Practical real-time intrusion detection using machine learning approaches.

[9] Tavallaee M, Bagheri E, Lu W, Ghorbani A A. Deep Neural network and Real-Time Intrusion detection system.

[10] Abdullah B, Abd-Alghafar I, Salama GI. The Machine Learning and Deep learning methods for intrusion detection system.

[11] Rong Wang, Yuansheng Dong, Juan He, P.R China. The Real-Time network intrusion detection using deferred decision and hybrid classifier.

[12] Valerio Morfino and Salvotore Ranpone, department of law, Economics, University of Sannio, I-82100 Benevento, Italy.

[13] Wang Peng, Xiangwei Kong, Guojin Peng, Xiaoya Li, Zhongjie Wang. Network Intrusion Detection Based on Deep Learning. 2019 International Conference on Communications, Information System and Computer Engineering (CISCE).

[14] Marzia Zaman,Chung-Horng. Evaluation of machine learning techniques for network Intrusion detection.

[15] Wathig Laftah AL-Yasena, Zulaiha Ali Othmana Mohd Zakree Ahmad. Multi-level Hybrid support vector machine and extreme learning machine based on modified k-means for intrusion detection system.