International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 03 | Mar 2024 www.irjet.net p-ISSN: 2395-0072

IDENTIFYING OF SECURITY THREAT IN THE NETWORK USING ML TECHNIQUES

Nelluri Raja Sekhar[2],Bandlamudi Sarath[3],Konakanchi Sai teja[4],Gokarla Syam Sundhar[5],
Mr.B.Kalyan Chakravarthy [6] M.Tech

[2,3,4,5] UG Students, Department of IT,
[6](Associate Professor)
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.

Abstract

Effective methods for identifying malicious activity in computer networks are in greater demand due to the complexity and diversity of cyberattacks becoming more and more complicated. In this paper, a unique machine learning approach to network intrusion detection is presented. We provide a multi-phase system that includes the steps of feature selection, extraction, and classification. The suggested framework analyse network traffic data and looks for patterns of suspicious behaviour using a variety of statistical and machine learning algorithms. Experiments carried out on a real-world dataset show how effective the suggested method is. The findings demonstrate that a variety of network assaults, such as Denial of Service (DoS), Remote to Local (R2L), User to Root (U2R), and probing attacks, may be reliably identified by our method. Additionally, our method performs better than some cutting-edge.

Keywords: XGBoost, LSTM, SMOTE, NSL-KDD, Network Attack, Machine Learning.

1.Introduction

The proliferation of cyber dangers and harmful actions can be attributed to the fast expansion of computer networks and the growing dependence on technology. In order to protect their networks and sensitive data, businesses are now very concerned with identifying and stopping these operations. Intrusion detection systems (IDS) and firewalls, two common forms of network security, are not very good at detecting and neutralizing. As a result, more sophisticated and effective methods of identifying and stopping harmful activity are required. A promising method for identifying and stopping harmful activity in computer networks is machine learning. Large amounts of network traffic data can be analysed by machine learning algorithms, which can then be used to spot trends and abnormalities that might point to possible malicious activity.

Next, network traffic is divided into types based on the model: malicious and regular. The efficacy of the suggested methodology is assessed using a dataset comprising diverse network assaults, showcasing the capability of machine learning to identify and avert malevolent actions within computer networks.

2.Objective

This paper's primary goal is to suggest a machine learning-based method for identifying harmful activity in computer networks. The strategy looks to analyse network traffic data using machine learning techniques in order to spot trends and abnormalities that might point to possible hostile activity. This research specifically aims to develop a machine learning model capable of reliably classifying network data into two categories: harmful and normal.
1. Assessing the suggested method's efficacy using a dataset made up of different network attacks.
2. Evaluating how well the suggested method performs in comparison to more established methods of network security, like intrusion detection systems and firewalls (IDS).
3. Providing information about how machine learning may be used to identify and stop dangerous activity in computer networks.

3.Related Work

Numerous investigations have been carried out to identify malevolent actions within computer networks using the utilization of machine learning methods and the NSL-KDD dataset. In this linked article, we review a few recent research that have employed the NSL-KDD dataset and the XGBOOST and LSTM algorithms to detect harmful activity in computer networks. One study suggested utilizing the XGBOOST algorithm in conjunction with machine learning to identify network assaults. The study trained the model using a variety of features taken from network traffic data. The XGBOOST model was then applied to categorize network traffic as harmful or legitimate.

Another study suggested utilizing the LSTM algorithm in conjunction with deep learning to identify network assaults. In order to train the LSTM model and model the

network traffic data, the study used a time series-based methodology.
Overall, by utilizing the NSL-KDD dataset, this research show how machine learning and deep learning methods can be used to identify harmful activity in computer networks. Future study can investigate the application of additional machine learning and deep learning algorithms for further enhancing the performance of network intrusion detection systems. Both the XGBOOST and LSTM algorithms have demonstrated promising results in detecting various forms of network intrusions.

4.Dataset Description

A benchmark dataset that is frequently used in studies assessing intrusion detection systems is the NSL-KDD dataset. It was developed in response to the shortcomings of the initial KDD Cup 1999 dataset, which had a number of problems such as duplicate entries, an unbalanced class distribution, and erroneous assumptions.
A modified version of the KDD Cup 1999 dataset, the NSL-KDD dataset contains a variety of network attack techniques, including DoS, probing, and user-to-root attacks. There are 41 features in all in the dataset: 7 nominal characteristics and 34 numerical features. With 125,973 instances in the training set and 22,544 instances in the testing set, the dataset is split into training and testing sets.

TABLE I:List of NSL-KDD dataset files and their descriptions

The NSL-KDD dataset consists of several files, including the following:

S.No.	File name	Description
1	KDDTrain+.txt	The training data, comprising 42 columns with 41 characteristics and one class label, and 125,973 instances overall, are contained in this file.
2	KDDTest+.txt	The testing data, comprising 42 columns, 41 characteristics, and one class label, total 22,544 occurrences, are contained in this file.
3	KDDTrain+_20 Percent.txt	For quicker experimentation, a 20% randomly sampled subset of the KDDTrain+.txt file with a total of 25,294 instances and 42 columns is included in this file.
4	KDDTest-21.txt	The testing data for 21 different attack types, including DoS, U2R, R2L, and probing attacks, is contained in this file. This file is used to assess how well intrusion detection algorithms work against different kinds of attacks.
5	KDDTest-10Percent.txt	For quicker testing, this file includes a 10% randomly picked portion of the KDDTest+.txt file, which has 2,255 occurrences and 42 columns overall.
6	KDDTest-21Percent.txt	This file includes a randomly selected 21% subset of the KDDTest+.txt file with 21 different attack types, such as probing, DoS, U2R, and R2L assaults.
7	KDDTest-10Percent-21.txt	For quicker testing, this file includes a 10% randomly picked portion of the KDDTest-21.txt file, which has 2,226 occurrences and 42 columns overall.

TABLE II:Mapping of Attack Class with Attack Type

The NSL-KDD dataset contains several types of attacks. According to the attack targets, can be divided into four categories:

Attack Class	Description
Denial-of-Service (DoS) attacks	By flooding networks with traffic or other kinds of demands, these attacks seek to interfere with the availability of network resources. The NSL-KDD dataset contains a variety of DoS attack types, including ICMP, UDP, and SYN floods.
User-to-Root (U2R) attacks	These exploits target user account vulnerabilities in order to obtain unauthorized access to a system. The NSL-KDD dataset consists of many types of U2R attacks, such as buffer overflow, loadmodule, and perl.
Remote-to-Local (R2L) attacks	These attacks are designed to take advantage of weaknesses in the remote user's account in order to obtain unauthorized access to a system. The R2L attacks in the NSL-KDD dataset include ftp_write, guess_passwd, and imap.

Probing attacks	By sending packets to different ports and protocols, these attacks seek to learn as much as possible about a system in order to find flaws. Numerous prodding attack types, including portsweep, nmap, and satan, are included in the NSL-KDD dataset.

There are twenty-three different types of assaults in the NSL-KDD dataset: four types of R2L attacks, three types of U2R attacks, two types of probing attacks, and fourteen types of DoS attacks. For the purpose of representing actual network activity, the collection also contains examples of typical traffic.

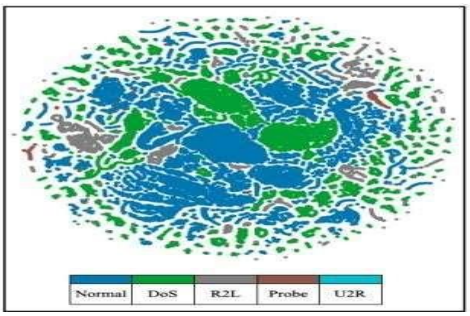

Fig 1. NSL-KDD Dataset

5.System Implementation

The system implementation for malicious activities detection in network typically involves the following steps:

1. **Data preprocessing:** In order to prepare the NSL-KDD dataset for machine learning methods, this stage entails cleaning and preparing it. This could entail turning categorical features into representations, eliminating superfluous features, and distributing the classes evenly.

2. **Feature Selection:** In order to train the machine learning models, this phase entails picking the most pertinent features from the preprocess dataset. This enhances the model's performance and lowers the dataset's dimensionality.

3. **Model training:** Using the preprocessed and chosen features, the machine learning models are trained in this step. XGBoost and LSTM are the two models employed for this system. While LSTM is a kind of recurrent neural network that can handle sequential input, XGBoost is a gradient boosting approach that makes use of decision trees.

4. **Model evaluation:** In this step, the testing dataset is used to assess how well the trained models perform. The models' performance is assessed using a variety of performance indicators, including accuracy, precision, recall, and F1 score. effectiveness in detecting malicious activities.

5. **Model tuning:** In order to maximize the models' performance, this stage entails adjusting their hyperparameters. Hyperparameters, such the learning rate, number of trees, and number of hidden layers, are those that are not discovered during training. To determine the optimal hyperparameters, using grid search or other optimization algorithms.

6. **System integration:** In this step, the trained models are integrated into a wider intrusion detection system. Real-time network traffic data analysis is possible with the models, which can also be used to notify security staff of any questionable activity.

The overall goal of this system implementation is to use machine learning techniques, XGBoost and LSTM, trained on the NSL-KDD dataset, to increase the precision and effectiveness of hostile activity identification in network traffic.

6.Prerequisites

The following are the prerequisites for implementing malicious activities detection in Network:

1. **Python Programming Language:** Python is a well-liked machine learning programming language. It offers a number of frameworks and tools, including keras, pandas, numpy, scikit-learn, and tensorflow. which are essential for implementing machine learning algorithms.

2. **Google Colab Notebook:** An open-source web tool called Google Colab Notebook allows users to create and share documents with live code, mathematics, graphics, and narrative text. It offers an interactive platform for machine learning and data analysis, which facilitates model implementation and NSL-KDD dataset exploration.

3. **Scikit-learn Library:** A Python machine learning toolkit called Scikit-learn offers a number of methods for dimensionality reduction, regression, clustering, and classification. In order to apply harmful activity detection using the NSL-KDD dataset, it also contains tools for feature selection, data preprocessing, and model evaluation.

4. **XGBoost Library:** The gradient boosting algorithm is efficiently implemented by the open-source

software library XGBoost. It can handle big datasets with millions of samples and thousands of attributes and is made to be extremely scalable.

5. **LSTM Architecture:** In order to use the NSL-KDD dataset to develop the LSTM model for malicious activity detection, it is necessary to comprehend the architecture of LSTM and its application in sequence modeling.

6. **Awareness with Machine Learning Concepts:** To perform malicious activity detection utilizing NSL-KDD dataset and machine learning methods, one must have a fundamental understanding of machine learning principles such as supervised and unsupervised learning, feature engineering, model selection, and evaluation.

7.Results

There is no need for preprocessing because all tests were carried out with Google Colab and the data have been cleansed. Eighty percent of the data is split, and SMOTE, XGBoost, and LSTM algorithms are used.

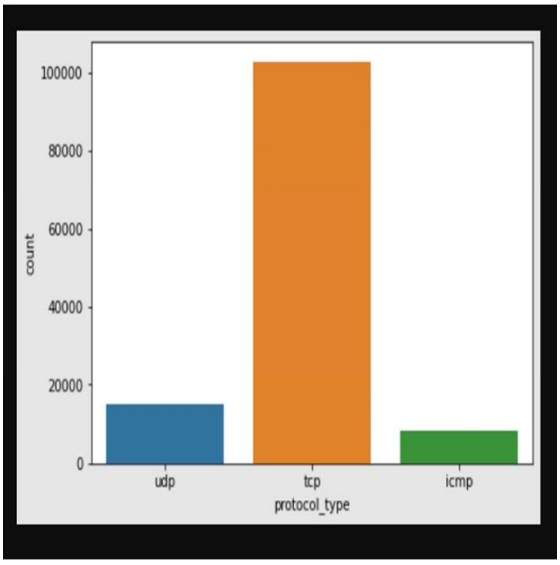

Fig.7.1.Shows the number of protocol types in the NSL-KDD dataset. The dataset consists of 3 different types of protocols: udp, tcp and icmp.

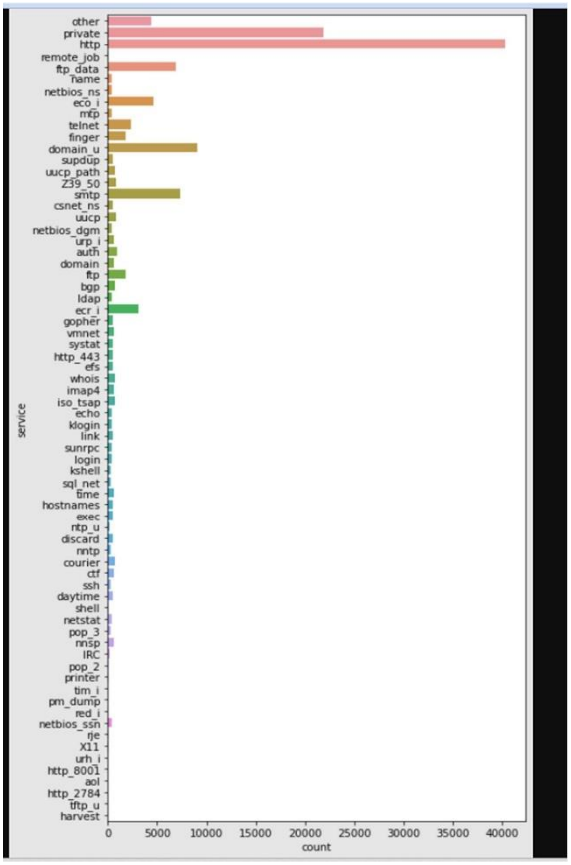

Fig.7.2.service_types of plots

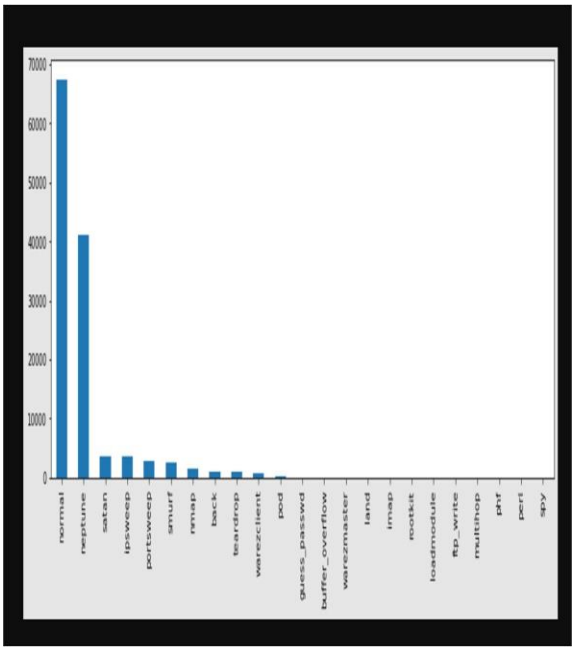

Fig.7.3.attack plot

Confusion Matrix and Classification Report for XGBoost classifier:

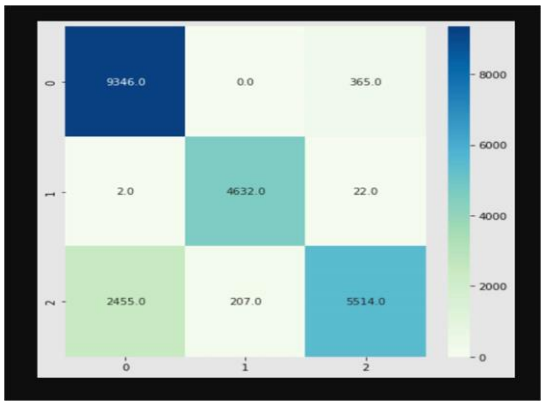

Fig.7.4.Confusion matrix of XG-Boost model

	precision	recall	f1-score	support
0	0.79	0.96	0.87	9711
1	0.96	0.99	0.98	4656
2	0.93	0.67	0.78	8176
accuracy			0.86	22543
macro avg	0.89	0.88	0.88	22543
weighted avg	0.88	0.86	0.86	22543

Fig.7.5.Classification report of XG-Boost model

Confusion Matrix and Classification Report for LSTM Classifier:

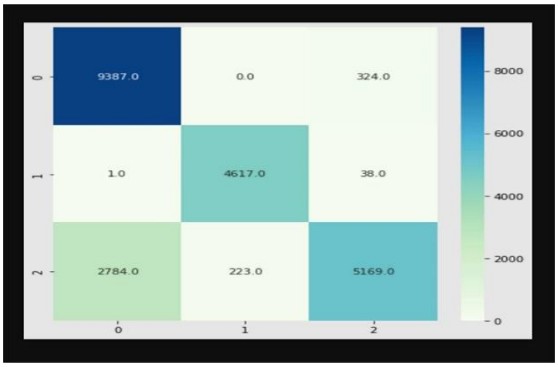

Fig.7.6. Confusion matrix of LSTM model

	precision	recall	f1-score	support
0	0.77	0.97	0.86	9711
1	0.95	0.99	0.97	4656
2	0.93	0.63	0.75	8176
accuracy			0.85	22543
macro avg	0.89	0.86	0.86	22543
weighted avg	0.87	0.85	0.84	22543

Fig.7.7 Classification report of LSTM model

Accuracy comparison of the models:

The model accuracy of XG-Boost and LSTM using the whole test and training set of the NSL-KDD data set is compared in Figure 6.15.

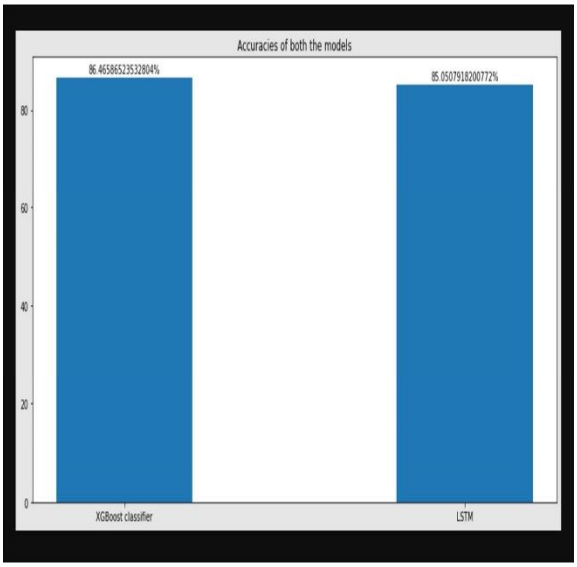

Fig.7.8 Accuracy comparison of models

8.Conclusion

In conclusion, the suggested approach to identify harmful activity in networks through the use of XGBOOST and LSTM machine learning algorithms exhibits encouraging outcomes. Methods for selecting In order to detect many forms of assaults, such as DoS, U2R, R2L, and probing, high accuracy, precision, recall, and F1-score are achieved through the use of features and two different types of models. When it comes to AUC-ROC score and computational efficiency, the XGBOOST model performs better than the LSTM model, however when it comes to identifying temporal dependencies in the data, the LSTM model performs better than the XGBOOST model. The outcomes demonstrate how the two models' complementary qualities can be used to further enhance detection performance.

In order to identify and stop harmful activity and enhance network security overall, the suggested strategy can be implemented in a real-time network environment. With an accuracy rate of higher than the LSTM method, the XGBoost algorithm has a superior classification effect.

9.Future Enhancements

Some potential future enhancements for malicious activities detection using NSL-KDD dataset and machine learning algorithms via XGBOOST and LSTM in network include:

1. **Incorporating real-time data streaming:** The dynamic nature of network traffic is not reflected in the static NSL-KDD dataset. Real-time data streaming integration can assist in identifying and addressing attacks in real-time, lowering the potential damage caused by malicious activities.
2. **Investigating other datasets:** Although the NSL-KDD dataset is frequently used for intrusion detection, other datasets, such UNSW-NB15 and CICIDS2017, can also be utilized to assess how well the suggested method works. Examining additional datasets can assist in verifying the detection system's robustness and generalizability.

10.References

[1] "Combating imbalance in network intrusion datasets," by D. A. Cieslak, N. V. Chawla, and A. Striegel, in Proc. IEEE Int. Conf. Granular Comput., May 2006, pp.732–737.

[2] M. Zamani and M. Movahedi, "Intrusion detection using machine learning techniques," arXiv:1312.2177, 2013. [Online]. http://arxiv.org/abs/1312.2177 is accessible.

[3] The paper "Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs" was presented by M. S. Pervez and D. M. Farid at the 8th International Conference on Softw., Knowledge, Inf. Manage. Appl. (SKIMA), December 2014, pages 1-6.

[4]H. Shapoorifard and P. Shamsinejad, "Introducing a novel hybrid approach integrating an enhanced KNN for intrusion detection," International Journal of Computer Applications, vol. 173, no. 1, pp. 5–9, September 2017.

[5] A novel PCA-firefly based XGBoost classification model for intrusion detection in networks utilizing GPU, S. Bhattacharya, P. K. R. Maddikunta, R. Kaluri, S. Singh, T. R. Gadekallu, M. Alazab, and U. Tariq, Electronics, vol. 9, no. 2, p. 219, Jan. 2020.

[6] A deep learning technique for network intrusion detection system, A. Javaid, Q. Niyaz, W. Sun, and M. Alam, Proc. 9th EAI Int. Conf. Bioinspired Inf. Commun. Technol.