Network Anomaly Detection Using Random Forest Algorithm

Abishek S[1] Sachin V[2] Sanjay Kumar S[3] Ahmed Mudassar Ali[4]
[1,2,3]UG Scholar [4]Assistant Professor
[1,2,3,4]Department of Information Technology
[1,2,3,4]S. A. Engineering College, Chennai, India

Abstract— Our research is to reconcile the complication and serious threat possessed by network intrusion. The term network intrusion or anomaly designates the serious threat created by hackers, high jackers, spoofers, and etc that means the intruders who try to access our system with counterfeit intention. This complexion and threat is faced by all the networks and so to consolidate this complication we make use of an anomaly detection system. The subsisting detection systems work on the basis of rules and limitations. Our exertion provides a detection system which analyses the data packet and improves the accuracy utilizing machine learning random forest algorithm. This examination is extensively superior as it can detect the network anomalies with better accuracy than the existing systems.

Keywords: Machine Learning, Random Forest Algorithm, Anomaly Direction, UNSW NB-15 Dataset

I. INTRODUCTION

This project addresses a broad of supervised machine learning tasks for classification of network anomalies, an approach that has been studied in security even as the volume and variety of cybersecurity threats and internet-based attack vectors proliferates. The more usage of computerized systems has given rise to critical threats such as zero-day attacks, mobile threats etc

The evolution of computer networks has highly computer security concerns, particularly internet security in today's networking surroundings and advanced computing facilities. Although Internet Protocols were not designed to place a more priority on security issues, network admins have to handle a more variety of anomaly attempts by both individuals with malicious intent and large botnet system.

According to internet security report, In 2018 there were more than 30million attacks per year. 2013 - hackers hacked a Yahoo's three billion email accounts gaining access to a sensitive customer information.2014 -Verizon's Data Breach Report, 63,437 security breaches stolen by hackers.2015 - Russian primarily based Cyber-attack cluster connected in scarf email and positive identification mixtures of over 117 million customers.2016 - Uber reportable that hackers stole the data of over 57 million riders and drivers.2017 - Hackers hacked a 412 million user accounts were stolen from Friendfinder's sites.2018 - Facebook code was exploited by hijackers and 50 million user accounts compromised.2019 - Captical One recently had over 106 million records taken containing personal and financial datas.2020 - Hackers stolen 500,000 Zoom passwords available for sale in dark web crime forums.2020 - Twitter breach scam created attackers swindle $121,000 in Bitcoin through nearly 300 transactions.2020 - 1001 cyber-attack cases are repored

II. RELATED WORK

Many Machine learning algorithms square measure develop network-based intrusion detection systems for classification and detection of cyber-attacks in network. As attacks square measure dynamical dynamically in term of volume and strategies, thus still existing system Researchers square measure ready to achieve good detection accuracy within the same, however the False Positive rate, low accuracy of those systems continues to be a serious drawback in network anomaly. They have not shown an in detailed performance analysis numerous of varied of assorted metric capacity unit algorithms on various data sets.

Hebatallah Mostafa Anwer et al used totally different methods for feature choice victimization filter and wrapper strategies. Naive Thomas Bayes and J48 machine learning algorithms were accustomed classify UNSW-NB15 dataset. They found half of 88% accuracy of the dataset.

Roshan, (2018) mentioned AN adaptational style of internet detection systems supported Extreme Learning Machines (ELM). The NSL-KDD dataset was applied for the analysis.it had been found that it will find novel attacks at the side of familiar attacks with a suitable rate of detection and false positives.

Mustapha Belouch, conducted AN experimental analysis to match performance of 4 commonly used classification algorithms like Naive Thomas Bayes, SVM, call Tree and Random Forest on Apache Spark huge information surroundings to guage detection time, building time and prediction time of classifiers, the task of classification algorithms was to classify network traffic information as traditional and attack. They used UNSW-NB15 information set for the analysis and over that Random Forest algorithmic program was outperformer with relevancy accuracy (97.49%), specificity (97.75%), sensitivity (93.53%) and execution time among all the opposite tested algorithms. Dipali Gangadhar Mogal, designed a central purpose algorithmic program with relevancy Association Rule Mining (ARM) as a way of feature choice to adopt the relevant options from the UNSW-NB15 and also the KDD-99 datasets. They used central purpose algorithmic program to cypher attribute values of central points, conjointly applied apriori algorithmic program to pick out options victimization association rule mining (ARM). supply Regression and Naïve Thomas Bayes was accustomed value datasets, they analyzed the lead to term of detection accuracy with relevancy time interval and located that CP with apriori is way higher than apriori alone and conjointly they compared the two datasets victimization each strategies. The performance on UNSW-NB15 was far better than KDDCup99.

III. PROPOSED WORK

All network detection system aims to provide zero-day attack. very low researchers has been done on the subject of network anomaly detection using Random Forest algorithm. In last year, researchers have developed many anomaly detection systems. which are able to provide good accuracy and quality Anomaly Detection. Researchers are able to achieve good detection accuracy in the same, but the False Positive rate, low accuracy of these systems is still a major problem in network detection. So the proposed system tries to increase the accuracy rate of the Anomaly detection by refining the tuning of parameters in Random Forest algorithm. Proposed system can achieve 80-90% of accuracy on Anomaly detection.

IV. SYSTEM ARCHITECTURE

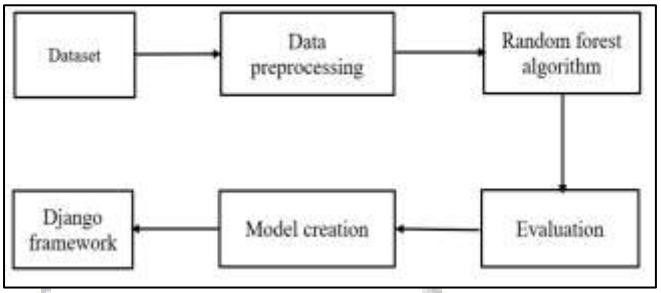

Fig. 1: System architecture diagram

V. OVERVIEW OF RANDOM FOREST ALGORITHM

In general, anomaly detection system ought to generate low false-positive rates and wear down (i) an outsized quantity of data for coaching and prediction; (ii) unbalanced information sets; (iii) an outsized range of features; and (iv) categorical and continuous options.

In this state of affairs, Random Forest models area unit appropriate as a result of they need the subsequent blessings. when compared with alternative usually known Machine-Learning methods: (i) low coaching time complexity, and quick prediction; (ii) resilience to wear down unbalanced datasets; (iii) embedded feature choice methodology and intrinsic metrics to rank options by importance; and (iv) able to deal natively with categorical and continuous features. These blessings area unit objectively discovered once the Random Forest models area unit submitted to comparative tests within the realm of anomaly detection system.

VI. RANDOM FOREST

It is planned by Breiman in 2001.This methodology is predicated on the proximity search and may be used each for regression and classification. Random Forest is AN ensemble methodology, that predict supported the results of a group of call Trees. it's a choice tree-based classifier. during this technique, The Random Forest algorithmic program combines the output of multiple (randomly created) call Trees to come up with the ultimate output. Random Forests offer smart accuracy and area unit appropriate to be utilized in feature choice approaches. We imply that the proximity perform is wont to implement a bunch Random Forest method for intrusion detection, equally to the network classification approach planned by Wange (2013). In such AN application, Random Forest models is used for categorical options and do not would like information standardisation and have choice, that area unit blessings over the usually used bunch approaches.

Panda et al. (2012) propose the utilization of Ensemble of Nested Dichotomies (END) (Dong et al. 2005)and a Random Forest model because the base classifier to find intrusions in networks with high detection rates and low false-positive rates. They evaluated the strategy on the NSL-KDD dataset and presented notable results for the planned approach compared to alternative thought-about ensembles. Bilge et al. (2012) propose the analysis of network flows (NetFlows) to spot botnet commandand management servers in massive scale. The used feature set springs from the NetFlows' variables andis the input of a Random Forest model used as a classifier. The alarm candidates area unit finally filtered by AN approach mistreatment some in public accessible informatics white-lists and black-lists to cut back the false positive rates. The authors performed tests mistreatment information from 2 real networks summing altogether1.6 billion flows per day and that they achieved values for the realm beneath the Curve (AUC) larger than0.6 in some cases. The work deals not solely with the detection methodology, however with the total resolution compounded by the subjects: information grouping, information transformation, detection methodology, processing architecture, performance analysis, and evasion resilience analysis.

Tesfahun and Bhaskari (2013) propose the utilization of SMOTE, to cut back the imbalance, and a Random Forest model as a classifier to find intrusions on networks

Zhang and Wang (2015) planned the utilization of Hadoop and driver (software) and a Random Forest model as a classifier to find intrusions on websites. The authors tested the strategy mistreatment the KDD99 information set and with data generated by a non-specified IDS. each datasets were used for training and testing.

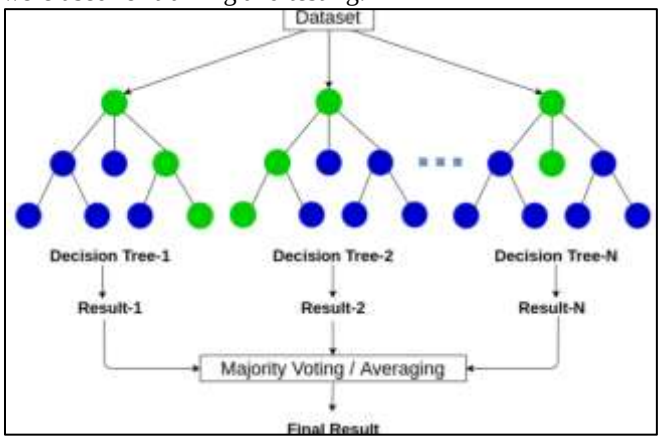

Fig. 2: Ranfom forest diagram

VII. METHODOLOGY

A. Dataset Description

There are many data sets available for research purposes, commonly used data set are KDD98, KDDCUP99, NSLKDD.

In this paper, we use UNSW-NB15 Dataset to detect network anomaly.UNSW-NB15 data set was created by the IXIA Perfect Storm tool in the ACCS. ACCS is Cyber vary work of the Australian Centre for Cyber Security. This knowledge set goal of generating this data set was to have a data set with a hybrid real trends events along with artificial contemporary attack behaviors. Tcpdump tool was utilized to capture the raw network traffic for capturing 100 GBs of raw data. Due to attack exploitation, the Ixia tool was designed in such the simplest way to simulate one attack per second throughout the primary simulation and ten attacks per second throughout the second simulation.

S.No	Type of attributes	Name of attributes	Sequence No.
1	Flow	Scrip, Sport, Dstip, Dsport, Proto	1-5
2	Basic	State, Dur, Sbytes, Dbytes, Stl, Dtl, Sloss, Dloss, Service, Sload, Dload, Spkts, Dpkts	6-18
3	Content	Swin, Dwin, Stcpb, Dtcpb, Smeansz, Dmeansz, trans_depth, res_bdy_len	19-26
4	Time	Sjit, Djit, Stime , Ltime, Sintpkt, Dintpkt, Tcprtt, Synack, Ackdat	27-35
5	General Purpose	is_sm_ips_ports, ct_state_ttl, ct_flw_http_mthd,is_ftp_logm, ct_ftp_cmd	36-40
6	Connection	ct_srv_src , ct_srv_dst , ct_dst_ltm, ct_src_ ltm ,ct_src_dport_ltm,ct_dst_sport_ltm,ct_dst_src_ltm	41-47
7	Labelled	attack_cat, Label	48-49

Table 1: Description of UNSW NB-15 Dataset

VIII. DATA PRE-PROCESSING

Data preprocessing is an integral step in machine learning classifiers as the good quality of data and the more information that can be delivered directly affects the ability of our model to learn, therefore it is extremely more important that we preprocess our data before feeding it our models. The process of cleaning raw data it to be used for machine learning activity is known as data preprocessing. All the null values and duplicate records in the dataset are handled. Categorical data is converted in to numerical form with the help of label encoder. Preprocessed data is separated as training and testing.

IX. EVALUATION AND METRICS

Evaluating a model is a core part of building effective machine learning model. Confusion matrixes are generated for each mechanism for machine learning classifiers. it shows classification problem and the size of table depends on the number of classes included in a particular data set. this matrix helps us to compare actual and predicted labels on dataset. The terms True Positive and True Negative implies correctly predicted conditions and similarly True positive rate and False positive rate. Accuracy, It is an ensemble method, meaning random forest model is made up of a large number of small decision trees estimators. which each produce their own predictions. this metric is used for reflecting overall success rate of an IDS.

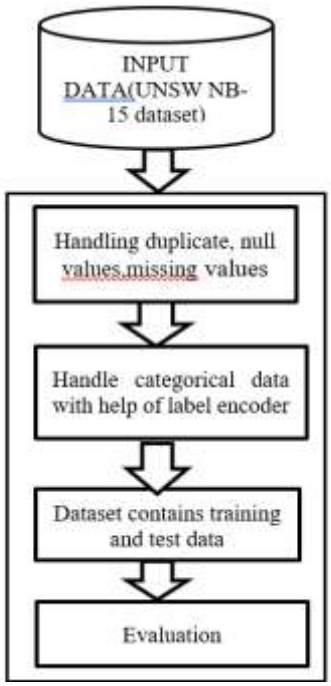

Fig. 3: Data pre-processing diagram

X. ANALYSIS

The selected ML classifiers namely Random forest algorithm are tested on UNSW-NB15 dataset, the dataset for intrusion detection. The experimental work is done on Intel Core i5, 8GB RAM using Python. After performing a data pre-processing, dataset is divided into two parts: training and testing. performance is evaluated on the basis of several parameters. It gives accuracy of selected classifiers. RF classifier out performs the other methods in terms of accuracy 82%, recall 0.98 and F1-score 0.84.it shows the macro avg 0.86 and weighted avg 0.87 in the selected group of classifiers**.**

XI. CONCLUSION

The experimental work has been carried out to evaluate the performance of the selected Machine learning classifiers based on proposed taxonomy namely Random forest algorithm for detection of Anomaly. These classifiers are tested on UNSW-NB15 data-set. The classifiers are compared on the basis of F1-Score, accuracy, TPR and FPR. The results of the algorithm rule shows that Random forest algorithm is better accuracy than other Machine learning algorithm. The accuracy of Random forest classifier comes out to be 82% accuracy. In selective parameters and multiclass classification for detection of Anomaly.

REFERENCES

[1] Mohammad future, this work can be extended for kazim hooshmand, Doreswamy (2019). Machine learning based network Anomaly detection. IJRTE,ISSN:2277-3878.
[2] Sarmah, A. (2001). Intrusion detection systems: Definition, need and challenges.
[3] Raouf, Mohammadsalahuddin, Noura limam, Sara ayoubi, Nashid Shahriar, Felipe Estrada- Solanoand

Oscar M. Caicedo. A comprehensive survey on machine learning for networking: Evolution applications (2018). Boutaba at. ol. journal Of internet service applications.

[4] Omer, K. A. A., & Awn, F. A. (2015). Performance Evaluation of Intrusion Detection Systems using ANN. Egyptian Computer Science Journal, 39(4).

[5] Diro, A. A., &Chilamkurti, N. (2018). Distributed attack detection scheme using deep learning approach for Internet of Things. Future Generation Computer Systems, 82, 761-768.

[6] Khan, J. A., & Jain, N. (2016). A survey on intrusion detection systems and classification techniques. Int. J. Sci. Res. Sci., Eng. Technol., 2(5), 202-208.

[7] Narudin, F. A., Feizollah, A., Anuar, N. B., &Gani, A. (2016). Evaluation of machine learning classifiers for mobile malware detection. Soft Computing, 20(1), 343-357.

[8] Belavagi, M. C., &Muniyal, B. (2016). Performance evaluation of supervised machine learning algorithms for intrusion detection. Procedia Computer Science, 89, 117-123.

[9] Rifkie primartha, Bayu adhi tama. Anomaly detection using random forest(2017).(ICODSE conference 2017).

[10] Geeta Kocher and Gulshan Kumar. Performance analysis of machine learning classifiers for intrusion detection using unsw--nb15 dataset(2020).AIRCC publishing(2020).

[11] Paulo Angelo alves, Andre costa(2019) A survey of random forest based methods for intrusion detection system. ACM Comput. Surv. 51, 3, Article 48 (May 2018)

