Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

Classification of Network Intrusion Attacks Using Machine Learning and Deep Learning

P. Roshni Mol [1*], Dr. C. Immaculate Mary[2]
[1] Ph.D. Research Scholar, Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu, India.

[2]H.O.D & Associate Professor, Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu, India.

[1]roshnimphil@gmail.com, [2]cimmaculatemary@gmail.com

Abstract

Modern era is loaded with data. Increase in usage of smart phones results in enormous amount of data generation. Highly sophisticated smart cities enabled with Internet of Things (IoT) devices produces large data. Sensitive data such as personal information, health information, financial information may be vulnerable and it's integrity could be lost. Highly voluminous data travelling to and fro through the network may encounter traffic. This network traffic can be either normal traffic or it may be intrusion created by the hacker to hack by introducing abnormal traffic over the network. Traditional Intrusion detection systems and firewalls may detect the attacks based on the signature pattern. This is not sufficient to detect the advance persistent threats or to detect unknown attacks. To identify and to classify various types of unknown attacks, it is essential to apply intelligent techniques. This paper aims to classify attacks like DoS (Denial of Service), Probe, R2L (Remote to Local), U2R (User to Root) which causes intrusion in the network. To identify and to analyze root cause of intrusion, a benchmark dataset named NSL-KDD (Network Security Laboratory- Knowledge Discovery and Data) is used. Detailed analysis of the NSL-KDD dataset is accomplished by using machine learning and deep learning. Four models are opted to perform comparative analysis. In the first model, Principal Component Analysis (PCA) is applied to minimize the dimension of data and machine Learning algorithms like logistic Regression, Random forest Classifier, Decision Tree Classifier are utilized to build the model. In the second model, algorithms like logistic Regression, Random forest Classifier, Decision Tree Classifier, Adaboost, and XGBoost are used. In the third model, multi-label classifier chains algorithm is applied to deploy the model. In the fourth model, deep neural network is used to accomplish deep learning model. The motive of this research is to find the best classifier that classifies data with high accuracy and to develop a model which serves the best for intrusion detection system. Comparative analysis of classifier algorithms is done and it is evident that, in the first model, Random forest algorithm produces 98.7% accuracy, in the second model, Adaboost algorithm produces 99.8% accuracy, in the third model, Multi label classifier chain based on random forest produces 99.6% accuracy and in the fourth model, deep neural network produces 99.2% accuracy. Among the four models, it is found that, Adaboost is the best algorithm which classifies and produces best results.

***Keywords*— NSL-KDD, Machine Learning, Deep Learning, Multi label classification, Intrusion detection**

1. INTRODUCTION

In this digital era, lots of data are transferred to and fro from source to destination. Due to COVID-19 (Coronavirus) pandemic situation, educational field as well as other fields are drawn to adapt online mode for learning and working. An article from Hindustan Times [1] reports that, Department of Telecommunications had collected data between 22 March 2020 and 28 March 2020 and states that Indians consumed internet on an average of 307 Petabytes of data. Most of the people started to watch videos and it has raised the traffic by 30%. It is essential to check the quality of the traffic. Technology evolves day by day to fulfil and to serve human needs but on the other end, dark side of the technology also evolves itself rapidly. Cyber security plays a major role for an organization or an enterprise to ensure its security within and outside the network. Intruders may intrude the network to collect sensitive information, to flood the network with malicious traffic. To handle these intruders, it is a pre requisite to set up an Intrusion

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

Detection System (IDS). James P. Anderson, of United States Air force [2] submitted a report on "Computer Security Threat Monitoring and Surveillance" in 1980. Based on this report, first IDS was developed. IDS collect traffic data and segregates to normal or malicious traffic. Data such as network traffic, system log, application log, security log can be collected for analysis. Proprietary or open source tools can be opted to gather streaming data. Proprietary tools like Fortinet, Sourcefire ,TippingPoint, etc.,are available. Bro-IDS, Samhain Labs, OpenDLP, OSSEC, Snort, Suricata,etc., are some of the freely available tools. Simulation tools can be used to mimic the network attacks and it is processed for analysis. Hackers or intruders may generate malicious traffic that leads to Denial of Service attack, probing attack etc. For a secured organization only Firewall and SEIM (Security Events and Information Management) is not sufficient. An IDS is required to handle intrusion in the network. After data acquisition, data analysis can be performed that is based on Anomaly-based IDS or Misuse-based IDS. Malicious traffic may result in denial of service attack, probing attack, privilege escalation attack, reconnaissance attack etc. IDS can be categorized into Network IDS, Host IDS, Network Node IDS, Protocol IDS, Wireless IDS, Network Behavior Analysis (NBA).There are various approaches for intrusion detection system. It can be statistics, rule, pattern, state and heuristic based detection. After the detection of intrusion or attack, alert will be generated. Existing Intrusion Detection System works well for already known attacks based on the signature and pattern of the particular attack. It is essential to develop IDS and IPS to handle advance persistent threats. This can be achieved through various machine and deep learning algorithms.

2. MACHINE LEARNING FOR INTRUSION DETECTION

Machine learning algorithms are categorized into supervised and unsupervised algorithms. If the label is available and if the target feature is known, then supervised algorithm can be applied. If the label is unknown and if the target feature is not available, unsupervised algorithms can be applied. Support Vector Machine (SVM), Artificial Neural Network (ANN), Navie Bayes, Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Random Forest are some of the supervised algorithms. Supervised algorithms perform classification. Unsupervised algorithms perform clustering. K-means clustering, Hierarchical clustering, K-NN (k nearest neighbors) are some of the unsupervised algorithms.

2.1. DATA ACQUISITION

Data can be collected from sources like Host and Network side. Host-based IDS detects the unusual traffic flow in the particular host system. Network-based IDS detects the malicious traffic among the systems which are in network. DARPA1998, KDD99, NSL-KDD, UNSW-NB15, CTU-13 etc., are some of the benchmark intrusion datasets which are available online [3]. Benchmark datasets for intrusion are inadequate. It is essential to create new datasets for research. In this paper to recognize and to categorize intrusion attacks like DoS, Probe, U2R and R2L, NSL-KDD [4] dataset is utilized.

2.2. EXPLORATORY DATA ANALYSIS

Once data is collected, it is essential to perform exploratory data analysis to find the correlation between the attributes or features. Outlier can be identified using visualization, so that it can be eliminated to get accurate results. Overall summary of the data can be acquired using the statistical functions like count, Count, Mean, Standard Deviation, Minimum Value, Maximum Value,25^{th} , 50^{th} and 75^{th} Percentile.

2.3. DATA PRE-PROCESSING

Data pre-processing is a necessary process in machine learning life cycle. Inconsistencies and duplicates in data are removed in this step. Feature scaling, feature transformation and feature selection is performed while pre-processing the data. Yuyang Zhou [5] proposed Bat algorithm based on correlation, as feature selection technique by applying C4.5, random forest classifiers. Senthilnayaki [6] used Information gain as feature selection for KDD cup 99 dataset. Shailendra Sahu[7] used PSO(Particle Swarm Optimization) with Gradient decision tree for feature selection. Selvakumar B et al. [8] used PCA for feature selection to process KDD cup 99.

In this paper feature scaling is done using Standard Scaler, Label Encoder and Label Binarizer. Standard scalar makes use of z-score normalization so that all values are converted to a specific range. Label Encoder is used to handle categorical attributes by assigning a value to the labels. One hot encoding is done for categorical features. Label Binarizer replaces categorical values with array of numbers. PCA is

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

used to minimize the dimension of the dataset.

2.4. MODEL GENERATION AND VALIDATION

Y.Bouzida [9] used SVM, multilayer perception classifier on NSL-KDD, ISCX and Kyoto 2006 datasets. Mohamoud M [10] used combined binary PSO, Standard PSO and SVM to process NSL-KDD dataset. S. Balakrishnan [11] applied SVM with Rule based classification on NSL-KDD and achieved better accuracy. Ripon Patgiri [12] used Random forest and SVM along with Recursive feature elimination technique.

In this paper, supervised algorithms like decision tree, adaboost, xgboost, logistic regression, random forest are used to create machine learning model which classifies different attacks into different classes. A classifier chain is applied to deal with multi label classification.

3. DEEP LEARNING FOR INTRUSION DETECTION

Deep learning can be used to detect anomalous traffic. Deep learning algorithm learns by itself and finds the anomalous traffic from normal traffic. Basic units of neural network are neurons, neuron weights, activation function, network of neurons, input, hidden and output layers. In this paper basic neural network is used for analysis. Keras is used to create deep neural network, relu and softmax are passed as the activation functions. Chuanlong Yin , Yuefei Zhu, Jinlong Fei, and Xinzheng He [13] proposes Recurrent Neural Network based Intrusion Detection System (RNN-IDS) and detected intrusion with minimal false positive count. DBN based IDS [14] was developed with 97.5% accuracy and it is compared with SVM, DBN accuracy.

4. RELATED WORKS

NSL-KDD dataset was analyzed using WEKA [15] in this paper. J48, SVM and Naive Bayes classification algorithms are implemented and have concluded that J48 classifier classifies with better accuracy.

Algorithms like Support Vector Machine, Random Forest, Gaussian Naive Bayes and Logistic Regression are applied for analysis [16]. Among them, Random Forest Classifier classifies NSL-KDD with 99% accuracy.

Hybrid model [17] was developed for NSL-KDD. After normalization, chi-square method was adapted for feature selection and features with low rank are removed. SVM classifier is used to train, test and validate and it produced results with 98% accuracy.

SVM based Intrusion detection model [18] was developed. Information gain feature selection with particle swarm optimization was adapted in order to propose FS-PSO-SVM model and obtained 99.8% accuracy.

Principal Component Analysis [19] was applied to minimize the size of the dataset and it was found SVM works well with high accuracy.

An intelligent intrusion detection system [20] with two-stage hybrid classification was developed. SVM and ANN were used to detect anomaly and misuse. SVM produced 98.72% result and ANN produced 86% result.

4.1. DATASET DESCRIPTION

KDD cup 99[21] dataset contains more number of duplicates. Redundant records will produce inaccurate results while processing. NSL KDD dataset is the processed version of KDD cup 99 dataset and it does not contain duplicates. This dataset is widely for research. There is a lack of availability in benchmark intrusion datasets. This dataset contains both train and test data. 42 features are available and 42nd feature is used to classify the data into normal or attack. Features can be categorized as basic, content, traffic and host [22]. Table I represents the features and its data type. Basic features are duration, protocol_type,service, src_bytes, dst_bytes, flag,land, wrong_fragment and urgent. Content features are hot,num_failed_logins, logged_in, num_compromised, root_shell, su_attempted, num_root, num_file_creations, num_shells, num_access_files, num_outbound_cmds,is_host_login and is_guest_login. Traffic features are count, serror_rate, rerror_rate, same_srv_rate, diff_srv_rate, srv_count, srv_serror_rate, srv_rerror_rate and srv_diff_host_rate. Host features are dst_host_count, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate and

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

dst_host_srv_rerror_rate.

Table I. Features and its Types of NSL-KDD Dataset

S.No	Features	Data Type
1	Duration	int64
2	protocol_type	object
3	Service	Object
4	Flag	Object
5	src_bytes	int64
6	dst_bytes	int64
7	Land	int64
8	wrong_fragment	int64
9	Urgent	int64
10	Hot	int64
11	num_failed_logins	int64
12	logged_in	int64
13	num_compromised	int64
14	root_shell	int64
15	su_attempted	int64
16	num_root	int64
17	num_file_creations	int64
18	num_shells	int64
19	num_access_files	int64
20	num_outbound_cmds	int64
21	is_host_login	int64
22	is_guest_login	int64
23	Count	int64
24	srv_count	int64
25	serror_rate	float64
26	srv_serror_rate	float64
27	rerror_rate	float64
28	srv_rerror_rate	float64

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

29	same_srv_rate	float64
30	diff_srv_rate	float64
31	srv_diff_host_rate	float64
32	dst_host_count	int64
33	dst_host_srv_count	int64
34	dst_host_same_srv_rate	float64
35	dst_host_diff_srv_rate	float64
36	dst_host_same_src_port_rate	float64
37	dst_host_srv_diff_host_rate	float64
38	dst_host_serror_rate	float64
39	dst_host_srv_serror_rate	float64
40	dst_host_rerror_rate	float64
41	attack_type	Object
42	difficulty_level	int64

Features like protocol_type, service,flag and attack_type are categorical data. Land,logged_in, root_shell, num_outbound_cmds,is_host_login and is_guest_login are binary data. Wrong_fragment, urgent, su_attempted, count,srv_count, serror_rate, srv_rerror_rate, same_srv_rate, diff_srv_rate, srv_diff_host_rate, dst_host_count, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate and dst_host_rerror_rate are discrete data. Duration,src_bytes,dst_bytes,hot,num_failed_logins,num_compromised,num_root, num_file_creations, num_shells and num_access_files are continuous data.

4.2. ATTACKS

Train data contains 22 types of attacks which are shown in table II. Test data contains 37 types of attacks which are shown in table III. Table IV represents the class of attacks. NSL-KDD dataset contains following types of attacks.
Denial of Service (DoS): Users will be restricted to access services.
Probing: Gaining access to the system through the weak point is probing.
User to Root (U2R): Illegal access to the root.
Remote to Local (R2L): Accessing local system's data remotely without authorization.
To develop effective IDS, it is essential to identify new attacks, and there is a need to prepare new attack signature.

Table II. Total Number of Attacks and its Types in Train data

normal	67343
neptune	41214
satan	3633
ipsweep	3599
portsweep	2931
smurf	2646
nmap	1493
back	956

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

teardrop	892
warezclient	890
pod	201
guess_passwd	53
buffer_overflow	30
warezmaster	20
land	18
imap	11
rootkit	10
loadmodule	9
ftp_write	8
multihop	7
phf	4
perl	3
spy	2

Table III. Total Number of Attacks and its Types in Test data

normal	9711
neptune	4657
guess_passwd	1231
mscan	996
warezmaster	944
apache2	737
satan	735
processtable	685
smurf	665
back	359
snmpguess	331
saint	319
mailbomb	293
snmpgetattack	178
portsweep	157
ipsweep	141
httptunnel	133
nmap	73
pod	41
buffer_overflow	20
multihop	18
named	17
ps	15
sendmail	14
xterm	13
rootkit	13
teardrop	12
xlock	9
land	7
xsnoop	4
ftp_write	3

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

perl	2
sqlattack	2
loadmodule	2
udpstorm	2
phf	2
worm	2
imap	1

Table IV. Classes of Attacks

Dos	**Probe**	**R2L**	**U2R**
Back	ipsweep	ftp_write	buffer_overflow
Land	nmap	guess_passwd	Loadmodule
Neptune	portsweep	Imap	Perl
Pod	satan	multihop	Rootkit
Smurf	mscan	Phf	Ps
Teardrop	saint	Spy	Xterm
apache2	worm	warezclient	Sqlattack
Mailbomb		warezmaster	Httptunnel
Processtable		snmpguess	
Snmpgetattack		Named	
Udpstorm		sendmail	
		Xlock	
		Xsnoop	
		Worm	

Table V. Frequency Check for Train Data

Normal	67343
DoS	45927
probe	11656
R2L	995
U2R	52

Table VI. Frequency Check for Test Data

Normal	9711
DoS	7636
probe	2574
R2L	2423
U2R	200

Frequency checks for both train and test data is performed. It is evident from table V and table VI, that DoS attack is high in both train and test data.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

5. MODEL SETUP

The analysis of data to categorize attacks is done using Python. As the first step, necessary library packages are imported; next NSL-KDD dataset is imported. Basic exploratory data analysis of dataset is done. Frequency table is constructed for continuous features. Histogram is generated for the features to find the distribution of attacks. Feature Transformation is done by feature encoding, feature scaling, label encoding and label binarizer. Figure 1 shows the importance of the features. All features are assigned with score based on the correlation of the particular feature with other features. Among the features, same_srv_rate contributes more with score 0.36 , service_ecr_i contributes with score 0.10 etc., Feature with highest score is the most important feature. Visualization of dataset is done using PCA. It is difficult to visualize the entire data because of its size. To overcome this problem PCA is applied and four PCA components are generated to visualize data using scatter 3D plot. Figure 2 represents data based on first three components PCA1,PCA2 and PCA3. Figure 3 represents data based on PCA2, PCA3 and PCA4 components. Each color in the 3D plot represents each attack. This paper deals NSL-KDD dataset with four models. From these four models, the best classifier which is suitable to handle NSL-KDD dataset is identified with high accuracy, precision and F1Score. Figure 6 shows four models which are used in this paper.

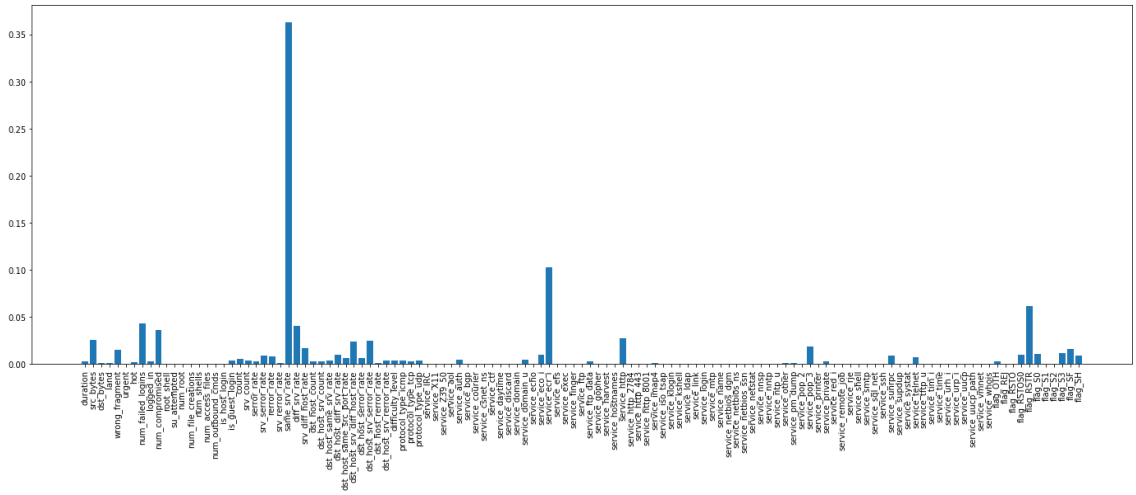

Fig. 1 Feature Importance Plot

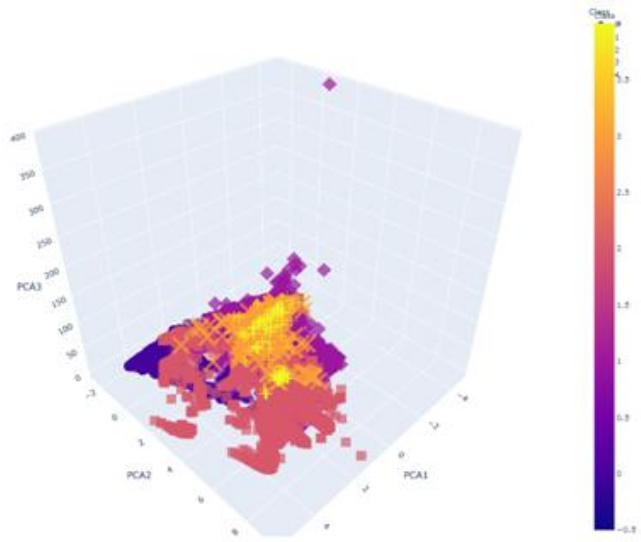

Fig. 2 Scatter 3D Plot of NSL-KDD Dataset for Principal Components 1, 2 and 3

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

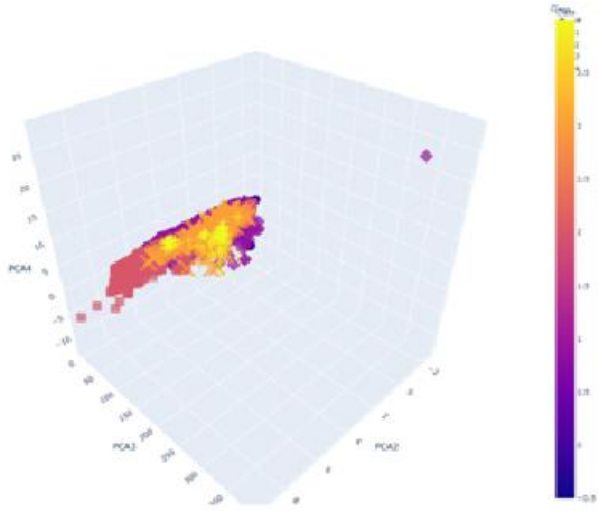

Fig. 3 Scatter 3D Plot of NSL-KDD Dataset for Principal Components 2, 3 and 4

5.1. LOGISTIC REGRESSION

The name logistic regression[23] may represents regression but it is a classification algorithm. Logistic Regression estimates discrete values. It can classify data into 0 or 1, yes or no, true or false. The output values always lies between 0 and 1. Classification can be either binary or multi class classification. If the target variable is categorical, logistic regression can be applied.

$Y=mx+c$ (1)

This is the equation of a straight line. Here m represents slope, x represents data points and c represents intercept. Slope is used for stretching the curve and intercept is used to move up and down through the graph.

The sigmoid or logistic function is denoted as

$$\sigma_{(z)} = \frac{1}{1+e^{-y}} \quad (2)$$

$$\sigma_{(z)} = \frac{1}{1+e^{-(mx+c)}}$$

If $-y=\infty$ then $\sigma(y) = 1$

If $y=-\infty$ then $\sigma(y) = 0$

If $y=0$ then $\sigma(y) = 0.5$

There are various methods for denoting equation for a straight lie

$h(\theta(x) = \theta_0 + \theta_1 x$ (3)

or

$h(\theta(x) = \beta_0 + \beta_1 x$ (4)

or

$y=w^T x+b$ (5)

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

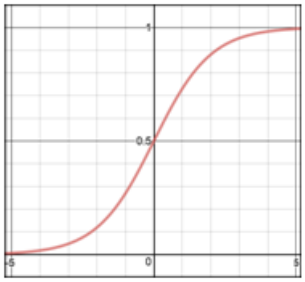

Fig. 4 Sigmoid Function [24]

Fig. 4 represents sigmoid function. Threshold value can be set to 0.5. If predicted value (P) is 0.8 and the threshold value is 0.5, then it belongs to class 1 or positive. If predicted value (P) is 0.2 and the threshold value is 0.5, then it belongs to class 0 or negative. Log loss function is used to calculate the error.

$$\frac{1}{N}\sum_{i=1}^{n}(y_i \log(P_1) + (1 - y_i)\log(1 - P_i)) \quad (6)$$

Log loss function is to minimize large negative number. Gradient descent is opted to optimize the loss.

Y_i=target variable

P_i= Probability

The advantage of using logistic regression is, it can be used to solve multiclass problems and it is resistant to over fitting.

5.2. DECISION TREE CLASSIFIER

Decision Tree classifier [25] is like tree structure which consists of branches and nodes. Branch serve as decision rules, nodes serves as outcome.

Root : This acts as the parent node in the tree.

Leaf : These nodes depend on root node and it is the outcome of other nodes.

Splitting: Subdivision of tree into sub nodes based on particular condition.

Branch Tree: After splitting, the separated tree is branch tree.

Pruning: Removal of unwanted branches.

Over fitting is possible in decision tree. Attribute Selection Measure (ASM) is essential to find the best feature in the dataset. Division of nodes based on best attribute is done recursively. Entropy is used to check whether it is pure set or not. Information Gain is used to measure the changes in entropy after subdivision of dataset. If entropy is less, then information gain will be higher.

5.3. RANDOM FOREST CLASSIFIER

Random Forest classifier [26] contains many decision trees and it takes the average of them to get accurate results. Over fitting is avoided in Random Forest classifier since, there are many trees. It works based on majority of votes. It can handle voluminous dataset. Over fitting is avoided, hence it produces accurate results.

5.4.ADABOOST

Yoav Freund and Robert Schapire proposed Adaboost (Adaptive Boosting) in 1996 [27]. Adaboost selects and trains the subset in a random manner. Higher weights are assigned to the data which is classified incorrectly. Classifier with higher weights is the most accurate classifier. This iteration will be done till it reaches maximum esitmators.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

5.5.XGBOOST

XGboost is also called as Extreme Gradient Boosting [28]. Overfitting can be avoided when using XGboost. Cross-validation is enabled by default in this algorithm. It was proposed by Tianqi Chen, Ph.D student, University of Washington. XGboost converts slow learners into fast learners. General parameters, booster parameters and learning task parameters are available for this algorithm.

5.6.MULTI LABEL CLASSSIFICATION

Multi label classification problem can be handled using classifier chains. It is the chain of binary classifiers. A chain of binary classifiers Classifier1, Classifier2, . . . , Classifier n is constructed. This chaining method is called classifier chains (CC) [29].This helps to classify the attacks.

5.7. MODEL GENERATION

5.7.1. MODEL 1:

Dimensionality reduction is done using Principal Component Analysis (PCA) [30]. After that, Logistic Regression, Random Forest Classifier and Decision Tree Classifier are applied to generate the model.

5.7.2.MODEL 2:

After feature transformation, the data is applied to classifiers like Logistic Regression, Random Forest Classifier, Decision Tree Classifier, Adaboost and XGBoost.

5.7.3.MODEL 3:

NSL-KDD has different kinds of attacks and it is a Multi-label classification problem. A classifier chain is applied to deal with multi label classification.

5.7.4.MODEL 4:

Deep learning algorithm is applied in this method. Basic neural network is used to serve the purpose.

Local Interpretable Model-agnostic Explanations (LIME) [31] is a python library that is used to explore the model. Fig. 5 depicts the prediction probabilities of the attacks which are represented in numbers, using LIME. Number 14 is the attack type. Features which contribute more to that attack are listed. Features like dst_host_serror_rate, dst_host_srv_serror_rate and flag_SF contributes more to perform attack number 14.

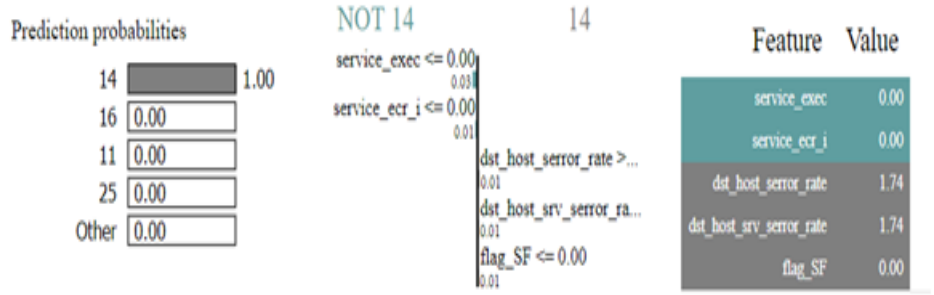

Fig. 5 LIME –Prediction Probabilities

LIME package explains the predictions done by various machine learning algorithms. Importance of features is identified using plot and LIME library. Based on this the model is generated.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

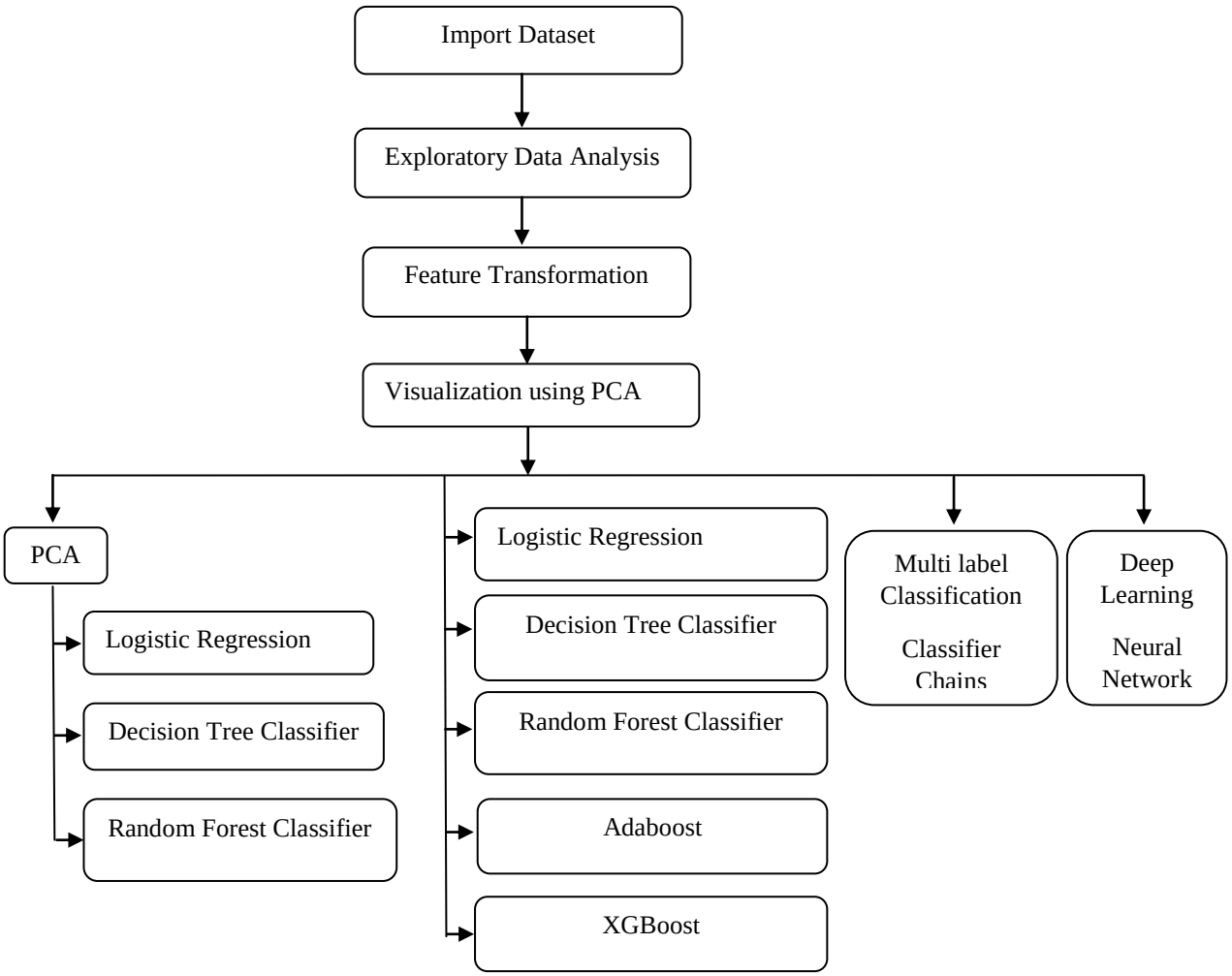

Fig. 6 Work flow of Machine Learning and Deep Learning Classifiers

6. RESULTS & DISCUSSION

Model 1: PCA with Classifiers

Table VII. Accuracy and Misclassification Rate of the Classifiers

Algorithm	Training Accuracy	Testing Accuracy	Misclassification Rate
Logistic Regression	88.8%	88.6%	11.3%
Random Forest	99.9%	98.7%	1.2%
Decision Tree	99.9%	97.9%	2.0%

Random forest algorithm produces 98.7% accuracy. Misclassification rate is 1.2% which is low when compared to logistic regression and decision tree algorithms.

Model 2: Transformed data with Classifiers

Table VIII. Accuracy and Misclassification Rate of the Transformed Classifiers

Algorithm	Training Accuracy	Testing Accuracy	Misclassification Rate
Logistic Regression	98.8%	98.7%	1.2%
Random Forest	99.9%	99.7%	0.2%

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

Decision Tree	99.9%	99.6%	0.3%
Adaboost	99.9%	99.8%	0.1%
XGBoost	99.7%	99.6%	0.3%

Adaboost algorithm performs better with 99.8% accuracy when compared to Random Forest, XGBoost, Decision Tree and Logistic regression. Misclassification rate of Adaboost is 0.1%.

Model 3: Transformed data with Multi-label classifier (Classifier Chain)

Table IX. Accuracy and Misclassification Rate of Multi Label Classifier

Algorithm	Training Accuracy	Testing Accuracy	Misclassification Rate
Random Forest	99.9%	99.6%	0.003%

99.6% accuracy is attained by applying Random forest Multi label classifier. Misclassification rate is 0.003%.

Model 4: Transformed data with Deep Learning (Deep neural network)

Table X. Accuracy and Misclassification Rate of Deep Neural Network

Algorithm	Training Accuracy	Testing Accuracy	Misclassification Rate
Deep Neural Network	99.4%	99.2%	0.7%

Basic network was applied as a deep learning approach and 99.2% accuracy is gained. 0.7% misclassification rate is attained. Fig. 7 shows the increasing level of accuracy of the model. Figure 8 shows the decreasing level of loss of the model.

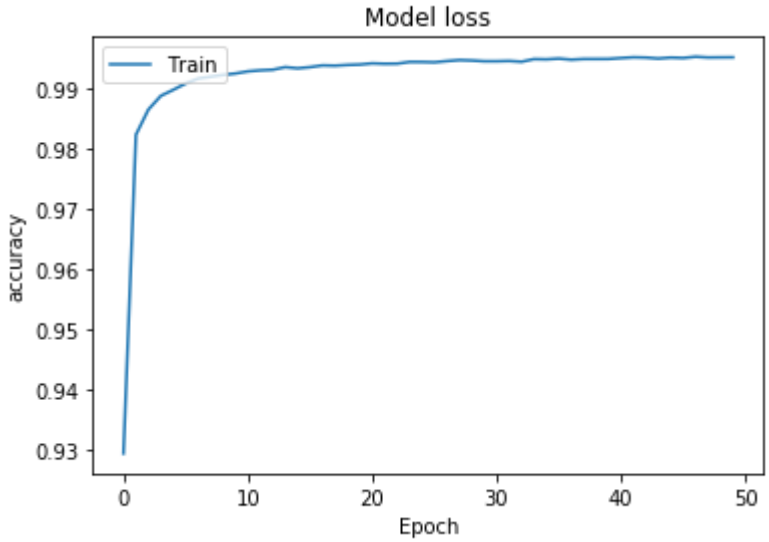

Fig. 7 Accuracy of Neural Network Model

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

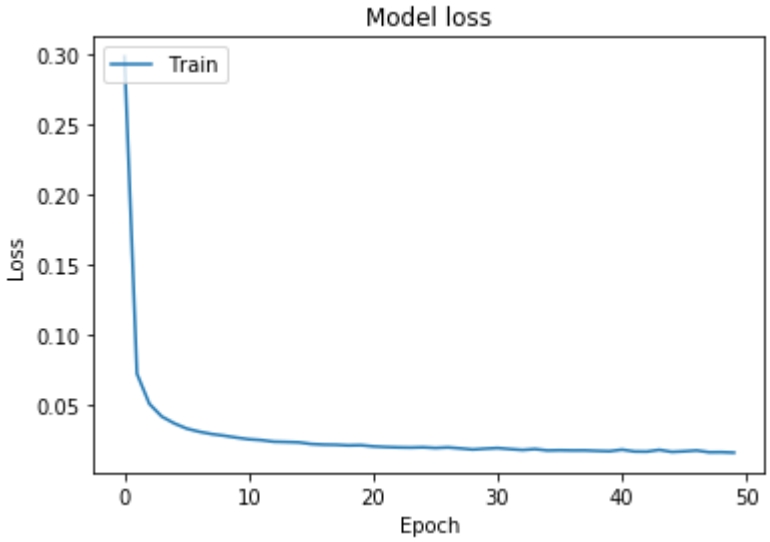

Fig. 8 Loss of Neural Network Model

Among the four methods, Adaboost algorithm works well with 99.8% accuracy and with less misclassification rate (0.1%)

7. PERFORMANCE METRICS & EVALUATION

After model generation, the evaluation of performance of the particular model must be measured. Performance of a model depends on number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Accuracy is a generic metric to determine the performance of the model. To handle imbalanced distribution of classes, performance metric like Recall, Precision and F1Score is required. Recall is also known as sensitivity. If the model has large number of False Negatives, then Recall is the best measure to handle it. Recall is used to calculate correctly classified positives from actual positives.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (7)$$

Precision is used to calculate number of positives which are correctly classified from predicted positives. If the model contains large number of False Positives, then Precision can be used.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

F1 score is used to provide a balance between precision and recall.

$$\text{F1 Score} = \frac{2*(Recall*Precision)}{(Recall+Precision)} \quad (9)$$

Model 1: PCA with Classifiers

Table XI. F1 Score, Precision and Recall of the Classifiers

Algorithm	F1 Score	Precision	Recall
Logistic Regression	91.4%	94.6%	88.6%
Random Forest	98.7 %	98.8%	98.7%
Decision Tree	97.9%	97.9%	97.9%

Among Logistic Regression , Random Forest and Decision Tree classifiers, Random Forest classifier has highest F1 Score (98.7%) and Precision (98.8%).

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

Model 2: Transformed data with Classifiers

Table XII. F1 Score, Precision and Recall of the transformed Classifiers

Algorithm	F1 Score	Precision	Recall
Logistic Regression	98.8%	99.0%	98.7%
Random Forest	99.7 %	99.7%	99.7%
Decision Tree	99.6%	99.6%	99.6%
Adaboost	99.8%	99.8%	99.8 %
XGBoost	99.6%	99.7 %	99.6 %

Among Logistic Regression , Random Forest and Decision Tree, Adaboost and XGBoost classifiers, Adaboost classifier has highest F1 Score (99.8%) and Precision (99.8%).

Model 3: Transformed data with Multi-label classifier (Classifier Chain)

Table XIII. F1 Score, Precision and Recall of Multi-label Classifier (Classifier Chain)

Algorithm	F1 Score	Precision	Recall
Random Forest	99.6 %	99.7%	99.6%

Multi-label classifier (Classifier Chain) which uses Random forest produces 99.6% F1 Score and 99.7% Precision.

Model 4: Transformed data with Deep Learning (Deep neural network)

Table XIV. F1 Score, Precision and Recall of Deep Neural Network

Algorithm	F1 Score	Precision	Recall
Deep Neural Network	99.3 %	99.4%	99.2%

Deep learning neural network produces 99.3% F1Score and 99.4% precision. Among the four methods Adaboost classifier has highest F1Score (99.8%) and Precision (99.8%).

8. CONCLUSION & FUTURE ENHANCEMENT

Best intrusion detection system must generate high accurate results and it should have lower false alarm rate. Adaboost algorithm performs well with 99.8% accuracy, 99.8% F1Score and 99.8% Precision. Hence the best classifier for NSL-KDD dataset is Adaboost. In future, intelligent IDS should be developed in order to identify active threats while data is streaming. Soft computing techniques like genetic algorithms, fuzzy and nature-inspired optimization can be used to provide better accurate results with less false rate. An effective response system should be build to respond to the attacks which are already happened

REFERENCES

[1] https://www.hindustantimes.com/india-news/india-s-internet-consumption-up-during-covid-19-lockdown-shows-data/story-ALcov1bP8uWYO9N2TbpPlK.html.

[2] https://www.threatstack.com/blog/the-history-of-intrusion-detection-systems-ids-part-1

[3] Markus Ring, Sarah Wunderlich, Deniz Scheuring, Dieter Landes and Andreas Hotho(2019), “A Survey of Network-based Intrusion Detection Data Sets”, arXiv:1903.02460.

[4] https://www.unb.ca/cic/datasets/nsl.html

[5] Shailendra Sahu, B.M. Mehtre(2015), “Network Intrusion Detection System Using J48, Decision Tree”, International Conference on Advances in Computing,Communications and Informatics (ICACCI), IEEE,.

[6] Senthilnayaki Balakrishnan, Venkatalakshmi, A. Kannan(2014), "Intrusion detection system using feature selection and classification technique", International Journal of Computer Science Applications 3

[7] Longjie Li, Yang Yu, Shenshen Bai, Jianjun Cheng, Xiaoyun Chen(2018), "Towards effective network intrusion detection: a hybrid model integrating gini index and GBDT with PSO", Journal of Sensors 1–9, https://doi.org/10.1155/2018/1578314.

[8] Selvakumar, B., & Muneeswaran, K. (2019). "Firefly algorithm based feature selection for network intrusion detection", Comput. Secur., 81, 148-155.

[9] Y. Bouzida, F. Cuppens, N.C. Boulahia, S. Gombault(2004), "Efficient intrusion detection using principal component analysis",Proc. SAR'04, La Londe, France.

[10] Mahmoud M. Sakr, Medhat A. Tawfeeq, Ashraf B. El-Sisi(2019), "Network Intrusion Detection System based PSO-SVM for Cloud Computing", International Journal of Computer Network Information Security 22–29.

[11] S. Balakrishnan, K. Venkatalakshmi, A. Kannan(2016), "A intrusion detection system using feature selection and classification technique", International Journal of Computer Science Applications 145–151.

[12] Ripon Patgiri, Udit Varshney, Tanya Akutota, Rakesh Kunde(2018), "An investigation on intrusion detection system using machine learning", IEEE Symposium Series on Computational Intelligence.

[13] Chuanlong Yin , Yuefei Zhu, Jinlong Fei, And Xinzheng He, (2017) "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks", IEEE Access, Vol. 5, pp 21954- 21961

[14] M. Z. Alom, V. Bontupalli and T. M. Taha(2015), "Intrusion detection using deep belief networks," 2015 National Aerospace and Electronics Conference (NAECON), Dayton, OH, 2015, pp. 339-344, doi: 10.1109/NAECON.2015.7443094.

[15] L.Dhanabal, Dr. S.P. Shantharajah(2015), "A Study on NSL-KDD Dataset for Intrusion

Detection System Based on Classification Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 6

[16] Manjula C. Belavagi∗ and Balachandra Muniyal(2016), "Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection", Twelfth International Multi-Conference on Information Processing, Procedia Computer Science 89

[17] Sumaiya Thaseen Ikram , Aswani Kumar Cherukuri (2017), "Intrusion detection model using fusion of chi-square feature selection and multi class SVM", Journal of King Saud University – Computer and Information Sciences , 29, 462–472

[18] Mostapha Chakir, Mohamed Moughit, Youness Idrissi Khamlichi, (2018) "An Effective Intrusion Detection Model Based on Svm with Feature Selection and Parameters Optimization", Journal of Theoretical and Applied Information Technology, Vol.96. No 12

[19] B. Subba, S. Biswas and S. Karmakar(2016), "Enhancing performance of anomaly based intrusion detection systems through dimensionality reduction using principal component analysis," 2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Bangalore, 2016, pp. 1-6, doi: 10.1109/ANTS.2016.7947776.

[20] Jamal Hussain, Samuel Lalmuanawma, Lalrinfela Chhakchhuak(2016), "A two-stage hybrid classification technique for network intrusion detection system", International Journal of Computational Intelligence Systems, Vol. 9, No. 5 , 863-875

[21] Gaurav Meena , Ravi Raj Choudhary(2017) ,"A Review Paper on IDS Classification using KDD 99 and NSL KDD Dataset in WEKA" International Conference on Computer, Communications and Electronics (Comptelix)

[22] Preeti Aggarwala, Sudhir Kumar Sharma(2015), "Analysis of KDD Dataset Attributes-Class wise for Intrusion Detection",3rd International Conference on Recent Trends in Computing, Procedia Computer Science 57.

[23] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[24] https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1927 - 1943
Received 20 January 2021; Accepted 08 February 2021.

[25] Hota H.S., Shrivas A.K. (2014) "Decision Tree Techniques Applied on NSL-KDD Data and Its Comparison with Various Feature Selection Techniques." In: Kumar Kundu M., Mohapatra D., Konar A., Chakraborty A. (eds) Advanced Computing, Networking and Informatics- Volume 1. Smart Innovation, Systems and Technologies, vol 27. Springer, Cham. https://doi.org/10.1007/978-3-319-07353-8_24

[26] Nabila Farnaaz and M. A. Jabbar,(2016) "Random Forest Modeling for Network Intrusion Detection System", Twelfth International Multi-Conference on Information Processing

[27] Arif Yulianto , Parman Sukarno and Novian Anggis Suwastika (2019), "Improving AdaBoost-based Intrusion DetectionSystem (IDS) Performance on CIC IDS 2017 Dataset", 2nd International Conference on Data and Information Science, Journal of Physics

[28] https://analyticssteps.com/blogs/introduction-xgboost-algorithm-classification-and-regression

[29] Jesse Read, Bernhard Pfahringer, Geoff Holmes , "Classifier Chains for Multi-label Classification", Machine Learning and Knowledge Discovery in Databases, 2009, Volume 5782 ISBN : 978-3-642-04173-0

[30] Krupa Joel Chabathula, Jaidhar C.D, Ajay Kumara M.A(2015), "Comparative Study of Principal Component Analysis Based Intrusion Detection Approach Using Machine Learning Algorithms", 3rd International Conference on Signal Processing, Communication and Networking (ICSCN)

[31] https://pythondata.com/local-interpretable-model-agnostic-explanations-lime-python/