IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

Network Intrusion Detection Using Machine Learning Techniques

Poonam Sapkale[1] and Rashmita Pradhan[2]
Students, Department of MCA
Late Bhausaheb Hiray S. S. Trust's Institute of Computer Application, Mumbai, India

Abstract: *In this work, I compare the utilization of assorted machine learning techniques for detecting intrusions in networks. The dataset that we used is NSL-KDD dataset, this dataset may be a refined version of KDD'99 dataset. Firstly, I analyze the dataset, to spot the quantity of examples and features that are present in it, to spot the amount of categorical features versus continuous features. Secondly, I perform preprocessing on the dataset, as a part of preprocessing step, I perform feature selection, followed by feature encoding, followed by feature scaling. I used Extra Trees Classifier feature selection technique, to extract those features which contribute the foremost. I used One Hot Encoder feature encoding technique, to encode the categorial features. I used Standard Scaler feature scaling technique, to scale the numerical features, so all the numerical features are within the same range. Thirdly, I apply various machine learning techniques like Decision Tree, Random Forest Classifier, Gaussian Naïve Bayes and KNN on this preprocessed dataset for training the models to classify normal and attack network traffic. Then i exploit these trained models on a test dataset, to perform classification of normal and attack network traffic. so we compare the accuracy achieved by each of those machine learning models, to spot the simplest machine learning model. Finally, compared the test accuracy achieved by the simplest machine learning model that's identified in previous step, and that we propose that, KNN classifier gives better accuracy for network intrusion detection.*

Keywords: Machine Learning, Network Intrusion, Intrusion Detection and Features.

I. INTRODUCTION

A network intrusion is an unauthorized penetration of your enterprise's network, or a non-public machine address in your assigned domain. Intrusions is additionally passive or active. Intrusions can come from inside your network structure or outside. Some intrusions are simply meant to permit you to understand the intruder was there by defacing your computer with various styles of messages or crude images. Others are more malicious and extract critical information. Some intruders implant carefully crafted code, like Trojan-type malicious software (malware), designed to steal passwords, record keystrokes, or open an application's "back door." An attacker can get into your system physically, externally or internally. The monitoring of network traffic for particular network segments or devices and analysis of network, transport, and application protocols to identify suspicious activity wont to guard against network intrusions. Network-based intrusion detection systems (NIDS) are devices intelligently distributed within networks which is used to inspect traffic flowing onto the devices on which they sit. NIDS are often software or hardware-based systems which are looking forward to the manufacturer of the system and can attach to numerous network mediums. Network intrusion detection systems operate at the network level and monitor traffic from all devices moving into and out of the network. NIDS performs analysis on the traffic trying to seem out patterns and abnormal behaviors upon which a warning is shipped.

II. PROBLEM STATEMENT

To distinguish the activities of the network traffic that the intrusion and normal is incredibly difficult and to wish much time consuming. An analyst must review all the information that enormous and wide to seek out the sequence of intrusion on the network connection. Therefore, it needs how that may detect network intrusion to reflect the present network traffics. Approach: during this study, a completely unique method to seek out intrusion characteristic for IDS using decision tree machine learning of knowledge mining technique was proposed. Method accustomed generate of rules is classification by ID3 algorithm of decision tree.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

2.1 Objective

The goal of a network intrusion detection system is to get unauthorized access to a network by analyzing traffic on the network for signs of malicious activity. The intrusion detection task is to create a predictive model capable of distinguishing between intrusions or attacks, and normal network connections. There are four main categories of attacks, namely: denial-of-service (DOS), unauthorized access from a far off machine, unauthorized access to local superuser (root) privileges, and probing.

III. LITERATURE REVIEW

In this section I describe about the previous work that was done by various researches in this field of network intrusion detection using machine learning and deep learning techniques.

[1] In the paper, author Mahbod Tavallaee et.al., peforms a statistical analysis was done on the KDD dataset and some potential issues in this dataset were identified. A new data set was proposed, that is NSL-KDD, in this newly proposed dataset the redundant records were removed from the training set and the 13 testing set. In this newly proposed dataset, the number of selected records from each difficulty level group is inversely proportional to the percentage of records in the original KDD data set. So, we also use this newly proposed dataset in our research

[2] In the paper, Machine Learning and Deep Learning Methods for Intrusion Detection Systems by Hongyu Liu * and Bo Lang they analyzes and refines the challenges and future trends within the field to supply references to other researchers conducting in-depth studies. Lacking of accessible datasets could also be the most important challenge. So the two learning approaches that is unsupervised learning and incremental learning approaches have broad development prospects.

[3] In the paper, Anomaly Detection with Robust Deep Autoencoders, by Chong Zhou et.al., they analyze the real-world problem of not having access to clean training data as required by standard deep denoising autoencoders. They propose model named as Robust deep autoencoder (RDA), which maintain all the capabilities of autoencoders and can also eliminate outliers and noise without access to any clean training data.

[4] In the paper, Hybrid Intelligent Intrusion Detection Schema, by Mostafa A. Salama et.al., various feature reduction techniques such as Principal component analysis, Gain ratio, Chi-Square, Deep belief network, were explored. Deep belief network and Support vector machine techniques were explored as classification techniques. It was described that using a Deep belief network for feature reduction, followed by using a Support Vector Machine technique for classification showed higher classification results compared to the other combinations.

[5] In the paper, A Deep Learning Approach for Network Intrusion Detection System, by Nathan Shone et.al., the weaknesses of using signature-based network intrusion detection techniques are discussed. And a novel deep learning classification model constructed using stacked nonsymmetric deep autoencoder is proposed. That new model has demonstrated high levels of accuracy, precision and recall together with reduced training time.

IV. RESEARCH METHODOLOGY

4.1 Analysis of NSL-KDD dataset

NSL-KDD dataset has 42 attributes for every connection record including class label containing attack types. The attack types are categorized into four attack classes as :

- **Denial of Service (DoS):** is an attack within which an adversary directed a deluge of traffic requests to a system so as to create the computing or memory resource too busy or too full to handle legitimate requests and within the process, denies legitimate users access to a machine.
- **Probing Attack (Probe):** probing network of computers to assemble information to be accustomed compromise its security controls.
- **User to Root Attack (U2R):** a category of exploit within which the adversary starts out with access to a standard user account on the system (gained either by sniffing passwords, a dictionary attack, or social engineering) and is in a position to take advantage of some vulnerability to achieve root access to the system.
- **Remote to Local Attack (R2L):** occurs when an attacker who has the flexibility to send packets to a machine over a network but who doesn't have an account thereon machine some vulnerability has been exploited to realize local access as a user of that machine.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

We have summarized our analysis on the NSL-KDD training dataset in the Table 1 and NSL-KDD testing dataset in Table 2.

Table 1. Details of the NSL-KDD training dataset

Details of Training dataset	Value	
Number of rows and columns	(25191, 42)	
Total number of features	41	
Number of categorical features	3	
Number of numerical features	38	
Are there any missing values	No	
Are there any duplicate records	No	
Number of different values for label	22	
Label distribution	normal	13448
	attack	11743

Table 2. Details of the NSL-KDD testing dataset

Details of Training dataset	Value	
Number of rows and columns	(11850, 42)	
Total number of features	41	
Number of categorical features	3	
Number of numerical features	38	
Are there any missing values	No	
Are there any duplicate records	No	
Number of different values for label	38	
Label distribution	normal	2152
	attack	9698

4.2 Pre-processing

For feature selection pre-processing step, where we wanted to identify the most important features that contribute the towards the label, so that we can eliminate the least contributing features, thereby improving the computational efficiency, we used the "Random Forest Classifier" algorithm. Here is the list of classification algorithms that we used

- DecisionTree
- LogisticRegression
- GuassianNaiveBayes
- K-Nearest neighbor classifier

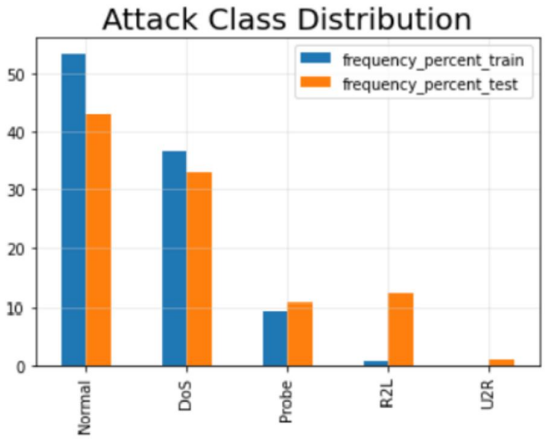
4.3 Exploratory Data Analysis

In statistics, an approach of analyzing data sets to summarize their main characteristics, using statistical graphics and other data visualization methods is done with help of exploratory data analysis. A statistical model is used or not, but primarily EDA is for seeing what the information can tell us beyond the formal modeling or hypothesis testing task.

Attack class distribution

	attack_class	frequency_percent_train	attack_class	frequency_percent_test
Normal	67343	53.46	9711	43.08
DoS	45927	36.46	7458	33.08
Probe	11656	9.25	2421	10.74
R2L	995	0.79	2754	12.22
U2R	52	0.04	200	0.89

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252



4.4 One-Hot Encoding

For categorical variables where no ordinal relationship exists, the integer encoding might not be enough, at best, or misleading to the model at the worst. during this case, a one-hot encoding may be applied to the ordinal representation. this can be where the integer encoded variable is removed and one new binary variable is added for every unique integer value within the variable. Each bit represents a possible category. If the variable cannot belong to multiple categories directly, then just one bit within the group is "on." this can be called one-hot encoding.

4.5 Implementation of ML Models

(I) KNN classifier model

K-Nearest Neighbor is one in every of the only Machine Learning algorithms supported Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that's most just like the available categories. K-NN algorithm stores all the available data and classifies a replacement information supported and similarly using K- NN algorithm suggests when new data appears then it are often easily classified into a well suite category. K-NN algorithm will be used for Regression similarly as for Classification but mostly it's used for the Classification problems. K-NN is non-parametric algorithm, which doesn't make any assumption on underlying data. it's also called a lazy learner algorithm because it doesn't learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that's much the same as the new data. The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbors
Step-2: Calculate the Euclidean distance of K number of neighbors
Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
Step-4: Among these k neighbors, count the number of the data points in each category.
Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
Step-6: Our model is ready.

(II) Guassian Naïve Bayes

Naïve Bayes algorithm is a supervised learning algorithm used for solving classification problems, which is predicated on Bayes theorem. It's mainly utilized in text classification that has a high-dimensional training dataset. Naïve Bayes Classifier is one in all the straightforward and handiest Classification algorithms which helps in building the fast machine learning models that may make quick predictions. It's a probabilistic classifier, which implies it predicts on the premise of the probability of an object. Some samples of Naïve Bayes Algorithm are spam filtration, sentimental analysis, and classifying articles. Gaussian: The Gaussian model assumes that features follow a standard distribution. This suggests if predictors take continuous values rather than discrete, then the model assumes that these values are sampled from the Gaussian distribution.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

(III) Logistic Regression

Logistic regression is one among the foremost popular Machine Learning algorithms, which comes under the Supervised Learning technique. it's used for predicting the explicit variable quantity employing a given set of independent variables. Logistic regression predicts the output of a categorical variable quantity. Therefore the result must be a categorical or discrete value. It may be either Yes or No, 0 or 1, true or False, etc. but rather than giving the precise value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is way just like the statistical regression except that how they're used. regression is employed for solving Regression problems, whereas Logistic regression is employed for solving the classification problems. In Logistic regression, rather than fitting a curve, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something like whether the cells are cancerous or not, a mouse is obese or not supported its weight, etc. Logistic Regression may be a significant machine learning algorithm because it's the power to produce probabilities and classify new data using continuous and discrete datasets. Logistic Regression may be wont to classify the observations using differing types of knowledge and might easily determine the foremost effective variables used for the classification.

(IV) Decision Tree

Decision Tree may be a Supervised learning technique that may be used for both classification and Regression problems, but mostly it's preferred for solving Classification problems. It's a tree structured classifier, where internal nodes represent the features of a dataset, branches represent the choice rules and every leaf node represents the end result. In an exceedingly Decision tree, there are two nodes, which are the choice Node and Leaf Node. Decision nodes are wont to make any decision and have multiple branches, whereas Leaf nodes are the output of these decisions and don't contain any more branches. The choices or the test are performed on the premise of features of the given dataset. It's a graphical representation for getting all the possible solutions to a problem/decision supported given conditions. It's called a choice tree because, just like a tree, it starts with the basis node, which expands on further branches and constructs a tree-like structure. So as to make a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. a call tree simply asks a matter, and supported the solution (Yes/No), it further split the tree into subtrees. The complete process can be better understood using the below algorithm:

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM)
Step-3: Divide the S into subsets that contains possible values for the best attributes.
Step-4: Generate the decision tree node, which contains the best attribute.
Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

4.6 Principal Component Analysis (PCA)

Principal Component Analysis, or PCA, could be a dimensionality-reduction method that's often wont to reduce the dimensionality of huge data sets, by transforming an outsized set of variables into a smaller one that also contains most of the knowledge within the large set. PCA is that the most generally used tool in exploratory data analysis and in machine learning for predictive models. Moreover, PCA is an unsupervised statistical technique want to examine the interrelations among a group of variables. It's also referred to as a general correlational analysis where regression determines a line of best fit.

HOW does one DO A PCA?

1. Standardize the range of continuous initial variables
2. Compute the covariance matrix to spot correlations
3. Compute the eigenvectors and eigenvalues of the covariance matrix to spot the principal components
4. Create a feature vector to make your mind up which principal components to stay
5. Recast the information along the principal component's axes

4.7 Recursive Feature Elimination (RFE)

Recursive Feature Elimination, or RFE for brief, could be a popular feature selection algorithm. RFE is popular because it's easy to configure and use and since it's effective at selecting those features (columns) during a training dataset that

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

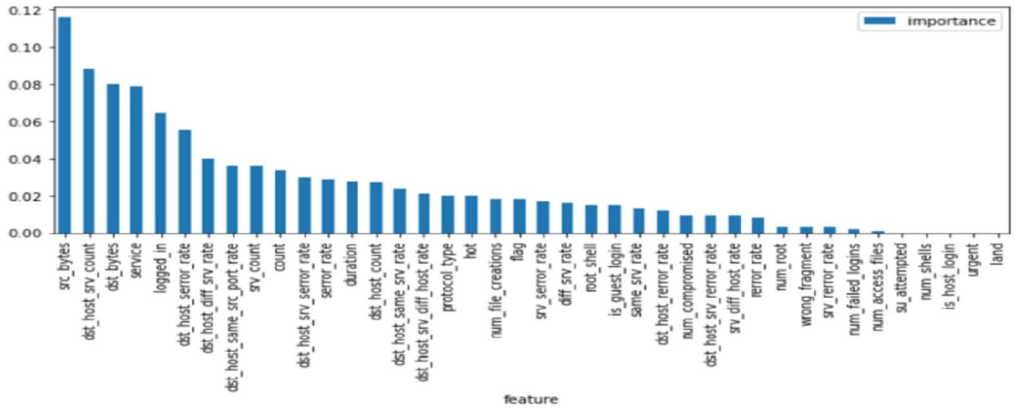
are more or most relevant in predicting the target variable. There are two important configuration options when using RFE: the selection within the number of features to pick and therefore the choice of the algorithm wont to help choose features. Both of those hyperparameters may be explored, although the performance of the tactic isn't strongly addicted to these hyperparameters being configured well.

4.8 Singular Value Decomposition (SVD)

Reducing the amount of input variables for a predictive model is spoken as dimensionality reduction. Fewer input variables may end up in a very simpler predictive model which will have better performance when making predictions on new data. Perhaps the more popular technique for dimensionality reduction in machine learning is Singular Value Decomposition, or SVD for brief. This can be a method that comes from the sphere of algebra and might be used as an information preparation technique to make a projection of a sparse dataset before fitting a model.

V. ANALYSIS

5.1 Feature Selection



5.2 Train Models

```
============================== Normal_DoS Naive Baye Classifier Model Evaluation ==============================

Cross Validation Mean Score:
 0.9737760413571536

Model Accuracy:
 0.9737686173767133

Confusion matrix:
 [[65346  1997]
 [ 1536 65807]]

Classification report:
              precision    recall  f1-score   support

         0.0       0.98      0.97      0.97     67343
         1.0       0.97      0.98      0.97     67343

    accuracy                           0.97    134686
   macro avg       0.97      0.97      0.97    134686
weighted avg       0.97      0.97      0.97    134686
```

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

```
============================== Normal_DoS Decision Tree Classifier Model Evaluation ==============================

Cross Validation Mean Score:
 0.9997698368976862

Model Accuracy:
 0.9999480272634127

Confusion matrix:
 [[67343     0]
 [    7 67336]]

Classification report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     67343
         1.0       1.00      1.00      1.00     67343

    accuracy                           1.00    134686
   macro avg       1.00      1.00      1.00    134686
weighted avg       1.00      1.00      1.00    134686

============================== Normal_DoS KNeighborsClassifier Model Evaluation ==============================

Cross Validation Mean Score:
 0.99656981084704

Model Accuracy:
 0.9977577476500898

Confusion matrix:
 [[67287    56]
 [  246 67097]]

Classification report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     67343
         1.0       1.00      1.00      1.00     67343

    accuracy                           1.00    134686
   macro avg       1.00      1.00      1.00    134686
weighted avg       1.00      1.00      1.00    134686
```

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

```
============================== Normal_DoS LogisticRegression Model Evaluation ==============================

Cross Validation Mean Score:
 0.9808072130256406

Model Accuracy:
 0.980836909552589

Confusion matrix:
 [[65532  1811]
 [  770 66573]]

Classification report:
              precision    recall  f1-score   support

         0.0       0.99      0.97      0.98     67343
         1.0       0.97      0.99      0.98     67343

    accuracy                           0.98    134686
   macro avg       0.98      0.98      0.98    134686
weighted avg       0.98      0.98      0.98    134686
```

5.3 Test Models

```
============================== Normal_DoS Naive Baye Classifier Model Test Results ==============================

Model Accuracy:
 0.8336536781408352

Confusion matrix:
 [[5487 1971]
 [ 885 8826]]

Classification report:
              precision    recall  f1-score   support

         0.0       0.86      0.74      0.79      7458
         1.0       0.82      0.91      0.86      9711

    accuracy                           0.83     17169
   macro avg       0.84      0.82      0.83     17169
weighted avg       0.84      0.83      0.83     17169
```

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

```
============================== Normal_DoS Decision Tree Classifier Model Test Results ==============================

Model Accuracy:
 0.8165880365775525

Confusion matrix:
 [[5591 1867]
 [1282 8429]]

Classification report:
              precision    recall  f1-score   support

         0.0       0.81      0.75      0.78      7458
         1.0       0.82      0.87      0.84      9711

    accuracy                           0.82     17169
   macro avg       0.82      0.81      0.81     17169
weighted avg       0.82      0.82      0.82     17169

============================== Normal_DoS KNeighborsClassifier Model Test Results ==============================

Model Accuracy:
 0.8666200710583027

Confusion matrix:
 [[5787 1671]
 [ 619 9092]]

Classification report:
              precision    recall  f1-score   support

         0.0       0.90      0.78      0.83      7458
         1.0       0.84      0.94      0.89      9711

    accuracy                           0.87     17169
   macro avg       0.87      0.86      0.86     17169
weighted avg       0.87      0.87      0.86     17169
```

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

```
============================ Normal_DoS LogisticRegression Model Test Results ============================

Model Accuracy:
 0.8418661541149747

Confusion matrix:
 [[5963 1495]
 [1220 8491]]

Classification report:
              precision    recall  f1-score   support

         0.0       0.83      0.80      0.81      7458
         1.0       0.85      0.87      0.86      9711

    accuracy                           0.84     17169
   macro avg       0.84      0.84      0.84     17169
weighted avg       0.84      0.84      0.84     17169
```

VI. CONCLUSION

Table 3: Accuracies measured for the four ML classification algorithms

ML model name	Training	Testing
'Naive Baye Classifier	0.97377	0.83365
'Decision Tree Classifier'	0.99994	0.81658
KNeighbors Classifier	0.99775	0.86662
Logistic Regression	0.98083	0.84186

We evaluated the varied machine learning techniques against the NSL-KDD dataset. On the training dataset, the accuracies achieved by the classic machine learning classification algorithms is 0.99994, whereas the accuracy on testing dataset achieved by the classic machine learning classification algorithms is 0.84186 PCA is way faster to coach and gave worst accuracy

RFE

- Feature: 6
- Accuracy: 0.9798537512

PCA

- Feature: 6
- Accuracy: 0.9737667829

SVD

- Feature: 6
- Accuracy: 0.9765543989

VII. FUTURE SCOPE

We can further experiment using Generative Adversarial Networks, type of neural networks and also perform experiments towards building an intrusion prevention system versus an intrusion detection system. As next step, planned

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

to built a fully connected deep neutral network, using the Auto Encoder in hidden layers this fully connected deep neutral network. Then measure the accuracies for the fully connected deep neutral network.

REFERENCES

[1]. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009).
[2]. Hongyu Liu * and Bo Lang, "Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey", State Key Laboratory of Software Development Environment, 17 October 2019.
[3]. Chong Zhou and Randy C. Paffenroth, "Anomaly Detection with Robust Deep Autoencoders", KDD 2017 Research Paper, KDD'17, August 13–17, 2017, Halifax, NS, Canada
[4]. Mostafa A. Salama, Heba F. Eid, Rabie A. Ramadan, Ashraf Darwish, and Aboul Ella Hassanien, "Hybrid Intelligent Intrusion Detection Scheme", Springer, pp 293-303.
[5]. Nathan Shone, Tran Nguyen Ngoc, Vu Dinh Phai, and Qi Shi, "A Deep Learning Approach for Network Intrusion Detection", IEEE Transactions on emerging topics in computational intelligence, vol. 2, no. 1, (2018)
[6] Donghwoon Kwon, Hyunjoo Kim, Jinoh Kim, Sang C. Suh, Ikkyun Kim, Kuinam J. Kim, "A survey of deep learning-based network anomaly detection", Springer Science+Business Media, LLC (2017).
[7] L. Dhanabal, Dr. S.P. Shantharajah, "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, June 2015.
[8] S. Revathi and Dr. A. Malathi, "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, vol. 2, issue 12, December – 2013.
[9] https://www.linkedin.com/pulse/k-nearest-neighborknn-algorithm-machine-learning-anusha-ranga
[10] https://techvaluetrends.com/knn-algorithm
[11] https://data-flair.training/blogs/machine-learning-classification-algorithms
[12] https://www.javatpoint.com/logistic-regression-in-machine-learning
[13] https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
[14] https://builtin.com/data-science/step-step-explanation-principal-component-analysis
[15] https://www.dataloco.com/574139/recursive-feature-elimination-rfe-for-feature-selection-in-python
[16] https://machinelearningmastery.com/singular-value-decomposition-for-dimensionality-reduction-in-python