COMPARISON OF CLASSIFICATION TECHNIQUES ON INTRUSION DETECTION SYSTEM

SadhanaKodali[1], M. Devi Sai Lakshmi, M.Sireesha, J.S.B.N. Malleswari, K. Boni Preetham.

Assoc Professor[1], LENDI INSTITUTE OF ENGINEERING & TECHNOLOGY, Jonnada, A.P.,India.
Student[2, 3, 4, 5, B.Tech] (CSE), LENDI INSTITUTE OF ENGINEERING & TECHNOLOGY, Jonnada,A.P.,India.

ABSTRACT:
With the evolution of Wireless Communication, there are many security threats over the internet. The Intrusion Detection System (IDS) helps find different types of intrusion and can also detect intruders. This project is proposed to develop an IDS by using various classification techniques like Linear Support Vector Machine (LSVM), Quadratic Support Vector Machine (SVM), K-Nearest-Neighbor (KNN), Multi-Layer Perceptron (MLP), and Auto Encoder. All the results of every classification technique are compared in terms of accuracy.

***KEYWORDS*: Linear Support Vector Machine (LSVM), Quadratic Support Vector Machine (SVM), K-Nearest-Neighbor (KNN), Multi-Layer Perceptron (MLP), and Auto Encoder.**

INTRODUCTION:

An Intrusion Detection System (IDS)[1] is a device or software-based programme that can be used to identify malicious or privacy-related network activity. Any illegal behaviour is immediately reported to the system administrator or forwarded directly to the central system via a SIEM mechanism.

NETWORK INTRUSION DETECTION SYSTEMS (NIDS)

Network Intrusion Detection Systems (NIDS) are installed at strategic locations throughout the network to monitor traffic to and from all connected devices. NIDS examines all presently passing traffic on the whole subnet and compares it to previously identified traffic generated by prior attacks. When unusual behavior is detected, an alarm is issued to the network administrator.

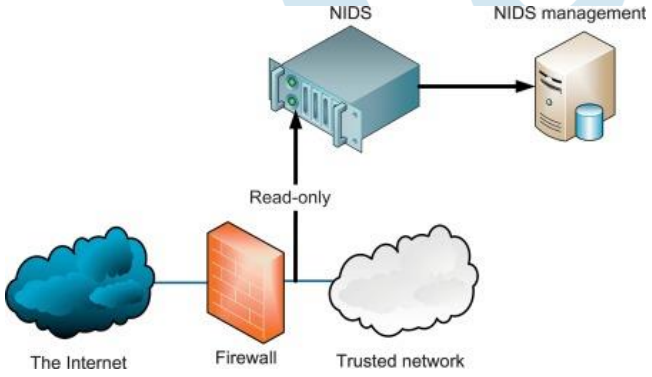


Figure (1)
Figure (1)describe about NIDS Architecture

1. CLASSIFICATION

In Machine Learning, Classification Techniques are further divided into different types of models, as shown in the Figure(2), as well as some of them listed below:-

1. Logistic Regression [2]
2. Naive Bayes Classifier [3]
3. nearest Neighbor [4]
4. Support Vector Machines [5]
5. Decision Trees [6]
6. Random Forest [7]
7. Neural Networks [8]

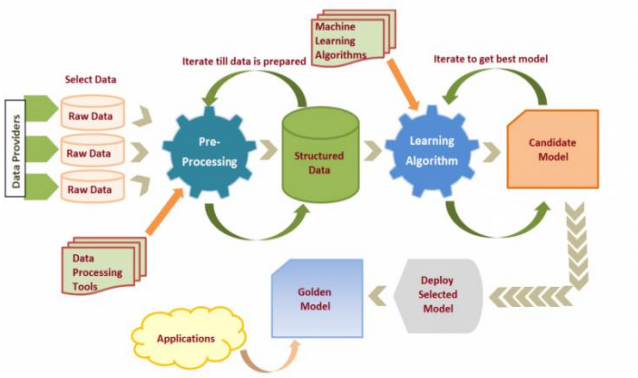
2. LITERATURE SURVEY

Network Intrusion Detection (Mukherjee, Heberlein, and Levitt 1994). Classification Techniques in Machine Learning: Applications and Issues (Soofi and Awan 2017) Network Traffic Classification Techniques and Comparative Analysis Using Machine Learning Algorithms (Shafiq et al. 2016). Survey of Classification Techniques in Data Mining (Phyu 2009).Intelligent Intrusion Detection Systems using Artificial Neural Networks (Shenfield, Day, and Ayesh 2018).

Network Intrusion Detection using Na¨ıve Bayes (Panda and Patra 2007. Random Forest Modeling for Network Intrusion Detection System (Farnaaz and Jabbar 2016.

3 PROPOSED SYSTEM

There are several security threats on the internet as a result of the expansion of wireless communication. The Intrusion Detection System (IDS) aids in the detection of many sorts of intrusions, as well as intruders. The goal of this project is to create an IDS by combining different categorization approaches such as Linear Support Vector Machine (LSVM), Quadratic Support Vector Machine (SVM), K-Nearest-Neighbor (KNN), Multi-Layer Perceptron (MLP), and Auto Encoder.Several steps are involved in performing a machine learning technique based on the classification method using a dataset—the basic steps involved in completing a machine learning process.



Figure(2).
Figure(2)describe about Machine Learning Architecture

4. DATA DESCRIPTION

The primary goal of this study is to analyse multiple classification algorithms and apply them to a dataset in order to determine which model provides the most accuracy and precision while reducing false positives. We used a dataset based on intrusion systems for our study on Network Intrusion Detection Systems (NIDS). A network connection is a series of TCP packets that begin and terminate at a set time interval, during which data flows from a source IP address to a target IP address, and each connection is classified as either normal or as an attack with a certain attack type. 1. Denial of Service (DoS): A DoS attack is one in which the intruder prevents authorised users from accessing the system or network. 2. Remote to Local (R2L): This is an attack that attempts to obtain access to a local account from a remote host or network.

5. SYSTEM DESIGN & IMPLEMENTATION

5.1 Data Preprocessing

Data Preprocessing is the first step in the machine learning process that must be completed before the process can be started.Raw and unfiltered data is transformed andconverted into a more acceptable and intelligible format via data preprocessing.
Data preprocessing is a method of converting raw data into a clean set of data. In other words, anytime data is acquired from various sources, it is obtained in raw format, which makes analysis impossible. Data Preprocessing consists of the following steps:
Libraries to Import
2. Importing Dataset
3. Verifying that no values are missing
4. Encoding Categorical Data
5. Splitting the Dataset into a Training set and Test Set
6. Scaling of Features

LINEAR SUPPORT VECTOR MACHINE

Support vector machines (SVMs) are supervised learning algorithms for classifying, predicting, and detecting outliers. The SVM algorithm's goal is to find a hyperplane in an N-dimensional space that distinguishes between data points. The hyperplane's size is determined by the number of features. If there are only two input characteristics, the hyperplane is just a line. When the number of input features reaches three, the hyperplane transforms into a two-dimensional plane. When the number of elements exceeds three, it becomes impossible to imagine. Consider two independent variables, x1 and x2, as well as one dependent variable, a blue or red circle.

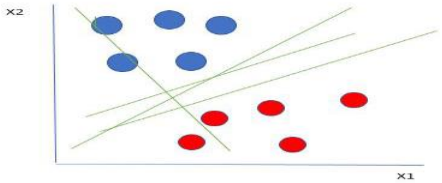


Figure (3)

Figure (3) describe about SVM Multiple lines (our hyperplane here is a line because we're just examining two input features, x1 and x2) separate our data points or classify them into red and blue circles. **Selecting the best hyper-plane:**The hyper plane that represents the greatest separation or margin between the two classes is a viable choice as the best hyperplane.

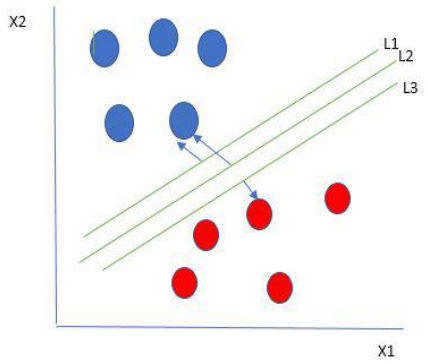


Figure (4)

Figure (4) describe about SVM Hyperplane

QUADRATIC SUPPORT VECTOR MACHINE

Quadratic support vector machines are offered as a way to circumvent problems like finding adequate kernel functions and setting hyper-parameters. Furthermore, data points from Universum that do not belong to any class can be used to integrate prior knowledge into the appropriate models, improving generalisation performance. This research proposes new Universum quadratic surface support vector machine models that are kernel-free. We also present an L1 norm regularised variant, which is useful for detecting potential sparsity patterns in the Hessian of a quadratic surface and reducing to standard linear models if the data points are (almost) linearly separable. Because the suggested models are convex, they can be solved using ordinary numerical solvers. Despite this, we develop a least squares version of the L1 norm regularised model and then create an efficient customised algorithm that only involves the solution of one linear system. Following that, certain theoretical features of these models are reported/proven. Finally, we run numerical tests on both manufactured and public benchmark data sets to show that the suggested models are feasible and effective.

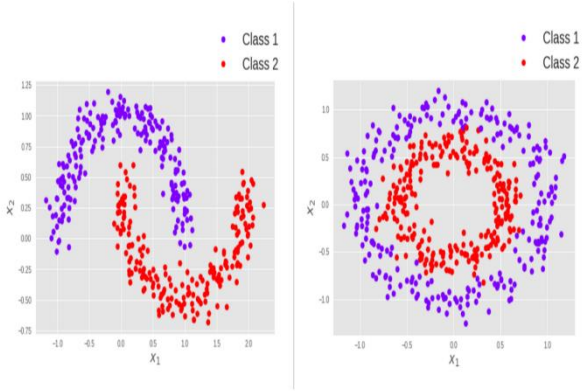


Figure (5)

Figure (5) describe about QSVM

K-Nearest Neighbor (KNN)

The K-Nearest Neighbour algorithm is based on the Supervised Learning technique and is one of the most basic Machine Learning algorithms.This algorithm assumes that the new case/data and existing cases are similar and places the new case in the category that is most similar to the existing categories. The K-NN algorithm saves all existing data and categorises additional data points based on their similarity. This implies that as fresh data comes in, the K-NN algorithm can quickly classify it into a good suite category.The K-NN algorithm is a non-parametric algorithm, which means it makes no assumptions about the data.It's also known as a lazy learnealgorithm since it doesn't learn from the training set right away; instead, it saves the dataset and uses it to classify the data.

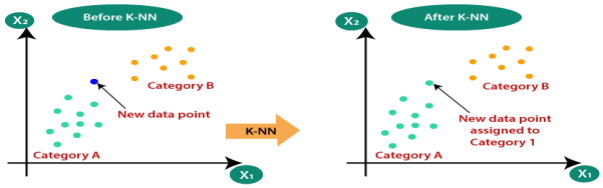


Figure (6)

This figure describe about KNN Classification

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 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.

To begin, we will choose k=5 as the number of neighbours.The Euclidean distance between the data points will then be calculated. The Euclidean distance is the distance between two points that we learned about in geometry class.
First, we'll decide on the number of neighbours, thus we'll go with k=5.
The Euclidean distance between the data points will then be calculated. The Euclidean distance is the distance between two points that we learned about in geometry class.

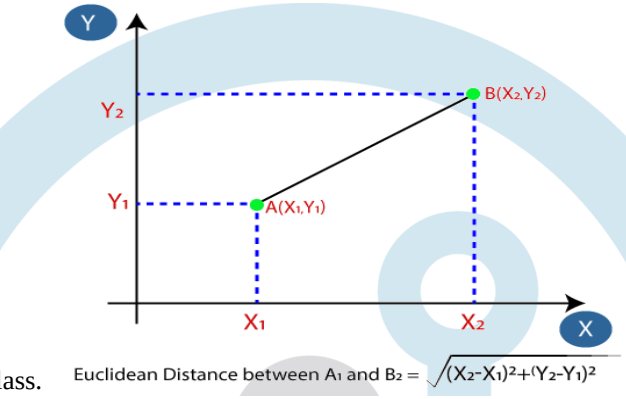
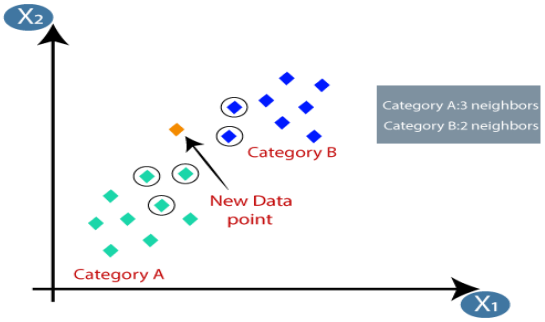


Figure (7)

figure (7) describe about Euclidean Distance

By calculating the Euclidean distance, we got the nearest neighbors as three neighbors in category A and two closest neighbors in category B. Consider the below image:



;

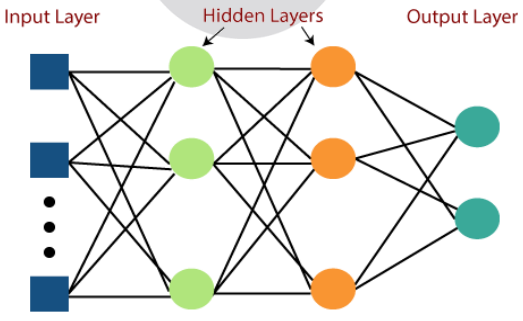
Figure (8)

Figure (8) describe about KNN Category

As we can see, the three nearest neighbors are from category A; hence this new data point must belong to category A.

MULTI-LAYER PERCEPTRON:

A multi-layer perceptron (MLP) is a feed-forward neural network augmentation. As demonstrated in, it has three layers: an input layer, an output layer, and a hidden layer. The input signal to be processed is received by the input layer.
The number of neurons in the input layer is determined by the size of the input vector, just as it is with all neural networks, while the number of classes to be learned determines the number of neurons in the output layer. The number of hidden layers to use and the number of neurons in each layer must be decided



.

Figure (9) describe about Multilayer Perceptron

MLP networks are used for supervised learning formats. A typical learning algorithm for MLP networks is also called a **back propagation's algorithm**.

AUTOENCODER:

An autoencoder is a type of neural network that is designed to try to duplicate its input into its output. It features a hidden layer on the inside that depicts a code that is utilised to represent the input.

Autoencoders are asymmetric ANNs with a middle layer that represents encoding of the input data. Autoencoder is ready to reassemble their input onto the output layer while confirming certain constraints that prevent them from replicating the data along with the network. Although autoencoder is the most commonly used term nowadays, they were previously known as auto-associative neural networks, diabolo networks, and neural replicator networks.

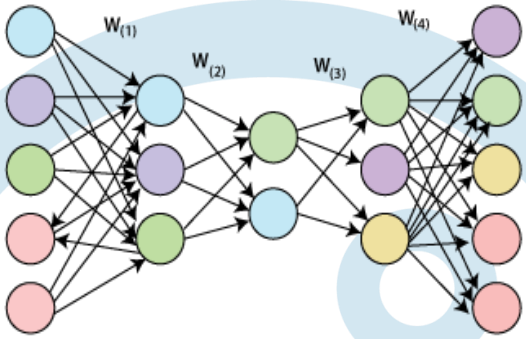


Figure (10)

Figure (10) describe about Auto Encoder

6. RESULTS:

```
In [17]: y_pred=knn.predict(X_test) # predicting target attribute on testing dataset
         ac=accuracy_score(y_test, y_pred)*100 # calculating accuracy of predicted data
         print("KNN-Classifier Binary Set-Accuracy is ", ac)

         KNN-Classifier Binary Set-Accuracy is  98.57433161872102

In [18]: print(classification_report(y_test, y_pred,target_names=le1_classes_))

                       precision    recall  f1-score   support

             abnormal       0.99      0.98      0.98     14720
               normal       0.99      0.99      0.99     16774

             accuracy                           0.99     31494
            macro avg       0.99      0.99      0.99     31494
         weighted avg       0.99      0.99      0.99     31494
```

KNN algorithm shows better accuracy results when compared to other techniques.

7. CONCLUSION

Intrusion Detection System (IDS) helps to find different types of intrusion and the intruders can also be detected. This project is proposed to develop an IDS by using the various classification techniques like Linear Support Vector Machine (LSVM), Quadratic Support Vector Machine (QSVM), K-Nearest-Neighbor (KNN), Multi Layer Perceptron (MLP), Auto Encoder. All the results of every classification technique are compared in terms of accuracy**.** By comparing all the accuracy results of every classification technique, K-Nearest Neighbor showed better accuracy result when compared to other.

REFERENCES:

1. Adetunmbi, Adebayo O et al. (2008). “Network intrusion detection based on rough set and k-nearest neighbour”.
2. MLCheatsheet(2017).LogisticRegression.
3. Survey of Classification Techniques in Data Mining (Phyu 2009).
4. Harrison, Onel (2018). Machine Learning Basics with the K-Nearest Neighbors Algorithm.
5. Bambrick, Noel (2016). Support Vector Machines.
6. Singh, Nagesh (2020a). Decision Tree Algorithm.
7. Random Forest Modeling for Network Intrusion Detection System (Farnaaz and Jabbar 2016).
8. Sharma, Sheetal (2017). Artificial Neural Network (ANN) in Machine Learning
9. Network Traffic Classification Techniques and Comparative Analysis Using Machine Learning Algorithms (Shafiq et al. 2016).