Development of Intrusion Detection System in Neural Network

P. Hemalatha
PG Scholar, Department of Computer Science and Engineering, Sardar Raja College of Engineering, Tamilnadu, India

Abstract-A device or software programme known as an intrusion detection system (IDS) analyses network or system activity to look for signs of hostile activity. The construction of IDS in a neural network is suggested in this research. The IDS classification is separated using a larger dataset. The idea of transforming unclean data into clean data is known as data pre-processing. Before running the method, the dataset is pre-processed to look for missing values, noisy data, and other abnormalities. Chi-square-based feature extraction is used during the extraction process. The Chi-Square approach is used to process the extraction areas in order to extract various characteristics and choose the essential features in order to enhance classification. The efficient Chi-square method is employed in this project to determine feature extraction and feature selection. The chosen characteristics are then used to accurately classify data using the gradient boosting classifier (GBC), cat boosting classifier (CBC), K-Nearest neighbour (KNN), and random forest classifier. Python software is used in the execution of this project.

***Keywords*- Intrusion Detection System, KNN, GBC, CBC, Random forest Classifier.**

1. INTRODUCTION

Deep learning methods are highly helpful since they operate from beginning to finish and can automatically identify feature representations from raw data. For the creation of high-performing, accurate ML applications, the availability of high-quality data is often a need. Data cleaning, also known as data replacement, data modification, and data deletion, is the act of correcting or erasing incorrect data from a data file. This includes processing erroneous values, missing values, and data rationality detection [1].The effectiveness of the categorization process is significantly influenced by the type of encoding technique used. A typical method for processing categorical data is one-hot encoding [2]. Categorical variables must be transformed into a format in order for ML models to be more effective at spotting instances of insider data leaking. To prevent incorrect interpretation of the correlations between independent variables, it only draws attention to the variables included in the features. Data normalisation is the process of proportionally scaling the data so that all values fall into the desired range [3].

A non-probabilistic binary classifier called M-SVM divides data into many groups. The binary division SVM accomplishes classification by classifying input data. SVM is a highly helpful tool for categorising undistributed and asymmetrically distributed data, which might include text, pictures, audio, and other types [4]. M-SVMs are based on the premise that two data classes can be separated by a margin on each side of a plane. An M-SVM is used to build a learning model based on supervised learning, which trains the model to categorise training data using prelabeled labels.

Additionally, assault strategies are developing daily, and the complexity of the mysterious offences that must be repelled is rising. In order to determine whether there is abnormal behaviour in the dataset and to offer trustworthy protection assistance for users or terminal equipment [5-7], IDs may detect and analyse network data .The development of deep learning offers a fresh approach to the Internet of Things' IDS study. In order to achieve real-time network status monitoring and the Internet of Things' current state is used as a network data set, and it is imported into the ML model for training and learning. Based on learnt normal and attack behaviour, machine learning-based IDS offers a learning-based approach to find attack classes [8-10].

This study makes a contribution of the development of intrusion detection system in neural network. The structure of this work is as follows: In the Section II, Proposed system is given. The study's results are summarized in Section III.

2. PROPOSED SYSTEM

The proposed development of intrusion detection system in neural network is shown in Figure 1. IDs may detect and analyse network data .The development of deep learning offers a fresh approach of IDS study. In order to achieve real-time network status monitoring and the Internet of Things' current state is used as a network data set, and it is imported into the ML model for training and learning.

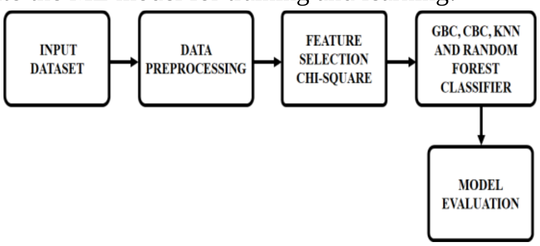

Figure 1. Block diagram for proposed system

The creation of an intrusion detection system using a neural network is suggested in this research. The IDS traffic categorization is separated using a larger dataset. The idea of transforming unclean data into clean data is known as data preparation. Before running the algorithm, the dataset is preprocessed to look for missing values, noisy data, and other abnormalities. Chi-square-based feature extraction is used during the extraction process. The Chi-Square approach is used to process the extraction areas in order to extract various characteristics and choose the essential features in order to enhance classification. The efficient Chi-square method is employed in this project to determine feature extraction and feature selection. For accurate classification, the chosen features are then put into a GBC, CBC, KNN, and random forest classifier.

A) Data Preprocessing

Preparing raw data for future processing by any sort of processing is known as data preprocessing, which is a subset of data preparation. It has long been considered the most important initial step in the data mining process. Data collection, cleansing, integration, transformation, reduction, discretization, normalisation or standardisation, feature selection, and data representation are the stages that make up data preparation.

B) Feature Selection CHI-Square

With the use of just pertinent data and the elimination of irrelevant data, feature selection is a technique for lowering the input variable for your model. It involves automatically selecting characteristics for your machine learning model that are pertinent to the problem you are attempting to solve. When choosing features, our goal is to choose those that depend heavily on the outcome. The observed count is close to the anticipated count when two characteristics are independent, hence the Chi-Square value will be lower. Therefore, a high Chi-Square score suggests that the independence hypothesis is false. The fact that chi-square is easier to calculate than other statistics is one of its biggest benefits. Additionally, it may be used with numerical (categorical) scale data. Whether or whether there is a "difference" between two or more participant groups can likewise be ascertained using this method.

C) KNN Classifiers

One of the most fundamental categories of machine learning algorithms, KNN is frequently employed for categorization. The classification of the data point is based on the classification of its neighbours. KNN categorises new data points according to how closely they resemble stored data points. The number of nearest neighbours to include in the majority voting process is indicated by the parameter 'k' in the KNN algorithm. KNN works by measuring the separations between a query and each occurrence in the data, choosing the sample size (K) that is closest to the query, and either selecting the label with the highest frequency (in classification) or averaging the labels (in regression) based on the results.

D) Gradient Boosting Classifier

Each prediction in gradient boosting aims to outperform the one before it by lowering the errors. Gradient Boosting's intriguing concept, however, is that it really fits a new predictor to the residual errors created by the preceding predictor, rather than fitting a prediction on the data at each iteration.

E) Cat Boosting Classifier

A recently released machine learning algorithm is called Cat Boost. It is simple to interface with deep learning frameworks such as Apple's Core ML and Google's TensorFlow. Cat Boost can work with several data formats to aid organisations today with a variety of issues. When utilising one_hot_max_size (Use one-hot encoding for any features with a number of distinct values less than or equal to the supplied parameter value), Cat Boost offers the flexibility to provide indices of categorical columns. By "generating random permutations of the dataset and computing the average label value for each sample with the same category value placed before the given one in the permutation," Catboost addresses categorical characteristics. Additionally, they offer discretisation into a predetermined number of bins (128 and 32) and GPU-accelerated data processing. The implementation of symmetric trees by CatBoost distinguishes it significantly from other boosting techniques. Although it may seem strange, this aids in reducing prediction time, which is crucial in low latency situations.

F) Random Forest Classifier

Regression and classification are only two of the many tasks that may be performed with the reliable machine learning algorithm Random Forest. It is an ensemble approach, which means that a random forest model is composed of several little decision trees, known as estimators, each of which generates a separate set of predictions. The estimators' estimates are combined by the random forest model to yield a more precise forecast.

3. RESULTS AND DISCUSSIONS

The work was conducted out using Phython. IDSs are intended to detect attacks, thus it's important to pick the best data source based on those features.ID System evaluates the data that is processed in order to classify IDSs.

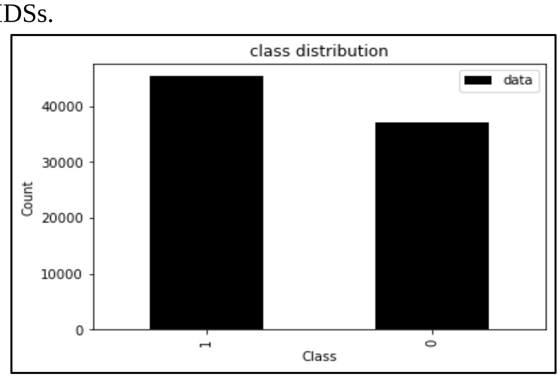

Figure 2. Class distribution

A class distribution may be thought of as a dictionary with the number of randomly generated samples to include in the dataset as the value and the key being the class value (for instance, 0 or 1).

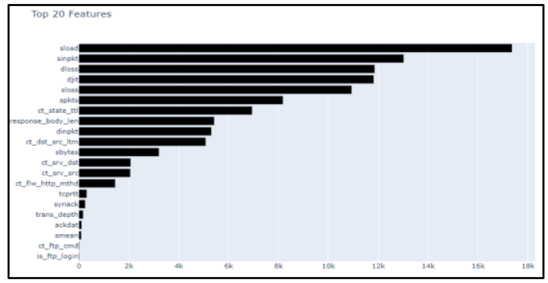

Figure 3. Top 20 Features

Figure 3 illustrates how to use a CHI-square to choose a clear visual picture. A statistical formula for comparing two or more statistical data sets is the chi-square formula.

```
Accuracy: 96.03%
Recall: 95.94%
Precision: 96.83%
F1-Score: 96.38%
time to train: 1.70 s
time to predict: 0.31 s
total: 2.01 s
```

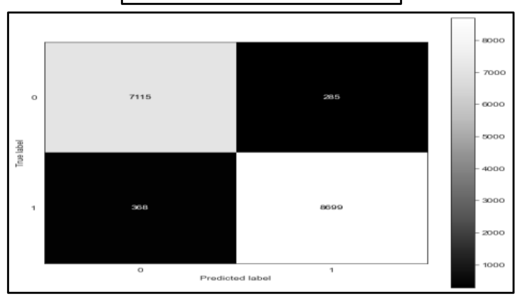

Figure 4. Cat boost Classifier

CatBoost is a supervised machine learning method that is used by the Train Using AutoML tool and uses decision trees for classification and regression. This figure shows the accuracy, recall, precision and F1-score. It is observed that an accuracy of 96.03%, recall of 95.94%, precision of 96.83%, F1-score of 96.38% is obtained.

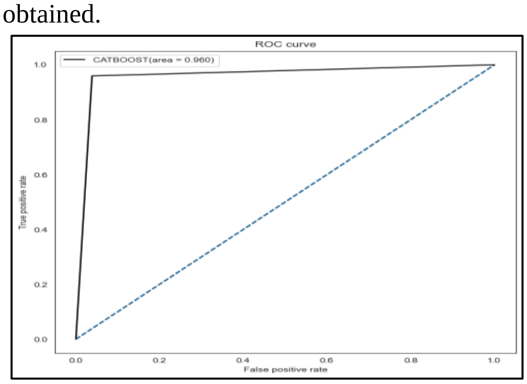

Figure 5. Roc curve for cat boosting classifier

Figure 5 represents that Roc curve for cat boosting classifier. In x axis false positive rates are shown and in y axis true positive rates are shown.

```
Accuracy: 95.87%
Recall: 96.05%
Precision: 96.43%
F1-Score: 96.24%
time to train: 52.54 s
time to predict: 0.05 s
total: 52.59 s
```

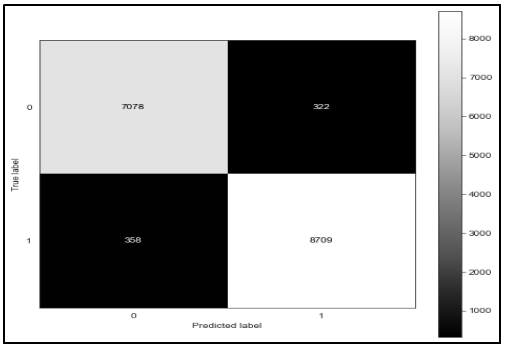

Figure 6. Gradient boosting classifier

In order to minimise a loss function, the functional gradient method known as Gradient Boosting continually chooses a function that points in the direction of a weak hypothesis or a negative gradient. The accuracy, recall, precision, and F1-score are displayed in this graph. It is noted that a 96.44% accuracy, 96.16% recall, 97.33% precision, and 96.74% F1-score are attained.

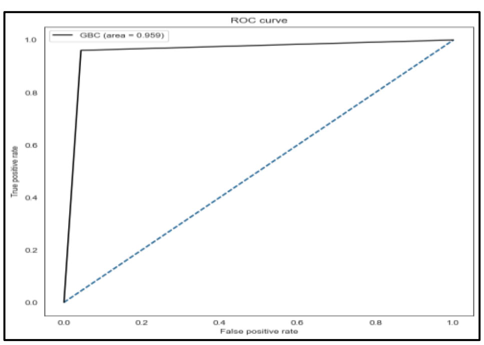

Figure 7. Roc curve for Gradient boosting classifier

Figure 7 shows the gradient boosting classifier's ROC curve. False positive rates are displayed on the x axis, and real positive rates are displayed on the y axis.

```
Accuracy: 96.44%
Recall: 96.16%
Precision: 97.33%
F1-Score: 96.74%
time to train: 0.01 s
time to predict: 7.89 s
total: 7.90 s
```

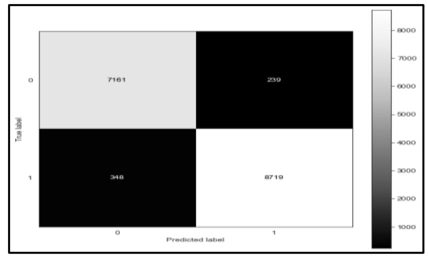

Figure 8. K-Nearest Neighbor

KNN classifier is supervised machine learning technique that is non-parametric. It is dependent on distance. The accuracy, recall, precision, and F1-score are displayed in this graph. It is noted that a 96.44% accuracy, 96.16% recall, 97.33% precision, and 96.74% F1-score are attained.

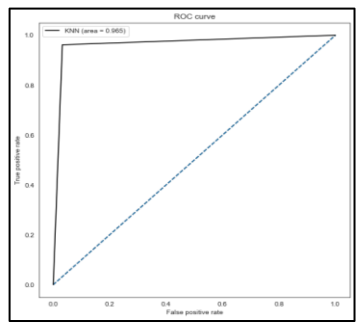

Figure 9. Roc curve for KNN

Figure 9 represents that Roc curve for gradient boosting classifier. In x axis false positive rates are shown and in y axis true positive rates are shown.

```
Accuracy: 97.73%
Recall: 97.41%
Precision: 98.46%
F1-Score: 97.93%
time to train: 8.49 s
time to predict: 0.15 s
total: 8.64 s
```

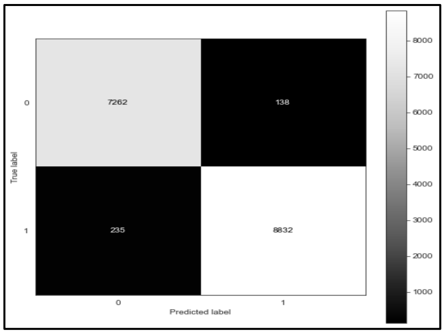

Figure 10. Random forest Classifier

Due to its excellent accuracy, resilience, feature significance, adaptability, and scalability, Random Forest is a well-known machine learning technique used for classification and regression problems. The accuracy, recall, precision, and F1-score are displayed in this graph. It is noted that a 97.73% accuracy, 97.41% recall, 98.46% precision, and a 97.93% F1-score are attained.

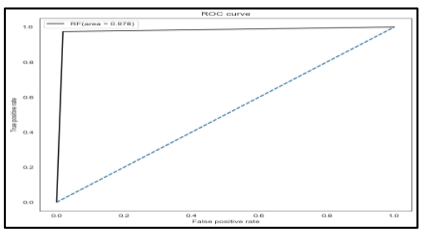

Figure 11. Roc curve for Random forest classifier

The Roc curve for the gradient boosting classifier is depicted in Figure 11. The x axis displays false positive rates, whereas the y axis displays real positive rates.

	Accuracy	Recall	Precision	F1-Score	time to train	time to predict	total time
CLF	96.03%	95.94%	96.83%	96.38%	1.8	0.3	2.1
GBC	95.87%	96.05%	96.43%	96.24%	51.1	0.1	51.2
KNN	96.44%	96.16%	97.33%	96.74%	0.0	6.7	6.7
RF	97.73%	97.41%	98.46%	97.93%	8.5	0.2	8.6

Figure 12. Model Performance

Figure 12 shows that predicted the accuracy, recall, precision, F1-score, time to train, time to predict, total time of CBC, GBC, KNN, RF.

4. CONCLUSION

In this project, it is suggested to classify an intrusion detection system using a machine learning classifier. To categorise intrusion detection systems, a larger dataset is utilised. Data preprocessing is the idea of turning unclean data into a clean data collection. The dataset is preprocessed to look for missing values, noisy data, and other anomalies before the algorithm is used. The Chi-square approach is used to extract features. The Chi-Square approach is used to process the extraction areas in order to extract a variety of features and choose the crucial traits to enhance categorization. In this project, feature extraction and feature selection are determined using the effective Chi-square method. The chosen characteristics are then added to a GBC, CBC, KNN, and random forest classifier for precise classification.

REFERENCES

[1] Vinayakumar, Ravi, Mamoun Alazab, K. P. Soman, Prabaharan Poornachandran, Ameer Al-Nemrat, and Sitalakshmi Venkatraman. "Deep learning approach for intelligent intrusion detection system." Ieee Access, Vol.7, pp: 41525-41550, 2019.

[2] Dahouda, Mwamba Kasongo, and Inwhee Joe. "A deep-learned embedding technique for categorical features encoding." IEEE Access, Vol. 9, pp: 114381-114391, 2021.

[3] H. Chen, J. Chen and J. Ding, "Data Evaluation and Enhancement for Quality Improvement of Machine Learning," in IEEE Transactions on Reliability, vol. 70, no. 2, pp. 831-847, June 2021.

[4] Ahmad, Iftikhar, Mohammad Basheri, Muhammad Javed Iqbal, and Aneel Rahim. "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection." IEEE access, Vol. 6, pp: 33789-33795, 2018.

[5] M. Wang, K. Zheng, Y. Yang and X. Wang, "An Explainable Machine Learning Framework for Intrusion Detection Systems," in IEEE Access, vol. 8, pp. 73127-73141, 2020.

[6] Bertoli, Gustavo De Carvalho, Lourenço Alves Pereira Júnior, Osamu Saotome, Aldri L. Dos Santos, Filipe Alves Neto Verri, Cesar Augusto Cavalheiro Marcondes, Sidnei Barbieri, Moises S. Rodrigues, and José M. Parente De Oliveira. "An end-to-end framework for machine learning-based network intrusion detection system." IEEE Access, Vol.9, pp: 106790-106805, 2021.

[7] Y. K. Saheed and M. O. Arowolo, "Efficient Cyber Attack Detection on the Internet of Medical Things-Smart Environment Based on Deep

Recurrent Neural Network and Machine Learning Algorithms," in IEEE Access, vol. 9, pp. 161546-161554, 2021.

[8] X. Gao, C. Shan, C. Hu, Z. Niu and Z. Liu, "An Adaptive Ensemble Machine Learning Model for Intrusion Detection," in IEEE Access, vol. 7, pp. 82512-82521, 2019.

[9] Goudjil, Mohamed, Mouloud Koudil, Mouldi Bedda, and Noureddine Ghoggali. "A novel active learning method using SVM for text classification." International Journal of Automation and Computing, Vol.15, pp: 290-298, 2018.

[10] M. A. Ferrag, L. Shu, O. Frida and X. Yang, "Cyber Security Intrusion Detection for Agriculture 4.0: Machine Learning-Based Solutions, Datasets, and Future Directions," in IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 3, pp. 407-436, March 2022.