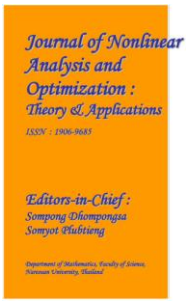
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1, No.15 : 2024
ISSN : **1906-9685**



Paper ID: ICRTEM24_133 **ICRTEM-2024 Conference Paper**

COMPARING DEEP LEARNING AND MACHINE LEARNING IN INTRUSION DETECTION SYSTEMS

[#1]K.CHANDRASENA CHARY, ***Research Scholar,***
[#2]Dr. ANOOP SHARMA, ***Guide,***
[#3]Dr.KISHOR KUMAR GAJULA, ***Co-Guide,***
Department of Computer Science & Engineering,
UNIVERSITY OF TECHNOLOGY, JAIPUR, RAJASTHAN
Corresponding Author: *K.Chandrasena Chary,* vemula.ramakrishna@yahoo.com

Abstract -Integrating blockchain technology into a cloud computing system improves the security and privacy of data and transactions. The term blockchain-based cloud computing refers to this type of cloud computing. Cloud computing enables you to store, manage, and access information online. In contrast, blockchain technology transparently and securely manages and stores data. Cloud computing and blockchain technologies may improve data security, transparency, and deter theft. The advantage of blockchain-based cloud computing is that it enables secure data management and storage on a public platform. Blockchain technology is used to store data in an independent web network. This makes it more difficult for unauthorized individuals to get access and modify data. Blockchain technology can create an unchangeable record of every transaction. This allows you to keep an eye on and verify the data. One of the most exciting aspects of blockchain-based cloud computing is its potential to improve data privacy. Blockchain technology provides a decentralized and secure method of processing and storing data, significantly reducing the danger of data breaches and intrusions. Blockchain technology can safeguard user data and enable private, secure communication among users. Integrating blockchain technology with cloud computing infrastructure might substantially alter how data is accessed, processed, and stored. Blockchain-based cloud computing enables users to manage their data in a secure, decentralized manner, potentially increasing openness, security, and privacy.

Key Words*:* IDS, Intrusion, KDD99, Logistic Regression, Naïve Bayes , Random Forest and CNN

1.INTRODUCTION

As cyberthreats have increased, so has the necessity of cyber security. The phrase cybersecurity refers to the procedures and instruments used to secure computer systems from risks to availability, confidentiality, and integrity. Internal and external intrusions are two types of security breaches: attacks from within and outside the business. An invader is someone who attempts to gain unauthorized entry, and any unwanted access is considered an intrusion. When suspicious or malicious activity is discovered in

network traffic, an intrusion detection system (IDS) notifies users. It makes a significant contribution to cybersecurity. Intrusion detection systems use two approaches to identify attacks: anomaly-based detection and signature-based detection. Signature-based detection compares data activity to a previously established signature or pattern. One disadvantage of signature-based detection is that it ignores any newly found malicious activity that is not already in the database. The behavior-based or statistical anomaly-based detection method detects any abnormality and generates an alarm, assisting in the identification of new sorts of dangers. Machines can mimic human behavior and attributes by utilizing artificial intelligence. Artificial intelligence enables machines to learn from their own experiences. Robots can perform human-like tasks such as evaluating massive volumes of data and recognizing patterns because they can adapt to new inputs. AI includes machine learning as a subset. It enables robots to learn from previous experiences and predict future events based on data. Deep learning, a type of machine learning, achieves tremendous strength and flexibility by teaching itself to represent the world as a hierarchical hierarchy of concepts or abstractions[1]. Data mining categorization entails dividing data examples into one or more categories. The process of categorizing a set of data into groups is known as classification. It supports both structured and unstructured data. The initial stage is to predict how the provided data points will be classified. Classes may also be referred to as labels, aims, or categories. Because the goal of attack detection is to discriminate between legitimate and malicious packets, it is considered a classification challenge. A lot of classification algorithms are designed to outperform each other. This study uses the KDD99 dataset to compare Deep Learning (CNN) with Naïve Bayes, Random Forest, and Logistic Regression. The algorithms are compared based on F1 score, accuracy, precision, and recall rate.

2. RESEARCH METHODOLOGY

Dataset Description

This study used the KDD99 dataset. The KDD99 compilers derived 41-dimensional features using DARAPA1998 data. The titles in KDD99 and DARPA1998 are identical. KDD99 divides characteristics into four categories: fundamental, content-based, time-based, and host-based statistics. The KDD99 dataset consists of 41 features per record. Each report is categorized as either "normal" or "specific type of intrusion." There are four types of intrusions: denial of service (DoS), illegal remote machine access (R2L), unauthorized superuser access (U2R), and probing attacks. Due to the large size of the KDD99 dataset, the project received 10% of it. The dataset was imported using pandas' read_csv() method.

Jupyter Lab

JupyterLab is a web-based interactive development environment designed exclusively for use with Jupyter notebooks, code, and data. JupyterLab's user interface is flexible and well-organized to facilitate a wide range of machine learning, scientific computing, and data exploration tasks. The platform provides a robust, versatile, and dependable environment for managing documents and tasks, including Jupyter notebooks, text editors, terminals, and customizable components.Python was used to develop the algorithms and evaluate the experimental data.

Data Preprocessing

Normalization is widely used in conjunction with feature scaling during data preprocessing for Machine Learning. The data was normalized using the Normalizer class from the sci-kit-learn toolkit. It normalizes the dataset's numerical columns to a uniform scale, preserving data integrity while reducing distortions caused by value range differences. This assignment requires binary classification to distinguish between a valid connection and a malicious attack. Standard connections are classified as one category, but all attacks are classified separately. After training the dataset, predictions are generated based on both the testing and trained data.

Implementation

Logistic Regression

The LR algorithm uses a parametric logistic distribution to determine the probabilities of various classifications. Developing and training LR models is straightforward and effective. LR's inability to handle nonlinear data limits its utility. It calculates the probability of an event with two outcomes. This method is fundamental but highly effective for addressing binary classification problems.We use the Scikit-learn library to build our logistic regression model.

The LogisticRegression class is defined and imported from the Sklearn library. After creating the model, the fit technique is used to train the data using the logistic regression model. Once the model has been trained, the data is predicted using the logistic regression model. The logistic regression model generates a normalized confusion matrix, as shown in Figure 1.

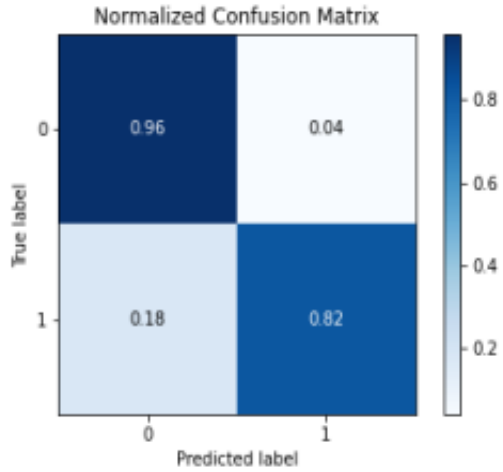


Fig -1: Confusion matrix for Logistic Regression

Random Forest Random

Forest Classifier: is one of the classification trees algorithms, the main goal of this algorithm is to enhance trees classifiers based on the concept of the forest. To implement this algorithm the number of trees within the forest should be figured because each individual tree within a forest predicts the expected output. Then next the voting technique is used to select the expected output that has the largest votes number [4]. Random Forest model is imported from sklearn and the model is instantiated, and an then we use the fit method on the model to train the data. After training the model, we predict the data usingpredict method on the random forest model. The obtained normalized confusion matrix of random forest is given in Fig-2.

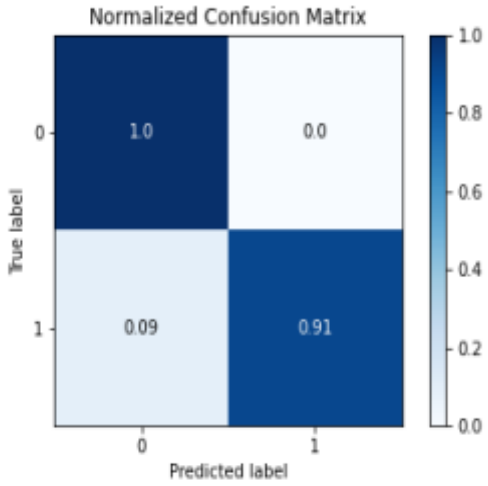


Fig -2: Confusion matrix for Random Forest

Naïve Bayes

Naïve Bayes is a supervised learning technique used for classification applications. It is supported by Bayes' theorem. This is primarily utilized in text classification when dealing with a training set including numerous variables. The Naive Bayes predictor accelerates the development of machine learning models by assisting with object categorization and giving fast prediction speeds.The predictor works as a probabilistic model, producing predictions based on the likelihood of a particular event occurring [5]. Strong individualistic principles imply that the chance of one attribute should not influence the probability of another. It is possible to accelerate the training or building of a classification model [6]. This strategy is used in predictive modeling to categorize items. It performs well in binary classification, as well as two- and multi-class classifications. To help comprehend this method, supply input in the form of categories or binary integers. The concept has proven effective in a variety of real-world scenarios, particularly in document categorization and spam avoidance. The code calls the GaussianNB function from the sklearn.naive_bayes library and imports the sklearn module. Fit is the process of training a logistic regression model using data once it has been created. Following model training, the speculation technique is applied to newly acquired

data. Figure 1 shows the corrected confusion matrix generated by the naïve Bayes model.

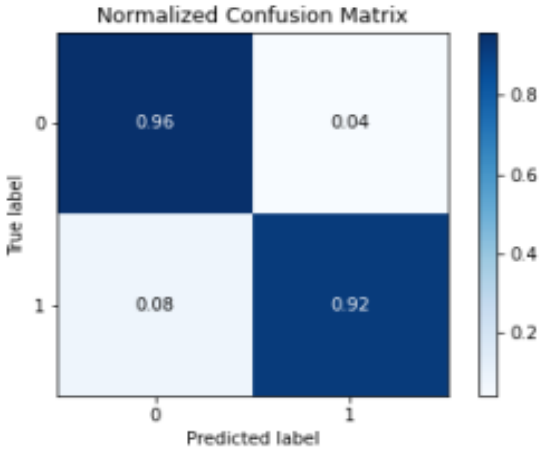


Fig -3: Confusion matrix for Naïve Bayes

CNN

A convolutional neural network is a sort of neural network that transforms high-resolution input into lower-resolution data with more detailed features.The model was built with the Keras toolbox's Sequential() method. This capability allows for the slow construction of models, one layer at a time. Before training a model, the input datasets must be changed to fit the desired format. The network has five secret layers: convolution, thickening, pooling, flattening, and dropout. The flatten layer is between the dense layer and Conv2D. The dropout layer's goal is to mitigate data overfitting. The Fit() method was called using the number of epochs, goal data, training data, and validation data to train the model. Epochs determine how many times the model runs through the data in a single training session. The model shows steady progress over time. After each iteration, the model stops evolving. We intend to run our model over ten epochs. By the first epoch, it was clear that the training of the five hidden layers had to achieve a minimum accuracy of 98.57%. In terms of equal levels, the training accuracy reached 99.73% at epoch 10. Figure 4 depicts the adjusted confusion matrix generated by the CNN.

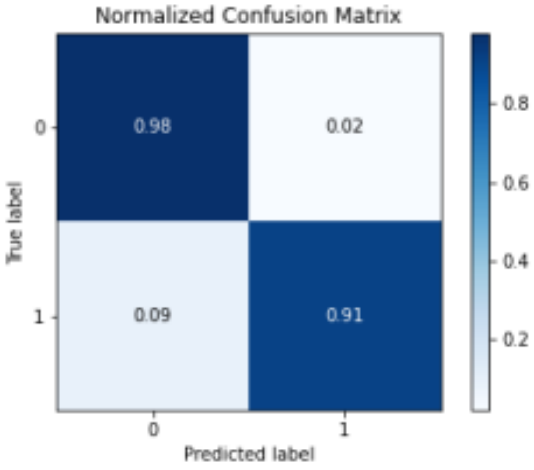


Fig -4: Confusion matrix for CNN

3. RESULTS

The algorithms were assessed and evaluated using a variety of performance metrics, including recall, accuracy, precision, and the f1-score. Someone was notified about categorization in a report. Figure 5 depicts the models' output as a bar graph. The most accurate model was Random Forest, with a weighted average precision of 0.95. The accuracy was higher than Naïve Bayes (0.94), CNN (0.94), and Logistic Regression (0.91). CNN, Naïve Bayes, and Random Forest all achieved a weighted average recall of 0.93. The grade for logistic regression was 0.85. Logistic Regression earned 0.86, whereas Random Forest, Naïve Bayes, and CNN achieved an accuracy of 0.93.Table 1 displays the accuracy statistics in a tabular format.

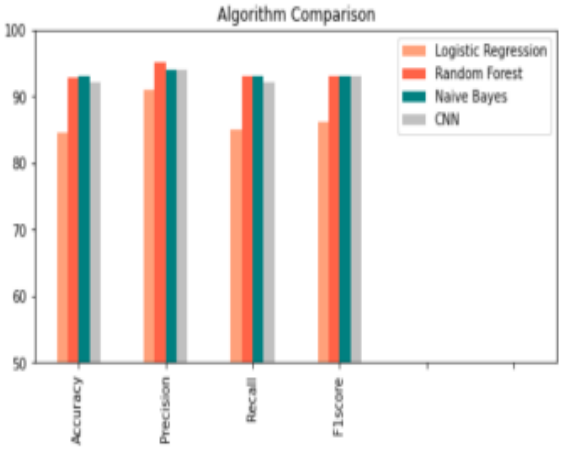


Fig -5: Algorithm Comparison

Table-1: Evaluation Metrics Table

Algorithm	Accuracy
Logistic Regression	84.6%
Random Forest	92.7%
Naïve Bayes	92.9%
CNN	92.1%

4. CONCLUSION

To detect intrusions, we compared CNN to machine learning methods Naïve Bayes, Random Forest Classifier, and Logistic Regression. The KDD99 dataset was utilized to complete the task. Because the KDD99 dataset is so huge, only 10% was used. The models were used to train and evaluate.The Naïve Bayes Classifier achieved an accuracy of 92.9%, surpassing CNN, Logistic Regression, and Random Forest Classifier.Compared to previous models, the Naïve Bayes model required less training time. After Nave Bayes, CNN and Random Forest are the next best models. The logistic regression model fared badly on our dataset when compared to other techniques, with an accuracy rate of 84.6%.

REFERENCES

1. W. S. A. Y. J. a. M. A. Quamar Niyaz, "A Deep Learning Approach for Network Intrusion Detection System," 2016.
2. V. K. P. a. S. R. K. Sharmila Kishor Wagh, "Survey on Intrusion Detection System using Machine Learning Techniques," 2013.
3. K. a. M. Dua, "Machine Learning Approach to IDS: A Comprehensive Review," 2019.
4. J. K. H. L. T. T. a. H. K. Jihyun Kim, "Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection," 2016.
5. Z. Wang, "Deep Learning-Based Intrusion Detection With Adversaries," 2018.
6. C.-H. R. L. Y.-C. L. a. K.-Y. T. Hung-Jen Liao, "Intrusion detection system: A comprehensive review," 2013.
7. T. N. N. V. D. P. a. Q. S. Nathan Shone, "A Deep Learning Approach to Network Intrusion Detection," 2018.
8. F. R. L. Y. X. C. L. Z. F. L. Chongzhen Zhang, "A Deep Learning Approach for Network Intrusion Detection Based on NSL-KDD Dataset," 2019.
9. M. A. S. K. P. P. A. A.-N. A. S. V. Vinayakumar R, "Deep Learning Approach for Intelligent Intrusion Detection System," 2018.
10. L. M. H. J. a. R. S. Mohamed Amine Ferrag, "Deep Learning Techniques for Cyber Security Intrusion Detection : A Detailed Analysis," 2019.
11. M. C. A. A. a. M. K. Usman Shuaibu Musa, "Intrusion Detection System using Machine Learning Techniques: A Review," 2020.