International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

# Comparative Analysis of Deep Learning Algorithm with Machine Learning Algorithm in Intrusion Detection System

**Hiba K.I[1], Dr. Anjana S Chandran[2]**

*[1]Student, Department of Computer Applications, SCMS School of Technology and Management, Muttom, Aluva, 683106*
*[2]Associate Professor, Department of Computer Applications, SCMS School of Technology and Management, Muttom, Aluva, 683106*

---

**Abstract -***Around the globe, cyber security is very vital as cyber-attacks are becoming an increasingly serious problem. Cyber security is the set of technologies and processes designed to protect computing systems against threats to confidentiality, integrity, and availability. An Intrusion Detection System (IDS) monitor for suspicious or malicious activity in the network traffic and it alerts when such an activity is discovered. It plays an important role in ensuring cyber security. This paper presents a comparative analysis of deep learning algorithm Convolutional Neural Network (CNN) with machine learning algorithms logistic regression, random forest and naïve bayes. To test the efficiency of the algorithms, data that is preprocessed is given as input to the algorithm and its performance is found from the observed results. The algorithms are trained and tested using KDD99 dataset and the evaluation metrics used to evaluate the experimental results include accuracy, precision, recall rate, and F1 score.*

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

## 1.INTRODUCTION

With the increasing cyber attacks, cyber security has become more vital. Cyber security is the set of technologies and processes designed to protect computing systems against threats to confidentiality, integrity, and availability. The security breaches include external intrusion and internal intrusion ie., attacks from outside the organization and attacks from inside the organization respectively. Any unauthorized access is called an intrusion and an intruder is a person who tries to gain unauthorized access. An Intrusion Detection System (IDS) monitor for suspicious or malicious activity in the network traffic and it alerts when such an activity is discovered. It plays an important role in ensuring cyber security. Intrusion detection systems use two types of methods to detect the attacks: signature-based detection and anomaly-based detection. Signature-based detection takes data activity and compares it to a signature or pattern. Signature-based detection has a constraint where a new malicious activity that is not in the database is ignored. Whereas anomaly-based detection method is the statistical anomaly-based or behavior-based detection, which detects any anomaly and gives alerts and hence it detects new types of attacks.

Artificial Intelligence is a technique which allows the machines to act like humans by replicating their behavior and nature. Artificial Intelligence makes it possible for the machines to learn from their experience. The machines adjust their response based on new inputs thereby performing human-like tasks by processing large amounts of data and recognizing patterns in them. Machine Learning is a subset of artificial intelligence. It allows the machines to learn and make predictions based on its experience(data). Deep learning is a particular kind of machine learning that achieves great power and flexibility by learning to represent the world as nested hierarchy of concepts or abstraction[1]. Classification is a data mining method of assigning data instances into one among the few categories. In classification, the process is to categorize a given set of data into classes. It can be performed on structured data or unstructured data. The process start by predicting the class of the given data points. The classes are also called target, label or categories. Attack detection is considered a classification problem as the target is to classify whether the packet is either normal or attack packet. Many classification algorithms are developed to outperform one another.

In this paper, Deep learning algorithm CNN is compared with machine learning algorithm Logistic Regression, Random Forest and Naïve Bayes using KDD99 dataset. The

Algorithms are being compared in terms of accuracy, precision, recall rate, and F1 score.

## 2. RESEARCH METHODOLOGY

### 2.1 Dataset Description

KDD99 dataset was used to carry out this project. KDD99 compilers extracted 41-dimensional features from data inDARPA1998. The labels in KDD99 are the same as the DARPA1998. There are four types of features in KDD99, i.e., basic features, content features, host-based statistical features, and time-based statistical features [2]. KDD99 consist of 41 features in total for each record. Each record is labeled as normal or the particular type of intrusion. There are 4 main intrusion types namely, dos, r2l, u2r and probe. Since there are many records, 10% of KDD99 was taken to carry out the project. To import the dataset, read_csv() function of pandas library was used.

### 2.2 Jupyter Lab

JupyterLab is a web-based interactive development environment for Jupyter notebooks, code, and data. JupyterLab is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning. It enables you to work with documents and activities such as Jupyter notebooks, text editors, terminals, and custom components in a flexible, integrated, and extensible manner [3] .The algorithms were implemented using jupyter lab and the experimental results were parsed and analyzed in Python.

### 2.3 Data Preprocessing

The normalization technique is often applied as part of feature scaling in data preparation for machine learning. So for data preprocessing, normalization method was used by importing the Normalizer class of the sci-kit-learn library. It is performed to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information. In this project, we are dealing with binary classification i.e., whether the connection is normal connection or is it an attack. So we denote normal connections as one class and all the attack as another class. After training the dataset, the testing data is taken and predicted with the help of trained data.

### 2.4 Implementation

#### 2.4.1 Logistic Regression

The LR algorithm computes the probabilities of different classes through parametric logistic distribution. An LR model is easy to construct, and model training is efficient. However, LR cannot deal well with nonlinear data, which limits its application[2]. It predicts the chance of an outcome which have dichotomy values i.e., two values. It is a simple but very powerful algorithm to solve binary classification problems.Scikit-learn library is used for our logistic regression model. LogisticRegression class is imported from sklearn library and defined. Once the model is defined, we use the fit method on the logistic regression model to train the data. After training the model, we predict the data using predict method on the logistic regression model. The obtained normalized confusion matrix of logistic regression model is given in Fig-1.

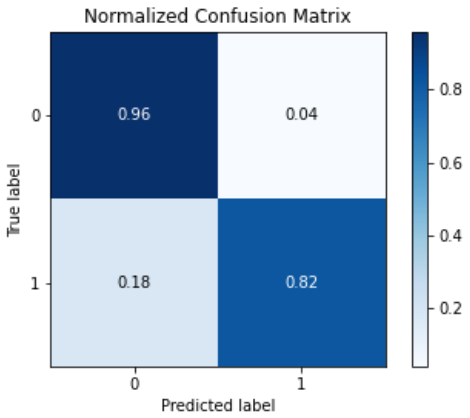


**Fig -1**: Confusion matrix for Logistic Regression

#### 2.4.2 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.

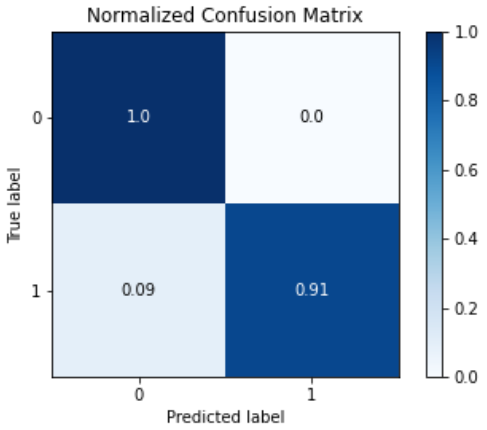


**Fig -2**: Confusion matrix for Random Forest

### 2.4.3 Naïve Bayes

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.It is a probabilistic classifier, which means it predicts on the basis of the probability of an object [5].

It is operated on strong individuality assumptions means that one attribute's probability should not affect that of other. It requires less time for training or modeling a model for classification [6]. It is a classification algorithm used for predictive modeling. It is used for binary i.e., two-class and multi-class classification problems. When the input values are described using binary or categorical input values, this technique is easiest to understand. It works well in many real-world situations, mainly document classification and spam filtering. GaussianNB from sklearn.naive_bayes is imported from sklearn library and defined. Once the model is defined, we use the fit method on the logistic regression model to train the data. After training the model, we predict the data using predict method on the model. The obtained normalized confusion matrix of naïve bayes is given in Fig-1.

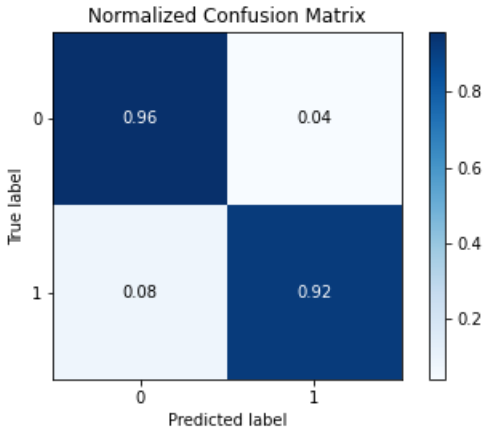


**Fig -3**: Confusion matrix for Naïve Bayes

### 2.4.4 CNN

A convolutional neural network is defined as a neural network that extracts features at a higher resolution, and then convert them into more complex features at a coarser resolution[7].To create the model we used Sequential() method from keras library. It allows building a model layer by layer. We need to reshape the dataset inputs to the shape that our model expects when we train the model. We have 5 hidden layers including convolution, dense, pooling, flatten, and dropout layer. In between the Conv2D layers and the dense layer, there is the flatten layer which serves as a connection between the convolution layer and dense layers. Dropout layer is used to reduce the overfitting of data. To train the model, fit() function was used with the parameters: training data, target data, validation data, and the number of epochs. The number of epochs defines the number of times that model will cycle through the data. The more epochs we run, it improves the model up to a certain point. After that point, the model stops to improve during each epoch. For our model, we will set the number of epochs to 10. For these 5 hidden layers, the minimum training accuracy was found to be 98.57% at epoch 1. For the same layers, the maximum training was found to be 99.73% at epoch 10. The obtained normalized confusion matrix of CNN is given in Fig-4.

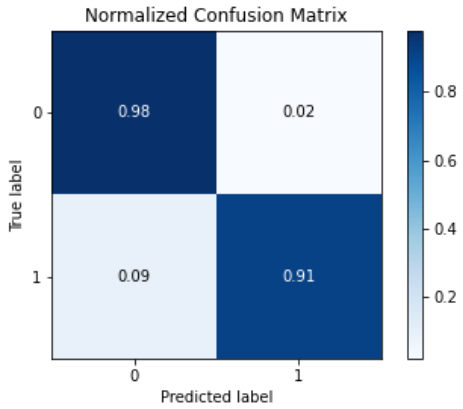


**Fig -4**: Confusion matrix for CNN

## 3. RESULTS

The comparative analysis of the algorithms was obtained by evaluation metrics which include accuracy, precision, recall and f1-score. It was obtained usingclassification report. The acquired results of the models are displayed in visualized format of bar graph in the fig.5. The weighted average of precision was highest for Random Forest with 0.95 , 0.94 for Naïve Bayes and CNN , and lowest for Logistic Regression with 0.91. The weighted average for recall was obtained as 0.93 for Random Forest and Naïve Bayes, 0.92 for CNN and 0.85 for Logistic Regression. Random Forest, Naïve Bayes and CNN obtained 0.93 whereas 0.86 for Logistic Regression.A tabulated result of accuracy is given in table-1.

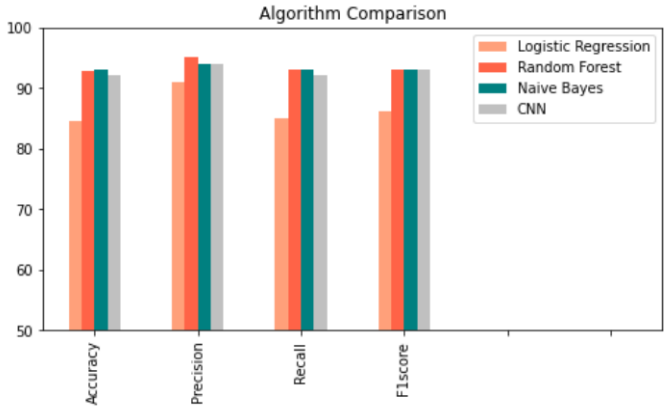


**Fig -5**: Algorithm Comparison

**Table-1**: Evaluation Metrics Table

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

## 4. CONCLUSION

In this paper, we conducted a comparative study of deep learning algorithm CNN and machine learning algorithm Logistic Regression, Random Forest Classifier, and Naïve Bayes algorithm for intrusion detection. The KDD99 dataset was used to carry out the project. 10% of actual KDD99 dataset was used as it has large amount of data. It was trained and tested with the models.It is observed that Naïve Bayes Classifier obtained more accuracy compared to Logistic Regression, Random Forest Classifier, CNN with 92.9%.It was observed that the time taken for training the Naïve bayes model was lesser compared to other models. After theNaïve Bayes model, the next better model is Random Forest and CNN. Logistic regression model showed poor performance with 84.6% accuracy compared to other algorithms on our dataset.

## REFERENCES

[1] "edureka," [Online]. Available: https://www.edureka.co/blog/ai-vs-machine-learning-vs-deep-learning/.

[2] H. L. a. B. Lang, "Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey," 2019.

[3] "Jupyter," [Online]. Available: https://jupyter.org/.

[4] M. A. S. K. a. M. A. Mohammad Almseidin, "Evaluation of Machine Learning Algorithms for Intrusion Detection System," 2017.

[5] "javapoint," [Online]. Available: https://www.javatpoint.com/machine-learning-naive-bayes-classifier.

[6] S. S. a. B. Bhushan, "Use of Machine learning algorithms for designing efficient cyber security solutions," 2019.

[7] L. M. H. J. a. R. S. Mohamed Amine Ferrag, "Deep Learning Techniques for Cyber Security Intrusion Detection : A Detailed Analysis," 2020.

[8] N. T. F. M. B.-A. a. A. Suad Mohammed Othman, "Survey on Intrusion Detection System Types," 2018.

[9] D. S. S. a. D. P. S. Aumreesh Ku. Saxena, "General Study of Intrusion Detection System and Survey of Agent Based Intrusion Detection," 2017.

[10] W. S. A. Y. J. a. M. A. Quamar Niyaz, "A Deep Learning Approach for Network Intrusion Detection System," 2016.

[11] V. K. P. a. S. R. K. Sharmila Kishor Wagh, "Survey on Intrusion Detection System using Machine Learning Techniques," 2013.

[12] K. a. M. Dua, "Machine Learning Approach to IDS: A Comprehensive Review," 2019.

[13] J. K. H. L. T. T. a. H. K. Jihyun Kim, "Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection," 2016.

[14] Z. Wang, "Deep Learning-Based Intrusion Detection With Adversaries," 2018.

[15] C.-H. R. L. Y.-C. L. a. K.-Y. T. Hung-Jen Liao, "Intrusion detection system: A comprehensive review," 2013.

[16] T. N. N. V. D. P. a. Q. S. Nathan Shone, "A Deep Learning Approach to Network Intrusion Detection," 2018.

[17] 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.

[18] M. A. S. K. P. P. A. A.-N. A. S. V. Vinayakumar R, "Deep Learning Approach for Intelligent Intrusion Detection System," 2018.

[19] L. M. H. J. a. R. S. Mohamed Amine Ferrag, "Deep Learning Techniques for Cyber Security Intrusion Detection : A Detailed Analysis," 2019.

[20] M. C. A. A. a. M. K. Usman Shuaibu Musa, "Intrusion Detection System using Machine Learning Techniques: A Review," 2020.

[21] I. G. P. V. a. J. K. Ansam Khraisat, "Survey of intrusion detection systems: techniques, datasets and challenges," 2019.

[22] V. V. R. P. a. K. M. P. V. Jyothsna, "A Review of Anomaly based Intrusion Detection Systems," 2011.

[23] M. a. Anna DrewekOssowicka, "A survey of neural networks usage for intrusion detection systems," 2020.

[24] A. L. B. a. E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," 2015.

[25] R. a. T. Thilagam, "A Review on the Effectiveness of Machine Learning and Deep Learning Algorithms for Cyber Security," 2020.

[26] L. K. ,. Z. L. Y. C. Y. L. H. Z. G. H. H. a. C. W. Yang Xin, "Machine Learning and Deep Learning Methods for Cybersecurity," 2018.

[27] U. B. a. M. Chachoo, "Intrusion Detection and Prevention System: Challenges & Opportunities," 2014.

[28] Y. N. I. a. F. J. Abdullayeva, "Deep Learning in Cybersecurity: Challenges and Approaches," 2020.