Performance evaluation of the learning models for Intrusion Detection using Machine Learning

Vyom Dubey
Student
School of computer science Engineering and information system Vellore institute of technology Vellore, Tamil Nadu, India
vyom.dubey2023@vitstudent.ac.in

Aastha Agarwal
Student
School of computer science Engineering and information system Vellore Institute of technology Vellore, Tamil Nadu, India
aastha.agrawal2023@vitstudent.ac.in

V MOHAMMED Hisham
Student
School of computer science Engineering and information system Vellore institute of technology Vellore, Tamil Nadu, India
mohamedhisham.v2023@vitstudent.ac.in

Siva Rama Krishnan S
Associate professor
School of computer science Engineering and information system
Vellore institute of technology
Vellore, Tamil Nadu, India
siva.s@vit.ac.in

Abstract- We live in the digital age, where most devices are interconnected, and information is in digital form. Security is a concern for individuals, organizations, and governments alike. With the rapid development of technology, Threats and intrusions are evolving as well, this situation demands advance development. Traditional Intrusion detection approaches rely on rule-based algorithm, As artificial intelligence and subdomains machine learning doing well in various fields this paper aims for comparative study between various algorithms for training and testing of Intrusion detection System (IDS).

1. Introduction

Technology is crucial in the modern era, but it has both positive and negative consequences. Cybercrime has grown rapidly due to tech innovation, with some society adopting advanced tactics for unlawful acts. Network security is crucial against malicious users, leading to the development of IDS to detect network and computer system attacks.

1.1 The Intrusion detection system

Intrusion are illegal actions that undermine data integrity, availability, or confidentiality in infrastructure or computers. As network assaults increase, intrusion detection systems (IDS) are crucial for protecting computer systems and networks from hackers. IDS monitors, identifies, and evaluates unusual activity, potentially breaching security regulations. The primary goal is to detect various forms of malware faster than a standard firewall.

1.2 Types of IDS

Signature-based intrusion detection systems (IDS) study specific patterns in network traffic, known as signatures, to detect attacks. However, they struggle to detect new malware attacks due to their unknown signatures. Anomaly-based IDS, introduced for finding unknown malware attacks, uses machine learning techniques due to their better generalized property. These models can be trained based on applications and hardware configurations. Anomaly-based IDS creates a trustful activity model, comparing any new threats to it, and if they are not found, it is considered suspicious. This approach is more effective in detecting new malware attacks.

2. Machine-learning-based Methods for Implementing IDSs

The number of threats necessitates the development of better intrusion detection systems (IDSs) capable of identifying complex malwares. Machine learning has been utilized to enhance intrusion detection by extracting information from large data sets, revealing patterns, and predicting actions, making it a widely used technique in intrusion detection systems.

2.1 Related work

Abu Al-Haija et al. [1] This paper enhances feature extraction and learning performance by combining deep convolutional neural networks with a UAV-IDS-2020 dataset two-class classifier. The UAV-IDS-ConvNet architecture provides a robust foundation for two-class prediction applications,

achieving detection accuracy of 99.50% and 90.00% in heterogeneous and homogeneous contexts.

CHUANLONG YIN et al. [2] The study presents a recurrent neural-network-based deep learning approaches for the intrusion detection systems, evaluating its performance in binary and the multiclass classifications, and comparing it to other machine learning approaches, revealing that the RNN-IDS is an excellent choice for intrusion detections.

Dr. D. William Albert's[3] Anomaly-based intrusion detection faces challenges in dealing with unique attacks without prior knowledge. This research applies SVM and ANN to the NSLKDD benchmark dataset to evaluate network intrusion performance.

R.Ravinder Reddy et al.[4] focus on improving the classification accuracy of intrusion detection in heterogeneous datasets, despite the effectiveness of SVM-based kernels. They aim to scale the dataset and enhance discriminant function performance.

Shistrut Rawat et al.[5] Intrusion detection is crucial for determining user behavior normality. SVM-based kernels categorize data, but heterogeneous data makes it difficult. Scaling datasets improves discriminant function performance for intrusion classification.

Min Du et al.[6] This study presents Deep Log, a deep learning-based approach for the anomaly detection in system logs. It identifies patterns in complex logs, ensuring system security and reliability. The model can distinguish regular behavior from anomalies, even in noise, enabling real-time responses to threats or failures. The report details its design, preprocessing, and validation.

3. Methodology

The technique that will be utilized in this study to identify network intrusions consists of these following steps:
(I) pre-processing and standardizing the data
(II) choosing features and
(III) classifying the results by analyzing different algorithm accuracy in DNN.

3.1 Dataset

The KDD99 dataset, widely used in network security, has been replaced by the NSL-KDD dataset, a publicly available dataset on Kaggle, due to its high number of identical records.

Traffic	Train	Test
Normal	67343	9711
Probe	11656	2887
U2R	52	67
Dos	45927	7458
R2L	995	2887

The NSL KDD dataset extracts 41 qualitative and quantitative characteristics for each TCP/IP connection, categorizing them as normal and aberrant, from 1,282,273 rows of training and test data.

3.2 Data preprocessing

The dataset underwent pre-processing and cleaning, removing null rows, scaling numerical columns, dropping categorical rows, and encoding some columns using hot encoding, with the outcome being a variable with 0,1 values.

3.3 Machine Learning Models

3.3.1 Naïve Bayesian

The naive Bayesian Classifier is a machine learning method based on the Bayes theorem, named after Thomas Bayes, which determines the likelihood of a class.

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

3.3.2 KNN

The K-nearest neighbor approach is a supervised learning classifier used for regression and classification problems, based on the concept of similarity. This KNN classifier is trained using 'n_neighbours = 20', with every data point in the immediate area assigned the same weight.

3.3.3 SVM

SVMs are supervised machine learning techniques used for classification and regression, dividing data into two groups. They can be used in multi-class situations, producing binary classification algorithms for each class.

3.3.4 Random Forest

A supervised learning algorithm, a random forest, generates decision trees trained using the "bagging" approach, improving classification and regression problems in machine learning systems and resisting overfitting, unlike decision trees.

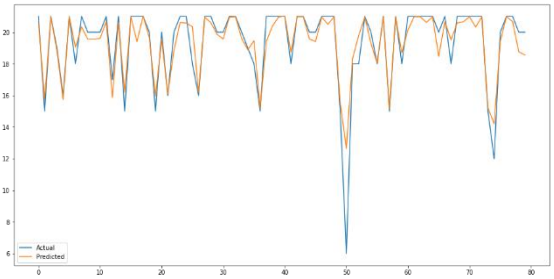
3.3.5 Neural Networks

Neural networks, or simulated neural networks (SNNs), often referred as artificial neural networks (ANNs) are crucial the in machine learning's and deep learning. Inspired by brain, they emulate biological neuron signaling using weights and thresholds. This architecture enhances learning capacity, enabling complex data processing and meaningful insights, supporting trend recognition, prediction, and decision-making.

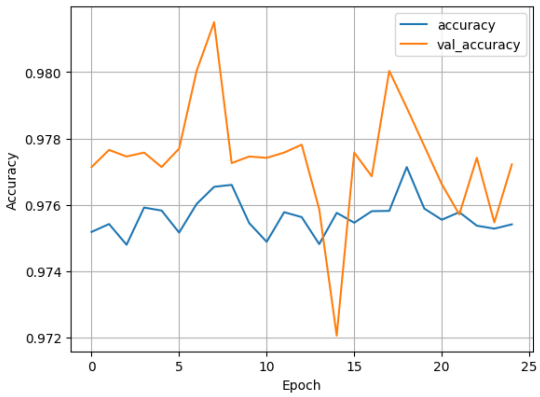
3.4 Evaluation and Results

3.4.1 XGboost

XGBoost is a efficient distributed gradient boosting approach used to develop machine-learning classifiers. It uses a tree-based strategy for rapid and reliable problem-solving, enhancing model performance by minimizing variance and enhancing training and testing speed.



XGBoost is a tree-based distributed gradient boosting technique used in machine-learning classifier development, enhancing model performance by minimizing variance and improving training and testing speed. for 25 Epoch the graph of accuracy and validation accuracy (val_accuracy) is shown in fig.



The study developed KNN, Nave Bayesian, Random forest, and SVC models for investigation, using PCA to reduce features and train them on NSL-KDD train data for evaluation.The random forest model has the highest training and test accuracy of 0.99, and the PCA with random forest has comparable results.

Training Algorithm	Training accuracy	Test Accuracy
KNN	0.99	0.98
Naïve Bayesian	0.91	0.91
Random forest	0.99	0.99
SVC (linear)	0.96	0.96
PCA + Random forest	099	099

4. Conclusion

In this article, many intrusion detection methods for networks were developed and evaluated. Every broader group has a large number of extra incursions, as the test dataset shows. In comparison to the test set accuracy, the model performed rather well when it was trained and tested utilizing the train-validation splits when new intrusions were seen. Random forest and PCA with Random Forest perform much better in terms of model fitting and accuracy on the test set, with a 0.993 accuracy, when compared to the other classifiers.

5. REFERENCES

[1] Abu Al-Haija, Qasem, and Ahmad Al Badawi, A. (2022). High-performance intrusion detection system for networked UAVs via deep learning. *Neural Computing and Applications*, *34*(13), 10885-10900.

[2] Yin, Chuanlong, Yuefei Zhu, Jinlong Fei, and Xinzheng He (2017). A deep learning approach for intrusion detection using recurrent neural networks. *Ieee Access*, *5*, 21954-21961.

[3] Taher, K. A., Jisan, B. M. Y., & Rahman, M. M. (2019, January). Network intrusion detection using supervised machine learning technique with feature selection. In *2019 International conference on robotics, electrical and signal processing techniques (ICREST)* (pp. 643-646). IEEE.

[4] Reddy, R. R., Ramadevi, Y., & Sunitha, K. N. (2016, September). Effective discriminant function for intrusion detection using SVM. In *2016 International conference on advances in computing, communications and informatics (ICACCI)* (pp. 1148-1153). IEEE.

[5] Rawat, S., Srinivasan, A., Ravi, V., & Ghosh, U. (2022). Intrusion detection systems using classical machine learning techniques vs integrated unsupervised feature learning and deep neural network. *Internet Technology Letters*, *5*(1), e232.

[6] Du, M., Li, F., Zheng, G., & Srikumar, V. (2017, October). Deeplog: Anomaly detection and diagnosis from system logs through deep learning. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security* (pp. 1285-1298).