World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

# Comparative analysis on intrusion detection system using machine learning approach

Venu Gopal Bitra [1], Ajay Kumar [1], Seshagiri Rao [1], Prakash [1] and Md. Shakeel Ahmed [2]

*[1] UG Students Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur, AP, India.*
*[2] Associate Professor, Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur, AP, India.*

World Journal of Advanced Research and Reviews, 2024, 21(02), 2555–2562

Publication history: Received on 19 February 2024; revised on 28 March 2024; accepted on 31 March 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.21.3.0983

## Abstract

The increasing popularity of online data storage and access has raised concerns about security and privacy in the face of growing online threats. However, with the rise of online threats, security and privacy have become major concerns. Intrusion detection systems (IDS) play an important role in protecting data integrity by identifying and quarantining records in the event of unexpected changes. Anomaly-based IDS, which uses machine learning-based approach and algorithms, is an effective way to detect known and unknown attacks, including zero-day attacks. The proposed project is used to create model to implement and analyze anomaly-based IDS to classify malicious attack types such as normal (non-intrusion), DoS, Probe, U2R and R2L. The analysis is conducted on KDDCup99 Dataset which consists of different attacks that a IDS go through. The Machine Learning Algorithms like KNN, SVM, Random Forest and LightGBM are used for the analysis. The Comparitive Analysis is made on KDDCup99 Dataset using the above Machine Learning Algorithms that uses the hybrid techniques and Ensemble techniques like Bagging and Boosting.

**Keywords:** Intrusion Detection System; Anomaly Based IDS; Analysis on KDDCup99; Machine Learning Approach.

## 1 Introduction

Intrusion detection system is a security mechanism designed to detect and prevent unauthorized access to computer networks and systems. The development of an effective intrusion detection system is crucial in ensuring the security of computer networks and systems against unauthorized access and attacks. With the increasing complexity of cyber threats, traditional rule-based systems are no longer sufficient in detecting and preventing security breaches. In this project, we selected the KDDCup99 dataset, a widely recognized bench mark dataset in the field of intrusion detection. Hence, the use of machine learning algorithms has become increasingly popular in IDS to enhance their accuracy and effectiveness by analyzing the KDDCup99 dataset. A Machine Learning approach is used for accurate prediction in the performance analysis.

This dataset contains comprehensive collection of network traffic data, including various types of network attacks and normal network activity. The system's primary objective was to identify and classify different types of network attacks, ranging from common attacks like DoS (Denial of Service) and probing attacks to more sophisticated intrusion attempts. The evaluation of the performance of machine learning algorithms is based on several metrics, such as accuracy, precision, recall and F1-score.The Machine Learning Algorithms like KNN, SVM, Random Forest and LightGBM are used for the analysis.

* Corresponding author: Venu Gopal Bitra

## 2 Methodology

The methodology employed in the research project for the comparative analysis on IDS using ML techniques consists of several key steps. It begins with collection of benchmark dataset, such as KDDCup99 that represents diverse network scenarios and intrusion types. The data undergoes through preprocessing which includes cleansing and normalization. The Machine Learning algorithms like KNN, SVM, Random Forest and Light Gradient Boosting are been selected for their capabilities in handling complex, imbalanced datasets. The models are trained and its performance is evaluated using metrics like accuracy, precision, recall and F1-score. The main goal is to analyze the KDDCup99 dataset and predict the best model.

Here's a detailed methodology broken down into steps

### 2.1 Data Collection

- Acquire relevant datasets, including the KDDCup99 which encompass different network traffic scenarios and intrusion types.
- Conduct data quality checks to ensure the datasets as reliable and free from errors, inconsistencies and missing values.

### 2.2 Data Preprocessing

- Perform data cleansing to handle issues like missing data, outliers and irrelevant features, ensure that data is consistent and suitable for analysis.
- Normalize or scale the data, if necessary to bring it into a consistent format model training.

### 2.3 Feature Selection

- The correlation matrix is built to detect the best feature that help to optimize the model and predict the accurate results.

### 2.4 Models Training

- Implement the KNN, SVM,Random Forest and LGBM algorithms and configure the parameters.
- Train the models using the pre-processed datasets, ensuring the data is divided into training and testing sets for evaluation.

### 2.5 Performance Evaluation

- Define a set of performance metrics, including accuracy, precision, recall,F1-score and confusion matrix to quantitatively assess the effectiveness of the models.
- Apply the metrics to the model's predictions to evaluate the performance.

### 2.6 Comparative Performance Analysis :

- The defined algorithms performance metrics are visualized and evaluated to predict the best model for the analysis of KDDCup99 dataset.

### 2.7 Purpose and Objectives of Proposed System

- The purpose of an Intrusion Detection System (IDS) is to enhance the security of computer systems and networks by identifying and responding to potential security incidents.
- The primary objectives and purpose of an IDS include:
- Threat or Anomaly Detection in the Network
- Attack Classifications such as DoS, Probe, U2R and R2L
- Helping organizations maintain the confidentiality
- Measure the ability of the intrusion detection system to accurately detect and classify different

## 2.8 Overall Architecture

### *2.8.1 Architecture*

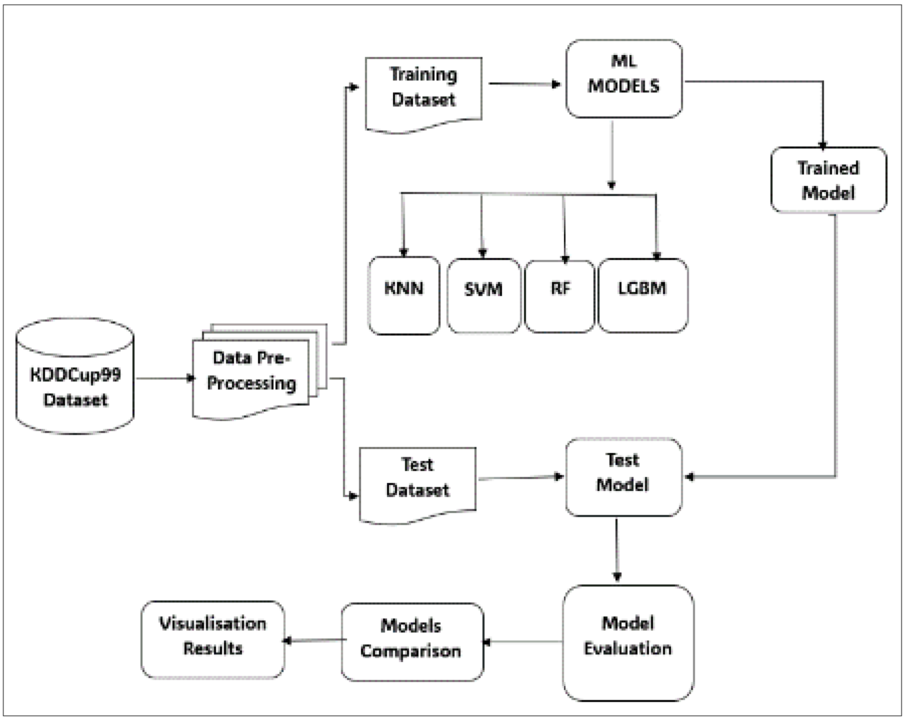


**Figure 1** System Architecture

## 2.9 Building of Correlation matrix

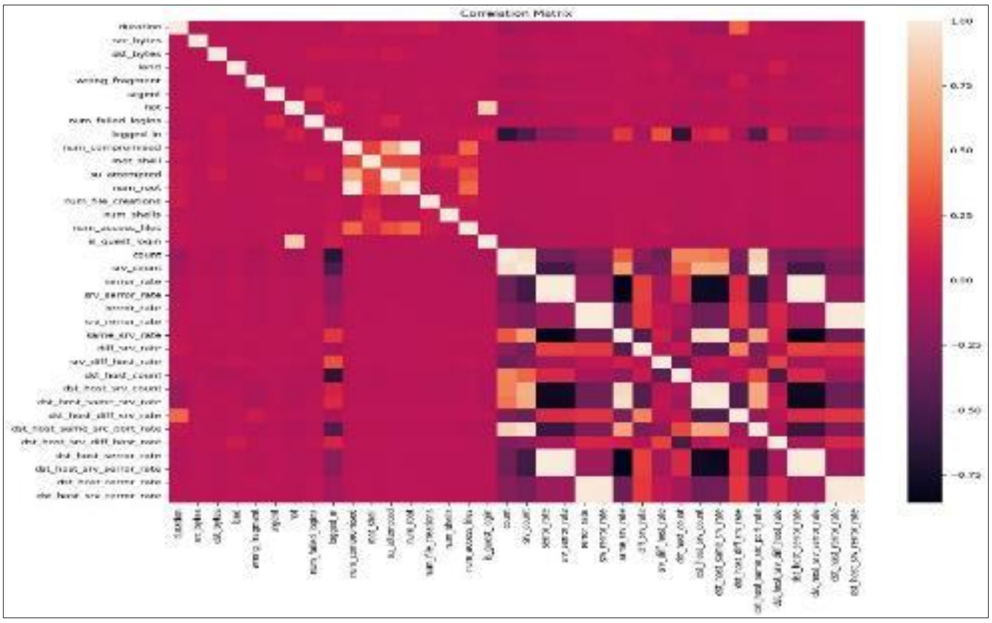


**Figure 2** Correlation matrix

The correlation matrices play a crucial role in data analysis, providing valuable insights into the relationships between variables and aiding in various analytical tasks such as feature selection, model building, and exploratory data analysis.

The main objective of using this correlation matrix includes:

- Understanding Relationships
- Feature Selection
- Multi Collinearity detection
- Data Exploration

## 3 Algorithms for IDS Analysis

### 3.1 K-Nearest Neighbors (KNN)

- KNN is a non-parametric classification algorithm that assigns a class label to a data point based on the majority class of its k-nearest neighbours in the feature space.
- KNN can be used in IDS to classify network traffic data into different categories (normal or attack) based on the similarity of network traffic patterns. It can effectively identify anomalies by comparing the behavior of network traffic with historical data.

### 3.2 Support Vector Machine (SVM) :

- SVM is a supervised learning algorithm that constructs a hyperplane in a high-dimensional space to separate data points into different classes, maximizing the margin between classes.
- SVMs are commonly used in IDS for binary classification tasks, where they can effectively classify network traffic as either normal or malicious based on features extracted from network packets. SVMs are particularly useful for detecting complex attack patterns.

### 3.3 Random Forest

- Random Forest is an ensemble learning method that builds multiple decision trees during training and combines their predictions through voting or averaging to improve accuracy and reduce overfitting.
- Here each decision tree is trained on a random subset of features and data points, and predictions are aggregated through voting or averaging.
- Random Forests are effective in IDS for both classification and anomaly detection tasks. They can handle high-dimensional data and capture complex relationships between network features, making them suitable for detecting various types of attacks with high accuracy.

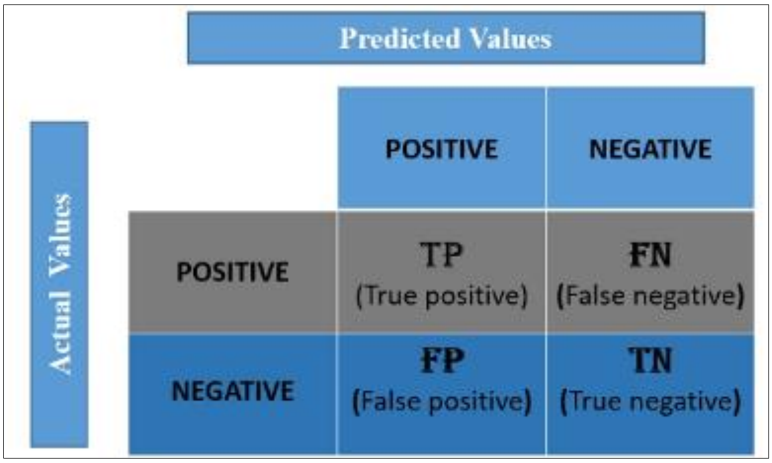
### 3.4 Light Gradient Boosting (LGBM) :

- Light Gradient Boosting Machine (LGBM) is a gradient boosting framework that uses tree-based learning algorithms and gradient boosting techniques to improve predictive accuracy.
- LGBM builds an ensemble of weak learners (decision trees) sequentially, where each new tree is trained to correct the errors of the previous ones, minimizing the loss function
- .LGBM is well-suited for IDS due to its ability to handle large-scale datasets efficiently and its high predictive performance. It can effectively identify complex attack patterns by combining multiple weak classifiers into a strong ensemble model, making it a popular choice for intrusion detection tasks in real-time network traffic analysis.

In this comparative analysis on the KDDCup99 Dataset, K-Nearest Neighbours (KNN) classifies instances based on similarity to neighbouring data points. Support Vector Machine (SVM) constructs hyperplanes to separate data into classes, maximizing the margin between them. Random Forest builds multiple decision trees and aggregates their predictions to improve accuracy. Light Gradient Boosting Machine (LGBM) sequentially trains ensemble models to minimize the loss function, achieving high predictive performance. Performance metrics such as accuracy, precision, recall, and F1-score are calculated to evaluate the effectiveness of each algorithm in intrusion detection anomalies.

# 4 Performance Metrics for Evaluation

## 4.1 Confusion Matrix



| | | Predicted Values | |
|---|---|---|---|
| | | POSITIVE | NEGATIVE |
| Actual Values | POSITIVE | TP<br>(True positive) | FN<br>(False negative) |
| | NEGATIVE | FP<br>(False positive) | TN<br>(True negative) |

**Figure 3** Confusion Matrix

- A confusion matrix is a tabular representation that helps to evaluate the performance of the models
- The terms "TP","FN","TN","FP" are commonly used in context of prediction to represent different outcomes in a confusion matrix.
    - True Positive (TP): The number of observations correctly predicted as positive for each class in a multi-class confusion matrix.
    - False Negative (FN): The number of observations incorrectly predicted as negative when they are actually positive for each class in a multi-class confusion matrix.
    - True Negative (TN): The number of observations correctly predicted as negative for each class in a multi-class confusion matrix.
    - False Positive (FP): The number of observations incorrectly predicted as positive when they are actually negative for each class in a multi-class confusion matrix.

## 4.2 Accuracy

- The proportion of correctly classified instances among all instances, regardless of class, in a multi-class confusion matrix.
- Proportion of correctly classified instances among all instances.

$$\text{Accuracy} = (TP + TN) / (TP+TN+FP+FN)$$

## 4.3 Precision

- The proportion of correctly predicted instances for a specific class out of all instances predicted as that class.
- Proportion of correctly identified positive cases out of all instances predicted as positive.

$$\text{Precision} = (TP) / (TP+FP)$$

## 4.4 Recall

- The proportion of correctly predicted instances for a specific class out of all instances belonging to that class.
- Proportion of correctly identified positive cases out of all actual positive cases.

$$\text{Recall} = (TP) / (TP+FN)$$

### 4.5 F1-Score

- The harmonic mean of precision and recall, providing a balanced measure of a model's performance across all classes in a multi-class confusion matrix.
- Harmonic mean of precision and recall, providing a balance between the two metrics.

F1-Score = (2*Precision*Recall) / (Precision + Recall)

## 5 Outputs and Results

**Table 1** Time and Scores Comparision

| MODEL | Training Time | Testing Time | Training Score | Testing Score |
|---|---|---|---|---|
| SVM | 198.045 | 132.095 | 99.871 | 99.876 |
| Random Forest | 0.944 | 0.051 | 99.969 | 99.917 |
| LGBM | 3.367 | 0.402 | 95.518 | 95.412 |
| KNN | 0.532 | 280.063 | 99.899 | 99.856 |

**Table** 2 Performance Metrics Comparision

| MODEL | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| SVM | 94.91 | 72.38 | 80.61 | 71.42 |
| Random Forest | 98.74 | 88.19 | 98.62 | 87.97 |
| LGBM | 95.41 | 79.48 | 95.42 | 80.35 |
| KNN | 92.98 | 69.28 | 83.82 | 79.36 |

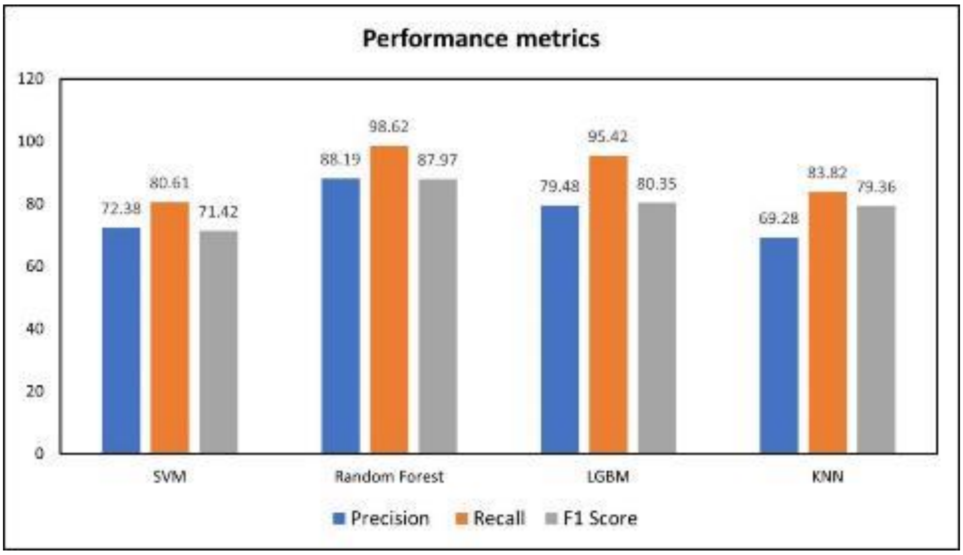


**Figure 4** Comparision of Performance Metrics

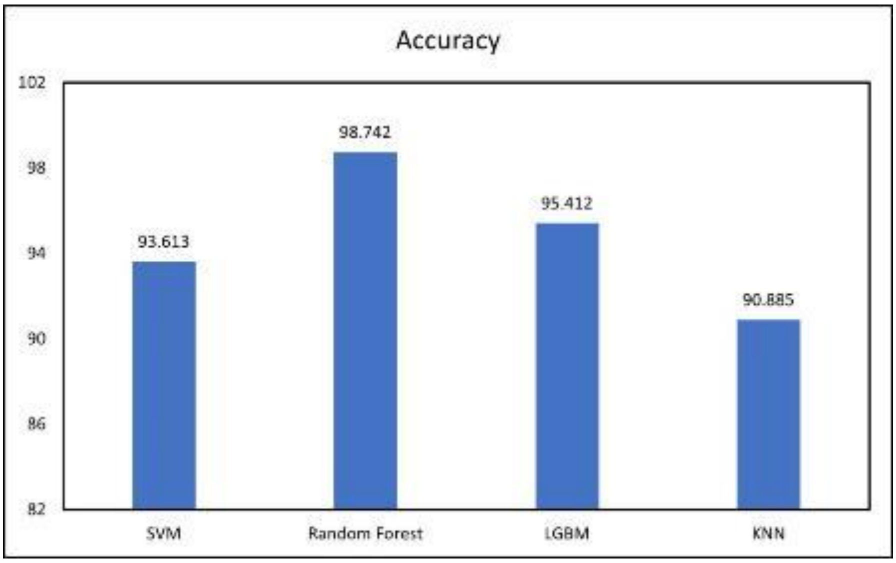


**Figure 5** Visualization results

## 6 Conclusion

The Comparative analysis on IDS KDDCup99 Dataset is performed to analyze the accurate performance of machine learning algorithms and ensemble techniques. The techniques like KNN, SVM, Random Forest and LGBM are used for the analysis. K-Nearest Neighbors (KNN) demonstrated simplicity and ease of implementation but showed lower performance compared to other algorithms. Support Vector Machine (SVM) showcased robustness in handling high-dimensional data and achieved competitive accuracy but with more training time and testing time. Light Gradient Boosting (LGBM) exhibited strong performance with its ensemble learning approach, effectively capturing complex relationships within the dataset. Random Forest outperformed other algorithms in terms of predictive accuracy and computational efficiency, making it a promising choice for intrusion detection tasks. Overall, Random Forest emerged as most efficient in terms of all metrics, followed by LGBM and SVM, while KNN lagged behind in terms of performance metrics. These findings provide valuable insights for selecting the most suitable algorithm for intrusion detection systems based on the scores, time and other metrics.

### 6.1 Future Scope

In future work, the scalability and performance of the IDS can be evaluated under different network conditions by testing it on a larger dataset. This could involve the collection of a larger dataset or using an existing publicly available dataset with a greater number of records. Additionally, alternative feature selection techniques such as wrapper methods or embedded methods could be explored and compared with the current correlation analysis-based method to determine the most effective approach for the IDS.

The IDS could be integrated with other security systems and technologies, such as firewalls and intrusion prevention systems, to provide a more comprehensive and robust security solution for modern networks. This could involve the development of a unified security platform that integrates multiple security components and provides a centralized view of the network security status.

## Compliance with ethical standards

*Disclosure of conflict of interest*

No conflict of interest to be disclosed.

## References

[1] Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset SM Kasongo, Y Sun - Journal of Big Data, 2020 – Springer

[2] Performance investigation of principal component analysis for intrusion detection system using different support vector machine kernels. MA Almaiah, O Almomani, A Alsaaidah, S Al-Otaibi… - Electronics, 2022

[3] An analysis of the KDD99 and UNSW-NB15 datasets for the intrusion detection system. MS Al-Daweri, KA Zainol Ariffin, S Abdullah… - Symmetry, 2020

[4] R. Ramakrishnan, C. Reddy, and S. R. Koduru. (2020). "A Review on Machine Learning Techniques for Intrusion Detection System." In 2020 IEEE 10th International Conference on Advanced Computing (ICoAC).

[5] D. Sharma, A. Jain, and A. Gupta. (2021). "Intrusion Detection System: A Review of Machine Learning Techniques and Datasets." In 2021 International Conference on Communication Systems, Computing and IT Applications (CSCITA).

[6] K. R. Prajapati and S. Shah. (2021). "A Comprehensive Review on Machine Learning Approaches for Intrusion Detection Systems." In 2021 11th International Conference on Cloud Computing, Data Science and Engineering (Confluence).

[7] Integrated security information and event management (siem) with intrusion detection system (ids) for live analysis based on machine learning. AR Muhammad, P Sukarno, AA Wardana - Procedia Computer Science, 2023

[8] Comparative analysis of intrusion detection systems and machine learning based model analysis through decision tree. Z Azam, MM Islam, MN Huda - IEEE Access, 2023

[9] Disha, R.A., Waheed, S. Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique. Cybersecurity 5, 1 (2022).

[10] Aboueata N, Alrasbi S, Erbad A, Kassler A, Bhamare D (2019) Supervised machine learning techniques for efficient network intrusion detection. In: 2019 28th international conference on computer communication and networks (ICCCN). IEEE, pp 1–8