IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# A Comparative Study Of Machine Learning Algorithms For Intrusion Detection System

Manasi panda

Department of Computer Science and Engineering,
C.V. Raman Global University, Bhubaneswar, India

**Abstract:** Intrusion Detection Systems (IDS) play a critical role in modern cyber security by monitoring network traffic to detect and prevent malicious activities. This study focuses on applying and comparing various ML algorithms Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF) to classify different types of network attacks such as DoS, Probe, R2L, and U2R. The CICIDS2018 dataset, known for its modern traffic features and attack diversity, was used as the benchmark. The dataset underwent preprocessing and feature selection to optimize model performance. Among the evaluated models, Random Forest outperformed others, achieving 98.76% accuracy, The results validate the effectiveness of ensemble learning methods in intrusion detection, offering a reliable solution for real-time cyber security monitoring.

keywords: Intrusion Detection System, CICIDS2018, Random Forest, Cyber security, Machine Learning, Network Traffic Analysis.

## I. INTRODUCTION

In today's increasingly connected world, securing network infrastructures against cyber threats has become a critical priority. Intrusion Detection Systems (IDS) serve as a vital component in cyber security frameworks, designed to monitor network traffic and identify unauthorized access, malicious activities, or policy violations. Traditional IDS methods, often based on predefined signatures or rule sets, struggle to detect novel or evolving attack patterns, leading to the need for more intelligent and adaptive solutions.

Machine learning (ML) has emerged as a powerful approach to enhance the effectiveness of IDS. By learning patterns from historical data, ML-based IDS can detect both known and unknown threats with greater accuracy and adaptability. Among the various ML algorithms, Random Forest (RF) has gained popularity due to its high classification accuracy, robustness against over fitting, and ability to handle large and complex datasets. This paper aims to review and analyze existing studies that have applied Random Forest algorithms to the CICIDS2018 dataset. It extracts key insights from these studies, identifies common challenges, and suggests future directions to improve the effectiveness of ML-based IDS.

## 2. DATASET OVERVIEW

The CICIDS2018 dataset was developed by the Canadian Institute for Cyber security (CIC) at the University of New Brunswick. It was designed to address the limitations of outdated intrusion detection datasets by providing a realistic and up-to-date benchmark for evaluating the performance of machine learning-based IDS models. The dataset includes around 80 features per network flow, capturing various statistical properties such as duration, packet sizes, flow rates, and protocol-level information. Data was collected over a five-day period and is available in multiple formats including CSV and PCAP, making it suitable for both traffic analysis and machine learning tasks**.** Many dataset are available publicly so some of them were created decades ago and may not be useful in detecting recent threats.UNSW-NB15 provides a middle ground but lacks the full range and realism offered by CICIDS2018.

Key benefits of CICIDS2018 include its labeled data, which facilitates supervised learning, the use of modern protocols (e.g., HTTPS, VoIP), and its ability to reflect current network behavior and threat vectors. As a result, CICIDS2018 is considered one of the most comprehensive and relevant datasets for developing and evaluating contemporary network intrusion detection systems.

table1 -dataset overview

| Dataset | Published By | Year | Features | Attack Types | Realism | Labeled | Modern Protocols | Remarks |
|---|---|---|---|---|---|---|---|---|
| **KDD Cup 99** | UCI/KDD DARPA | 1999 | 41 | 4 main categories (DoS, Probe, R2L, U2R) | Low | Yes | No | Outdated, redundant records |
| **NSL-KDD** | University of New Brunswick | 2009 | 41 | Same as KDD, reduced redundancy | Low–Moderate | Yes | No | Improved KDD, but lacks modern attacks |
| **UNSW-NB15** | University of New South Wales (Cyber Range Lab) | 2015 | 49 | 9 attack types (e.g., Fuzzers, Analysis) | Moderate–High | Yes | Partial (FTP, HTTP) | Better diversity, more balanced |
| **CICIDS2018** | Canadian Institute for Cyber security (CIC) | 2018 | 80 | 15+ attacks (DDoS, Brute Force, Botnet, etc.) | High | Yes | Yes | Most realistic, suitable for ML and deep learning |

## 3. RELATED WORKS

In Vikash Kumar (2019) ,researcher worked with the RTNITP18 dataset and developed an integrated classification-based Intrusion Detection System (IDS). It got 83.8% accuracy, which is okay, but the model was only tested on known types of attacks—it might struggle with new or unknown threats**.** In [2] Ferrag et al. They used the CIC-IDS-2018 dataset and applied deep learning models to detect intrusions. The models performed really well, achieving 97.1% accuracy. But there's a catch—they require a lot of computing power, which makes them hard to use in real-time or on systems with limited resources. In [3] Khan et al.

(2019) , This study used the UNSW-NB15 dataset and experimented with traditional machine learning models like Support Vector Machine (SVM) and Random Forest. They achieved a decent 89.3% F1-score, but their analysis didn't dive deep into feature importance or selection, which limits how well the model can be fine-tuned or understood. In [8] Vinaya kumar et al. (2019) , This team took it up a notch by using deep neural networks across multiple datasets to test the generalize ability of their models. They got results ranging from 91.2% to 97.6% accuracy. However, training these models took a long time, making them less practical for quick deployment or updates. In [7] Kumar et al. (2021), They combined Convolution Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks and tested this hybrid model on the CIC-IDS-2018 dataset. It achieved an impressive 98.7% accuracy, showing strong detection capabilities. But the model's architecture is quite complex, which means implementation and fine-tuning require expert-level knowledge and resources.

table 2 literature survey

| Research | Dataset | Methods | Performance | Limitations |
|---|---|---|---|---|
| Ferrag et al. | CIC-IDS-2018 | Deep Learning models | 97.1% accuracy | \|High computational cost |
| Khan et al. (2019) | UNSW-NB15 | SVM, Random Forest | 89.3% F1-score | Limited feature analysis |
| Vikash kumar (2019) | RTNITP18 | Integrated classification based IDS | 83.8% accuracy | Limited to known attacks |
| Vinayakumar et al. (2019) | Multiple datasets | Deep neural networks | 91.2-97.6% accuracy | High training time |

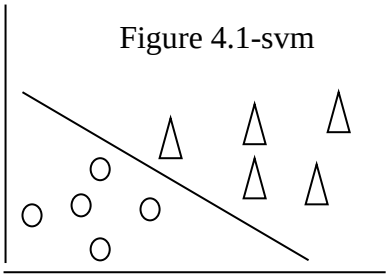
## 4. CLASSIFICATION ALGORITHMS

One of the fundamental techniques in machine learning used for this purpose is classification. Classification algorithms help sort unlabelled data into meaningful categories. In this study, the following algorithms were used:

- A SVM is a supervised learning algorithm that constructs a hyper plane in high-dimensional space to separate different classes. SVM performs well in binary classification and is known for high accuracy and robustness.

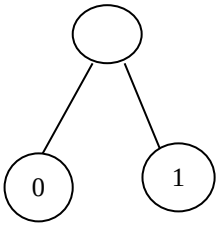
.

Figure 4.1-svm



- A Decision Tree uses a tree-like structure where internal nodes represent features, branches represent decision rules, and leaves represent outcomes. However, it may suffer from over fitting, especially with noisy data or deep trees.

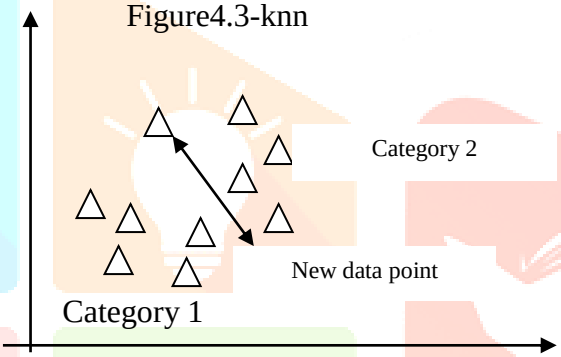
Figure 4.2–decision tree




- Random Forest is an ensemble of multiple decision trees trained on bootstrapped subsets of data with random feature selection. It improves accuracy, reduces over fitting, and is suitable for high-dimensional datasets. It also provides feature importance, making it.

- KNN (K Nearest Neighbour) is a supervised machine learning algorithm used for classification and regression tasks. It is known as a lazy learner because it doesn't learn from the training data immediately but stores the dataset and performs calculations at the time of prediction.

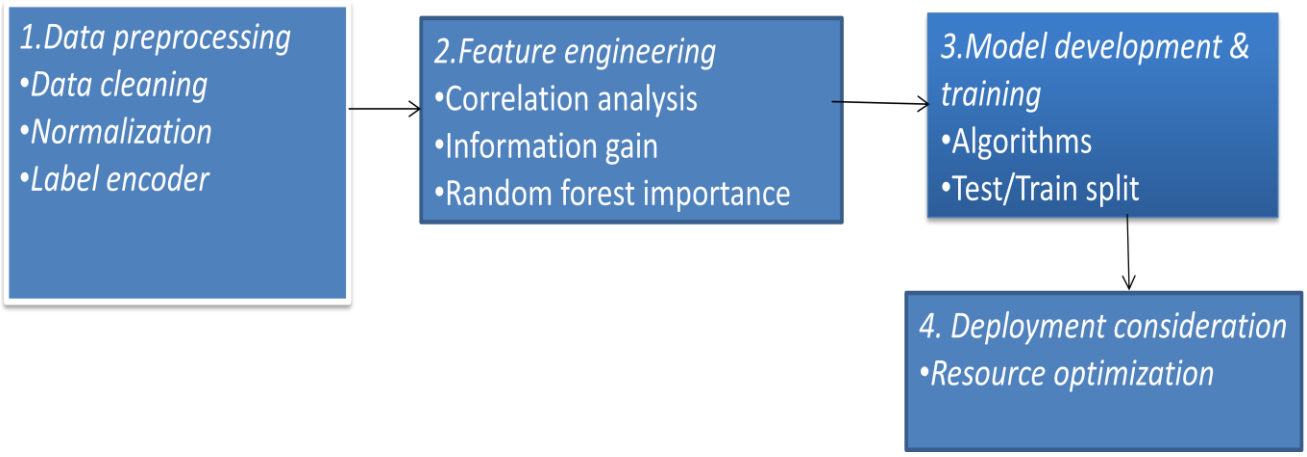
Figure4.3-knn




## 5. METHODOLOGY

A comparative analysis was conducted between various machine learning algorithms—SVM, KNN, and Random Forest—for classifying network traffic to detect intrusions. The evaluation focused on accuracy and detection performance. The CIC-IDS-2018 dataset was used as the primary data source. It contains network traffic data labeled with different types of attacks. Initially, the raw dataset was cleaned and normalized. The class attribute included various attack types.

figure 5.1 - methodology




## 6. EXPERIMENTAL RESULT AND ANALYSIS

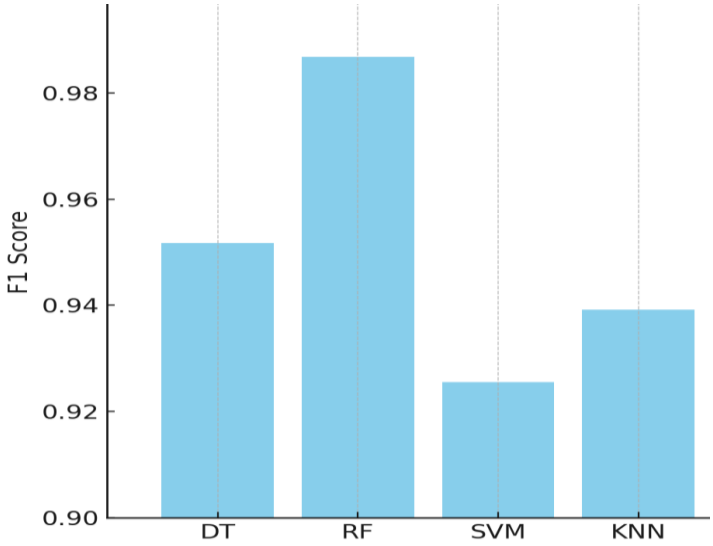
In this section, we are going to mention about proposed techniques and which model provide more accuracy.

- The beginning step include in cleaning the dataset which consist missing value in case of pre-processing clean dataset and handle missing values , replace infinity values with NaN ,Handle missing values - replace with median and Label encoding refers to converting categorical text data into a numerical format
- The feature selection process help to identify the top 20 most important features from the original 80 features. The feature importance ranking based on Random Forest.
- Model training and evaluation function that train the model and make predictions and calculate metrics and generate classification report.
- Various machine learning algorithms are implemented to build predictive models. These models are trained using the processed and feature-optimized dataset. Trained models are rigorously evaluated using performance metrics such as accuracy, precision, recall, F1-score, tuned to enhance the predictive performance of the models.

table 3 comparison between models

| Model | Accuracy | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|
| Decision Tree | 0.9574 | 0.9483 | 0.9574 | 0.9517 | 0.9782 |
| Random Forest | 0.9876 | 0.9863 | 0.9876 | 0.9868 | 0.9983 |
| SVM | 0.9231 | 0.9306 | 0.9231 | 0.9255 | 0.9615 |
| KNN | 0.9419 | 0.9385 | 0.9419 | 0.9392 | 0.9709 |

figure6.1 comparison according to f1 score




## 7. CONCLUSION

This research designed, implemented, and evaluated a machine learning-based Network Intrusion Detection System (NIDS) using the CIC-IDS-2018 dataset. The dataset is treated by the four algorithms SVM, KNN, DT and RF. Random forest model achieved state-of-the-art performance, with an F1-score of 0.9868 and 98.76% accuracy when trained on selected features, highlighting the effectiveness of ensemble methods. A comparative analysis of multiple machine learning models for performance evaluation. Future research is encouraged to explore hybrid models, advanced feature engineering, and real-time deployment to further improve IDS effectiveness in dynamic environments.

### 7.1 Limitations

- The model's performance is measured on a specific dataset and may not perform to all network environments.
- While efficient, the system still requires substantial computational resources for initial training.

## 8. REFERENCES

1. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018) (pp. 108-116).

2. Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. Journal of Information Security and Applications, 50, 102419.

3. Khan, M. A., Karim, M., & Kim, Y. (2019). A scalable and hybrid intrusion detection system based on the convolutional-LSTM network. Symmetry, 11(4), 583.

4. Zhou, Y., &Pezaros, D. (2019). Evaluation of machine learning classifiers for zero-day intrusion detection. In 2019 International Conference on Machine Learning and Cybernetics (ICMLC) (pp. 1-6). IEEE.

5. Liu, H., & Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. Applied Sciences, 9(20), 4396.

6. Kanimozhi, V., & Jacob, T. P. (2019). Artificial intelligence based network intrusion detection with hyper-parameter optimization tuning on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing. In 2019 International Conference on Communication and Signal Processing (ICCSP) (pp. 0033-0036). IEEE.

7. Kumar, P., Gupta, G. P., & Tripathi, R. (2021). An ensemble learning and fog-cloud architecture for DDoS attack detection in IoT environment. IEEE Transactions on Network Science and Engineering.

8. Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. IEEE Access, 7, 41525-41550.

9. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4765-4774).

10. Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

11. Chen, T., &Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

12. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (pp. 3146-3154).

13. Chawla, N. V., Bowyer, K. W., Hall, L. O., &Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

14. Abdulhammed, R., Musafer, H., Alessa, A., Faezipour, M., &Abuzneid, A. (2019). Features dimensionality reduction approaches for machine learning based network intrusion detection. Electronics, 8(3), 322.

15. Ahmim, A., Maglaras, L., Ferrag, M. A., Derdour, M., & Janicke, H. (2019). A novel hierarchical intrusion detection system based on decision tree and rules-based models. In 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS) (pp. 228-233). IEEE.

16. Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. Journal of Big Data, 7(1), 1-29.

17. Hindy, H., Brosset, D., Bayne, E., Seeam, A., Tachtatzis, C., Atkinson, R., &Bellekens, X. (2020). A taxonomy of network threats and the effect of current datasets on intrusion detection systems. IEEE Access, 8, 104650-104675.

18. Chadza, T., Kyriakopoulos, K. G., & Lambotharan, S. (2019). Contemporary sequential network attacks prediction using hidden Markov model. In 2019 17th International Conference on Privacy, Security and Trust (PST) (pp. 1-3). IEEE.

19. Ring, M., Wunderlich, S., Scheuring, D., Landes, D., &Hotho, A. (2019). A survey of network-based intrusion detection data sets. Computers & Security, 86, 147-167.