2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON)
29-30 November, IRAB, BUET, Dhaka, Bangladesh

2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON) | 979-8-3315-3440-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/RAAICON64172.2024.10928505

Enhancing Cybersecurity with an Investigation into Network Intrusion Detection System Using Machine Learning

Md Shohanur Rahman
Department of CSE
Daffodil International University
Dhaka, Bangladesh
hello.msrahman@outlook.com

Wahid Tausif Islam
Department of CSE
Daffodil International University
Dhaka, Bangladesh
tausif3244@gmail.com

Md. Rifat Ahmed Khan
Department of CSE
American International University
Bangladesh
Dhaka, Bangladesh
23-51796-2@student.aiub.edu

***Abstract*—The growing dependence on networked systems underscores the importance of robust security measures to fend off potential intrusions. This research undertakes the task of assessing different machine-learning models concerning network intrusion detection. Following an in-depth examination and comparison, models such as Decision Tree, Logistic Regression, XGBoost, and Random Forest exhibit promising aptitude in precisely identifying potential intrusions. Notably from our proposed model, Random Forest emerged as the top performer among these models, attaining the highest accuracy scores (99%). Its proficiency lies in delivering precise predictions and effectively capturing the intricacies of network intrusion detection. However, the Support Vector Machine (SVM) displayed limitations in its performance throughout this research. These discoveries illuminate the positive and negative aspects of these patterns, providing valuable knowledge about their efficiency in detecting intrusions. The study underscores the necessity for further fine-tuning of successful models and the exploration of advanced techniques to fortify network security. This research, as a whole, signifies an initial and crucial stride in the direction of formulating network intrusion detection systems that are more effective and reliable within the realm of network security.**

Keywords—Network Intrusion Detection System, Intrusion Detection System, Machine Learning, Intrusion Detection Using Machine Learning.

I. Introduction

In today's digitally reliant world, protecting computer networks from malicious intrusions is crucial. As cyber threats grow in complexity, Network Intrusion Detection Systems (NIDS) play a vital role in enhancing information security [1]. Traditional detection methods often fall short in addressing modern threats, highlighting the need for adaptive solutions. This study examines the application of machine learning algorithms in NIDS, aiming to improve detection accuracy and adaptiveness to both known and emerging threats [2]. By evaluating various algorithms, this research contributes empirical insights to cybersecurity, emphasizing machine learning's potential in real-world network defense.

II. Literature Review

Li et al. [3] introduce DAFL, an intrusion detection system using federated learning with dynamic filtering and weighting to improve detection while reducing communication overhead. Dao et al. [4] propose integer linear programming (ILP) and heuristics for efficient task allocation in multi-level IoT NIDS. K. A. Taher et al. [5] utilize machine learning and feature selection on the ASNM dataset to build a more accurate NIDS. Hossen et al. [6] examine ensemble classifiers, including BNBDT and Random Forest, using the NSL-KDD dataset. Mehmood et al. [7] present a method combining SVM and ANFIS for intrusion detection. Rincy et al. [8] propose NID-Shield, a hybrid IDS with the CAPPER feature selection strategy. Acharya et al. [9] address challenges in NIDS such as class imbalance, feature selection, normalization, and overfitting.

III. Methodology

Data is fundamental to machine learning, powering algorithms for predictive analysis. Though we used regression analysis, machine learning typically employs various techniques like classification and regression. The workflow is shown in Fig. 1 below.

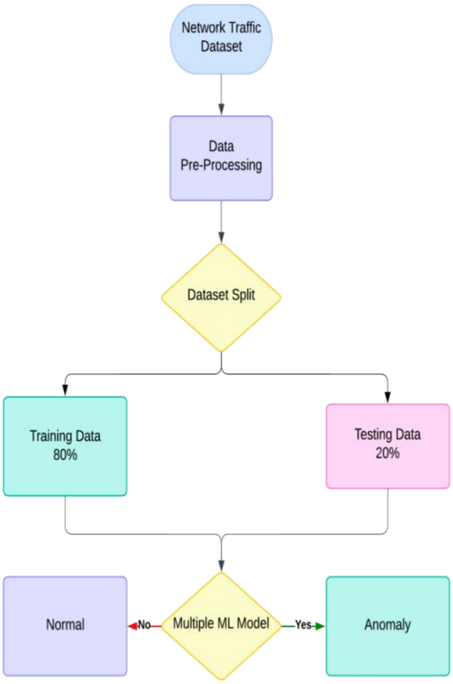

Fig. 1. Overview of the steps involved

A. Dataset

The dataset, sourced from Kaggle [10], includes 25,192 rows and 39 columns of network traffic data, covering both normal and anomalous activities. Its diversity makes it ideal

979-8-3315-3440-0/24/$31.00 ©2024 IEEE

for training an intrusion detection model and is provided in CSV format.

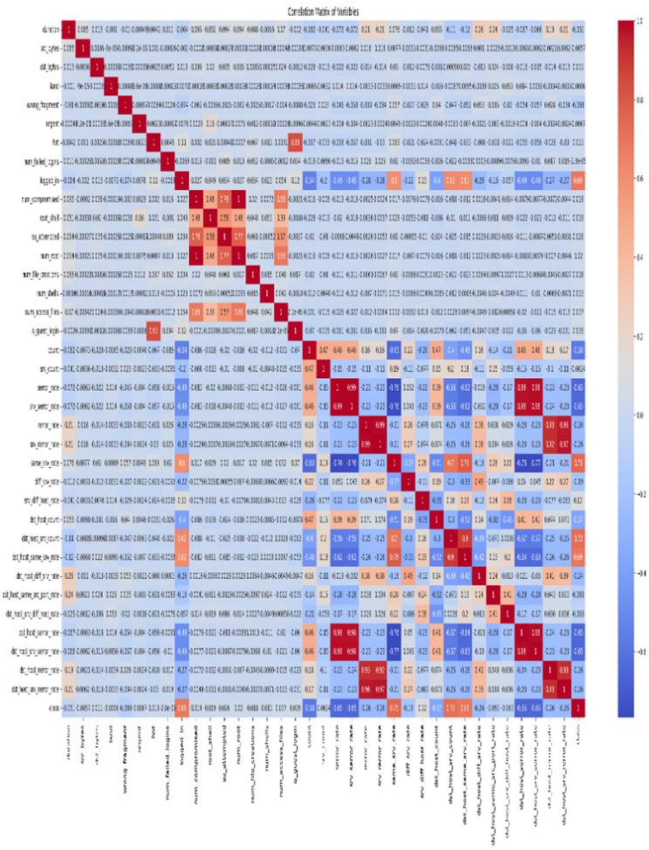

Fig. 2. Correlation Matrix visualization for the dataset

The heatmap in Fig. 2 visually highlights patterns and relationships within the dataset, enhancing comprehension of feature interactions. Such visual aids improve our understanding, helping identify key elements critical for evaluating machine learning models in Network Intrusion Detection Systems (NIDS).

B. Data Pre-processing

Efficient data pre-processing is essential for robust machine learning analysis. This section details key steps to prepare the dataset for intrusion detection. The pre-processing techniques used in this study are as follows:

- Handling Missing Values: Missing values in the dataset were addressed through imputation and elimination, ensuring smooth learning for machine learning algorithms [11].
- Transforming Categorical Variables: Categorical variables are prepared for machine learning using one-hot encoding, which converts categories into binary vectors (1 for presence, 0 for absence) [12].
- Dropping Unnamed Column: The unnamed column, likely irrelevant and automatically generated, was removed to simplify the dataset and reduce noise, ensuring the model focuses on relevant information.

C. Machine Learning Models

Machine Learning models are essential tools for enhancing Network Intrusion Detection Systems (NIDS). Each model offers unique strengths and complexities, enabling more effective cybersecurity analysis and intrusion detection.

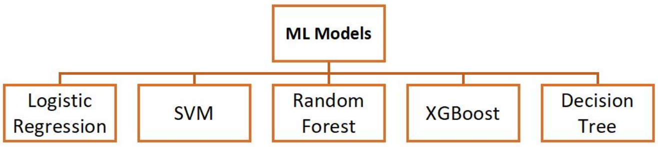

Fig. 3. Tabulation of ML Models

a) Logistic Regression (LR): Logistic Regression, key in binary classification, underpins our intrusion detection framework by effectively modeling probabilities to distinguish normal from anomalous network activities.

b) Decision Tree (DT): Decision Tree algorithms use a hierarchical structure to partition datasets, making complex decision-making transparent and interpretable, essential for NIDS.

c) Random Forest (RF): Random Forest, a robust ensemble algorithm, combines multiple Decision Trees to improve accuracy and reduce overfitting, effectively detecting complex intrusion patterns.

d) Gradient Boosting (XGB): Gradient Boosting, including variants like XGBoost, builds sequential models to enhance accuracy and adaptability, effectively detecting subtle intrusion patterns.

e) Support Vector Machine (SVM): SVM, a key classification model, defines optimal hyperplanes to separate normal and intrusive network behaviors, supporting both linear and nonlinear distinctions.

Each algorithm uniquely strengthens our NIDS. Rigorous testing and hyperparameter tuning helped identify the best-performing solution for our intrusion detection task.

D. Classification Metrics

To assess our machine learning-based NIDS, we conducted an in-depth evaluation using key metrics: Accuracy, Recall, Precision, F1-Score, and the Confusion Matrix to analyze prediction details.

a) Confusion Matrix: The Confusion Matrix detais true positive (TP), true negatives (TN), false positives (FP), and false negatives (FN), providing insights into prediction accuracy and error types. This analysis guides model refinement by highlighting classification errors.

b) Accuracy: Accuracy measures the overall correctness of our models, representing the proportion of true positives and true negatives to all cases.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{1}$$

c) Precision: Precision reflects the accuracy of our models' positive predictions, calculated by dividing true positive predictions by total positive predictions.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

d) Recall (Sensitivity): Recall, or sensitivity, measures the model's ability to capture all actual positive cases, calculated as the proportion of true positives to all actual positives.

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

e) F1-Score: The F1-Score is a reasonable measure of the model's capacity to identify both positive and negative events.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

IV. Result Analysis

A comprehensive accuracy analysis of machine learning algorithms for NIDS revealed that Logistic Regression achieved 87.00%, Decision Trees excelled at 99.38%, Random Forest reached 99.68%, and XGBoost performed well with 99.25%. In contrast, SVM lagged at 53.60%, indicating a need for further optimization. This analysis provides valuable insights into each model's strengths, aiding in selecting the most effective algorithm for enhanced cybersecurity.

TABLE I. CLASSIFICATION METRICS OF ML ALGORITHMS (FOR ANOMALY)

Model	Accuracy	Precision	Recall	F1-Score
LR	0.8700	0.86	0.86	0.86
DT	0.9935	0.99	0.99	0.99
RF	0.9972	1.00	1.00	1.00
XGB	0.9927	0.99	0.99	0.99
SVM	0.5360	0.67	0.00	0.00

XGBoost, Random Forest, Decision Tree, and Logistic Regression show strong accuracy and recall, while Support Vector Machine (SVM) struggles with low precision and recall, resulting in a negligible F1 score.

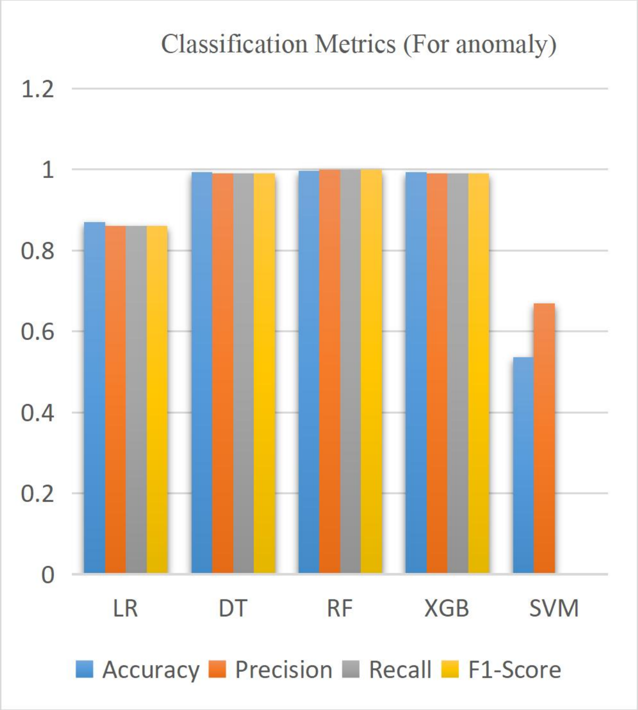

Fig. 4. Comparison of Model Performance (For anomaly)

Fig. 5 compares F1-score, accuracy, precision, and recall across models, showing a strong performance for XGBoost, Logistic Regression, Random Forest, and Decision Tree, while Support Vector Machine (SVM) is less effective.

TABLE II. CLASSIFICATION METRICS OF ML ALGORITHMS (FOR NORMAL)

Model	Accuracy	Precision	Recall	F1-Score
LR	0.8700	0.88	0.88	0.88
DT	0.9935	0.99	0.99	0.99
RF	0.9972	1.00	1.00	1.00
XGB	0.9927	0.99	0.99	0.99
SVM	0.5360	0.54	1.00	0.70

Table II, shows that Decision Tree and Random Forest exceed 99% accuracy, with strong results from XGBoost and Logistic Regression. Support Vector Machine shows trade-offs, with high recall but lower accuracy.

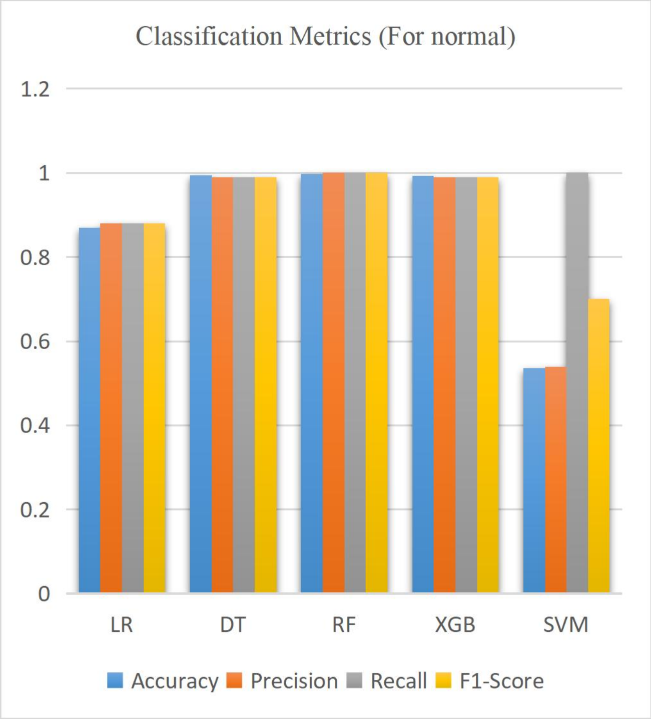

Fig. 5. Comparison of Model Performance (For normal)

Fig. 5 compares accuracy, recall, and F1-score, showing a strong performance for Logistic Regression, Decision Tree, Random Forest, and XGBoost, while SVM lags.

The confusion matrix is key for evaluating classification models, showing true positives, true negatives, false positives, and false negatives. Our analysis found Random

Forest to be the most accurate model, showcasing its effectiveness in classification tasks.

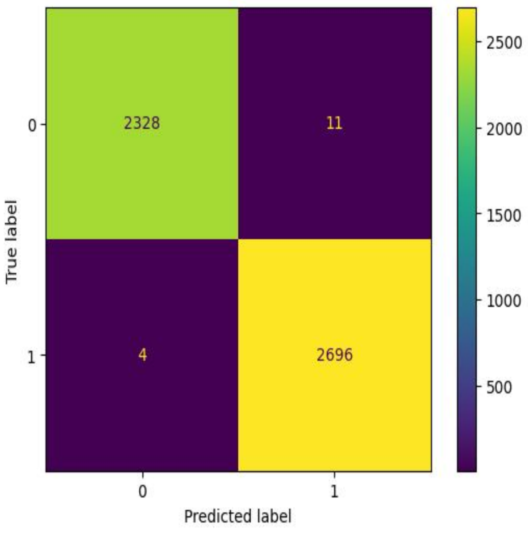

Fig. 6. Confusion Matrix of Decision Tree

Decision Tree, Logistic Regression, XGBoost, and Random Forest excel in detecting intrusions, while SVM lags. Prioritizing these stronger models could improve detection accuracy.

V. Conclusion

The assessment of diverse models in network intrusion detection underscores the proficiency of models like Decision Tree, Logistic Regression, XGBoost, and Random Forest in accurately recognizing potential intrusions. Nonetheless, the observed performance limitations within the Support Vector Machine (SVM) signal the necessity for further investigation and potential enhancements. Subsequent efforts will revolve around refining and optimizing the more successful models to amplify their precision, recall, and overall efficacy in detecting network intrusions.

VI. Future Work

In order to strengthen the development of more robust intrusion detection systems, we intend to investigate cutting-edge methodologies or include ensemble techniques. This study forms the basis for future research endeavors that endeavor to devise more trustworthy and efficacious strategies for strengthening network security against breaches.

References

[1] M. D. Young, “Electronic Surveillance in an Era of Modern Technology and Evolving Threats to National Security,” Stanford Law & Policy Review, vol. 22, p. 11, 2011, Accessed: Jan. 05, 2024. [Online]. Available: https://heinonline.org/HOL/LandingPage?handle=hein.journals/stanlp22&div=5&id=&page=.

[2] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, “A Detailed Investigation and Analysis of Using Machine Learning Techniques for Intrusion Detection,” IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 686–728, 2019, doi: https://doi.org/10.1109/comst.2018.2847722.

[3] “An Efficient Federated Learning System for Network Intrusion Detection | IEEE Journals & Magazine | IEEE Xplore,” ieeexplore.ieee.org. https://ieeexplore.ieee.org/abstract/document/10032055 (accessed Jan. 05, 2024).

[4] T.-N. Dao, D. V. Le, and X. N. Tran, “Optimal network intrusion detection assignment in multi-level IoT systems,” Computer Networks, vol. 232, p. 109846, Aug. 2023, doi: https://doi.org/10.1016/j.comnet.2023.109846.

[5] K. A. Taher, B. Mohammed Yasin Jisan, and Md. M. Rahman, “Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection,” 2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST), Jan. 2019, doi: https://doi.org/10.1109/icrest.2019.8644161.

[6] “Anomaly Based Network Intrusion Detection Using Ensemble Classifiers | IEEE Conference Publication | IEEE Xplore,” ieeexplore.ieee.org. https://ieeexplore.ieee.org/abstract/document/10103301 (accessed Jan. 05, 2024).

[7] M. Mehmood et al., “A Hybrid Approach for Network Intrusion Detection,” Computers, Materials & Continua, vol. 70, no. 1, pp. 91–107, 2022, doi: https://doi.org/10.32604/cmc.2022.019127.

[8] T. Rincy N and R. Gupta, “Design and Development of an Efficient Network Intrusion Detection System Using Machine Learning Techniques,” Wireless Communications and Mobile Computing, vol. 2021, pp. 1–35, Jun. 2021, doi: https://doi.org/10.1155/2021/9974270.

[9] “Efficacy of Heterogeneous Ensemble Assisted Machine Learning Model for Binary and Multi-Class Network Intrusion Detection | IEEE Conference Publication | IEEE Xplore,” ieeexplore.ieee.org. https://ieeexplore.ieee.org/abstract/document/9495864 (accessed Jan. 05, 2024).

[10] “Network Intrusion Detection,” www.kaggle.com. “Network Intrusion Detection,” www.kaggle.com. https://www.kaggle.com/datasets/sampadab17/network-intrusion-detection/.

[11] M. S. Rahman, Akteruzzaman Ashik, and M. S. Rahman, “ Comprehensive Analysis of Stress Levels Using Ensemble Learning and Neural Networks for Insightful Understanding,” vol. 3, pp. 1344 - 1349, May 2024, doi: https://doi.org/10.1109/iceeict62016.2024.10534534.

[12] M. S. Rahman, M. A. Rahman, T. A. Nipa, and M. Asif, “A Machine Learning Approach to Predictive Modeling for Breast Cancer Prediction,” Apr. 2024, doi: https://doi.org/10.1109/icaeee62219.2024.10561811.