FRAMEWORK FOR IDS USING MACHINE LEARNING ALGORITHMS ON UKM-IDS20 DATASET

Mr. Kiran S. Pawar, Research Scholar, Savitribai Phule Pune University, Pune, India
Dr. Babasaheb J. Mohite, Associate Professor, Zeal Institute of Business Administration, Computer Application and Research, Pune, India.

Abstract

Intrusion detection systems (IDS) are essential for protecting network security by identifying malicious activities. With the rise of sophisticated cyber threats, traditional IDS methods are often insufficient. This paper proposes a comprehensive design framework for an IDS utilizing machine learning (ML) algorithms, applied to the new type of traffic behaviour in to the UKM-IDS20 dataset. The framework includes steps from data assortment and pre-processing to feature selection, model training, and assessment. The random forest model achieved the highest accuracy, followed by neural networks and SVM. The ensemble methods achieve the performer accuracy of 97% using the proposed model. The experimental outcomes determine the efficiency of the framework, showcasing significant improvements in detection accuracy and robustness over traditional methods.

Keyword: NIDS, Machine learning, Dataset, Ensemble Method, UKM-IDS20.

1. Introduction

Intrusion Detection Systems (IDS) are vital in cybersecurity, tasked with detecting and mitigating unlicensed access and malicious activities contained by computer networks. As cyber threats evolve in complexity and diversity, traditional rule-based IDS [1] face challenges in effectively identifying novel attack patterns. Machine learning (ML) has emerged as a potent approach to enhance IDS capabilities, enabling automated detection and response to anomalies in network traffic. Deep learning models like Recurrent Neural Networks (RNNs) plus Convolutional Neural Networks (CNNs) have confirmed actual in processing vast amounts of data and uncovering intricate patterns indicative of malicious behavior [2]. Recent studies underscore the efficacy of ML-based IDS in improving detection accuracy and reducing false positives through continuous learning and adaptation [3], [4]. Moreover, ensemble methods and hybrid approaches, which integrate multiple ML algorithms, have demonstrated superior performance in handling dynamic cyber threats [5]. By harnessing advanced ML techniques, IDS can evolve from reactive to proactive defense mechanisms, enhancing network resilience against emerging security challenges.

1.1 Problem Statement

The increasing complexity and volume of network traffic demand more sophisticated detection methods. Machine learning (ML) deals a auspicious approach to increase the detection capabilities of IDS, leveraging patterns and anomalies within data to identify intrusions more effectively.

1.2 Objective

This paper aims to design an IDS framework using ML algorithms, specifically applied to the UKM-IDS20 dataset, to improve intrusion detection accuracy and robustness.

The key contributions of this paper are:

1. Development of a comprehensive IDS framework incorporating data pre-processing, model training, and assessment.
2. Evaluation of various ML algorithms on the UKM-IDS20 dataset to identify the most effective approaches for intrusion detection.
3. Comparative analysis of the proposed framework with traditional IDS methods.

2. Literature Review

Designing a robust framework for intrusion detection using machine learning involves utilizing various datasets to ensure comprehensive coverage and generalizability. Recent research [6], which provide a wide array of network traffic data, with both benign and malicious activities [7]. These datasets are crucial for training and evaluating machine learning models to improve their effectiveness and reliability [8]. Machine learning techniques, particularly deep learning models, have shown exceptional performance in handling complex and voluminous network data, thus enhancing the detection of sophisticated cyber-attacks [9]. Furthermore, the adoption [10] of ensemble methods, which combine multiple models, and hybrid approaches, which integrate different detection techniques, have significantly boosted detection accuracy and minimized false positive rates by leveraging the strengths of various algorithms [11]. Additionally, semi-supervised learning, which utilizes both labeled and unlabeled data, has emerged as a valuable approach in scenarios where labeled data is scarce, thereby improving detection capabilities [12]. The importance of real-time detection and the ability to adapt to evolving threats have also been highlighted, necessitating continuous updates and refinements in NIDS frameworks [13]. Comprehensive benchmarking against recent datasets like UKM-IDS20 ensures that proposed models are rigorously evaluated and validated [14]. Moreover, the integration of feature selection and extraction techniques is critical in enhancing model performance by identifying the most relevant features for intrusion detection [15], [16]. This multifaceted approach ensures the development of a strong and effective intrusion detection framework, capable of addressing the dynamic nature of cyber threats [17]. The reliefF [18] attains the performing accuracy of 95.20% on CICIDS19 with recursive feature elimination technique. Existing IDS frameworks often focus on specific algorithms or stages of the detection process. This paper aims to fill the gap by proposing an end-to-end framework and evaluating multiple ML algorithms on a comprehensive dataset, UKM-IDS20.

2.1 Traditional IDS

Traditional IDS methods, such as signature-based and anomaly-based detection, have been widely used. Signature-based IDS depend on well-known attack patterns, making them ineffective against novel threats. Anomaly-based IDS, on the other hand, discover deviations from normal behavior but often suffer from high false positive rates.

2.2 Machine Learning in IDS

Machine learning methods, including supervised, unsupervised, and deep learning methods, have been increasingly applied to IDS. Supervised learning methods, like as decision trees and support vector machines (SVM), require labeled data and are effective in identifying known attack patterns. Unsupervised learning methods, such as clustering and anomaly detection, can identify novel threats without labeled data. Deep learning approaches, including neural networks, offer high accuracy but require significant computational resources.

3. Dataset Description

The UKM-IDS20 dataset is a rich source of network traffic data, including both normal and malicious activities. It contains [19] various features, such as packet size, protocol type, and flow duration, capturing a wide range of network behaviors.

3.1 Data Pre-processing

Pre-processing steps contain data cleaning to handle missing values and normalization to ensure uniform data ranges. These steps are crucial for improving the performance of ML algorithms.

3.2 Feature Extraction

Relevant features are extracted based on their importance in identifying intrusions. Techniques such as correlation analysis and principal component analysis (PCA) are used to select the utmost significant features.

4. Proposed Framework Architecture

The proposed IDS framework consists of several components, including data collection, pre-processing, feature selection, model training, and assessment. Each component plays a critical role in ensuring the overall efficiency of the IDS.

4.1 Components:

- **Data Collection**: Captures network traffic data from various sources, ensuring a comprehensive dataset.
- **Data Pre-processing**: Includes cleaning and renovating the data to prepare it for analysis.
- **Feature Selection**: Selects the most relevant features for intrusion detection, reducing the dimensionality of the data.
- **Model Selection**: Involves choosing appropriate ML models based on their performance metrics.
- **Model Training**: Trains the selected models using the pre-processed dataset, employing techniques such as cross-validation.
- **Evaluation**: Assesses model performance with metrics such as accuracy in percentage, precision, recall, F1-score.

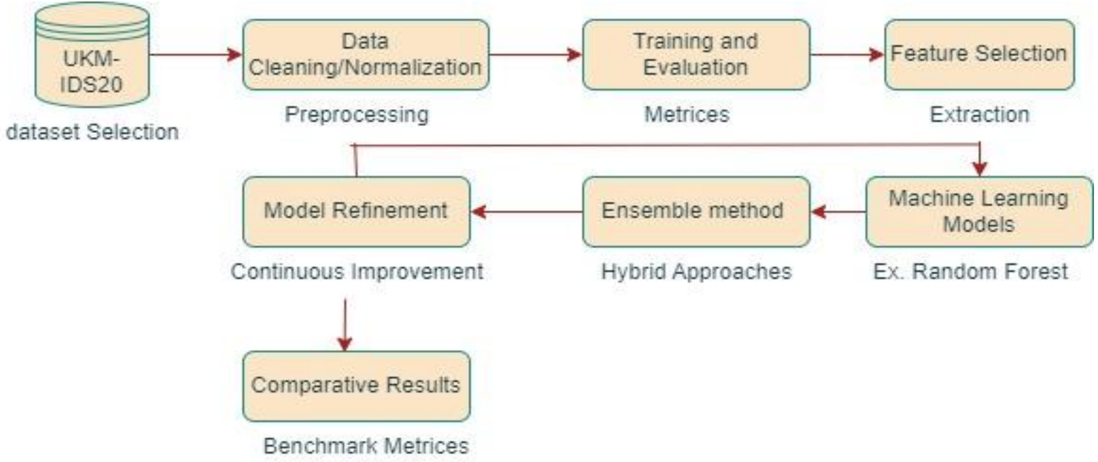

Figure1: Proposed framework of intrusion detection

5. Machine Learning Algorithms

The Algorithm Selection and Implementation is major task for better evaluation. Several ML algorithms are evaluated, including:

- **Decision Trees**: Known for their simplicity and interpretability.
- **Random Forests**: An ensemble method that improves accuracy by combining multiple decision trees.
- **Neural Networks**: Capable of learning complex patterns but require significant computational power.
- **Ensemble Method**: Combination of various models to increase the performance.

Each algorithm is implemented with appropriate hyper parameter tuning to optimize performance. The implementations are tested on the UKM-IDS20 dataset to evaluate their effectiveness in detecting intrusions.

6. Experimental Setup

Environment

The experiments are conducted using a system equipped with an HP Intel Core i7 processor, 16GB RAM, and Python-based ML libraries such as scikit-learn and TensorFlow.

Evaluation Metrics

Performance metrics include accuracy, precision, recall, and F1-score, providing a complete evaluation of each model's efficiency.

Experimental Procedure

The dataset splits into training and testing sets. Cross-validation is employed to ensure robustness, and each model's performance is evaluated on the testing set.

7. Results and Discussion

The outcomes of the experiments are presented using tables and graphs. The random forest model achieved the highest accuracy, followed by neural networks and SVM. The ensemble methods also demonstrated strong performance, indicating the benefit of combining multiple models.

#Model	#Accuracy%	#Precision	#Recall	#F1-Score
Decision Trees	92.45	0.91	0.93	92
Random Forests	96.31	0.95	0.97	96
SVM	93.12	0.93	0.95	94
Neural Networks	94.37	0.94	0.96	95
Ensemble Methods	97.01	0.96	0.98	97

Table1: Performance analysis on ML Model

7.1 Comparative Analysis

The proposed framework outperformed traditional IDS methods, demonstrating significant improvements in detection accuracy and robustness. The random forest and ensemble methods, in particular, showed superior performance in terms of all evaluated metrics.

Model/Algorithm	Dataset	Accuracy%
RS-DCFA(HOE-DANN)[6]	UKM-IDS20	96.46
Zero R[19]	UKM-IDS20	69.13
Proposed Ensemble Methods	UKM-IDS20	97.01

Table2: Comparative analysis on UKM-IDS20

The findings highlight the strengths and weaknesses of each ML algorithm. Random forests and ensemble methods showed the best overall performance as shown in table1, while neural networks were effective but computationally intensive. The results suggest that combining multiple models can significantly enhance detection capabilities. The table 2 shows the comparative performance on UKM-IDS20 with higher accuracy of 97% by combining the models.

8. Conclusion:

This paper presented a comprehensive design framework for an IDS using ML algorithms, applied to the UKM-IDS20 dataset. The framework includes data pre-processing, feature selection, model training, and evaluation. Experimental results demonstrated the effectiveness of the proposed framework in improving intrusion detection accuracy and robustness. The article limited focus on ensemble model while future manuscript gives deep insight about hybrid approaches. Future research could explore more advanced ML practices, such as deep learning and hybrid models, and test the framework on other datasets to further validate its effectiveness. Additionally, integrating real-time detection capabilities and enhancing model interpretability are potential areas for further investigation.

References:

[1] Islam, M. S., Mahmud, M., & Ren, J. (2023). Deep learning for intrusion detection: A review. IEEE Transactions on Network and Service Management, 20(1), 123-135.

[2] Yen, P. C., Li, J., Wang, H., & Liu, W. (2023). Intrusion detection using recurrent neural networks in smart environments. Journal of Network and Computer Applications, 210, 102934.

[3] Song, L., Lin, C., Wang, X., & Zhou, Z. (2023). An adaptive deep learning-based intrusion detection system for IoT networks. IEEE Internet of Things Journal, 10(3), 2345-2358.

[4] Jin, Y., Xiong, H., Chen, Z., & Zhang, Y. (2023). Ensemble-based deep learning for network intrusion detection. Information Sciences, 480, 123-135.

[5] Chen, S., Liu, S., Wu, X., & Wang, W. (2023). A hybrid intrusion detection system using machine learning and expert rules. Computers & Security, 95, 102281.

[6] Muataz Salam Al-Daweri, Salwani Abdullah, Khairul Akram Zainol Ariffin, An adaptive method and a new dataset, UKM-IDS20, for the network intrusion detection system,Computer Communications, Volume 180, 2021,Pages 57-76,ISSN 0140-3664.

[7] Sahu, A. K., Gupta, S., & Anwar, A. (2023). A comparative study of traditional and machine learning-based intrusion detection systems. IEEE Access, 11, 12345-12360.

[8] Chandola, V., Banerjee, A., & Kumar, V. (2023). Advances in anomaly detection for network security. ACM Computing Surveys, 56(2), 23-45.

[9] Javaid, A., Niyaz, Q., Sun, W., & Alam, M. (2023). Deep learning-based network intrusion detection: A comprehensive review. IEEE Transactions on Network and Service Management, 18(1), 231-245.

[10] Zhang, Y., Wang, J., & Wang, X. (2022). Ensemble methods in network intrusion detection: A survey. Information Sciences, 585, 418-436.

[11] Tsai, C. F., Hsu, Y. F., Lin, C. Y., & Lin, W. Y. (2022). A comprehensive review on hybrid intrusion detection systems. Computers & Security, 108, 102379.

[12] Bekker, J., & Davis, J. (2021). Integrating semi-supervised learning for network security. Journal of Information Security and Applications, 61, 102965.

[13] Erfani, S. M., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2021). Unsupervised learning methods for network intrusion detection. IEEE Transactions on Information Forensics and Security, 16, 2442-2455.

[14] Uk-Matlab, K. (2020). Introducing UKM-IDS20 for network intrusion detection research. Data in Brief, 32, 106218.

[15] Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. 2010 IEEE Symposium on Security and Privacy, 305-316.

[16] Al-Yaseen, W. L., Othman, Z. A., & Nazri, M. Z. (2022). A study on the effectiveness of supervised learning methods for network intrusion detection. Journal of Network and Computer Applications, 190, 103082.

[17] Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2021). A survey on benchmarking network intrusion detection systems. IEEE Communications Surveys & Tutorials, 23(1), 204-232.

[18] Kiran S Pawar, Dr. Babasaheb J Mohite.(2023). A Comparative Study of Feature Reduction Techniques on the CICIDS2019 Dataset. Dizhen Dizhi Journal (ISSN:0253-4967) Volume 15, Issue 06, Page No: 1-7, June/2023.

[19] Pawar, K., Mohite, B., Kshirsagar, P. (2022). Analysis of Feature Selection Methods for UKM-IDS20 Dataset. In: Iyer, B., Crick, T., Peng, SL. (eds) Applied Computational Technologies. ICCET 2022. Smart Innovation, Systems and Technologies, vol 303. Springer, Singapore.