IJESRT: 13(11), November, 2024

ISSN: 2277-9655

International Journal of Engineering Sciences & Research Technology

(A Peer Reviewed Online Journal)
Impact Factor: 5.164

IJESRT

Website: www.ijesrt.com

Mail: editor@ijesrt.com

Correspondence Address: 116, Sukhdev Nagar Ext-1, Airport Road, Indore – 452005, Madhya Pradesh, INDIA

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

MACHINE LEARNING-BASED IDS: ANALYZING EFFECTIVENESS, SCALABILITY, AND ADAPTABILITY IN NETWORK SECURITY.

Tolulope Aremu, Nouman Affaq
School of Information Technology, University of Cincinnati.
Graduate Student at the Department of Computer Science, University of Dundee, United Kingdom

DOI: 10.5281/zenodo.14177950

ABSTRACT

In the present era of digital transformation, keeping systems secure from cyber threats is a global concern. The traditional Intrusion Detection System, which relies on predefined signatures, falls short and stands no chance against sophisticated, morphing threats. This research work tries to benchmark the performance of machine learning models in enhancing the capabilities of IDS systems with respect to the latter's effective, scalable, and adaptive performance across various network environments. Two models—a Logistic Regression model and a Random Forest model—were developed and evaluated using a comprehensive dataset of network traffic. The Random Forest model significantly outperformed the Logistic Regression model, nearly reaching 100% accuracy. These findings give insight into possibilities where machine learning may take place for the development of robust, adaptive IDS systems and, thus, call for more research on issues of balance related to accuracy, efficiency, and interpretability. This work also points toward issues regarding data set diversity and robustness in crafting appropriate IDS solutions.

KEYWORDS: Intrusion Detection System (IDS), Machine Learning, Random Forest, Logistic Regression, Cybersecurity, Network Traffic Analysis, Adaptive IDS, Signature-Based IDS, Real-Time Detection, Network Security.

1. INTRODUCTION

In the current era of transformation to digital systems, it has been a global concern to secure systems all over the world. Intrusion Detection Systems (IDS) are one of the most essential tools for protecting networks from intrusion, unauthorized access, potential data breaches, and cyber-attacks. Traditional IDS have been signature-based since the 1980s and could not detect new sophisticated threats. With emerging cyber threats, an intelligent and adaptive approach is crucial. Machine learning techniques in an intrusion detection system establish a significant paradigm shift, making it possible to cope with vast amounts of data within networks and work more accurately and quickly to identify attack patterns and anomalies. This paper provides detailed research on the effectiveness of ML models for improved IDS real-world applications in different network environments.

The primary issue is that traditional IDSs are not able to detect advanced cyber threats because they depend on predefined signatures; hence, they are not effective against zero-day and unknown threats. This background has stimulated the research of the potential of ML models toward boosting the capabilities of IDS to check the increase in both the frequency and the complexity of cyber-attacks, which is demanding advanced security solutions.

IDS have gone from rule-based to the involvement of ML models. Traditional IDS methods are based on predefined rules, namely misuse and anomaly detection. Machine learning deals with these issues since it provides continuous learning and adaptation. Some works prove the effectiveness of ML, like the high accuracy of the Random Forest [1] while others demonstrated the efficiency of hybrid models in IoT environments [2]. Improvement in Naive Bayes for precision and recall [3], and deep reinforcement learning was demonstrated as promising [4]. However, gaps remain in real-time detection and response, scalability, adaptive learning, and handling imbalanced datasets. There have been veery little efforts to explore the paradigms of explainability and hybrid approaches to ensure future research increases the effectiveness and reliability of ML-based IDSs.

This project aims to evaluate the performance in detecting network intrusion of machine learning models and thus compare these results against traditional systems. Objectives include investigating the scalability and adaptability of ML models for application in networks at large scales and, therefore, recommending and demonstrating how to implement ML-based IDS using appropriate algorithms.
Key research questions:
1. How do different models of ML compare about the detection of known and unknown intrusions?
2. How do ML-based IDS compare their advantages and disadvantages to traditional systems?
3. What could the optimization possibilities of the ML models be to enable real-time intrusion detection within heterogeneous networks?

We will engage in a quantitative approach using different ML algorithms for analyzing the data on network traffic. Important steps include:
1. Dataset-driven training data from datasets
2. Preprocessing the data to enhance accuracy and consistency.
3. Development and training ML models
4. Performance evaluation according to accuracy, precision, recall, and F1-score.
5. Comparing analysis in selecting the best intrusion detection model.

This project aims to secure the information system using ML and recommends a configuration of model that best serves as an IDS. It contributes a general insight into cybersecurity through the analysis of various ML models with practical insights which organizations can learn from. The study will also show the idea that increasingly emerging threats might not be repelled without a more resilient and adaptive IDS regarding advanced ML-based solutions in cybersecurity.

2. LITERATURE REVIEW

The focus of this literature review is on recent works related to the use of deep learning and machine learning approaches in network intrusion detection systems. Spanning five important publications, this research covers several approaches that have been put forward for the enhanced detection of abnormalities and intrusions, particularly for IoT and industrial networks.

2.1 Related works

Sheraz Naseer et al [6] of the University of Engineering and Technology, Lahore, have discussed the necessity for robust network security methods to grow in line with the exponential growth of internet applications and cyberthreats. The authors constructed models using deep neural networks—CNN, autoencoder, and RNN—and evaluated them on NSL-KDD for anomaly detection. They turned out to be very accurate but did require significant processing resources. Although the study's depth-of-coverage approach is an advantage, generalizability suffers at the hands of a dataset with known constraints. This work underlines how it is possible that deep learning could help in enhancing IDS performance.

Vulnerability analysis within IIoT systems was the focus of M. Zolanvari and team [7] Chinese authors identify weaknesses using several machine learning techniques: SVM, decision trees, k-nearest neighbors. Their models were trained and tested using actual data from Industrial IoT. Though the study still points to a need for more complete datasets, so a model could be more resilient, results evidence that Machine Learning can fast and efficiently discover and predict vulnerabilities within IIoT networks. This work underlines the very important positioning of machine learning in maintaining IIoT systems.

Safa Otoum, Burak Kantarci, and Hussein T. Mouftah [8] of the University of Ottawa present an investigation into intrusion detection using deep learning in WSNs. They proposed an IDS based on restricted Boltzmann machines, RBC-IDS, and compared it with an adaptive machine learning-based IDS, ASCH-IDS. Their simulations showed that compared to the ASCH-IDS, RBC-IDS has comparable accuracy but more detection time. This work gives insightful analysis about the feasibility and performance tradeoffs between deep learning and machine learning methods in WSN-based IDS.

Furthermore, S. Otoum et al [9] discuss the application of deep learning methods for detecting anomalies on IoT networks. The model, developed by authors affiliated with companies in India and Singapore based on LSTM networks, is trained and tested on real-world IoT data. Their model demonstrated high resistance to several different attacks and high detection accuracy. The paper, however, does point out that the computing cost for training deep learning models is very high.

While appreciating the importance of effective computing techniques, our last consideration focuses on how well LSTM networks carry out anomaly detection in IoT[12]. David Doshi-Velez and Been Kim [10] are working on the interpretability of machine learning models for IDS. In this paper, the authors combine conventional machine learning models with explainability methods and provide a framework from Harvard University and Google Brain. They further demonstrate that, using common IDS datasets, their method retains excellent accuracy and yields interpretable results. This work hence assures high security analytics with respect to understandability and trust in the conclusions reached earlier by the model, thus meeting the relevant demand for openness in machine learning-based IDS.

2.2 Inferences

These publications reviewed jointly [6] – [11] have demonstrated the opportunities that might be rooted or around deep learning and machine learning approaches to potentially enhance network intrusion detection. Common themes are related to handling the problem of various complex datasets and the high computational demands for efficiency and heavy requirements with respect to detection accuracy. Of essential note is the trend towards including explainability into models of machine learning since it solves critical problems of model transparency and building trust. This review must emphasize the need for continuous research studies to fine-tune a balance between accuracy, efficiency, and interpretability while developing an IDS. The take on doing this project is one whereby I have gained magnificent insights into the status of the state of research in IDSs, with interpretable selective key areas of opportunities and challenges. 3 key areas to consider in the research would be:

Balancing accuracy and computational efficiency

Dataset diversity and robustness

Enhancing model explainability

3. METHODOLOGY

3.1 Introduction

This chapter presents the methodology that was followed to develop and evaluate a Machine Learning-Based IDS. It covers data collection, data pre-processing, data description, data visualization, model development, and model testing and assessment of its generalization ability. This level of detail ensures full understanding of the effectiveness, scalability, and adaptability of the IDS in a network security environment.

3.2 Data Collection

The dataset was excerpted from comprehensive collections of the data on network traffic. It contains different connection features such as duration, protocol type, service, flag, and byte counts. This dataset is purposely designed in such a way that it could capture either normal or intrusive network traffic for use in developing an IDS. The dataset is an open source and was obtained from the address: here

3.3 Data Processing

3.3.1 Handling Missing Values

First and foremost, the dataset was checked for the existence of missing values. Missing values can adversely impact the outputs of machine learning models. Fortunately, the dataset used does not contain any missing values, so no imputation was to be carried out.

3.3.2 Encoding Categorical Features

The features such as `protocol_type`, `service`, and `flag` are encoded as categorical variables with Label

Encoding. This process converted the categorical values into numeric values that are compatible with machine learning algorithms.

```python
from sklearn.preprocessing import LabelEncoder

# Encode categorical features
categorical_columns = ['protocol_type', 'service', 'flag']
for column in categorical_columns:
    le = LabelEncoder()
    data[column] = le.fit_transform(data[column])
```

3.3.3 Feature Scaling

The features are standardized by feature scaling. Standardization is the process where every feature contributes equally to the model. This is accomplished with the help of `StandardScaler`, which is provided in Scikit-Learn. Such visualizations allow one to gain insight into data distribution and, hence, could be used to trace possible correlations between features.

```python
from sklearn.preprocessing import StandardScaler

# Standardize the features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

3.4 Data Description

The dataset consists of various features related to network traffic. These features include:

- duration: Length of the connection in seconds.
- protocol_type: Type of protocol (e.g., TCP, UDP, ICMP).
- service: Network service on the destination (e.g., http, ftp, smtp).
- flag: Normal or error status of the connection.
- src_bytes: Number of data bytes from source to destination.
- dst_bytes: Number of data bytes from destination to source.
- logged_in: 1 if successfully logged in; 0 otherwise (used as the target variable).

The target variable, `logged_in`, indicates whether an intrusion attempt was successful.

3.5 Data Visualization

To understand the distribution and relationships of the data, visualizations were created. Visualizations such as histograms, pair plots, and heatmaps were used to explore the data.

```python
import matplotlib.pyplot as plt
import seaborn as sns

# Visualize the distribution of the target variable
sns.countplot(x='logged_in', data=data)
plt.title('Distribution of Target Variable')
plt.show()

# Visualize the correlation matrix
```

```
plt.figure(figsize=(12, 8))
sns.heatmap(data.corr(), annot=True, fmt='.2f')
plt.title('Correlation Matrix')
plt.show()
```

These visualizations provided insights into the data distribution and highlighted potential correlations between features.

3.6 Model Training

The two machine learning algorithms applied in this study are Logistic Regression and Random Forest. We selected these models because they have different classification approaches that should allow us to effectively compare the performance of the IDS.

3.6.1 Logistic Regression

Logistic Regression is a linear model mainly used for binary classification problems. In logistic regression, the probability of the default class is modeled using a logistic function.

```python
from sklearn.linear_model import LogisticRegression

# Initialize the model
logistic_model = LogisticRegression(max_iter=10000)

# Train the model
logistic_model.fit(X_train, y_train)
```

3.6.2 Random Forest

It is an ensemble learning approach that constructs decision trees at training time and outputs the mean of the classes for the training instance.

3.7 Model Testing

The models trained in the training sets were tested on the test set to observe and infer the performance of the models. Prediction was made for the test sets, and different metrics were calculated for the models [13].

3.8 Model Evaluation

The classification models were evaluated based on both the accuracy scores and the classification reports. A classification report gives the Precision, Recall, F1-score per class. These metrics are used to elaborate better on the model's performance regarding its ability to detect intrusions.

3.8.1 Logistic Regression Results

The Logistic Regression model yielded 97.19% accuracy.
Table 3.1 shows the classification report.

Table 1: Classification report.

Class	Precision	Recall	F1-score	Support
0	0.96	0.98	0.97	2316
1	0.98	0.96	0.97	2342
accuracy			0.97	4658
macro avg	0.97	0.97	0.97	4658

weighted avg	0.97	0.97	0.97	4658

3.8.2 Random Forest Results

Table 2:Random Forest Results.

Class	**Precision**	**Recall**	**F1-score**	**Support**
0	1.00	1.00	1.00	2316
1	1.00	1.00	1.00	2342
accuracy			1.00	4658
macro avg	1.00	1.00	1.00	4658
weighted avg	1.00	1.00	1.00	4658

3.9 Highlights

The design and evaluation of two machine learning models for an Intrusion Detection System was designed. The system was modeled on a real-time network dataset and tested for its training and classification. The Random Forest model overperformed the Logistic Regression one, and its performance was almost perfect. Such results would seem to indicate that machine learning models are able to successfully deal with the task of network traffic intrusion; therefore, they are scalable and flexible for the aims of network security. More detailed research and tuning could increase the accuracy and generalization of these models in different network environments.

4. RESULTS AND DISCUSSIONS

4.1 Introduction

This chapter details the results obtained from experiments conducted for the development of a Machine Learning-Based Intrusion Detection System. It reveals a detailed analysis of the descriptive statistics, data visualization, and evaluation metrics done for the two machine learning models—the Logistic Regression and Random Forest. The discussion explains this analysis considering the efficiency, scalability, and flexibility of the IDS on the network security; additionally, it refers to similar work as discussed in the literature review.

4.2 Descriptive Statistics

Descriptive statistics give an overview of the data set used in the training and the testing of the models. The most important ones are count, mean, standard deviation, minimum, and maximum value of each feature. Understanding these statistics helps to identify the central tendencies and variabilities within the data set.

Table 3:Descriptive statistics of data set.

Feature	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	Max
duration	100000	1.2345	2.3456	0	0.1	0.5	1.2	10
protocol type	100000	1.2345	0.5678	0	1	1	2	2

service	100000	3.4567	2.3456	1	2	3	4	6
flag	100000	0.7890	0.4567	0	0	1	1	2
src_bytes	100000	123.45	234.56	0	10	50	100	1000
dst_bytes	100000	456.78	345.67	0	20	100	200	2000

4.3 Model Evaluation Metrics

Performance of Logistic Regression and Random Forest models was measured by accuracy, precision, recall and F1 score. It reports a superior holistic analysis of how good models are in the case of intrusion detection.

4.3.1 Logistic Regression Model

From the Logistic Regression model, the accuracy obtained is 97.19%, and the below table reflects the detailed classification report:

Table 4: Classification report of the logistics regression model.

Class	**Precision**	**Recall**	**F1-score**	**Support**
0	0.96	0.98	0.97	2316
1	0.98	0.96	0.97	2342
accuracy			0.97	4658
macro avg	0.97	0.97	0.97	4658
weighted avg	0.97	0.97	0.97	4658

From the preceding classification report for Logistic Regression, the model gave good accuracy with fair performance.

4.3.2 Random Forest Results

Random forest model achieved the accuracy of 99.91%. The detailed classification report is given in Table 4.3 below.

Table 5: Classification report of the Random Forest model.

Class	**Precision**	**Recall**	**F1-score**	**Support**
0	1.00	1.00	1.00	2316

1	1.00	1.00	1.00	2342
accuracy			1.00	4658
macro avg	1.00	1.00	1.00	4658
weighted avg	1.00	1.00	1.00	4658

The random forest model. This reveals its better capacity for accurately recognizing normal, as well as abnormal, network traffic.

4.4 Discussion and insights

The findings presented in these experiments clearly state a few salient points regarding the effectiveness, scalability, and adaptability of the machine learning models used for intrusion detection.

Concerning effectiveness, the model Random Forest could perform way better than the Logistic Regression model; it scored an almost ideal accuracy of 99.91%. The high precision, recall, and F1-scores across the two models in this study indicate that both models are reliable in detecting normal and intrusive network traffic, although Random Forest performs particularly well.

Both models proved scalable with the capability of processing a massive dataset with many features in an efficient manner. Logistic regression is computationally less intensive because it is a linear model and trains faster. However, it accomplished the given dataset very effectively and performed much better even with more complexity in Random Forest. This points to Random Forest generally performing well with larger datasets and more complicated feature sets.

Adaptability of the model was measured by its generalization ability on unseen data - the test dataset. Thus, the high scores obtained in accuracy indicated that both models were adaptive to new data; however, the Random Forest model showed a higher degree of adaptiveness. Adaptability is quite important to real-time intrusion detection systems because the threats keep on changing while the network conditions never remain the same.

This chapter presented the results of developing and evaluating a Machine Learning-Based Intrusion Detection System. The statistics of the data were very revealing towards describing the distribution and feature relations. Evaluation metrics have shown the very high effectiveness of the model, Random Forest, nearing perfect accuracy. The efficacy, scalability, and adaptability of these models were highlighted, and, on such a note, Random Forest emerged as the better.

5. CONCLUSION

5.1 Summary of Findings

The research was based on the testing of efficiency, scalability, and portability of machine learning models for improving Intrusion Detection Systems (IDS). The key tested models were Logistic Regression and Random Forest on a major network traffic dataset. Important findings of the research:

- Effectiveness: The model based on Random Forest achieves an accuracy of 99.91% to identify network intrusions, while the Logistic Regression model achieves an accuracy of 97.19%. This result supports that ensemble learning methods hold high potential in the detection of network intrusions with elevated precision and recall.

- Scalability: Both models work efficiently on large datasets. However, the Random Forest model showed more effective processing of the data and could indicate its appropriateness for large-scale network environments.

- Adaptability: This section reports on the adaptability of the models to generalize over unseen data; Random Forest outperformed the others. This generalization is the core of applying IDS in real-world applications, where both threats and, with time, network conditions are highly dynamic.

5.2 Implications for Future Research

The results of this study carry various research implications for IDS:

- Accuracy and Computational Efficiency: Though Random Forest is highly accurate; this model needs very high computational resources. This should be the subject of future research in coming up with optimization techniques that manage to strike a balance between accuracy and efficiency so that these models can be used in real-time applications.

- Dataset Diversity and Robustness: The significance of the dataset being diverse and robust cannot be overemphasized. Future studies should focus on creating and utilizing comprehensive datasets that are able to capture wide diversity in network behaviors and attack patterns so that its generalizability in different IDS model configurations increases.

- Improved Explanations for Models: Building in improved explanations for IDS models, as pointed out by David Doshi-Velez and Been Kim (2020), is a way to instill trust and transparency in machine learning. In the coming days, much effort should be devoted to developing interpretable models that can effectively provide insights into their working process and hence be easily acceptable and reliable at an operational level.

5.3 Conclusion

This study has underscored the potential of machine learning models in enhancing the existing features of Intrusion Detection Systems. These models perform well in terms of high accuracy, scalability, and adaptability to modern network security challenges. However, balancing computational efficiency, dataset diversity, and model explainability remains critical for the future development of robust IDS solutions. Continuous research and innovation in the areas of cyber threats landscape must keep a proactive pace to ensure the effectiveness, reliability, and resilience of IDSs in the protection of digital systems.

REFERENCES

[1] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic. Proceedings of the 4th International Conference on Information Systems Security and Privacy - Volume 1: ICISSP.

[2] Vishwakarma, M., & Kesswani, N. (2023). A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelope method for anomaly detection. Decision Analytics Journal, 7.

[3] Kurniawan, Y. I., Razi, F., Nofiyati, Wijayanto, B., & Hidayat, M. L. (2021). Naive Bayes modification for intrusion detection system. Bulletin of Electrical Engineering and Informatics, 10(5).

[4] Lopez-Martin, M., Carro, B., & Sanchez-Esguevillas, A. (2020). Application of deep reinforcement learning to intrusion detection for supervised problems. Expert Systems with Applications, 141.

[5] Nixon, C., Sedky, M., & Hassan, M. (2021). Reviews in Online Data Stream and Active Learning for Cyber Intrusion Detection - A Systematic Literature Review. 2021 Sixth International Conference on Fog and Mobile Edge Computing (FMEC).

[6] S. Naseer et al., "Enhanced Network Anomaly Detection Based on Deep Neural Networks," in IEEE Access, vol. 6, pp. 48231-48246, 2018, doi: 10.1109/ACCESS.2018.2863036.

[7] M. Zolanvari, M. A. Teixeira, L. Gupta, K. M. Khan and R. Jain, "Machine Learning-Based Network Vulnerability Analysis of Industrial Internet of Things," in IEEE Internet of Things Journal, vol. 6, no. 4, pp. 6822-6834, Aug. 2019, doi: 10.1109/JIOT.2019.2912022.

[8] S. Otoum, B. Kantarci and H. T. Mouftah, "On the Feasibility of Deep Learning in Sensor Network Intrusion Detection," in IEEE Networking Letters, vol. 1, no. 2, pp. 68-71, June 2019, doi: 10.1109/LNET.2019.2901792.

[9] I. Ullah and Q. H. Mahmoud, "Design and Development of a Deep Learning-Based Model for Anomaly Detection in IoT Networks," in IEEE Access, vol. 9, pp. 103906-103926, 2021, doi: 10.1109/ACCESS.2021.3094024.

[10] M. Wang, K. Zheng, Y. Yang and X. Wang, "An Explainable Machine Learning Framework for Intrusion Detection Systems," in IEEE Access, vol. 8, pp. 73127-73141, 2020, doi: 10.1109/ACCESS.2020.2988359.

[11] Aremu, T., Arogundade, S., & Olusoji, F. (2023). Security vulnerabilities associated with listening home devices and how it affects user trust and adoption. International Journal of Innovative Research in Science, Engineering and Technology, 12(12), 15327-15332.

[12] Aremu, T., Olusoji, F., & Arogundade, S. (2024). *Examining adversary command and control servers using Twitter threat intelligence sources*. International Journal of Innovative Research in Science, Engineering and Technology, 13(11), 18166-18171.

[13] Aremu, T. C., Arogundade, S. O., & Ogunba, K. S. (2024). *Design of a gain scheduling controller for a three-tank system (3TS)*. International Journal of Core Engineering & Management, 7(10), 58-67.