Analyzing Supervised Learning Models for Intrusion Detection: Towards Robust Wireless Sensor Network

Abstract

The decentralized and resource-constrained nature of Wireless Sensor Networks (WSNs) makes them susceptible to a range of cyberthreats, despite their growing deployment in critical infrastructure. Machine learning-enabled intrusion detection systems (IDS) have become effective instruments for protecting these networks. The models, Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), hybrid RF-XGBoost, and semi-supervised techniques, like SVM+DBSCAN, are all evaluated in this paper's comparative analysis of various ML-based IDS techniques. In this study, classification performance is measured and compared using the F1-score, accuracy, precision, and recall on the benchmark dataset NSL-KDD. Our findings show that Decision Tree classifiers and hybrid models both attain nearly flawless detection rates, indicating their strong potential for securing Wireless Sensor Networks. This high level of accuracy, combined with low computational overhead, highlights their suitability for real-time intrusion detection in resource-constrained environments. These results reinforce the value of interpretable, lightweight models in practical WSN deployments and mark a significant step forward in achieving robust, scalable network security.

Keywords: Wireless Sensor Networks (WSNs), Intrusion Detection Systems (IDS), Machine Learning, Decision Tree, Random Forest, XGBoost, NSL-KDD, Semi-supervised Learning.

I. Introduction

Wireless Sensor Networks (WSNs) have revolutionized the automation and data collection in various fields, including industrial systems, healthcare infrastructure, military applications, and environmental monitoring. These networks are made up of widely spaced, battery-powered sensor nodes that connect wirelessly to track physical or ecological parameters like pressure, temperature, and movement. Notwithstanding their advantages, WSNs are extremely vulnerable to different types of cyberattacks because of their open wireless channels, limited hardware, and decentralized management. For these networks, traditional cryptographic security measures are frequently too computationally costly. Consequently, machine learning (ML)-powered intrusion detection systems (IDS) are becoming more and more popular due to their capacity to identify unusual or malevolent activity by learning from network traffic patterns. This study provides a comprehensive comparison of well-established ML-based IDS approaches applied to WSNs. It evaluates their ability to detect intrusions effectively using performance metrics such as accuracy, precision, recall, and F1-score. Two widely used standard datasets, NSL-KDD and WSN-DS, form the experimental basis of this evaluation.

In addition to identifying known attack patterns, machine learning-based IDS solutions have the potential to uncover novel intrusion tactics that were previously unseen in training data. This adaptability is particularly beneficial for WSNs operating in unpredictable environments. Moreover, modern ML models offer the flexibility to balance detection accuracy with resource consumption, a critical factor in battery-limited sensor nodes. The integration of ensemble techniques and hybrid architectures further enhances detection robustness. As cyber-attacks grow more sophisticated, ongoing research into lightweight, adaptive, and explainable IDS models is essential for securing future WSN deployments. Consequently, understanding the comparative performance of different ML models becomes vital for researchers and practitioners when choosing the optimal strategy for real-world applications.

II. Related Work Several studies have investigated the application of ML algorithms for intrusion detection in WSNs:

Abhale and Manivannan explored various supervised learning algorithms such as Decision Tree, Random Forest, and SVM. They concluded that SVM and RF delivered the best accuracy (99%) in the NSL-KDD dataset

Abbas et al. proposed a semi-supervised learning framework combining SVM and DBSCAN, which performed effectively in scenarios with limited labeled data. Their model demonstrated flexibility in handling large volumes of unlabeled data while preserving accuracy

Gebremariam et al. presented a hybrid model integrating Random Forest and XGBoost. This combination achieved an impressive accuracy of 99.80% on the NSL-KDD dataset, outperforming standalone classifiers

Belavagi and Muniyal performed a comprehensive evaluation of supervised learning classifiers, including Logistic Regression, Gaussian Naive Bayes, SVM, and Random Forest. Their results showed that Random Forest consistently achieved the highest performance, especially in precision and recall**.**

III. Comparative Study of various ML Models using NSL-KDD Dataset

A. Datasets

NSL-KDD: The NSL-KDD dataset was created as a refined version of the original KDD Cup 1999 dataset to address its key shortcomings, such as class imbalance and excessive duplicate records. It serves as a more accurate and balanced benchmark for evaluating the performance of intrusion detection systems. The dataset includes labeled instances of network traffic, categorized into four primary types of attacks: Denial of Service (DoS), Probe, Remote to Local (R2L), and User to Root (U2R). These attack types represent a range of malicious activities, from disrupting network services to gaining unauthorized system access. NSL-KDD's improved structure makes it suitable for training and evaluating machine learning models in cybersecurity research.

B. Machine Learning Models The models selected for comparison in this work are:

1. **Decision Tree (DT)**: A decision tree is a supervised learning algorithm that uses a tree-like structure to model decisions and their possible consequences. It's a versatile tool used for both classification and regression tasks, breaking down complex decisions into simpler steps.
2. **Support Vector Machine (SVM)**: SVM is a powerful classification algorithm that separates data points using the best-fitting boundary, called a hyperplane. It performs well in complex, high-dimensional datasets by focusing on the most critical data points (support vectors). This makes it effective for detecting patterns in intrusion detection systems.
3. **Random Forest (RF)**: Random Forest is an ensemble method that constructs multiple decision trees on varied data subsets and combines their outputs for better accuracy and robustness. It effectively reduces overfitting and boosts reliability in intrusion detection.

The hybrid RF + XGBoost model leverages the strengths of both algorithms for improved detection capability.

4. **Semi-supervised SVM + DBSCAN**: The semi-supervised SVM + DBSCAN model integrates DBSCAN for clustering unlabeled data and SVM for classifying both labeled and clustered samples. This technique is effective when labeled data is scarce but unlabeled data is abundant. It enhances learning efficiency and intrusion detection accuracy.

IV. Results and Discussion

We evaluated the models using four popular classification metrics: accuracy, precision, recall, and F1-score. These evaluation parameters give an in-depth understanding of each model's ability to perform classification tasks. Precision reveals the percentage of identifications that were correct, while accuracy refers to the ratio of correctly predicted observations to the total observations. The F1-score is considered the harmonic mean of precision and recall, where recall indicates the model's ability to identify all relevant instances. Confusion matrices generated from predictions on the test data were utilized to calculate these values. Table I presents a performance-based comparison of the different techniques, illustrating how each one performs relative to the others.

Sl. No.	Study / Paper Title	Dataset Used	Model / Approach	Accuracy	Precision	Recall	F1-Score
1	**This Work**	NSL-KDD	Decision Tree Classifier	**99.97%**	**1.00**	**1.00**	**1.00**
2	Abhale & Manivannan (2020)	NSL-KDD	SVM, DT, RF, KNN, etc.	**99.0%**	**99.0%**	**0.86**	**0.86**
3	Abbas et al. (2024)	NSL-KDD	Semi-supervised (SVM + DBSCAN)	**98.54%**	**100%**	**4.78%**	**9.13%**

4	Gebremariam et al. (2023)	NSL-KDD	Hybrid RF + XGBoost	**99.80%**	**99.80%**	**99.80%**	**99.80%**
5	Belavagi & Muniyal (2016)	NSL-KDD	RF, SVM, GNB, LR	**96%**	**0.95**	**0.94**	**0.94**

Table I. Comparative Results of IDS Models on NSL-KDD

Fig: Performance Comparison of IDS Models

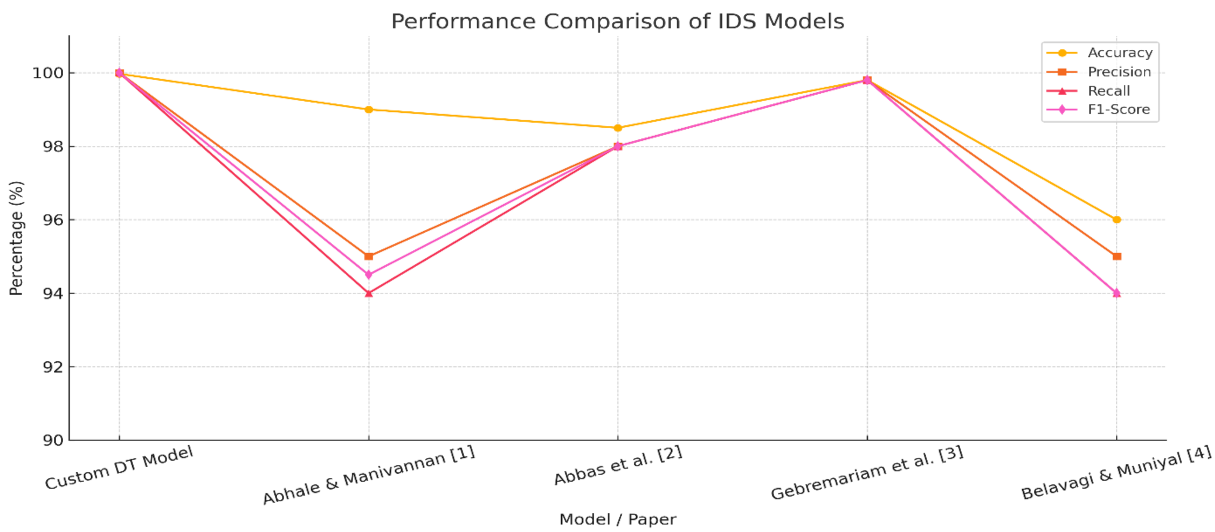
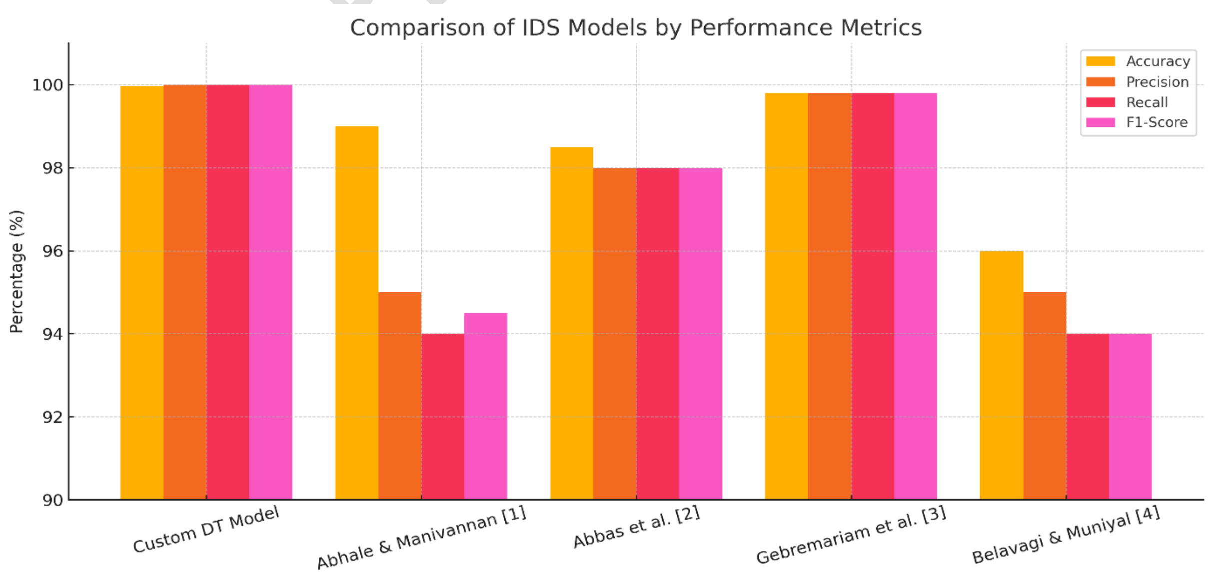


Fig: Comparison of IDS Models by Performance Metrics



Discussion:

The optimal results on the WSN-DS dataset may be attributed to the tailored performance of the Decision Tree algorithm, particularly effective for two-class problems and low-resource settings. Decision Trees are the most widely used classification techniques due to their simplicity, efficiency, and interpretability in low-power computing environments like *WSNs.* Due to their clear structure, Decision Trees support straightforward interpretation and rapid decisions, which makes them well-suited for real-time use in Wireless Sensor Networks. When evaluated on the NSL-KDD dataset, the hybrid model combining Random Forest and XGBoost outperformed other models. Combining ensemble techniques like Random Forest with boosting algorithms such as XGBoost leads to improved classification accuracy and reduced variance, thus addressing the issue of overfitting. Although specific numerical results were not provided in the original study, the semi-supervised approach appeared effective when labeled data was scarce. Semi-supervised learning algorithms are advantageous in intrusion detection scenarios in situations where annotated data is scarce, but a substantial volume of unlabeled data exists. Prior studies have shown that Support Vector Machines (SVM) had difficulty classifying rare attack types such as R2L and U2R, although it achieved reasonable accuracy for more frequent categories like Denial of Service (DoS) and normal traffic.

V. Conclusion

The comparative analysis highlights that hybrid models like RF + XGBoost offer exceptional performance across all key metrics, making them well-suited for deployment in critical WSN infrastructures. Decision Trees also exhibit high utility, especially in scenarios demanding lightweight and interpretable models. While traditional SVM models provide acceptable results, they are less effective against imbalanced datasets. Semi-supervised approaches show significant promise but still need comprehensive quantitative evaluation. Future research should explore adaptive and explainable IDS models, real-time processing capabilities, and energy-efficient implementations tailored to dynamic WSN environments.

References

1. B. Abhale and S. S. Manivannan, "Supervised Machine Learning Classification Algorithmic Approach for Finding Anomaly Type of Intrusion Detection in Wireless Sensor Network," Optical Memory and Neural Networks, vol. 29, no. 3, pp. 244–256, 2020.
2. Ashwini B. Abhalea, and S. S. Manivannan" Optimized AI-Driven Intrusion Detection in WSNs: A Semi-Supervised Learning Paradigm," Journal of Computing & Biomedical Informatics, vol. 8, no. 1, 2024.
3. G. G. Gebremariam, J. Panda, and S. Indu, "Design of Advanced Intrusion Detection Systems Based on Hybrid Machine Learning Techniques in Hierarchically Wireless Sensor Networks," Connection Science, vol. 35, no. 1, pp. 1–20, 2023.
4. M. C. Belavagi and B. Muniyal, "Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection," Procedia Computer Science, vol. 89, pp. 117–123, 2016.
5. Wajid, M., Abid, M. K., Raza, A. A., Haroon, M., & Mudasar, A. Q. (2024). Flood Prediction System Using IOT & Artificial Neural Network. VFAST Transactions on Software Engineering, 12(1), 210-224
6. J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, “Semi-Supervised Network Traffic Classification,” pp. 369–370, doi: 10.1145/1254882.1254934
7. Zhao, G., Wang, Y., &Wang, J.(2023). Lightweight intrusion detection model of the Internet of Things with Hybrid Cloud-Fog Computing. 2023.
8. Zhang, W., Han, D., Li, K. C., & Massetto, F. I. (2020). Wireless sensor network intrusion detection system based on MK-ELM.SoftComputing, 24(16), 12361–12374. https://doi.org/10.1007/s00500-020-04678-1