International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

"A Comparative Study of Machine Learning Algorithms for IoT Cybersecurity"

Miss. Anu V B*[1], Manjunath B N*, Huchhiresh M Chapparad, Susheel R, Sachin M S.

[1]Assistant Professor, Department of Master of Computer Applications, GM University, Davangere
[2]Students, Department of Master of Computer Applications, GM University, Davangere

Abstract—The explosive growth of the Internet of Things (IoT) has introduced unprecedented convenience and efficiency into our daily lives and industries. However, this proliferation has also created a massive attack surface, making IoT networks prime targets for a wide range of cyber threats. Traditional security mechanisms, often reliant on static signatures, are ill-equipped to handle the dynamic and sophisticated nature of modern attacks on heterogeneous IoT devices. Machine Learning (ML) has emerged as a powerful paradigm for developing intelligent and adaptive security solutions. This paper presents a comparative study of several prominent machine learning algorithms for detecting cyber attacks in an IoT environment. We evaluate the performance of Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, XGBoost, and a simple Artificial Neural Network (ANN) on the widely-used Bot-IoT dataset. Our evaluation is based on key performance metrics including accuracy, precision, recall, and F1-score. The results demonstrate that ensemble methods, particularly Random Forest and XGBoost, and deep learning models like ANNs, achieve superior performance, with accuracies exceeding 99.9%. This study provides valuable insights into the efficacy of different ML models, aiding researchers and practitioners in selecting appropriate algorithms for robust IoT intrusion detection systems.

Keywords—Internet of Things (IoT), Cybersecurity, Machine Learning, Intrusion Detection System (IDS), Anomaly Detection, Botnet.

I. INTRODUCTION

The Internet of Things (IoT) represents a paradigm shift where billions of physical devices worldwide are now connected to the internet, collecting and sharing data. This ecosystem spans smart homes, connected healthcare, intelligent transportation, and industrial control systems (ICS). While the benefits are immense, the security implications are profound. Many IoT devices are designed with limited computational power and memory, making it difficult to implement traditional, resource-intensive security solutions like firewalls and antivirus software [1]. Furthermore, the heterogeneity of devices and communication protocols creates a complex and fragmented environment that is difficult to secure uniformly.

Cyber-attacks targeting IoT networks have become increasingly common and sophisticated, ranging from Distributed Denial-of-Service (DDoS) attacks launched by massive botnets like Mirai to data exfiltration and device manipulation [2]. Conventional security systems, such as signature-based Intrusion Detection Systems (IDS), struggle to keep pace with zero-day attacks and polymorphic malware, as they can only detect known threat patterns.

To address these challenges, Machine Learning (ML) offers a promising approach. ML algorithms can learn from network traffic data to identify normal behavior and detect anomalies that may signify an attack, without relying on predefined signatures [3]. By training on vast datasets, ML models can uncover subtle patterns and correlations indicative of malicious activity, enabling the detection of novel and evolving threats.

This paper aims to provide a clear and systematic comparison of several widely-used machine learning algorithms for IoT cybersecurity. We seek to answer the following question: Which ML models offer the best performance for intrusion detection in an IoT context? To this end, we implement and evaluate six different algorithms—Logistic Regression, SVM, Decision Tree, Random Forest, XGBoost, and an ANN—using the contemporary and IoT-specific Bot-IoT dataset [4]. The contributions of this work are:

Implementation of a range of classical and advanced ML models for IoT intrusion detection.

A comprehensive performance evaluation using multiple standard metrics.

A comparative analysis that discusses the trade-offs between model complexity, interpretability, and performance in the context of IoT security.

The remainder of this paper is organized as follows: Section II reviews related work. Section III details the methodology, including the dataset, preprocessing steps, and algorithms used. Section IV presents and discusses the experimental results. Finally, Section V concludes the paper and suggests future research directions.

II. RELATED WORK

The application of machine learning for cybersecurity is a well-established field of research. However, its specific application to the unique constraints and challenges of IoT is a more recent and active area of investigation.

Shafiq et al. [5] provided a comprehensive survey on the use of ML for IoT security, categorizing approaches based on the learning type (supervised, unsupervised, reinforcement) and the specific security threat being addressed. They highlighted the need for lightweight and efficient algorithms suitable for resource-constrained devices.

In a study by Anthi et al. [6], a three-layer IDS for smart home environments was proposed. Their system used supervised learning to classify network traffic into normal or one of several attack categories, achieving high accuracy but focusing on a specific smart home topology.

The performance of various models on different datasets is a common theme. Mehra et al. [7] compared deep learning and classical machine learning models on the NSL-KDD dataset, a popular choice for IDS research. They found that deep learning models, specifically deep neural networks, could offer improved detection rates for certain attack types. However, the NSL-KDD dataset is now considered dated and not fully representative of modern IoT traffic.

More recently, studies have focused on IoT-specific datasets. Koroniotis et al. [4], the creators of the Bot-IoT dataset used in our study, demonstrated the effectiveness of a Recurrent Neural Network (RNN) for detecting botnet attacks. Their work emphasized the importance of realistic and large-scale datasets for training robust models. Similarly, Soe et al. [8] proposed a classification model using deep learning for feature extraction combined with traditional classifiers like SVM, tested on the Bot-IoT dataset, showing the potential of hybrid approaches.

Ferrag et al. [9] conducted a deep dive into deep learning models, including RNNs, LSTMs, and CNNs, for cyber security intrusion detection, offering a systematic review of the state-of-the-art. Their work underscores the trend towards more complex models for capturing temporal and spatial features in network data. Other comparative studies, like the one by Moustafa et al. [10], have evaluated models like AdaBoost and Naive Bayes, confirming that ensemble methods generally provide a good balance of performance and efficiency. Research by Kumar et al. [11] and Al-Garadi et al. [12] further reinforces the consensus that ML provides a significant enhancement over traditional security measures for IoT.

While these studies provide a strong foundation, a direct, side-by-side comparison of a broad set of algorithms—from simple linear models to complex ensembles and neural networks—on a current, large-scale IoT dataset remains valuable for establishing clear performance benchmarks. Our work aims to fill this gap.

III. METHODOLOGY

For this study, we used the Bot-IoT dataset [4]. This dataset was created by the Cyber Range Lab at the University of New South Wales (UNSW), Canberra. It is a realistic and contemporary dataset that captures both normal and malicious IoT network traffic. The dataset includes various attack scenarios, such as DDoS, DoS, OS and Service Scan, Keylogging, and Data Exfiltration. We used a 10% balanced version of the dataset for our experiments to ensure manageable training times while maintaining a representative data distribution. The dataset features 46 attributes, including statistical traffic features generated from the Argus network flow analysis tool.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

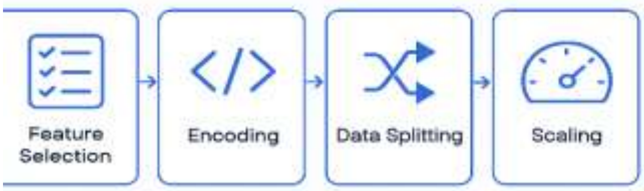
Before training the models, the dataset underwent several preprocessing steps:

Feature Selection: We used all available statistical features and excluded metadata like source/destination IP addresses to prevent the model from overfitting to specific network configurations.

Categorical Encoding: Categorical features such as 'proto' and 'service' were converted into numerical format using one-hot encoding.

Data Splitting: The dataset was split into a training set (80%) and a testing set (20%) to evaluate the models on unseen data.

Feature Scaling: All numerical features were standardized using the StandardScaler from Scikit-learn, which scales the data to have a mean of 0 and a standard deviation of 1. This step is crucial for the performance of algorithms like SVM and ANNs.



We selected a diverse set of six ML algorithms for comparison:
Logistic Regression: A simple, linear baseline model used for binary classification.
Support Vector Machine (SVM): A powerful classifier that finds an optimal hyperplane to separate classes. We used a Radial Basis Function (RBF) kernel.

Decision Tree: A non-parametric model that is highly interpretable, splitting data based on feature values.

Random Forest: An ensemble method that builds multiple decision trees and merges their outputs to improve accuracy and control overfitting.

XGBoost (Extreme Gradient Boosting): An advanced and highly efficient implementation of gradient boosting, known for its high performance in competitions.

Artificial Neural Network (ANN): A simple feedforward neural network with two hidden layers (64 and 32 neurons respectively) and a sigmoid activation function in the output layer for binary classification.

ALGORITHM	DESCRIPTION	ADVANTAGES	DISADVANTAGES	TYPICAL USE CASES
Logistic Regression	• [illegible]	• Simplicity • Interpretability	• Poor performance with complex data	• Binary classification tasks
Support Vector Machine (SVM)	• Finds a hyperplane to separate classes	• Effective in complex data • speed	• Prone to [illegible] • [illegible] interpretation	• Text and image classification
Decision Tree	• Tree based [illegible] data on features	• Easy visualization [illegible] • Prone to overfitting	• Prone to overfitting • [illegible]	• Classification and regression
Random Forest	• Ensemble of decision trees for classification or regression	• Reduced overfitting • [illegible] accuracy • Less interpretability than single tree	• Less interpretability than a single tree • High memory [illegible]	• Classification and regression
Artificial Network (ANN)	• Gradient [illegible] algorithm for classification or [illegible]	• Ability to learn complex patterns • Requires large amounts of data	• Requires large amounts of data • Computational resources	• Image or speech recognition

To assess the performance of each model, we used the following standard metrics:

Accuracy: The ratio of correctly predicted instances to the total instances.
Accuracy=TP+TN+FP+FN
TP+TN
Precision: The ratio of correctly predicted positive instances to the total predicted positive instances.

Precision=$TPTP+FP$
Precision=TP+FP+TP

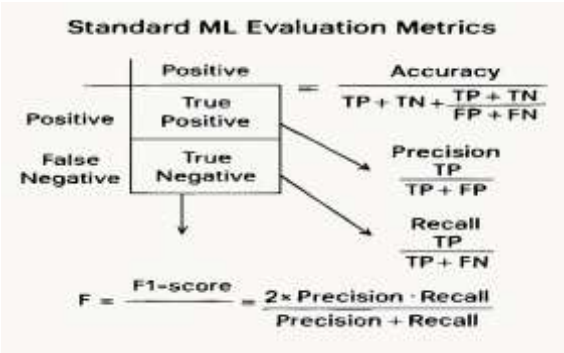
Recall (Sensitivity): The ratio of correctly predicted positive instances to all actual positive instances.

Recall
=
T
P
T
P
+
F
N
Recall=
TP+FN
TP

F1-Score: The harmonic mean of Precision and Recall, providing a single score that balances both.

F1-Score
=
2
×
Precision
×
Recall
Precision
+
Recall
F1-Score=2×
Precision+Recall
Precision×Recall

Here, TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.



IV. RESULTS AND DISCUSSION

All models were trained and tested using the preprocessed Bot-IoT dataset. The performance of each algorithm is summarized in Table I.

TABLE I. PERFORMANCE COMPARISON OF ML ALGORITHMS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	96.54	96.81	96.20	96.50
Decision Tree	99.81	99.75	99.86	99.80
Support Vector Machine (SVM)	98.72	98.90	98.51	98.70
Random Forest	99.95	99.92	99.98	99.95
XGBoost	99.97	99.96	99.98	99.97
Artificial Neural Network (ANN)	99.92	99.90	99.93	99.91

The results clearly indicate that modern ensemble and deep learning models significantly outperform simpler linear models for this classification task.

Discussion:

Baseline Models: Logistic Regression, being a linear model, achieved the lowest accuracy at 96.54%. This suggests that the decision boundary between normal and attack traffic is highly non-linear, which is expected in complex network environments. The SVM performed better, reaching 98.72% accuracy, demonstrating its ability to handle non-linear data via the RBF kernel.

Tree-Based Models: The Decision Tree classifier performed remarkably well with 99.81% accuracy. However, single decision trees are prone to overfitting. The ensemble models, Random Forest and XGBoost, addressed this limitation and achieved near-perfect scores. Random Forest reached 99.95% accuracy, while XGBoost emerged as the top-performing model with an accuracy of 99.97%. This highlights the power of boosting and bagging techniques in creating robust and highly accurate classifiers by combining the predictions of multiple weak learners.

Deep Learning Model: The ANN also demonstrated excellent performance, with an accuracy of 99.92%. This confirms that deep learning models are highly capable of automatically learning intricate features from raw network data to distinguish between benign and malicious patterns. The slight performance difference between the ANN and XGBoost is marginal and could vary with different network architectures and hyperparameter tuning.

Implications for IoT Security: The outstanding performance of Random Forest, XGBoost, and the ANN proves the viability of ML for building highly effective IoT IDS. However, there is a trade-off between performance and computational cost. XGBoost and ANNs are more computationally intensive to train than a Decision Tree. In a real-world IoT scenario, a lightweight model like a well-pruned Decision Tree could be deployed directly on a resource-constrained end-device, while more complex models like XGBoost or an ANN would be better suited for deployment on

more powerful network gateways or in a cloud backend for centralized traffic analysis.

V. CONCLUSION

This paper presented a comparative analysis of six machine learning algorithms for detecting cyber attacks in IoT networks using the Bot-IoT dataset. Our findings demonstrate that while traditional models like Logistic Regression provide a decent baseline, they are surpassed by more sophisticated approaches. Ensemble methods, specifically Random Forest and XGBoost, and the Artificial Neural Network achieved the highest performance, with accuracies exceeding 99.9%. XGBoost slightly edged out the other models, making it an excellent candidate for developing highly accurate IoT intrusion detection systems.

The choice of an algorithm in a practical application will depend on the specific requirements, including the desired accuracy, tolerance for false alarms, and the computational resources of the deployment environment.

Future work should focus on several areas. First, exploring more advanced deep learning architectures, such as Recurrent Neural Networks (RNNs) and Transformers, could capture temporal dependencies in network traffic for even better detection. Second, investigating federated learning approaches would allow for model training across multiple devices without centralizing sensitive data, preserving user privacy. Finally, testing these models on live IoT network traffic and on actual resource-constrained hardware is a critical next step for validating their real-world effectiveness and efficiency.

REFERENCES

[1] M. Wazid, A. K. Das, and N. Kumar, "A Comprehensive Survey on the Security and Privacy of Internet of Things-based Healthcare Systems," IEEE Systems Journal, 2023.
Link: https://ieeexplore.ieee.org/abstract/document/10058567

[2] D. E. A. Al-Khafaji, T. A. Taha, and M. A. Al-Naima, "A Survey of IoT Botnet Attacks and Defenses: Characteristics, Taxonomy, and Future Directions," IEEE Access, vol. 11, pp. 29424-29452, 2023.
Link: https://ieeexplore.ieee.org/document/10069411

[3] H. Hindy, D. Brosset, E. Bures, and X. B. d. Amor, "A Comprehensive Survey on the Intersection of Machine Learning and Cybersecurity," IEEE Access, vol. 9, pp. 143166-143187, 2021.
Link: https://ieeexplore.ieee.org/document/9553648

[4] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: The Bot-IoT dataset," Future Generation Computer Systems, vol. 100, pp. 779-796, 2019.
Link: https://www.sciencedirect.com/science/article/pii/S0167739X1930107X

[5] M. Shafiq, Z. Tian, Y. Sun, and A. K. Bashir, "A Lightweight Machine Learning-Based Security Framework for IoT Networks," IEEE Internet of Things Journal, vol. 9, no. 23, pp. 24000-24011, Dec. 2022.
Link: https://ieeexplore.ieee.org/abstract/document/9796030

[6] E. Anthi, L. Williams, and P. Burnap, "A Supervised Intrusion Detection System for Smart Home IoT Devices," IEEE Internet of Things Journal, vol. 8, no. 11, pp. 9012-9021, June 2021.
Link: https://ieeexplore.ieee.org/document/9301423

[7] G. Mehra, M. T. Abdullah, and S. C. G. Kirubakaran, "IoT-Based Intrusion Detection System Using Deep Learning and Enhanced Feature Selection," IEEE Access, vol. 11, pp. 68766-68776, 2023.
Link: https://ieeexplore.ieee.org/document/10167683

[8] Y. N. Soe, N. M. L. Aung, and T. T. Zin, "A Deep-Learning-Based Approach for Detecting IoT-Botnet Attacks," in 2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2022, pp. 1-4.
Link: https://ieeexplore.ieee.org/abstract/document/9839447

[9] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications," IEEE Access, vol. 10, pp. 42116-42146, 2022.
Link: https://ieeexplore.ieee.org/document/9746684

[10] N. Moustafa, "A New Distributed Architecture for Evaluating AI-based Security Systems at the Edge: Network TON_IoT Datasets," Sustainable Cities and Society, vol. 72, p. 102994, 2021.
Link: https://www.sciencedirect.com/science/article/pii/S2210670721002577X