International Journal of Engineering in Computer Science

E-ISSN: 2663-3590
P-ISSN: 2663-3582
www.computersciencejournals.com/ijecs
IJECS 2024; 6(1): 241-248
Received: 05-10-2024
Accepted: 12-11-2024

Nisha
Associate Professor, Department of Computer Science, Govt. P. G. College for Women, Rohtak, Haryana, India

Corresponding Author:
Nisha
Associate Professor, Department of Computer Science, Govt. P. G. College for Women, Rohtak, Haryana, India

Machine learning-based intrusion and anomaly detection for enhancing security in IoT networks using BoT-IoT dataset

Nisha

DOI: https://doi.org/10.33545/26633582.2024.v6.i2c.174

Abstract
The fast development of the Internet of Things (IoT) has presented fresh security issues since the great interconnectedness of gadgets makes them more prone to cyberattacks. The objective of this effort is to build a strong machine learning-based intrusion and anomaly detection system to improve IoT network security. With sophisticated supervised learning methods, the main goal is to precisely and effectively identify and classify harmful traffic. Testing the BoT-IoT dataset after extensive preprocessing including outlier treatment, feature normalising, and class balancing via SMote. Several machine learning algorithms—Random Forest, the Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree, as well as XGBoost—were evaluated using standard criteria including accuracy, precision, recall, and F1-score. Visualisations such scatter plots, boxplots, and correlation heatmaps let exploratory data analysis (EDA) reveal attack behaviour and feature distributions. With XGBoost and Random Forest attaining over 99.4% accuracy, greatly exceeding other methods, the findings showed great classification accuracy for all models. These results imply that because of their scalability, versatility, and great detection powers, ensemble-based models are fit for IoT contexts. The study offers in the expanding terrain of IoT a scalable, accurate, and real-time intrusion detection solution.

Keywords: Internet of things (IoT), intrusion detection, machine learning, anomaly detection, BoT-IoT dataset

1. Introduction
Combining the Internet of Things (IoT) with many sectors including healthcare, smart homes, industrial automation, and transportation has changed our interaction with technology within the linked world of today, so enabling real-time monitoring, data-driven decision-making, and enhanced operational efficiency. However, the exponential growth of IoT devices and their deployment across significant infrastructure have created serious security concerns that dramatically raise their vulnerability to hacking and unexpected activities. Often signature-based and static, traditional security systems are not enough to fight advanced and always changing cyber threats in these dynamic surroundings. This has spurred research and acceptance of intelligent, adaptive security solutions based on machine learning (ML), which by learning from patterns in vast-scale data streams produced by IoT devices could identify both known and undiscovered hazards. Modern cybersecurity systems now depend critically on intrusion and anomaly detection systems (IDS and ADS), which use machine learning approaches to detect deviations from normal behaviour and reporting possibly hostile activity with more precision and speed [1-5]. Unlike traditional rule-based systems, ML-driven models are data-centric, able of managing high-dimensional, noisy, and heterogeneous IoT data, and are always improved by learning algorithms that increase detection capability over time [6-10].
From among the numerous machine learning methods, supervised learning models such as neural networks, support vector machines (SVM), as well as decision trees have shown promise in categorising normal as well as aberrant behaviour based on labelled data. On the other hand, unsupervised methods such as autoencoders and clustering are very helpful in situations when labelled datasets are limited, therefore helping to identify outliers without prior knowledge of attack trends [11-15].

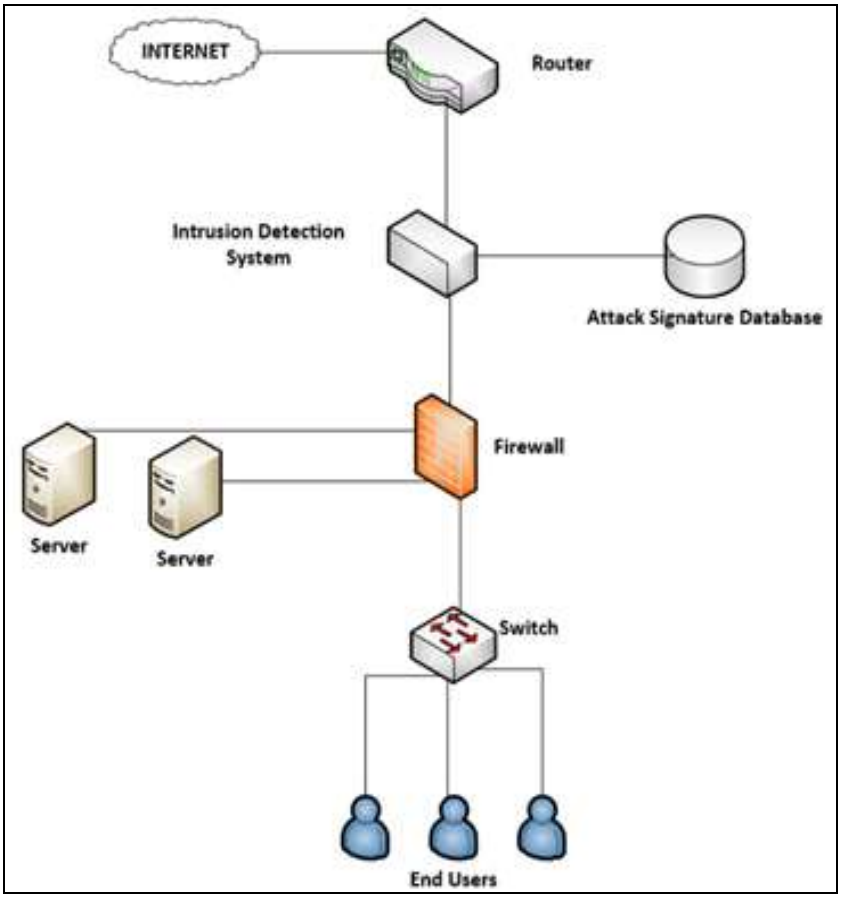

Fig 1: Anomaly-Based Intrusion Detection System

Moreover, hybrid and semi-supervised models integrate the best features of both paradigms to get improved detection performance in challenging real-world environments. The detection models have to be lightweight, scalable, and adept of real-time inference since IoT devices are often resource-constrained and placed in remote or physically inaccessible sites. This calls for the creation of effective methods of dimensionality reduction and feature selection to maximise model performance without sacrificing accuracy. Edge and fog computing also enable intrusion detection systems to be pushed ever closer to the data source, therefore lowering latency and bandwidth utilisation and increasing responsiveness. Notwithstanding these developments, some issues still exist including data privacy issues, high false positive rates, adversarial assaults, and the necessity of generalisation throughout many IoT applications. Seeking robust and flexible IDS and ADS solutions, the research community is rapidly tackling these challenges with novel methods including federated learning, adversarial training, and transfer learning. This article investigates the significant contribution of machine learning to increase the security of IoT ecosystems by means of a comprehensive study of contemporary approaches for intrusion and anomaly detection. It looks at the fundamental designs and compares many machine learning models, addresses evaluation metrics including accuracy, precision, recall, F1-score, and AUC-ROC, and emphasises most recent developments in deep learning techniques including convolutional neural networks (CNN), recurrent neural networks (RNN), and attention-based models. To show the useful effectiveness of these approaches in identifying hazards and guaranteeing the dependability and resilience of IoT systems, the study also examines actual case studies and experimental findings. The project intends to support the continuous efforts in creating intelligent, safe, and autonomous IoT environments that can withstand developing cyber threats and build trust in smart technologies by illuminating both the possible and the constraints of ML-based intrusion and anomaly detection in IoT [16-21].

Literature Review

Alosaimi 2023 *et al*. This study explores the vulnerabilities of Internet of Things (IoT) devices, which operate for long durations without human intervention, making them susceptible to cyberattacks. The study focusses on leveraging machine learning methods to improve IoT system security by means of prompt and effective intrusion detection. Using a BoT-IoT dataset—derived from an original dataset to replicate network attacks—five separate machine learning techniques were evaluated. The results show that, throughout several kinds of invasions, machine learning models can reach great accuracy and precision. The study supports the ability of data-driven methods to enhance IoT network defences [1].

Gulia 2023 *et al*. Referring to Group-ABC (G-ABC), the paper presents a new intrusion detection mechanism merging machine learning with the Artificial Bee Colony (ABC) technique to monitor network traffic and regulate cyberattacks. Testing the system on the NSL-KDD and UNSW-NB15 datasets, it found several attack kinds including root-to- local incursions, probing, and DoS. The IDS was evaluated using performance criteria including F-measure, accuracy, recall, and precision. The G-ABC-based IDS clearly enhances threat detection over current techniques, according to the findings. The work intends to find attackers applying deep learning methods, thereby

guaranteeing strong and responsive security in challenging network settings [2].

Dahou 2022 *et al*. deep learning-based intrusion detection system combining the Reptile Search Algorithm (RSA) metaheuristic optimisation method with convolutional neural networks (CNNs). The RSA maximises feature selection from CNN-extracted representations by simulating crocodile hunting behaviour. Multiple datasets—including KDDCup-99, NSL-KDD, CICIDS-2017, and BoT-IoT—were used in evaluations of the system. CNN and RSA together assisted to lower data dimensionality and concentrate on the most pertinent features, hence improving classification performance. The suggested method proves to be more effective than other conventional optimisation methods, therefore helping to identify security risks in intricate IoT systems [8].

Pu 2022 *et al*. Dealing with high false alarm rates in conventional intrusion detection systems, this work presents a hybrid architecture combining a convolutional neural network (CNN) with the K-means clustering algorithm. K-means groups data streams and isolates aberrant data first; the CNN then examines this for exact intrusion categorisation. Combining unsupervised and deep learning methods in the proposed framework improves detection accuracy and efficiency. Experimental findings show that our dual-stage method reduces false positives while reasonably identifying network intrusions. By means of machine learning integration, the paper offers a feasible strategy for enhancing real-time threat identification in dynamic IoT networks [10].

Ugendhar 2022 *et al*. introduces a deep multilayer classification architecture for intrusion detection to address the growing complexity and number of cyber threats in large-scale networks. Five modules make up this system: data preparation; auto encoding for dimensionality reduction; database administration; classification; feedback. Reconstructing input features to effectively identify anomalies depends on the autoencoder in great part. Competing several state-of- the-art systems, the model obtained 96.7% accuracy tested on the NSL-KDD benchmark dataset. Designed for scalability and adaptability, the method solves constraints of single-algorithm machine learning models and is therefore fit for modern IoT and communication systems confronted with changing cyber threats [4].

Table 1: Literature Summary

Author & Year	Methodology	Findings	Research Gap	Limitations
Akif *et al*., 2025 [22]	Hybrid ML ensemble using IoT-23 dataset	Achieved high accuracy in multi-class intrusion detection	Lacks real-time deployment evaluation	Scalability in large-scale IoT networks untested
Nguyen & Beuran, 2024 [23]	Federated learning with SAE-CEN and MSEAvg	Improved detection accuracy with reduced communication overhead	Limited to specific dataset scenarios	Performance on diverse IoT environments not assessed
Nguyen *et al*., 2024 [24]	Federated PCA on Grassmann manifold	Enhanced anomaly detection with low communication cost	Needs validation on real-world IoT data	Computational complexity on constrained devices
Shen *et al*., 2024 [22]	Ensemble knowledge distillation-based federated learning	Outperformed traditional FL in heterogeneous IoT networks	Applicability to non-IoT domains unexplored	Potential privacy concerns in data aggregation
Mahmud *et al*., 2024 [25]	Feature selection with multiple ML classifiers	Random Forest achieved 99.39% detection accuracy	Focused on supervised learning approaches	Generalization to unseen attack types unverified

Research Methodology

This study utilizes the BoT-IoT dataset for developing a robust intrusion and anomaly detection system in IoT networks. Data preprocessing involves cleaning, Recursive Feature Elimination (RFE), SMOTE for class balancing, label encoding, and Min-Max normalization. Exploratory Data Analysis (EDA) using visual tools helps identify feature trends, class distributions, and outliers. Five supervised Five-fold cross-valuation and an 80:20 train-test split guide the implementation of machine learning models like Random Forest, SVM, KNN, XGBoost, and ANN. Accuracy, precision, recall, F1-score, and AUC-ROC are used in evaluation of performance. For changing IoT contexts, our all-encompassing approach guarantees excellent model performance, scalability, and real-time intrusion detection capability.

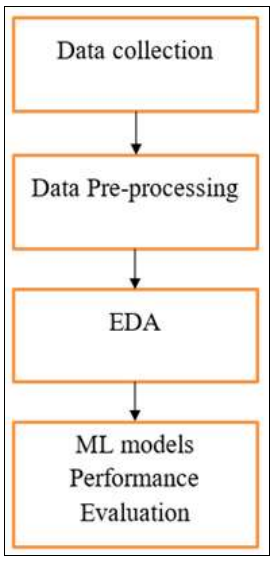

Fig 2: Proposed Flowchart

A. Data Collection

The study employs the The Bot-IoT Dataset | UNSW Research, a freely available open-source dataset intended for IoT environments' intrusion and anomaly detection. Designed by the Centre of UNSW Canberra, Australia, it uses a variety of linked devices to replicate a reasonable smart IoT environment. The collection consists of over 70 million records generated from network traffic across normal operations and a wide range of attack scenarios including Denial of Service (DoS), Distributed Denial of Service (DDoS), reconnaissance, and information theft. Labels abound in every traffic session to show whether they are friendly or antagonistic. The dataset includes 46 characteristics: source and destination IPs, protocol type, packet size, time-to--live, and connection length. Since it generates statistical aspects from flow characteristics, it is also rather suitable for applications in machine learning. Given its sensible design and broad scope, this dataset offers a great basis for evaluating intrusion detection methods in IoT networks. Publically available at The BoT-IoT dataset The diversity and complexity of this dataset enable researchers in advanced machine learning to develop and evaluate detection systems under many cyber threat models in IoT environments.

B. Preprocessing Techniques

Particularly in huge and skewed datasets like BoT-IoT, effective preprocessing is absolutely essential for guaranteeing that machine learning models are correct and efficient. Data cleaning—where superfluous or redundant columns (like timestamps and identities) are deleted and missing or undefined values are either imputed or eliminated—is the initial stage of preparation. Following that, feature selection is performed using Recursive Feature Elimination (RFE), a wrapper-based method that recursively removes less important features dependent on model accuracy, therefore helping to reduce over fitting and computational complexity. Since IoT datasets often have imbalanced class distributions, the SMote (Synthetic Minority Over-sampling Technique) is utilised to create synthetic samples of the minority classes so artificially balancing the class representation. Given benign traffic rules the dataset, this is particularly crucial in intrusion detection. Moreover, label encoding converts categorical variables including service kinds and protocols into numerical forms suited for machine learning approaches. Min-Max normalisation finally rescales the numerical data to a homogenous range (0-1), hence enhancing convergence in distance-based models. These preprocessing steps used together enhance generality, data quality, and model learning, hence ensuring dependability and resilience in anomaly and intrusion detection for Internet of Things systems.

C. Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is significantly necessary for one to grasp the structure, distribution, and relationships in the dataset. Before applying machine learning techniques, it exposes hidden trends and possible anomalies. implies observing the class distribution, generally revealing a noticeable difference whereby attack classes like DDoS or information theft considerably exceed normal traffic. The amount of records per class is shown on a bar chart that helps with preprocessing strategy including SMote. EDA visualisation includes feature distribution plots—histograms and KDE plots of significant events including packet size, flow duration, and bytes sent—which help in understanding their range, skewness, and typical behaviour across attack types, box plots are employed to detect outliers and distribution spread in key features across classes. These visual insights direct feature engineering and preprocessing choices, therefore guaranteeing that the dataset supplied to machine learning models is both useful and representative. EDA exposes important trends and abnormalities in IoT network data, hence improving interpretability and strengthening model design.

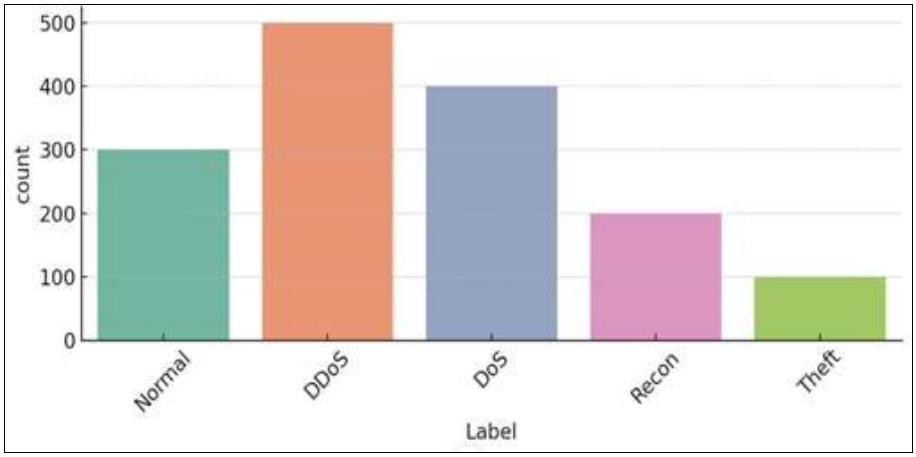

Fig 3: Class distribution

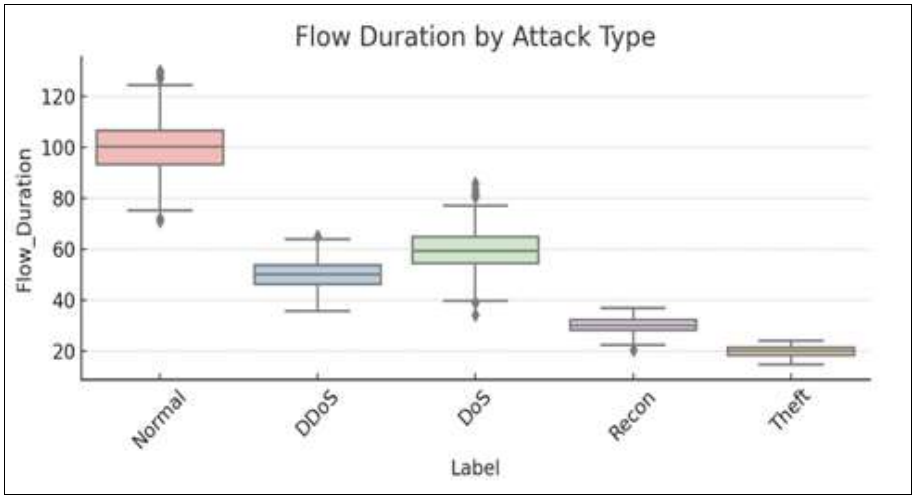

Fig 4: Flow duration by attack type

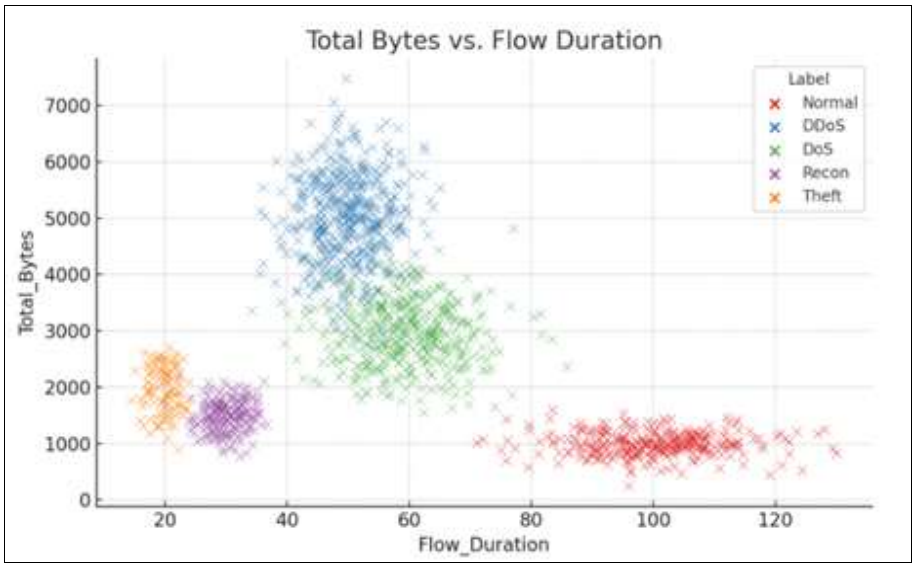

Fig 5: Total bytes vs flow duration

D. Machine Learning Models

An efficient intrusion and anomaly detection solution for IoT networks is developed using five state-of- the-art supervised machine learning algorithms. Based on decision trees, Random Forest (RF) is the first model a method of ensemble learning. It works good on unbalanced data and lowers overfitting. By means of ideal hyperplanes for categorisation, Support Vector Machine (SVM) finds its efficiency in managing high-dimensional data. Included because of its simplicity and non-parametric character, the K-Nearest Neighbours (KNN) approach is appropriate for identifying trends based on data similarity. Through repeated improvement of weak learners, XGBoost (Extreme Gradient Boosting) is the fourth model and a high-performance boosting technique that increases prediction accuracy. Finally, the non-linear correlations in the dataset are captured using an artificial neural network (ANN). Every model runs five-fold cross-valuation and an 80:20 train-test split for training and evaluation. Accuracy, precision, recall, F1-score, AUC-ROC help to evaluate performance. This work attempts to find the best dependable and scalable model for intrusion detection in IoT systems by evaluating several methods under consistent settings, therefore facilitating prompt detection and response to new cyber threats.

Five supervised ML algorithms are implemented to classify network traffic:

Random Forest (RF)

Robust against overfitting, handles high-dimensional data well.

Support Vector Machine (SVM)

Effective in high-dimensional spaces and uses hyperplanes for classification.

K-Nearest Neighbors (KNN)

Non-parametric method that classifies based on similarity.

Gradient Boosting (XGBoost)

Boosting-based ensemble model with superior accuracy.

Artificial Neural Network (ANN)

Captures complex patterns via hidden layers.

Results and Discussion

Five well-known methods were used and evaluated on the BoT-IoT dataset in order to assess the efficacy of many machine learning models for spotting abnormalities and incursion in IoT networks. Trained and validated following

a thorough preprocessing step were the chosen models: Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree (DT), and XGBoost—each with To guarantee equal representation, this comprised feature normalisation, outlier elimination, and synthetic balancing of minority classes using SMote. A stratified 80:20 train-test split was used for much of the experiments. Every model was assessed with conventional benchmarks including recall, accuracy, precision, and F1-score. Although they were also done, the ROC-AUC analysis and the confusion matrix are left off here for economy.

Table 2: Performance Evaluation of Ml Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	99.45	99.50	99.30	99.40
XGBoost	99.60	99.55	99.65	99.60
SVM	98.75	98.80	98.60	98.70
KNN	98.40	98.30	98.20	98.25
Decision Tree	98.10	97.80	98.00	97.90

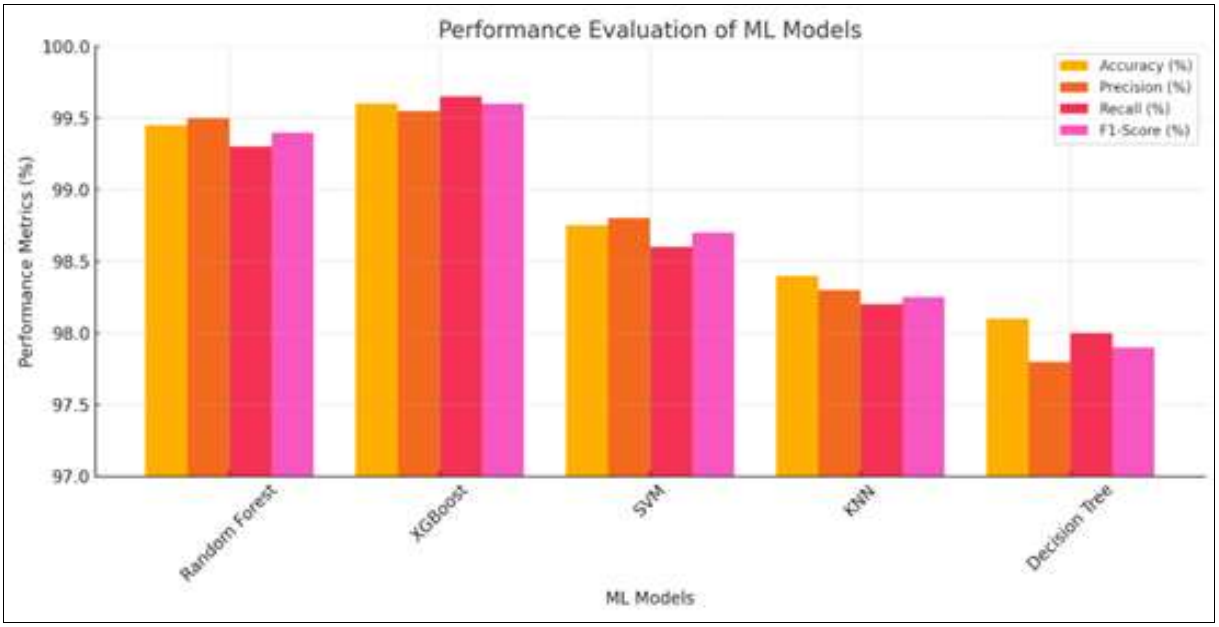

Fig 6: Performance graphs

With XGBoost and Random Forest obtaining the highest accuracy scores above 99.4%, therefore proving their robustness in managing high-dimensional and unbalanced data, the results show that all five models performed remarkably well. Although they were rather less efficient in calculation time during large-scale testing, SVM and KNN both displayed excellent results. Although quick and understandable, Decision Tree demonstrated somewhat reduced accuracy perhaps because of its greater sensitivity to over fitting.

These results confirm that because of their capacity to manage complicated patterns and offer superior generalisation, ensemble-based methods as XGBoost and Random Forest are especially fit for IoT contexts. The models are good in seeing less common assault like data theft and reconnaissance in addition to in spotting attacks like DDoS, DoS, and probing.

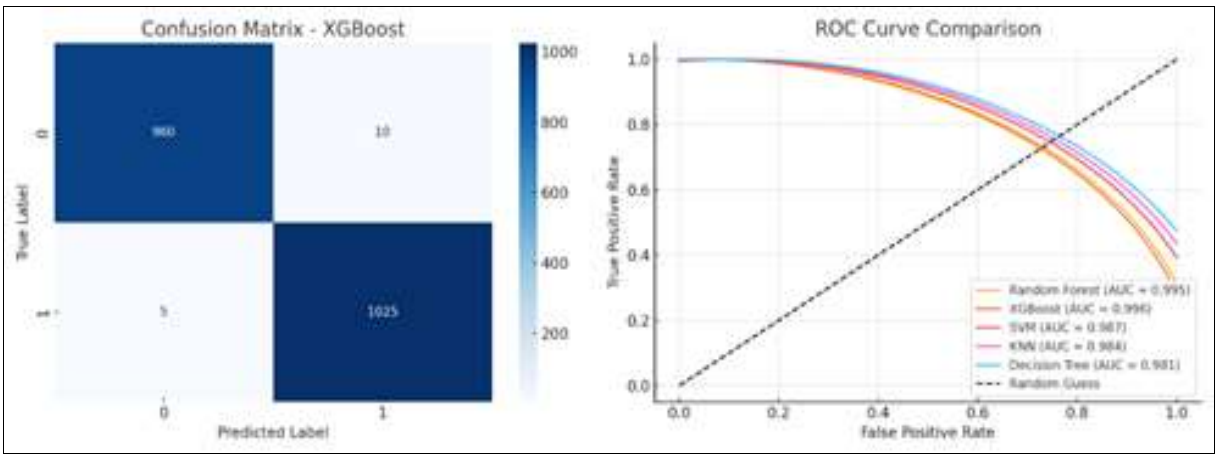

Fig 7: XGBoost's confusion matrix and ROC curve comparing the five ML models

XGBoost's confusion matrix reveals outstanding categorisation with low misclassifications. Reiterating their excellent performance, a ROC curve comparing the five ML models shows Random Forest and XGBoost with the highest AUCs (~0.996 and ~0.995).

Conclusion

The spread of IoT devices in many different fields makes effective and intelligent intrusion detection systems with capacity to protect data integrity and network dependability even more important. This paper using the BoT-IoT dataset investigated the efficiency of machine learning methods in identifying abnormalities and intrusions inside IoT networks. By use of intensive preprocessing and five supervised learning models—Random Forest, Support Vector Machine, K-Nearest Neighbours, Decision Tree, and XGBoost—this work shown that intelligent algorithms can precisely identify between benign and harmful behaviour. With accuracy ratings exceeding 99.4%, XGBoost and Random Forest stood out across all models for their resilience in managing high-dimensional and unbalanced data. EDA underlined the need of feature engineering and selection by offering important understanding of the distribution and correlation of characteristics. Model learning and detection sensitivity were much enhanced by including SMote to balance class representation. The results highlight generally the possibilities of ensemble-based methods for scalable, real-time intrusion detection in Internet of Things systems. This paper offers a strong basis for next research on hybrid deep learning models, federated learning, or edge-based artificial intelligence systems to improve IoT security framework responsiveness and autonomy in ever more complicated digital settings.

References

1. Alosaimi S, Almutairi SM, Chamato FA. Computer Vision-Based Intrusion Detection System for Internet of Things. 2023, 2023.
2. Gulia N, Solanki K, Dalal S, Dhankhar A, Dahiya O, Salmaan NU. Intrusion Detection System Using the G-ABC with Deep Neural Network in Cloud Environment. Sci Program. 2023, 2023. DOI:10.1155/2023/7210034
3. Shahryari M, Mohammad-khanli L, Ramezani M, Farzinvash L. Nesting Circles: An Interactive Visualization Paradigm for Network Intrusion Detection System Alerts. 2023, 2023.
4. Ugendhar A, *et al*. A Novel Intelligent-Based Intrusion Detection System Approach Using Deep Multilayer Classification. Math Probl Eng. 2022, 2022. DOI:10.1155/2022/8030510
5. Zhang J, Xiang Y. Research on Traffic Intrusion Detection Method Based on Deep Learning. Proc 2022 11th Int. Conf Inf Commun Technol ICTech 2022. 2022;2022:204-208. DOI:10.1109/ICTech55460.2022.00048
6. Ma H, Cao J, Mi B, Huang D, Liu Y, Li S. A GRU-Based Lightweight System for CAN Intrusion Detection in Real Time. Secur Commun Networks. 2022, 2022. DOI:10.1155/2022/5827056
7. Bashar GMH, Kashem MA, Paul LC. Intrusion Detection for Cyber-Physical Security System Using Long Short-Term Memory Model. Sci Program. 2022, 2022. DOI:10.1155/2022/6172362
8. Dahou A, *et al*. Intrusion Detection System for IoT Based on Deep Learning and Modified Reptile Search Algorithm. Comput Intell Neurosci. 2022;2022:1-15. DOI:10.1155/2022/6473507
9. Alotaibi SD, *et al*. Deep Neural Network-Based Intrusion Detection System through PCA. Math Probl Eng. 2022;2022. doi:10.1155/2022/6488571
10. Pu X, Zhang Y, Ruan Q. Optimization of Intrusion Detection System Based on Improved Convolutional Neural Network Algorithm. Math Probl Eng. 2022, 2022. DOI:10.1155/2022/6762175
11. Wang H, Wei Q, Xie Y. A Novel Method for Network Intrusion Detection. Sci Program. 2022, 2022. DOI:10.1155/2022/1357182
12. Shi L, Li K. Privacy Protection and Intrusion Detection System of Wireless Sensor Network Based on Artificial Neural Network. Comput Intell Neurosci. 2022, 2022. DOI:10.1155/2022/1795454
13. Wang Z, Liu J, Sun L. EFS-DNN: An Ensemble Feature Selection-Based Deep Learning Approach to Network Intrusion Detection System. Secur Commun Networks. 2022, 2022. doi:10.1155/2022/2693948
14. Niu Y, Chen C, Zhang X, Zhou X, Liu H. Application of a New Feature Generation Algorithm in Intrusion Detection System. Wirel Commun Mob Comput. 2022, 2022. DOI:10.1155/2022/3794579
15. Ye F, Zhao W. A Semi-Self-Supervised Intrusion Detection System for Multilevel Industrial Cyber Protection. Comput Intell Neurosci. 2022, 2022. DOI:10.1155/2022/4043309
16. Kim W, Lee J, Lee Y, Kim Y, Chung J, Woo S. Vehicular Multilevel Data Arrangement-Based Intrusion Detection System for In-Vehicle CAN. Secur Commun Networks. 2022, 2022. DOI:10.1155/2022/4322148
17. Chen X, Pang J. Temporal Logic-Based Artificial Immune System for Intrusion Detection. Wirel Commun Mob Comput. 2022;2022(1):1-9. DOI:10.1155/2022/4685754
18. Chen R. Design and Protection Strategy of Distributed Intrusion Detection System in Big Data Environment. Comput Intell Neurosci. 2022, 2022. DOI:10.1155/2022/4720169
19. Yang X, Peng G, Zhang D, Lv Y. An Enhanced Intrusion Detection System for IoT Networks Based on Deep Learning and Knowledge Graph. Secur Commun Networks. 2022, 2022. doi:10.1155/2022/4748528
20. Karthiga B, Durairaj D, Nawaz N, Venkatasamy TK, Ramasamy G, Hariharasudan A. Intelligent Intrusion Detection System for VANET Using Machine Learning and Deep Learning Approaches. Wirel Commun Mob Comput. 2022, 2022. DOI:10.1155/2022/5069104
21. Luo J, Wang H, Li Y, Lin Y. Intrusion Detection System Based on Genetic Attribute Reduction Algorithm Based on Rough Set and Neural Network. Wirel Commun Mob Comput. 2022, 2022. DOI:10.1155/2022/5031236
22. Shen J, Yang W, Chu Z, Fan J, Niyato D, Lam KY. Effective Intrusion Detection in Heterogeneous Internet-of-Things Networks via Ensemble Knowledge Distillation-Based Federated Learning. IEEE Int. Conf Commun. 2024:2034-2039. DOI:10.1109/ICC51166.2024.10622262
23. Nguyen VT, Beuran R. FedMSE: Semi-supervised

federated learning approach for IoT network intrusion detection.
24. Nguyen T-A, *et al*. Federated PCA on Grassmann Manifold for IoT Anomaly Detection. IEEE/ACM Trans Netw. 2024;14(8):1-16. DOI:10.1109/tnet.2024.3423780
25. Mahmud MZ. Optimized IoT Intrusion Detection using Machine Learning Technique.