A Comparitive study on IOT security using Machine Learning Techniques

Eldo Ephraim joseph

School of Sciences

Christ(Deemed to Be) University

Bengaluru ,India

[eldojoseph2002@gmail.com](mailto:eldojoseph2002@gmail.com)

Vandana Sharma

School of Sciences

Christ(Deemed to Be) University,

Bengaluru ,India

[vandana.juyal@gmail.com](mailto:vandana.juyal@gmail.com)

Preety Shoran

School of Sciences

Christ(Deemed to Be) University

Bengaluru,India

[preety@christuniversity.in](mailto:preety@christuniversity.in)

*Abstract*— The security of IoT devices has been one pressing concern in light to how they make use of networks. The IoT system can be harmed by DDoS attack and C&C unit to compromise. This makes it difficult to receive accurate signals on the security breaches unless the field has powerful classification algorithms in their armoury. The existing algorithms to be utilized for classification and identification of the attacker's malicious IoT network traffic are Decision Tree, Random Forest and Histogram-Based Gradient Boosting this paper compares all three above-mentioned machine learning algorithm as a task based on performance []\*. IoT-23 dataset: The training and testing of the trained model implemented a subset of attacks — it was mainly accuracy oriented, i.e., most reliable attack evaluation based on average means. Therefore, it is apparent from the obtained results that all three operators perform well in terms of detecting attacks and Histogram-Based Gradient Boosting gives the best overall accuracy thus one can use this model for further IoT security developments. This study is especially helpful in assessing the efficacy of machine learning as an applicable extent for IoT network protection against novelty attacks.

Keywords—Internet of things (IoT), Machine Learning, Artificial Intelligence, Cybersecurity

# Introduction

Due to the fact that there are an exorbitant amount of IoT devices in market, and they make up such a notable part of technology industries from healthcare, manufacturing transportation or...smart homes. The IoT gathers by orders of magnitude more data than it can collect on its own because devices talk to each other and are networked throughout the world like they never have been before. But this connectivity has created a new dimension to security threats. As the count of devices skyrockets to billions worldwide, this enlarged attack surface enables ill-intentioned intruders to take advantage and infiltrate your IOT systems through cyber-attacks. The researchers as well as government has understood this and because of which most or our industries also have integrated security apparatus into these systems.

There are many things that need to be done in context of security for IoT networks, one such problem space is to identify and stop needles hidden by hackers’ activities carried out on the Network like Distributed Denial of Service (DDoS) attacks, Command-and-Control (C&C operations), Data Breaches etc. These attacks can carry wide-ranging repercussions, anywhere from service interruption to intellectual property theft and also function as a jumping off point for more expansive accessibility contrived cyber threats. Spot the fact that they take advantage of basically non-existent compute power and barely any security checks on many IoT devices. But these approaches are limited when it comes to the Internet of Things (IoT): one-size-fits-all network security solutions like firewalls and anti-virus software may be wholly incapable in protecting decentralized distributed systems. When it comes to IoT phenomenon, so the more advanced machine learning (ML) algorithms are used in detecting or identifying potential malicious network traffic.

Machine learning algorithms can process data on a massive scale using historical information to detect patterns and make predictions, which is expected solution for the security issues in an IoT environment. An example for the former would classification algorithms that help locate malicious network activities as a fixed number of class (e.g., benign classes and attack traffic) where once an activity features is extracted, its label has to be given out by comparison with current labelled data. assess the performance of ML models under classification accuracy, which is crucial for identifying and stopping IoT security threats. Among the models most commonly employed in this type of cybersecurity work, Decision Trees, Random Forests and Gradient Boosting have shown good results.

Decision Tree - Which we can consider as a reverse tree... Its primary advantage is interpretability... Yet, as Matinkhah et al. [1] point out, Decision Trees are prone to overfitting, especially with the noisy data often encountered in IoT environments. The drawback is this result in performance drops for new or unseen data and it especially occurs more frequently in high-dimensional datasets with high noise such as IoT networks. [2]

Aside from that, overfitting might occur with Decision Trees but Random Forest mitigates this by creating an ensemble of trees, as described by Hussain et al [5]. This method helps to prevent overfitting and makes your model generalizes on the test set, this is why Random Forest become a great algorithm for classification tasks. Random Forest models do work by each tree selection random features on random sample of data points and this description somehow lowers variance of the model but also keep bias in normal levels. Because of the random modelling nature and also to be able working on large, unbalanced datasets Random Forest found extensive use cases applying cyber security necessitated scenarios. [4]

The Hist Gradient Boosting Classifier, gives good results in various classification tasks. It builds stacking models on top of others to correct the errors that its earlier waves made, begetting a final model with extremely high accuracy. By fitting the model on examples that are hardest-to-classify using an iterative optimization strategy as it is implemented in performing boosting, and hence by making our half-trained learning operator also learn many non-half-assed features. The histogram version of Gradient Boosting as explored by Tahir et al. [3], enhances efficiency by classifying continuous feature values. This algorithm is particularly suitable for highly imbalanced class distributions seen in security datasets — and more so in relevant IoT applications. [6]

Large scale security model that systems accurate and performs well is in absolutely as must have for IoT network architecture, this research goes with evaluating & performance comparison among three dominant machine learning algorithm Decision Tree (DT), Random Forest (RF)and Histogram-based Gradient Boosting (GBM). Among all these algorithms the goal is to see which algorithm suits right with a perfect machine learning model that can be put into practical security use, giving us maximum accuracy in classifying IoT network traffic as benign or malicious.

We used the IoT-23 dataset, a publicly available benchmark data-set that has been created to assess how well machine learning models protect an IOT environ (just through those ML features). Dataset: IoT Audit attacks include different malware attack modes such as DDoS, C&C communications and benign traffics on IoT devices with the network traffic data labelled. The diversity in data types, originality and representation is what makes this dataset more suitable for training/testing of models intended to detect activity as malicious one.

IoT Deployment Challenges the IoT-23 dataset presents challenges often found in practical deployments of the IoT. One of the key reasons is that benign traffic highly overwhelms malicious one as there exists a large imbalance among our dataset. It skews models and enables a flawed disciplinary model wherein crucial exposition of impending attack is not followed. The second challenge was, because the dataset is so messy and has many columns representing a variety of network protocols and features (this also made some sense for all types of records) that converting them could potentially occur greater or less noise to our classification problem. Therefore, careful pre-processing and informative feature selection are essential to develop models that learns from the rich data what true patterns look like. The dataset multiple attack types hence a model has to perform well over different kinds of attacks for it to be robust towards an unseen image.

.

# **RELATED WORK**

There are various studies that have took place in the field of IoT security to secure and manage data.

### [1]" Exploring Impact of Data Noise on IoT Security: a Study using Decision Tree Classification in Intrusion Detection Systems." Decision tree classifiers have an intuitive, easy to interpret tree-like structure that allows complex decisions be broken down into more interpretable rules they perform well in security issues where understanding the process is crucial. But they are prone to overfitting, even in the case of high-dimensional or noisy IoT data, which can constrain their generalization ability and real-world performance. In practice, the computational demands of complex trees can tax resource-constrained IoT devices so design and optimization are critical to make this practical for deployment.

### [2]"Using machine learning algorithms to enhance IoT system security " This paper is to explore the effects of data noise in cases based on decision tree for intrusion detection systems (IDS) in IoT networks. Experiments on four benchmark dataset show clean data has highest accuracy and any noise in use cases reduced its efficacy drastically. Although the study provide an invaluable information for IoT based IDS construction, it only focus on decision tree and is oblivious to other ML approaches.

### [3] "Enhancing IoT Security through Machine Learning-Driven Anomaly Detection."  The paper investigates machine learning algorithms for the improvement of IoT security, where it analyses different anomaly detection techniques based on the Random Forest, Decision Treee, SVM and Gradient Boosting. Decoding attack traffic: Of all the twelve models, Gradient Boosting achieved the best precision of 89.34% in recognizing an attacker. Nonetheless, drawbacks to consider are the possibility of overfitting and further related works on ensemble learning approaches for compensation of this limitation.

### [4] "Improving the Security of Internet of Things (IoT) Applications Based on a New Machine Learning Technique."This paper presents a novel method for enhancing IoT security because by coupling Salp Swarm Algorithm (SSA) with Decision Tree only and Random Forest, which are separated models in the context of comparison. SSA improves the training of those models, making it easier to discover intricate correlations that present in IOT data. The limitation of this study is that it is not for general purposes due to the focus on a specific data set and further study would be needed or improvements in order to take mechanism to other types of IoT applications.

### [5]"Machine Learning in IoT Security: Current Solutions and Future Challenges" This paper evaluates the performance of a Decision Tree (DT) and Random Forest (RF) model with Salp Swarm Algorithm(SSA), as an enhancement to secure IoT applications. The SSA model achieves best training performance of 95.54% accuracy for DT, and 96.19%accuracy with RF on an IoT dataset using the CPU-GPU platform until M =24 where it also starts to optimizes resource utilization swiftly then stabilizing as we increase resources from there see (Fig4). Nevertheless, it is tied to a cognitive dataset and more research needs precisely be executed on its generalization applicability for other IoT security challenges

### [6]" An improved anomaly detection model for IoT security using decision tree and gradient boosting." In this paper, a gradient boosting with decision trees-based algorithm has been proposed for IDS in the security of IoT via Catboost open sourcing. It is said that the accuracy of this method has been improved and has made GPU faster to detect their presence. Note, that the research concentrates on network-based intrusion detection; thus it only covers securing IoT devices. It does not provide comparison with other ensemble learning state-of-the-art methods, which could limit the possibilities of finding alternative effective solutions.

### [7] “Anomaly detection in IoT-based healthcare: machine learning for enhanced security” In this paper, we aim to explore the application of machine learning for detecting anomalies in IoT Healthcare by using a well-known dataset namely CICIoT2023. The study does not compare with deep learning models, and it conducts a shallow analysis on attack types individually (Recon and Spoofing). It also relies on a manipulated dataset, which raises questions regarding generalization to the real world.

### [8] “Malicious URL Detection using optimized Hist Gradient Boosting Classifier based on grid search method.”The paper studies detection of the malicious URLs, to get beyond feature extraction cascade process and more variants with class-imbalance solution; in this sense This article is a short version for Histogram-based Gradient Boosting Classifier (HGBC) optimizing together by grid search also SMOTE due to number features at ABCdef . It is able to achieve 96% accuracy, which suggests that HGBC performs well for detection of malicious URLs. Nevertheless, the study has a few limitations which include using an incomplete dataset for training, potential overfitting by SMOTE and not being tested in live traffic scenarios hence possibly non-generalization to new attacks. Further validation of these results with different datasets and scenarios in the real-world is needed.

# **research gap**

Even with an increasing number of the research work to make use of machine learning for IoT security, a deep review in current literatures shows critical gaps needing further investigation. One of the general patterns that emerges is that many studies are narrowly focused on a particular dataset, algorithm or type attack. Although decision trees ([1], [2]) provide interpretation, the possibility of overfitting — especially with noisy IoT data is imperfect— shall require designing proper optimization approaches and implementations like-constrained Competing Assertions(s) The same applies to the research on boosting ([3], [6]) — while it demonstrates many improvements in classification results over existing ensemble learning methods, these improvements are not smooth; there is still no rigorous evidence for whether emerging deep-learning methodologies bring more significant improvement.

This focus is very limited in its scope and reaches only a tiny fraction of all security threats. Although necessary (and complex) on its own, network intrusion detection is but one of many components needed for a complete IoT security stack which also involves device-level vulnerabilities and secure authentication mechanisms and data privacy issues [6]. In addition, the widespread use of gamed or underspecified datasets [[4], [5], [7]] calls into question the generalizability of results to real-world IoT settings where devices are diverse and attack patterns constantly evolve.

The reliance on curated datasets means our current batch of research evidence is potentially very biased, which only emphasises the need for more thoughtful work in evaluating models within real-world scenarios. Presence of network variability, resource constraints of IoT devices, and constantly changing cyber threats require solutions that adapt effectively to different scenarios. We call for future work to focus on the evaluation of these solutions in real-life setting using different use cases and data, as well as operationalizing their deployment on resource-constrained IoT devices commonly used nowadays.

# **problem statement**

Internet of Things (IoT) devices, although transformative across many industries, have increased the scope for cyber threats manifold by exposing large amount sensitive data and vital infrastructure. Machine learning (ML) holds great promise in improving IoT security but existing methodologies are as of yet restricted by critical constraints. There was a trade-off because decision tree-based intrusion detection systems have good interpretability but they suffer from performance drops when exposed to noisy and high-dimensional IoT data. In addition, most current ML models are not generalizable to new datasets, unseen attack strategies or real-world deployment.

In addition, the lack of explain ability (which decreases trust in a model) does not help much. Moreover, the computational requirements of high-level ML algorithms are typically outside of what resource-constrained IoT devices can handle. Consequently, due to the emergence of rapidly evolving and versatile security threats against the ubiquitous IoT environment including end systems and wide area networks (WANs), there is a vital necessity for highly generic secure ML-based solutions that are not only explainable but also lightweight in terms of computational complexity.

# **methodology**

##### **DATA COLLECTION**

Data collection and pre-processing are very important for any machine learning project because the result you will get, depends on your data. The dataset used in this research is iot23 csv, which is labelled network traffic data from the IoT-23 dataset. This dataset consists of plenty number of network traffic logs including benign and malignant activities such as various attacks namely DDoS, C&C, horizontal port scans. This data is necessary for building machine learning models to detect the good and bad traffic.

This study utilizes data from network capture logs collected in the context of IoT (Internet of Things) devices. The proliferation of IoT devices and the lack of security in most — if not all, have made them enticing targets for cyber attackers. Since the benign traffic is accompanied with different types of attacks that these devices encounter, this dataset provides a good source for creating models to address security issues using machine learning.

1. **DATA PREPROCESSING**

The collection of data results in raw form that are usually full with noises, irrelevant features or missing values. This raw data requires cleaning and transforming that makes it suitable for modelling which is one of the most important steps in any machine learning process. This step is very important as the sort of data you get, in large part affects how well your ML model can perform. [*Figure 1*.]

Initial data processing involved a careful selection of 6 million rows from the original dataset, which contained over a billion entries. This selection process ensured a balanced representation across categorical features, mitigating potential biases due to data imbalances. Subsequently, data quality was enhanced by removing null values, duplicate entries, and unnamed columns. Column names were then standardized for improved clarity and interpretability. Finally, factorization techniques were employed to consolidate rows with synonymous labels, further streamlining the dataset for analysis.

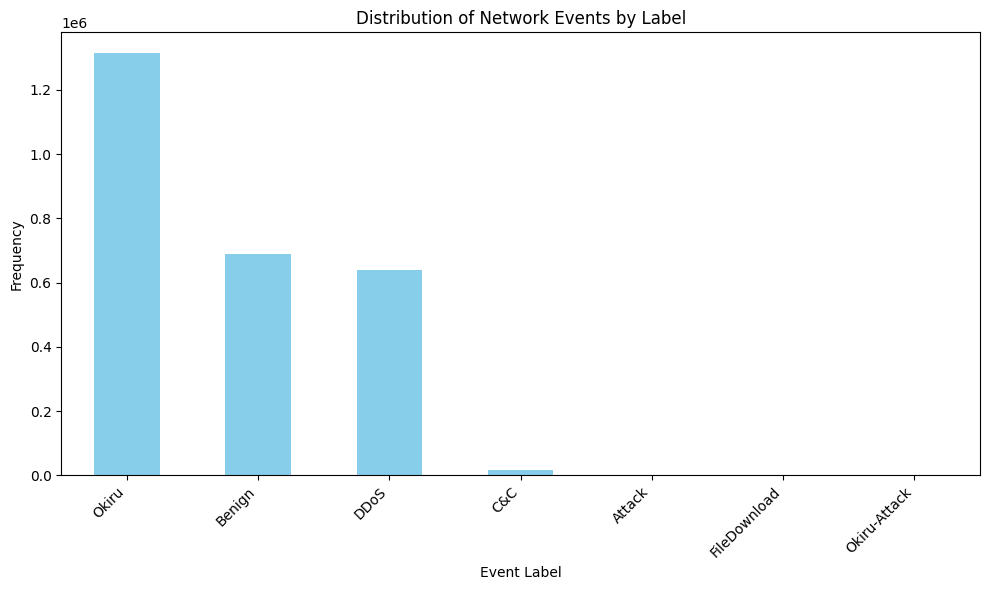


Figure 1. Distribution of Network Events

1. **DATA ANALYSIS**

Following data cleaning, exploratory data analysis was conducted using visualizations to gain insights into feature relationships and distributions. Textual columns deemed irrelevant to the analysis, such as IP addresses, were removed. Categorical features 'connection\_state' and 'connection\_duration' were encoded numerically to facilitate model training. A correlation heatmap, grouped by labels, was generated to identify feature dependencies. Based on this analysis, only features exhibiting either positive or negative correlations with the target variable were retained, excluding those with zero correlation.[ *Figure 2.*]

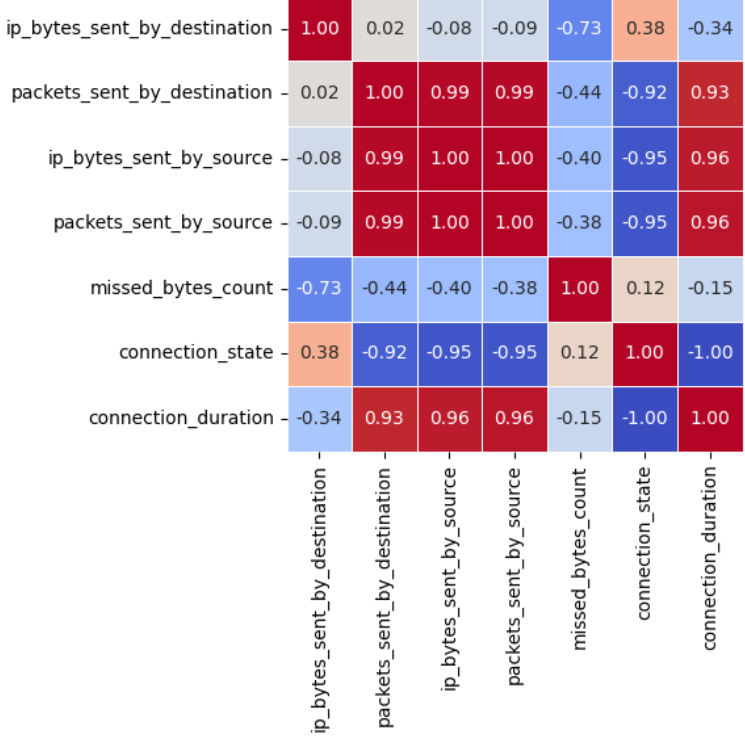


Figure 2. Correlation Heatmap for relationship between different network feature

The dataset was then partitioned into training and testing sets using an 80/20 split, employing train\_test\_split with a random\_state of 42 for reproducibility. This resulted in training and testing sets with shapes (2126069, 7) and (531518, 7) for features and (2126069,) and (531518,) for labels, respectively.

Three different classification algorithms were evaluated: RandomForestClassifier, DecisionTreeClassifier, and HistGradientBoostingClassifier. Each model was trained on the training data and its performance assessed using accuracy as the evaluation metric on the held-out test data.

# **Results**

The performance of the three classifiers on the test set is summarized below:

|  |  |
| --- | --- |
| Classifier | Accuracy |
| RandomForestClassifier | 0.8850 |
| DecisionTreeClassifier | 0.8850 |
| HistGradientBoostingClassifier | 0.8777 |

Both RandomForestClassifier and DecisionTreeClassifier achieved comparable accuracy scores of approximately 88.50%, demonstrating strong predictive capabilities. HistGradientBoostingClassifier, while still performing well, exhibited a slightly lower accuracy of 87.77%.

# **Discussion**

The results indicate that both RandomForestClassifier and DecisionTreeClassifier are highly effective in classifying the target variable based on the selected features. Their similar performance suggests that the dataset may be well-suited for tree-based models, potentially due to the presence of non-linear relationships or complex interactions between features. The slightly lower accuracy of HistGradientBoostingClassifier could be attributed to its sensitivity to hyperparameter tuning or the specific characteristics of the dataset.

Further investigation could involve exploring alternative evaluation metrics beyond accuracy, such as precision, recall, and F1-score, to gain a more comprehensive understanding of the models' performance across different classes. Additionally, hyperparameter optimization and feature engineering techniques could be employed to potentially improve the accuracy of all models, particularly HistGradientBoostingClassifier.

The insights gained from this study can be valuable for [mention the specific application or domain]. The high accuracy achieved by the chosen classifiers suggests the feasibility of developing a robust predictive model for [mention the specific task or goal]. Future work could focus on deploying the model in a real-world setting and evaluating its performance on new, unseen data.

##### .References

1. Matinkhah, S. Mojtaba, Roya Morshedi, and Akbar Mostafavi. "Exploring Impact of Data Noise on IoT Security: a Study using Decision Tree Classification in Intrusion Detection Systems." *Journal of AI and Data Mining* 11.4 (2023): 609-626.
2. El-Sofany, Hosam, et al. "Using machine learning algorithms to enhance IoT system security." *Scientific Reports* 14.1 (2024): 12077.
3. Tahir, Usama, et al. "Enhancing IoT Security through Machine Learning-Driven Anomaly Detection." *VFAST Transactions on Software Engineering* 12.2 (2024): 01-13.
4. Hussein, Alia A., et al. "Improving the Security of Internet of Things (IoT) Applications Based on a New Machine Learning Technique." *BIO Web of Conferences*. Vol. 97. EDP Sciences, 2024.
5. Hussain, Fatima, et al. "Machine learning in IoT security: Current solutions and future challenges." *IEEE Communications Surveys & Tutorials* 22.3 (2020): 1686-1721.
6. Douiba, Maryam, et al. "An improved anomaly detection model for IoT security using decision tree and gradient boosting." *The Journal of Supercomputing* 79.3 (2023): 3392-3411.
7. Khan, Maryam Mahsal, and Mohammed Alkhathami. "Anomaly detection in IoT-based healthcare: machine learning for enhanced security." *Scientific Reports* 14.1 (2024): 5872.
8. Maftoun, Mohammad, et al. "Malicious URL Detection using optimized Hist Gradient Boosting Classifier based on grid search method." *arXiv preprint arXiv:2406.10286* (2024).
9. https://www.kaggle.com/datasets/engraqeel/iot23preprocesseddata