DESIGNING AND SIMULATING A MACHINE LEARNING-BASED BOTNET DETECTION MODEL FOR IOT SYSTEMS

Done By:

Mohammad Abdalaziz

Supervisor :

Dr. Anastassia Gharib

Princess Sumaya University for Technology
جامعة الأميرة سميّة للتكنولوجيا

Content

- Introduction
- Objectives
- Background
- Design Requirements and Constraints
- Methodology
- System Architecture and Implementation Details
- Results
- Challenges and Solutions
- Future Directions
- Conclusion

Introduction

- Our project focuses on critical issues facing IoT, as they become target for malicious attacks which compromise both privacy and functionality.
- Our project aims to enhance IoT security by designing and simulating a machine learning-based botnet detection model with the application of Federated learning.

Objectives

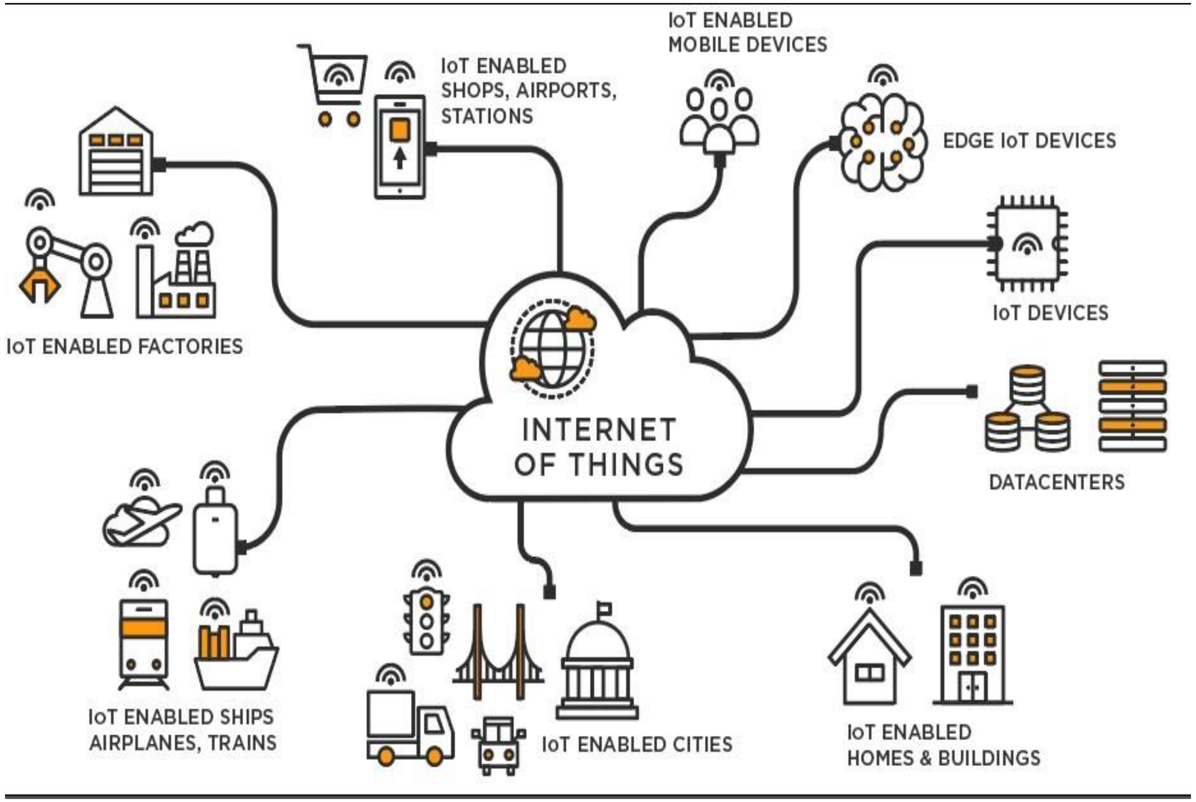
- Our primary objective is to explore and compare the effectiveness of two distinct approaches Centralized random forest and neural network model against Federated random forest
- Intend to evaluate not only accuracy and efficiency but also identifying security threats and their capability to handle data privacy
- The comparative analysis will help us determine most suitable machine learning strategy

Internet Of Things (IoT)

- the **Internet of Things** (**IoT**) represents all computing devices that are connected to the internet.

- IoT enables these interconnected devices to communicate and share data with each other through the internet.

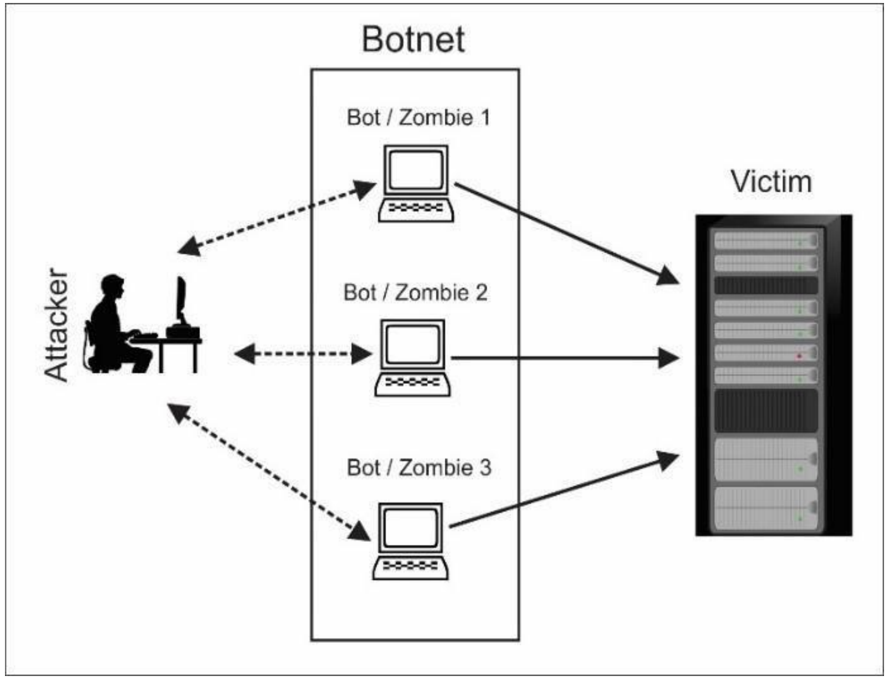
- This data can be collected, analyzed, and utilized to make informed decisions, improve efficiency in various applications.
- These devices are vulnerable to attacks


IoT ENABLED MOBILE DEVICES
IoT ENABLED SHOPS, AIRPORTS, STATIONS
EDGE IoT DEVICES
IoT DEVICES
IoT ENABLED FACTORIES
INTERNET OF THINGS
DATACENTERS
IoT ENABLED SHIPS AIRPLANES, TRAINS
IoT ENABLED CITIES
IoT ENABLED HOMES & BUILDINGS

“What is the Internet of Things (IoT)?,” *TIBCO Software*.
https://www.tibco.com/referencecenter/what-is-the-internet-of-things-iot

Botnet Attacks

- Botnet attacks are cybersecurity threat
- Botnet attack refers to malicious effort where the bots, a network of compromised computers that is controlled by botmaster
- These bots that can be devices connected to internet and are infected with malware allowing attacker to control them
- DDoS attacks, spam and phishing ,data theft and Brute force attack are malicious activities associated with botnet attacks



https://www.researchgate.net/figure/A-Typical-Botnet-AttackStructure_fig1_298788691

Machine Learning

- It is a subject of Artificial intelligence(AI) that focuses on using data and algorithms to learn from data and make predictions and decisions
- Create system that automatically learn , improve and solve complex problems
- Data, Algorithms, Training ,Testing and evaluation are components of machine learning

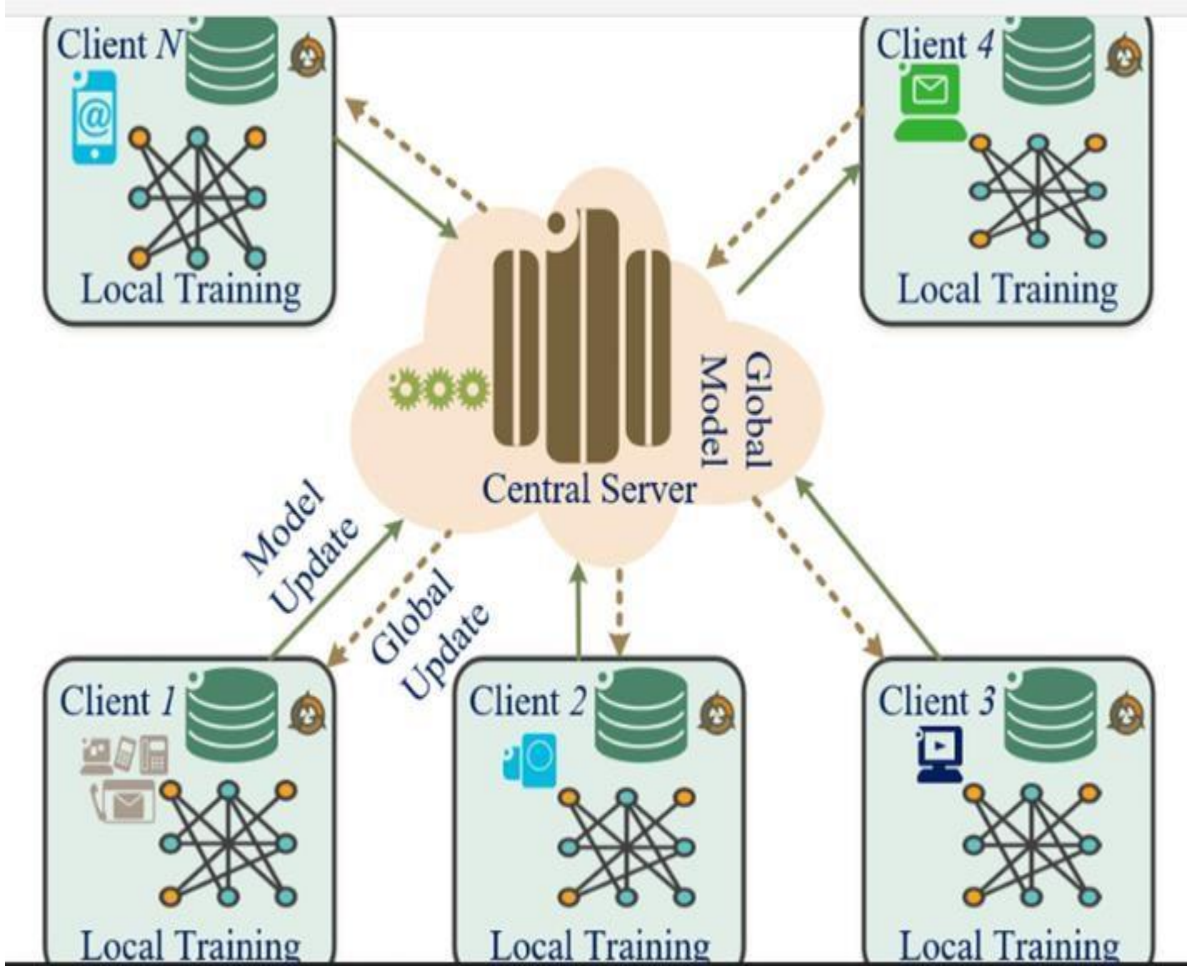
Federated Learning

- It Is **decentralized** form of Machine Learning
- Was developed to address privacy concerns in traditional centralized data processing

- Federated learning works by :

1. Training a central model across decentralized devices or servers
2. Instead of moving all data to a central location, the model is trained locally on each device
3. and only the model updates are shared.

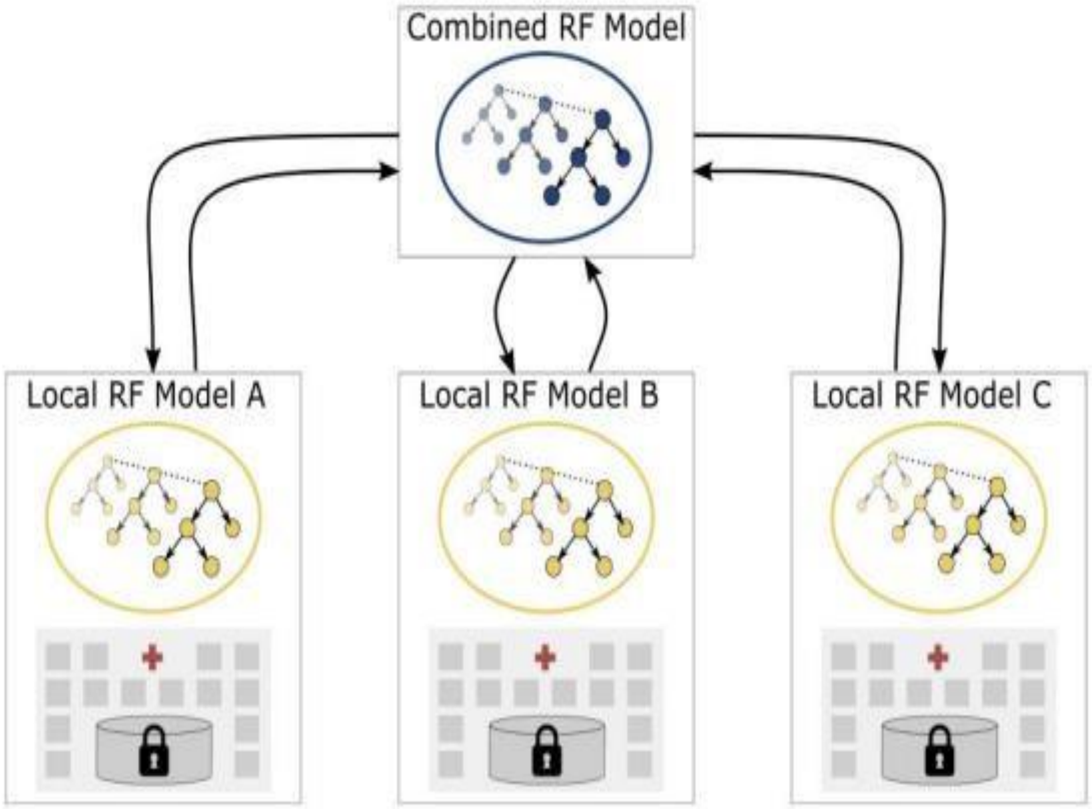
- This method not only protect users privacy but also minimize data transmission cost and speed up the learning process



Federated Learning for Intrusion Detection System: Concepts, Challenges and Future Directions - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/General-Federated-Learning-Architecture_fig1_352505433 [accessed 16 Jan, 2024]

Random Forest

- Random Forest is a popular ML algorithm
- It is an ensemble learning method that builds multiple decision trees and merges their predictions to obtain a more accurate and stable result.
- It stands out for its efficiency in training, it takes less training time as compared to other algorithms.



A.-C. Hauschild et al., "Federated Random Forests can improve local performance of predictive models for various healthcare applications," vol. 38, no. 8, pp. 2278–2286, Feb. 2022, doi: https://doi.org/10.1093/bioinformatics/btac065.

Design Requirements and Constraints

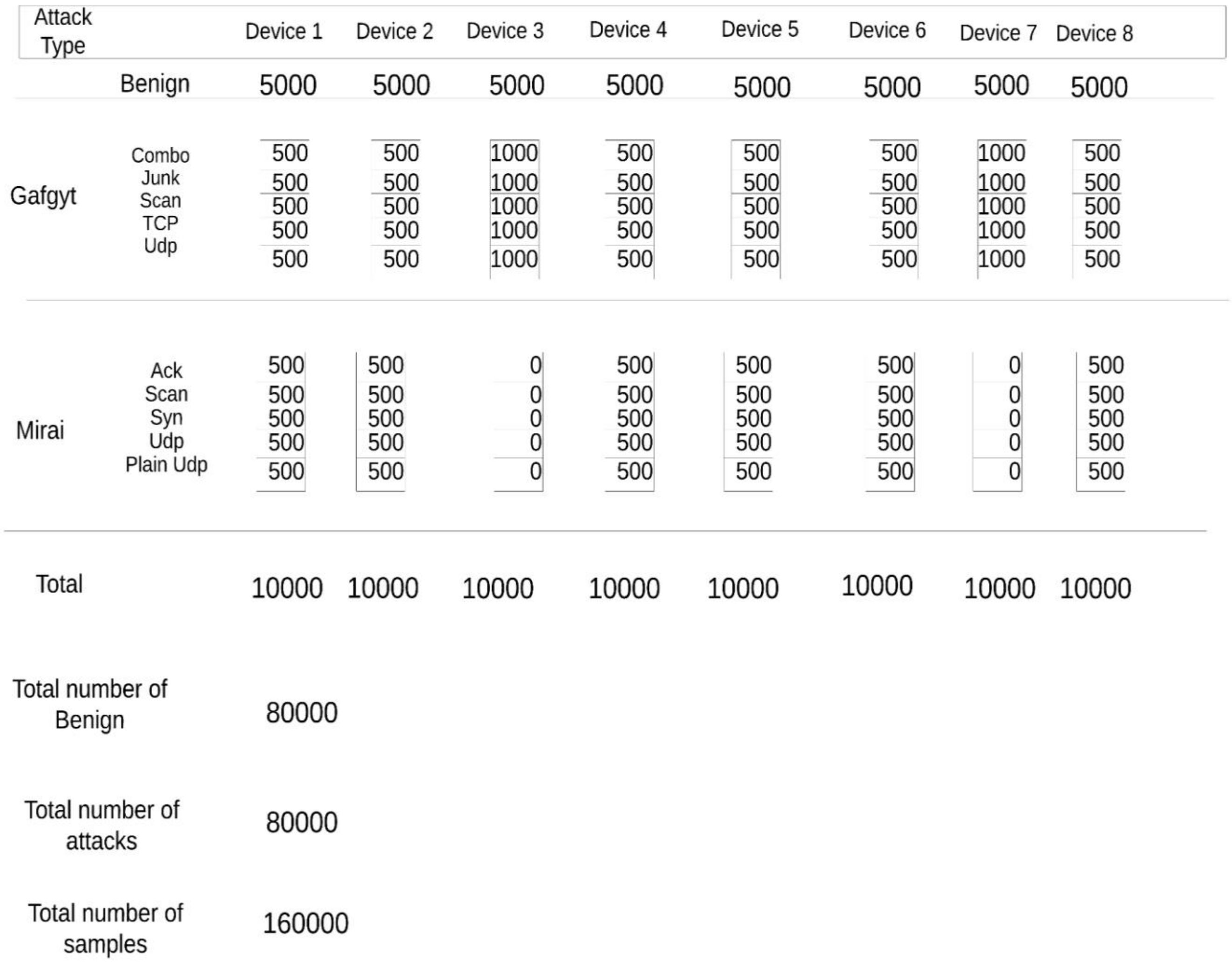
- Works effectively across various IoT devices
- Scalable
- Jupyter Notebook
- N-BaIoT Dataset

N-Balot Dataset

- Data from several devices, each vulnerable to attacks, is combined and analyzed

- This combined dataset is shuffled to randomize the row order to ensure no bias based on the data entry order.

- The final, shuffled dataset is exported a CSV file, combining samples from various attack kinds and devices into a comprehensive
  dataset that we used for our model



Attack Type		Device 1	Device 2	Device 3	Device 4	Device 5	Device 6	Device 7	Device 8
	Benign	5000	5000	5000	5000	5000	5000	5000	5000
Gafgyt	Combo	500	500	1000	500	500	500	1000	500
	Junk	500	500	1000	500	500	500	1000	500
	Scan	500	500	1000	500	500	500	1000	500
	TCP	500	500	1000	500	500	500	1000	500
	Udp	500	500	1000	500	500	500	1000	500
Mirai	Ack	500	500	0	500	500	500	0	500
	Scan	500	500	0	500	500	500	0	500
	Syn	500	500	0	500	500	500	0	500
	Udp	500	500	0	500	500	500	0	500
	Plain Udp	500	500	0	500	500	500	0	500
Total		10000	10000	10000	10000	10000	10000	10000	10000

Total number of Benign	80000
Total number of attacks	80000
Total number of samples	160000

as

Data Visualization

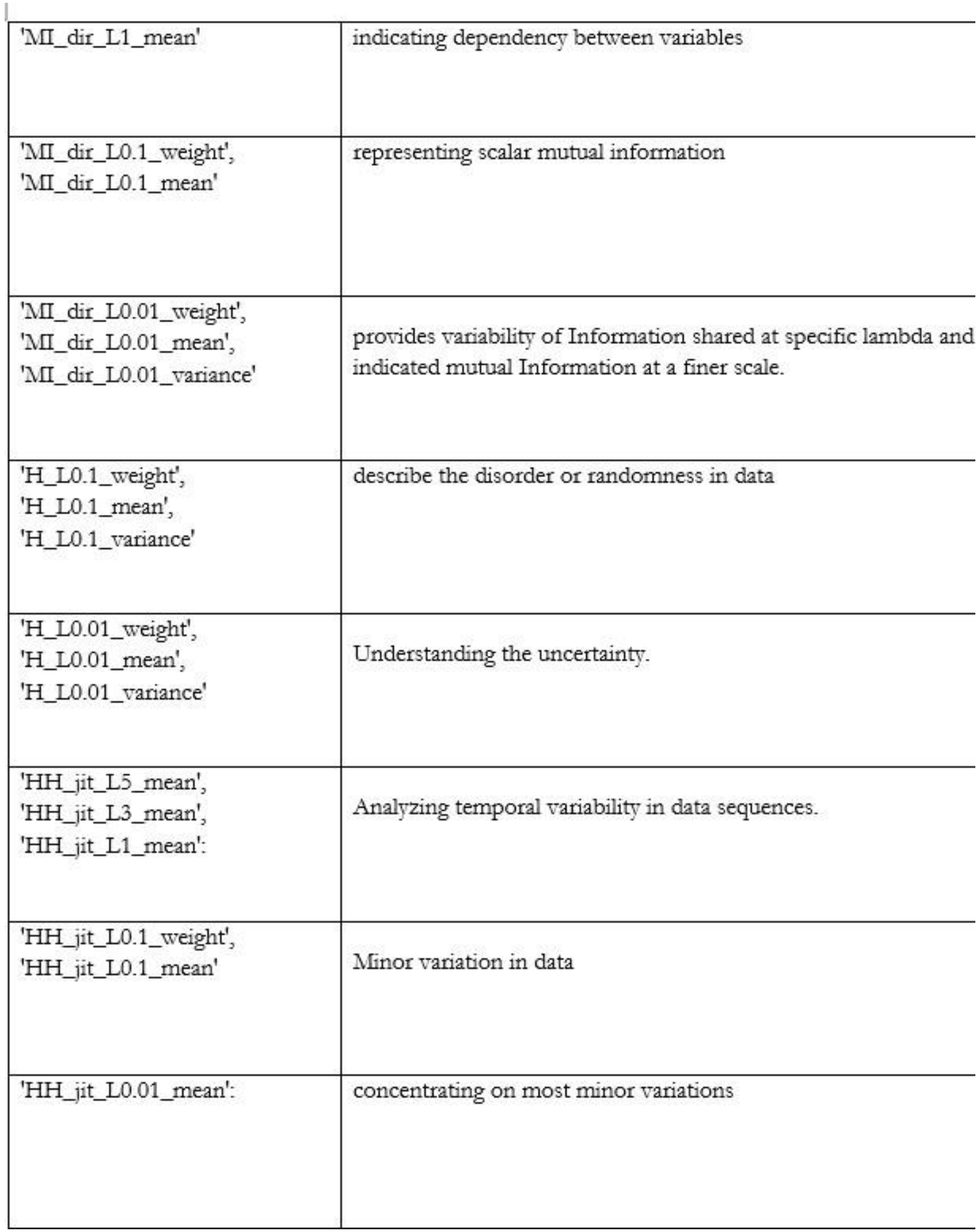
- The **analyze_and_clean_data**function is designed to perform data analysis and cleaning on a list of datasets.
- It iterates through each dataset, identifying missing and infinite values, determining columns to drop or blame, and analyzing rows for missing values proportion. It is crucial to ensure data quality before moving forward with data analysis or ML tasks
- **Analysis for Datasets**
- **Identification of Missing and Infinite Values**
- **Threshold Setting**
- **Column Dropping**
- **Columns for Imputation**
- **Rows to Drop**
- **Detailed Count of Missing and Infinite Values**.

Feature selection

- The feature selection approach uses the *SelectFromModel* class and a *Random Forest Classifier*. The threshold of 0.01 is a crucial criterion for assessing the significance of individual features.

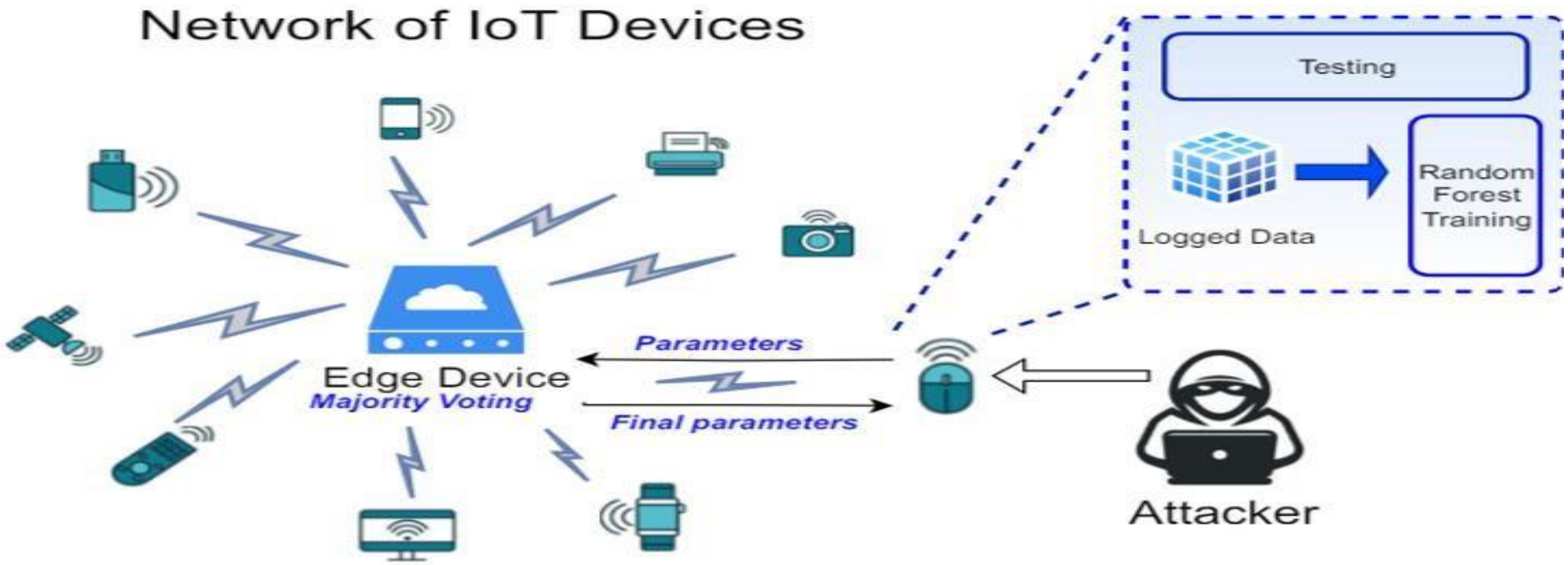
- After that, these essential scores are contrasted with the threshold. To ensure that only the most influential features, as judged by their contribution to model accuracy, are kept, only those with scores that match or surpass the 0.01 criterion are chosen for additional examination.

- The selected features, identified through the process described above, include a variety of metrics primarily focused on mutual information (MI), entropy (H), and jitter (HH_jit) at different scales and conditions.

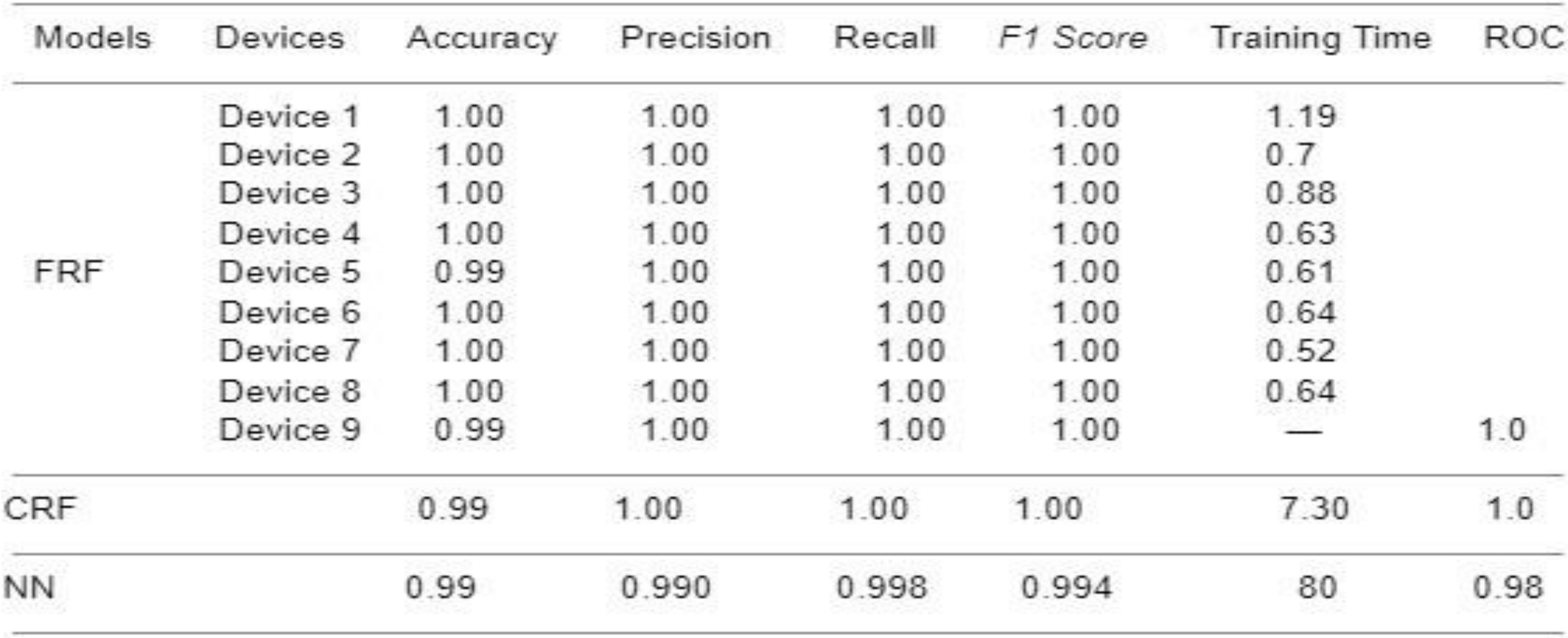


'MI_dir_L1_mean'	indicating dependency between variables
'MI_dir_L0.1_weight', 'MI_dir_L0.1_mean'	representing scalar mutual information
'MI_dir_L0.01_weight', 'MI_dir_L0.01_mean', 'MI_dir_L0.01_variance'	provides variability of Information shared at specific lambda and indicated mutual Information at a finer scale.
'H_L0.1_weight', 'H_L0.1_mean', 'H_L0.1_variance'	describe the disorder or randomness in data
'H_L0.01_weight', 'H_L0.01_mean', 'H_L0.01_variance'	Understanding the uncertainty.
'HH_jit_L5_mean', 'HH_jit_L3_mean', 'HH_jit_L1_mean':	Analyzing temporal variability in data sequences.
'HH_jit_L0.1_weight', 'HH_jit_L0.1_mean'	Minor variation in data
'HH_jit_L0.01_mean':	concentrating on most minor variations

System Architecture and Implementation Details



Results



Models	Devices	Accuracy	Precision	Recall	*F1 Score*	Training Time	ROC
FRF	Device 1	1.00	1.00	1.00	1.00	1.19	
	Device 2	1.00	1.00	1.00	1.00	0.7	
	Device 3	1.00	1.00	1.00	1.00	0.88	
	Device 4	1.00	1.00	1.00	1.00	0.63	
	Device 5	0.99	1.00	1.00	1.00	0.61	
	Device 6	1.00	1.00	1.00	1.00	0.64	
	Device 7	1.00	1.00	1.00	1.00	0.52	
	Device 8	1.00	1.00	1.00	1.00	0.64	
	Device 9	0.99	1.00	1.00	1.00	—	1.0
CRF		0.99	1.00	1.00	1.00	7.30	1.0
NN		0.99	0.990	0.998	0.994	80	0.98

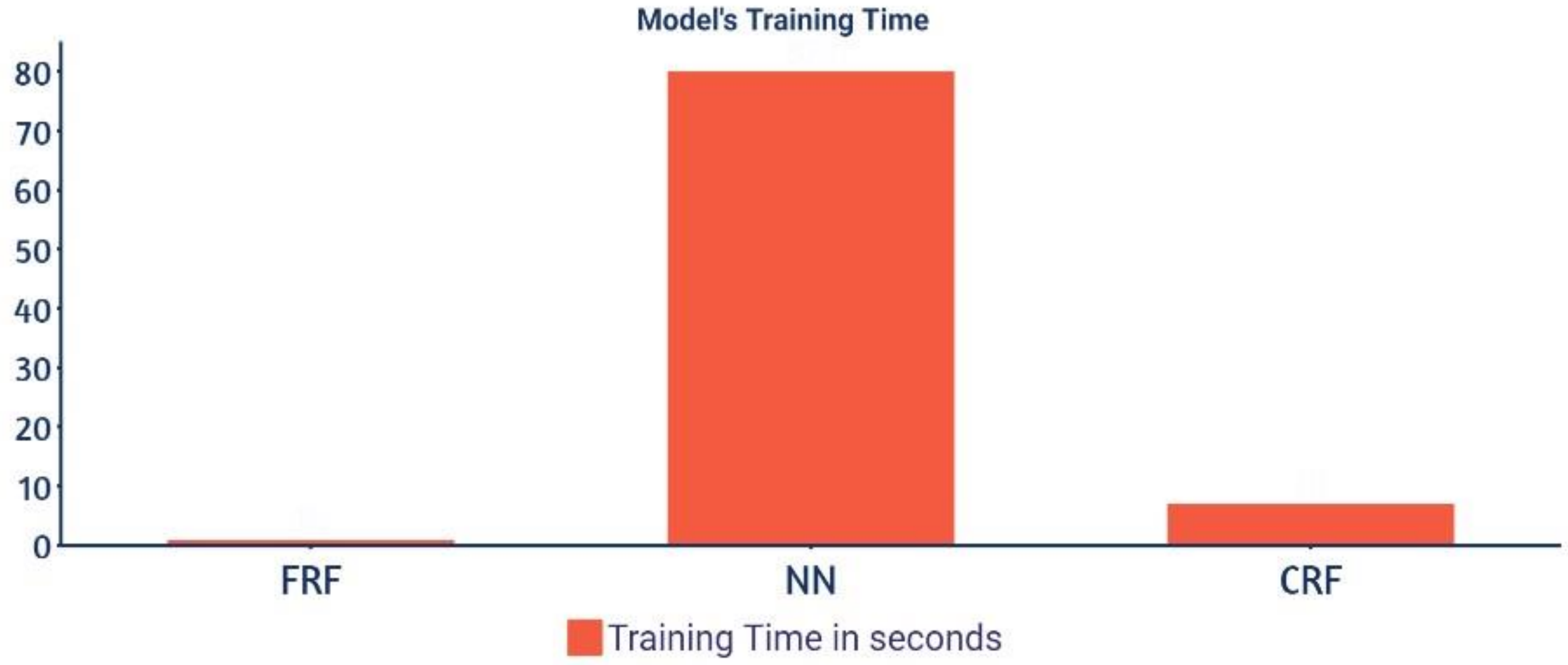
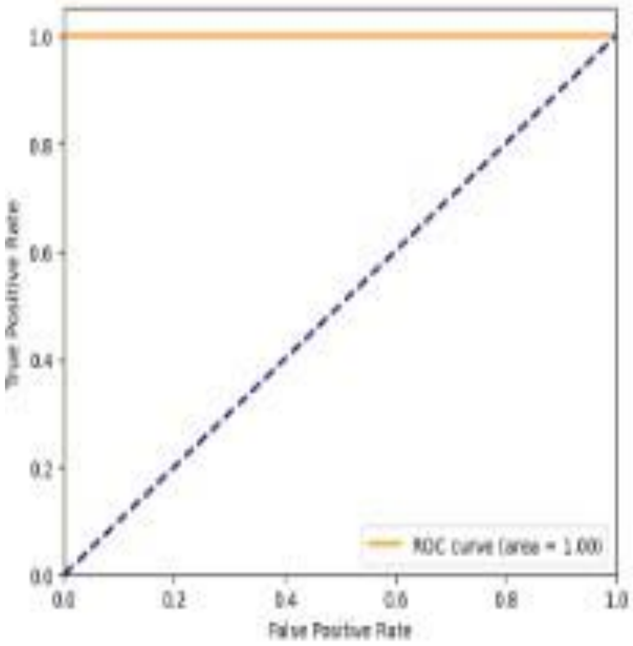
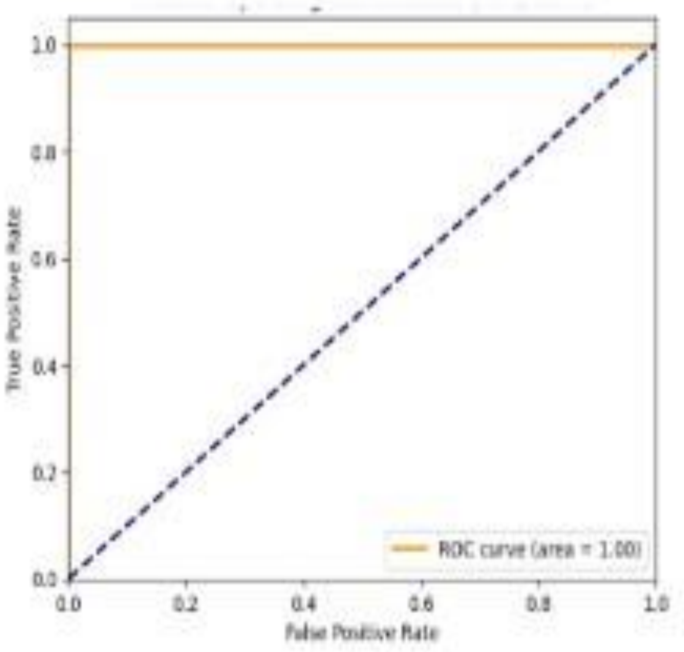


Figure 9: Model Training Time

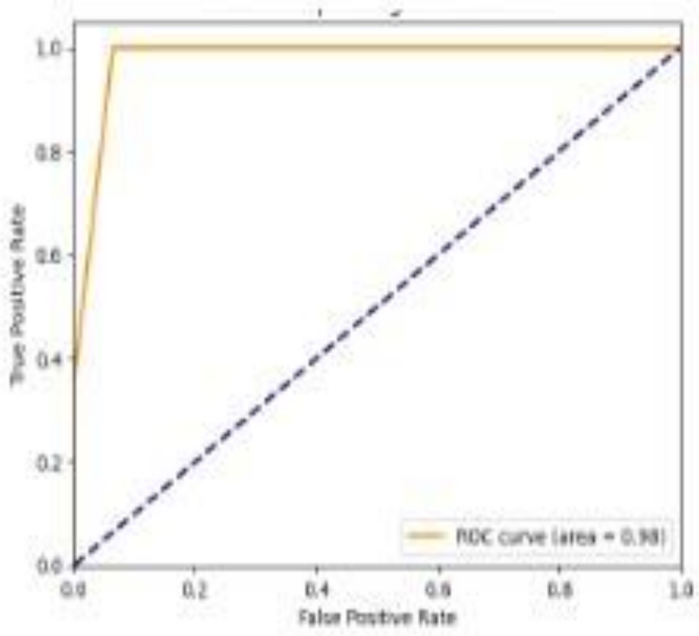
ROC



(a) FRF Model

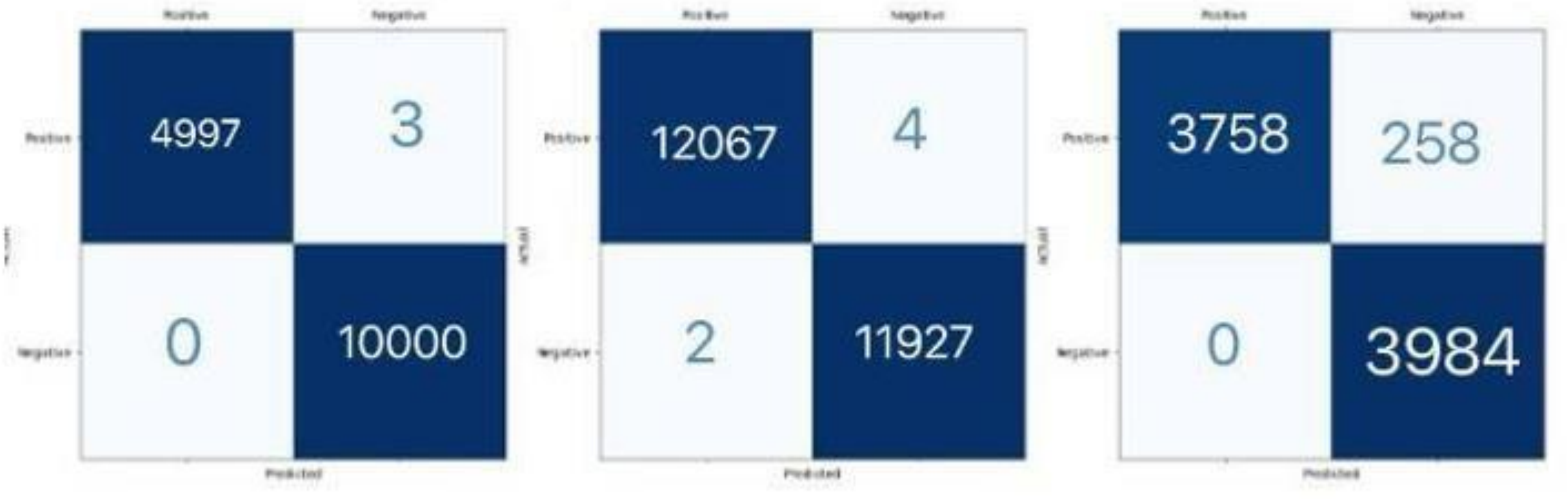


(b) CRF Model



(c)NN Model

Confusion Matrix



	Positive	Negative
Positive	4997	3
Negative	0	10000

(a)FRF Confusion Matrix

	Positive	Negative
Positive	12067	4
Negative	2	11927

(b)CRF Confusion Matrix

	Positive	Negative
Positive	3758	258
Negative	0	3984

(c)NN Confusion Matrix

Challenges and Solutions

- Challenges faced during the work in this project, In managing data privacy while maintaining high detection accuracy.

- One major challenge was ensuring that the Federated Learning model could effectively combine insights from multiple devices without accessing the actual data, to protect user privacy.
- To address this, we refined the data aggregation techniques, ensuring that only model updates were shared, not the raw data

Future Directions

1. Extend our model to detect other types of cybersecurity threats in IoT systems

2.Explore advanced machine learning techniques to boost the accuracy and efficiency of our models further

3.Enhance the scalability of our Federated Learning model to support larger networks of IoT devices

Conclusion

- In conclusion, our project has successfully demonstrated that using machine learning, specifically Federated Random Forest, can significantly enhance the detection of botnet attacks in IoT devices while also protecting user privacy.
- Our models, especially the Federated Random Forest, showed high accuracy and efficiency in detecting threats.
- Going forward, this work lays a strong foundation for further research and development in the field of IoT security, offering a promising path towards safer and more reliable IoT ecosystems.

Princess Sumaya جامعـــة
University الأميــرة سميّــة
for Technology للتكنولوجيا

Questions?