ARES conference

Benevento, Italy
August 29 - September 01, 2023

MINDS

Breaching the Defense: Investigating FGSM and CTGAN Adversarial Attacks on IEC 60870-5-104 AI-enabled Intrusion Detection Systems

D. Asimopoulos, P. Radoglou-Grammatikis, Ioannis Makris, Valeri Mladevov, Konstantinos E. Psannis, Sotirios Goudos and Panagiotis Sarigiannidis

MetaMind Innovations P.C, Greece

AI4CYBER – Artificial Intelligence for next generation CYBERsecurity

Funded By the European Union

Authors & Contributors

MINDS

Dimitrios-Christos Asimopoulos
Ioannis Makris

UNIVERSITY OF WESTERN MACEDONIA

Panagiotis Radoglou Grammatikis
Panagiotis Sarigiannidis

ARISTOTLE UNIVERSITY OF THESSALONIKI

Sotirios Goudos

UNIVERSITY OF MACEDONIA

Konstantinos E. Psannis

Technical University of Sofia

Valeri Mladenov

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101070450 (AI4CYBER) .

Presentation Structure

Introduction

1

Introduction

Related Work

Contributions

Main Part

2

Proposed Intrusion Detection System

FGSM and CTGAN Attack WorkFlow

FGSM Attack and Evaluation

CTGAN Attack and Evaluation

Conclusions

3

Sum up important key points

Future Work

Q/A

Introduction, Relevant Work & Contributions

Introduction

Technology Enablers for Critical Infrastructures: IoT, 5G, and AI empower Critical Infrastructures like the smart grid, fostering sustainability, efficiency, and real-time control.

Cybersecurity Challenges and Consequences: The integration of advanced technologies exposes Critical Infrastructures to multi-step attacks, potentially causing extensive outages, financial losses, and safety hazards.

AI's Role in Cybersecurity and Detection: AI-driven detection systems offer adaptive defense mechanisms, enabling the identification of novel threats and anomalies while enhancing response capabilities.

Challenges in AI-Powered Security: AI-powered security solutions face challenges such as false alarms and susceptibility to adversarial attacks, requiring ongoing refinement to maintain accuracy and reliability.

Related Work

2019

Adel Alshamrani et al.

- A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities. IEEE Communications Surveys & Tutorials

2019

Panagiotis I Radoglou Grammatikis, Panagiotis G Sarigiannidis, and Ioannis D Moscholios

- Securing the Internet of Things: Challenges, threats and solutions. Internet of Things

2021

Panagiotis Radoglou-Grammatikis et al.

- Modeling, detecting, and mitigating threats against industrial healthcare systems: a combined software defined networking and reinforcement learning approach. IEEE Transactions on Industrial Informatics

2023

Ke He, Dan Dongseong Kim, and Muhammad Rizwan Asghar

- Adversarial Machine Learning for Network Intrusion Detection Systems: A Comprehensive Survey. IEEE Communications Surveys & Tutorials

Contributions

AI-Powered Intrusion Detection System (IDS) against IEC 60870-5-104 Attacks:

- An AI-powered IDS is provided in terms of detecting and mitigating various cyberattacks against the IEC 60870-5-104 protocol. For this purpose, four ML/DL models (Decision Tree, XGBOOST, Random Forest and MLP) are used and compared with each other.

Investigating FGSM Adversarial Attacks:

- We investigate how FGSM evasion adversarial attacks can affect the detection performance of the previous ML/DL models.

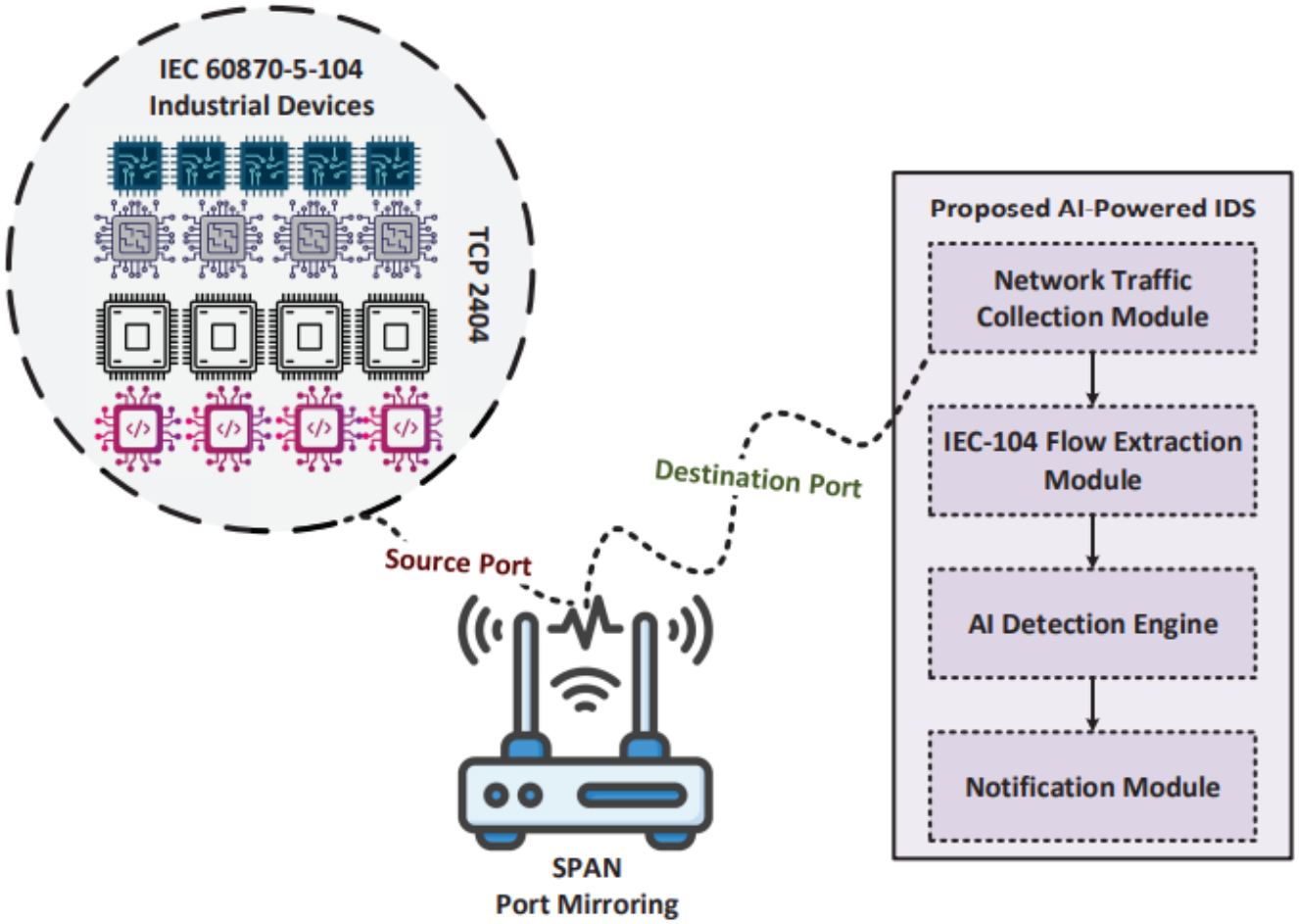
Development of CTGAN Adversarial Attack Generator:

- We implement an adversarial attack generator that takes full advantage of FGSM and CTGAN

Proposed Intrusion Detection System (IDS)

AI Detection Engine

- Decision Tree
- Random Forest
- XGBOOST
- Custom Multilayer Perceptron



Training Dataset

IEC 60870-5-104 Intrusion Detection Dataset

IEEE: [IEC 60870-5-104 Intrusion Detection Dataset]

Zenodo: [IEC 60870-5-104 Intrusion Detection Dataset]

Includes flow statistics related to the following IEC 60870-5-104 attacks:

MITM

traffic sniffing

C_RD_NA_1

C_CI_NA_1

C_RP_NA_1

C_SE_NA_1

M_SP_NA_1_DOS

C_CI_NA_1_DOS

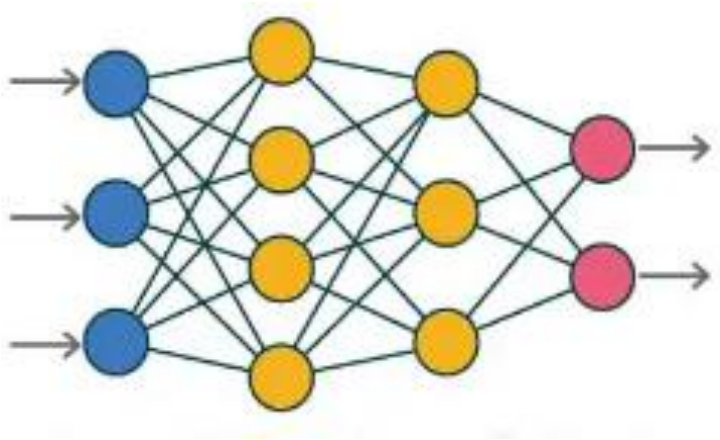
C_SE_NA_1_DOS

C_RD_NA_1_DOS

C_RP_NA_1_DOS

Custom Multilayer Perceptron Model

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 128)	3328
layer_1 (Dense)	(None, 256)	33024
layer_2 (Dense)	(None, 256)	65792
layer_3 (Dense)	(None, 256)	65792
layer_4 (Dense)	(None, 256)	65792
layer_5 (Dense)	(None, 128)	32896
layer_6 (Dense)	(None, 128)	16512
layer_7 (Dense)	(None, 128)	16512
layer_8 (Dense)	(None, 64)	8256
layer_9 (Dense)	(None, 64)	4160
layer_10 (Dense)	(None, 64)	4160
layer_11 (Dense)	(None, 12)	780



Evaluation Metrics

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate

$$FPR = \frac{FP}{FP + FN}$$

F1 Score

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN}$$

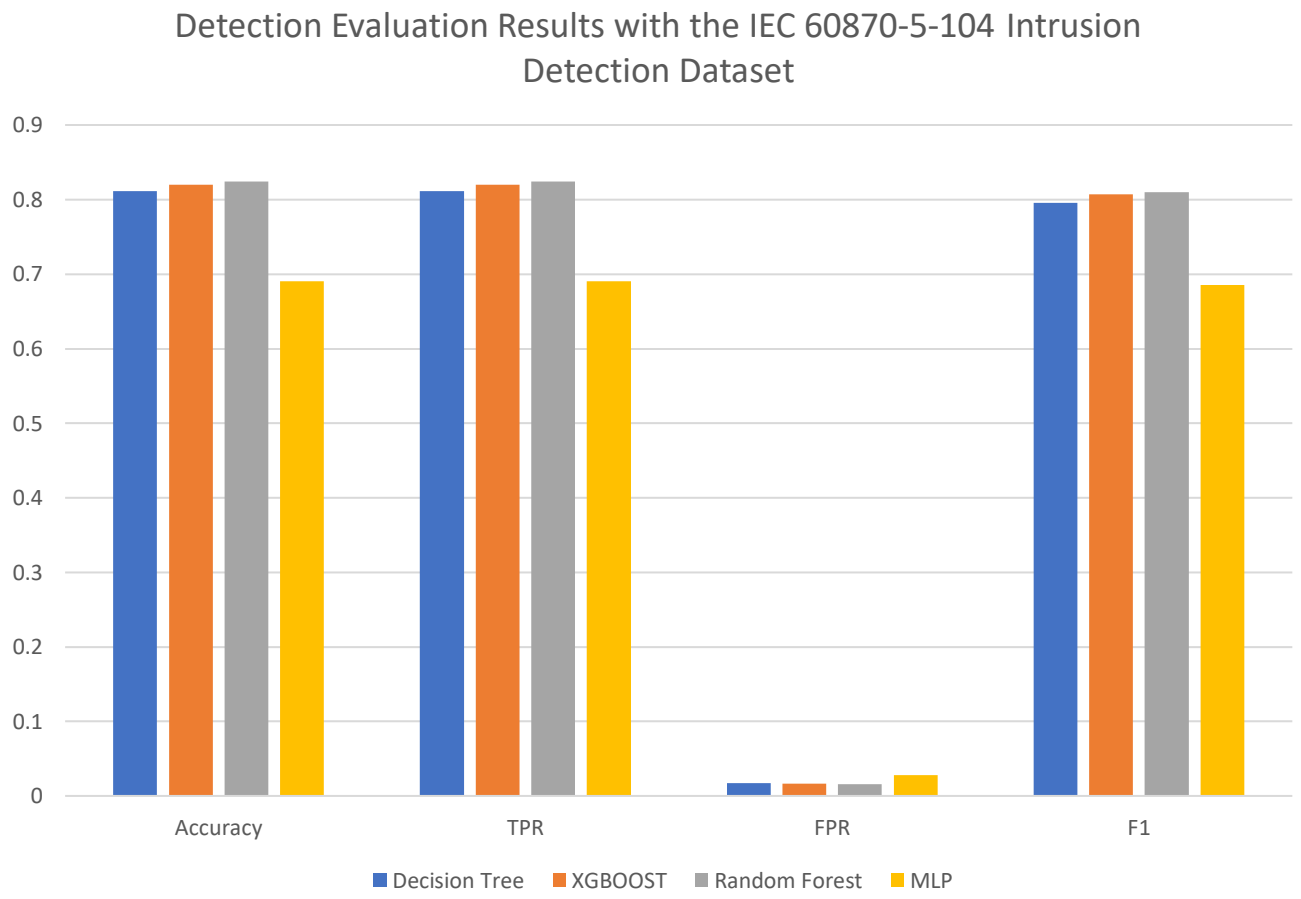
$TP \rightarrow$ True Positives

$TN \rightarrow$ True Negatives

$FP \rightarrow$ False Positives

$FN \rightarrow$ False Negatives

Model Evaluation Results

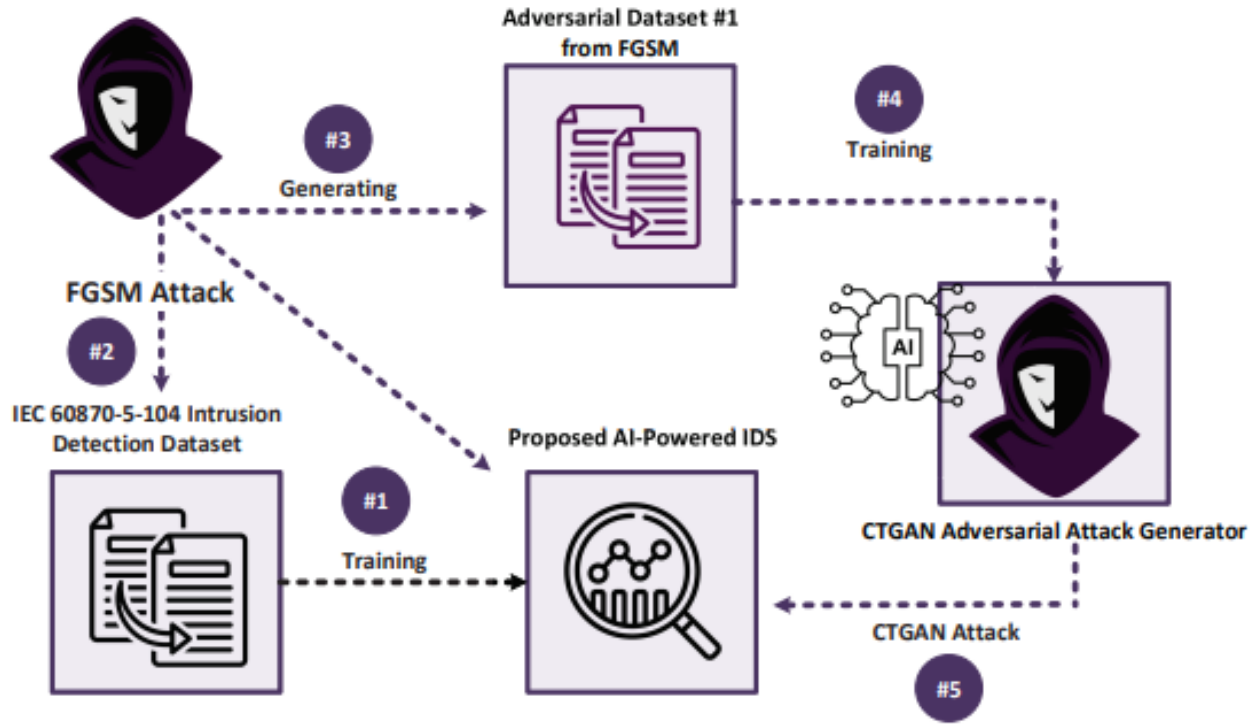


The best results are achieved by Random Forest where:

$Accuracy$ = 0.8244,
$T\ PR$ = 0.8244,
FPR = 0.0159 and
$F\ 1$ = 0.8098

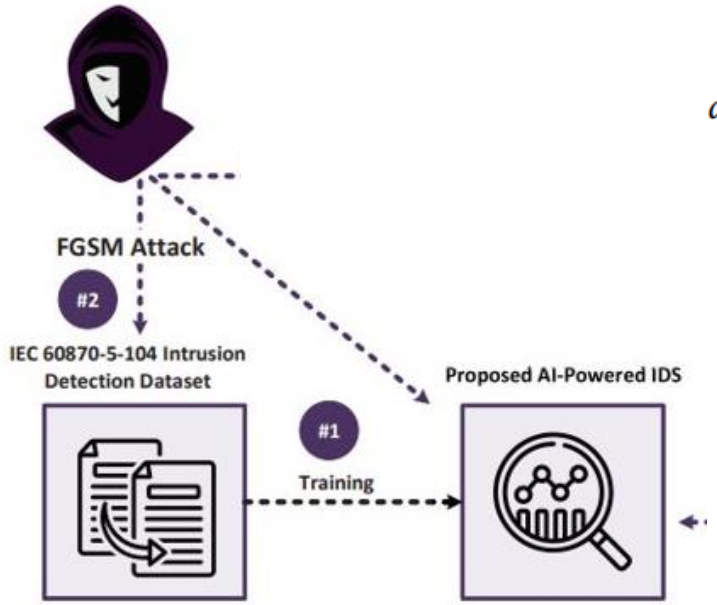
FGSM and CTGAN Attack WorkFlow

WorkFlow Scheme



Fast Gradient Sign Method (FGSM) Attack

FGSM ADVERSARIAL ATTACKS



$$adv\_x = x + \epsilon \cdot \text{sign}\left(\nabla_x J(\theta, x, y)\right)$$

adv_x → Adversarial data

x → Original data

y → Original input label

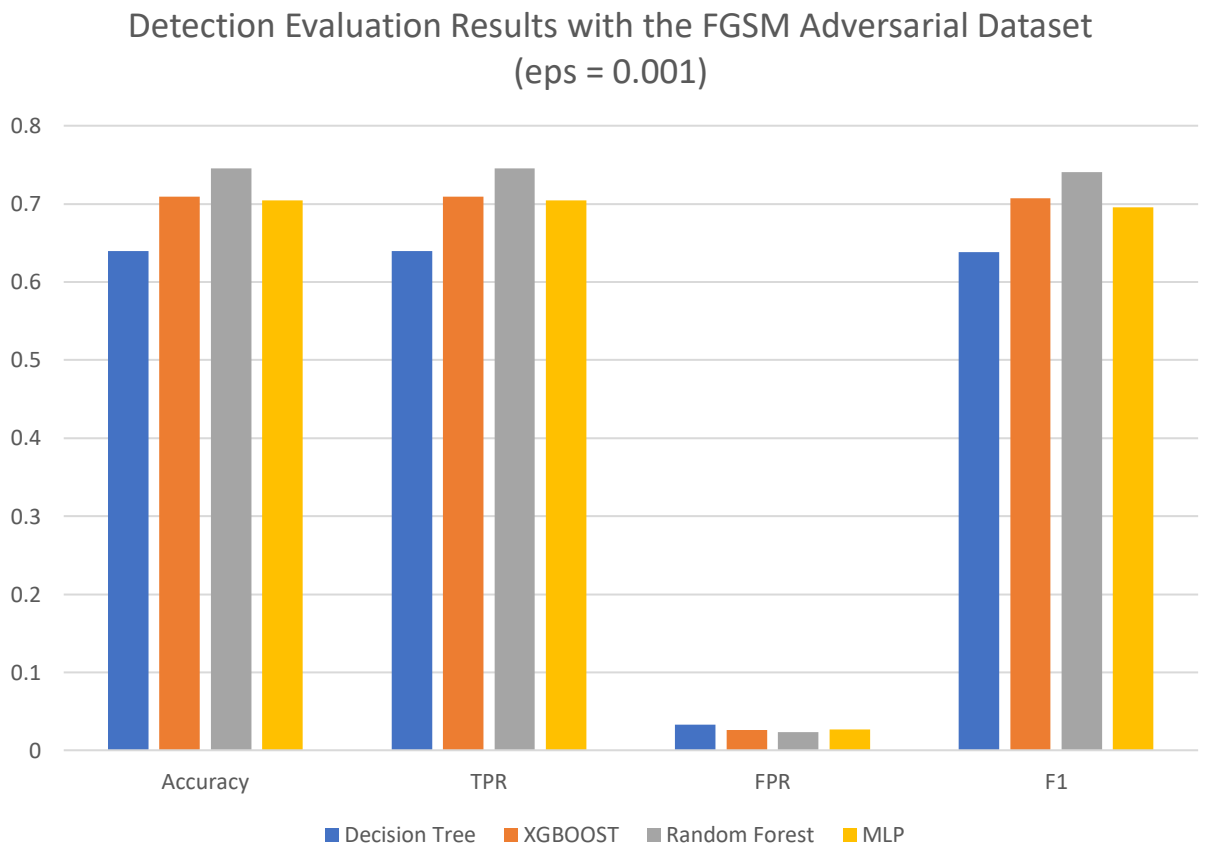
ϵ → Multiplier to ensure the perturbations are small

θ → Model parameters

J → Loss function. In our case, the CrossEntropy function

Three FGSM Datasets were created using epsilon = 0.001, 0.003, 0.01

FGSM Evaluation results



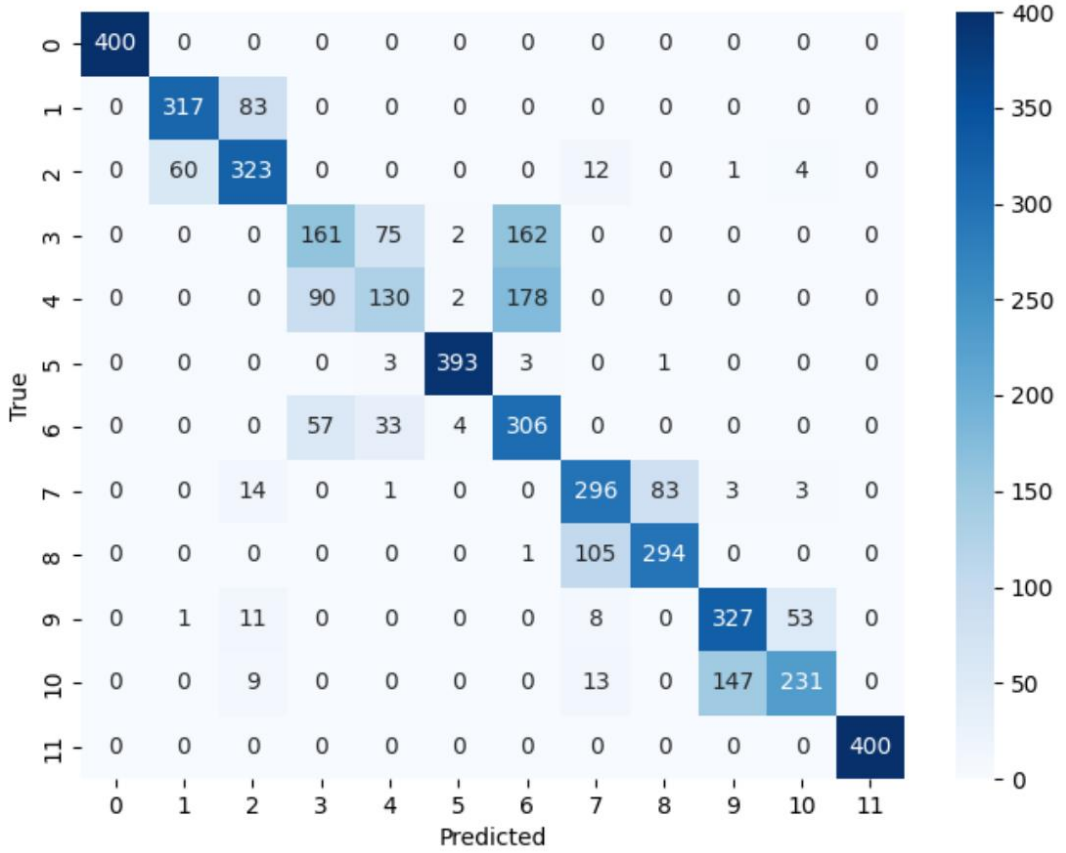
ML/DL Method	Accuracy	TPR	FPR	F1
Decision Tree	0.6395	0.6395	0.0327	0.6382
XGBOOST	0.7095	0.7095	0.0264	0.7072
Random Forest	0.7454	0.7454	0.0231	0.7405
MLP	0.7047	0.7047	0.0268	0.6955

The best performance is achieved by Random Forest where:

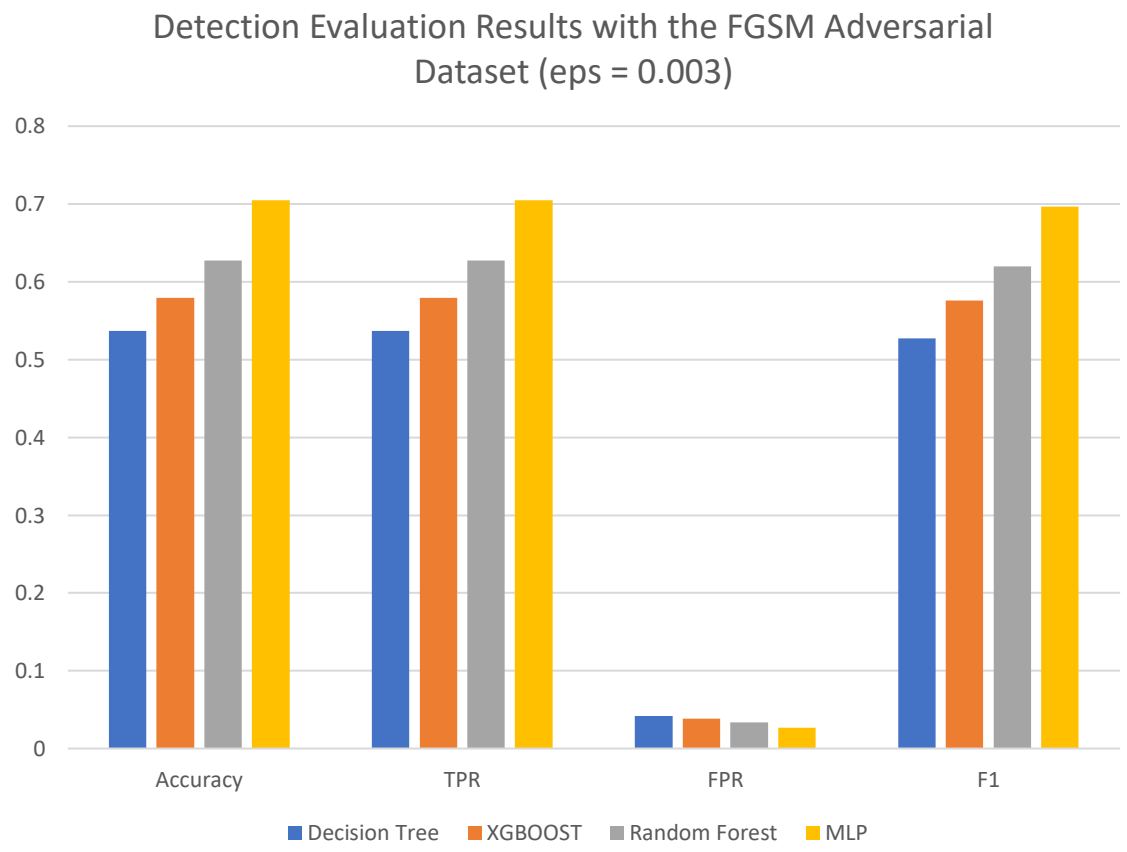
Accuracy = 0.7454,
TPR = 0.7454,
FPR = 0.0231 and
F1 = 0.7405

FGSM Evaluation results

Confusion Matrix of Random Forest with the FGSM Adversarial Dataset (eps = 0.001)



FGSM Evaluation results

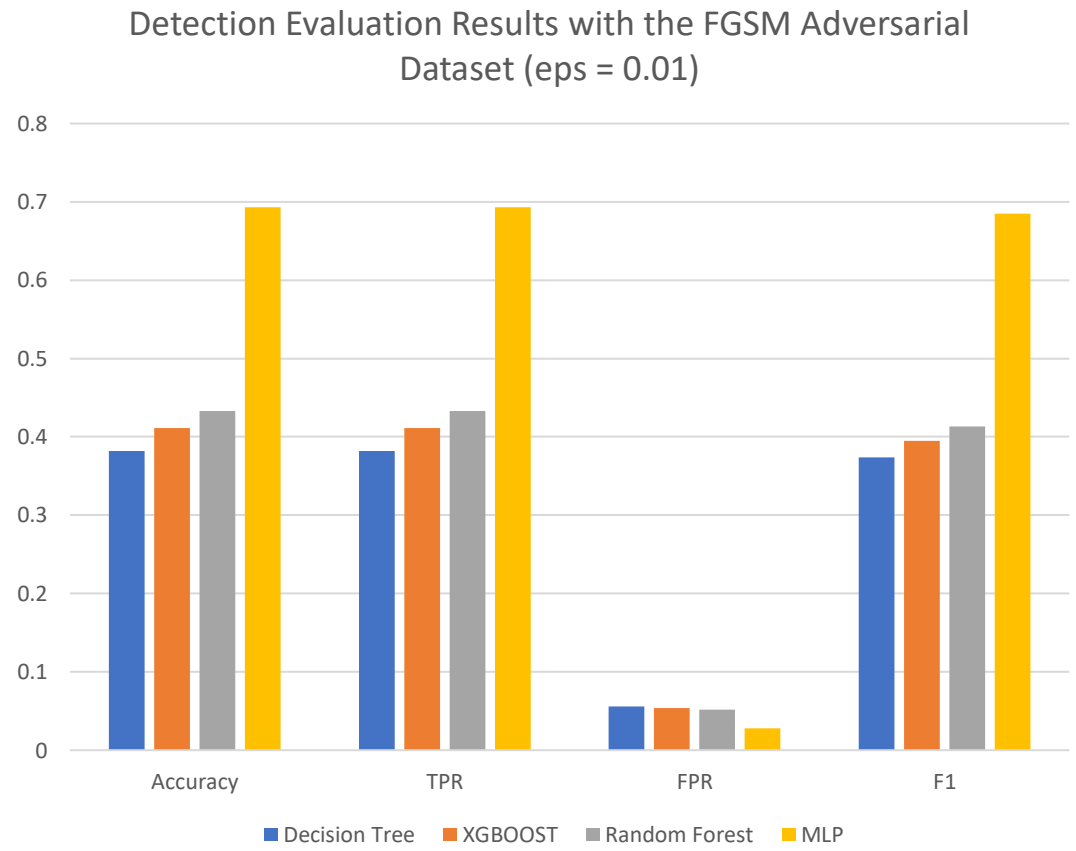


ML/DL Method	Accuracy	TPR	FPR	F1
Decision Tree	0.5370	0.5370	0.0420	0.5273
XGBOOST	0.5797	0.5797	0.0382	0.5762
Random Forest	0.6272	0.6272	0.0338	0.6202
MLP	0.7052	0.7052	0.0267	0.6966

The best performance is achieved by MLP where:

Accuracy = 0.7052,
TPR = 0.7052,
FPR = 0.0267 and
F1 = 0.6966.

FGSM Evaluation results



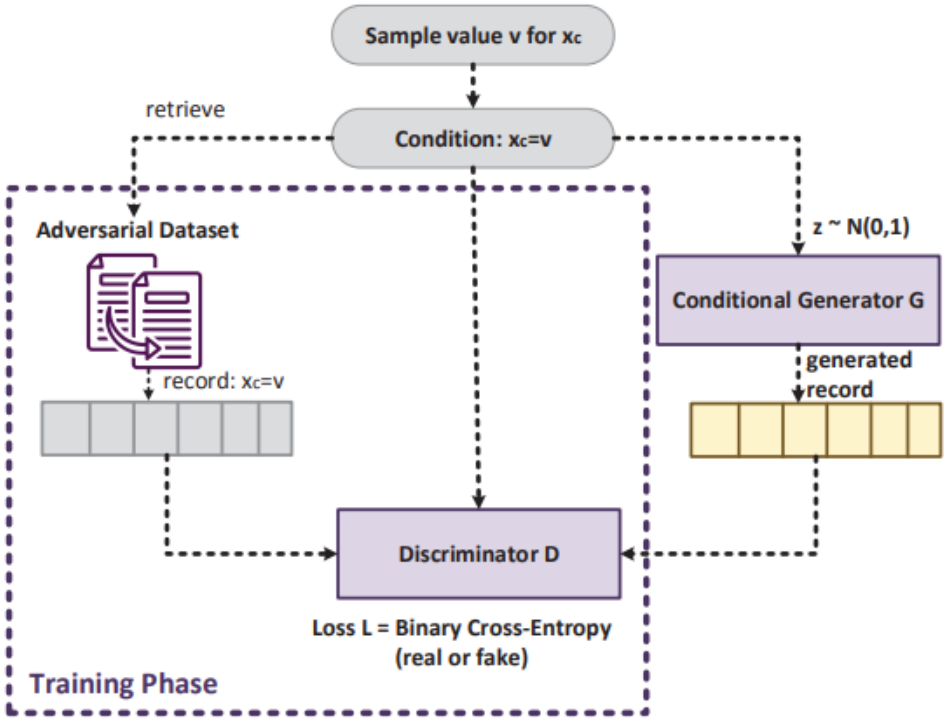
ML/DL Method	Accuracy	TPR	FPR	F1
Decision Tree	0.3818	0.3818	0.0561	0.3738
XGBOOST	0.4110	0.4110	0.0535	0.3947
Random Forest	0.4331	0.4331	0.0515	0.4133
MLP	0.6933	0.6933	0.0278	0.6851

The best performance is achieved by MLP where:

Accuracy = 0.6933,
TPR = 0.6933,
FPR = 0.0278 and
F1 = 0.6851

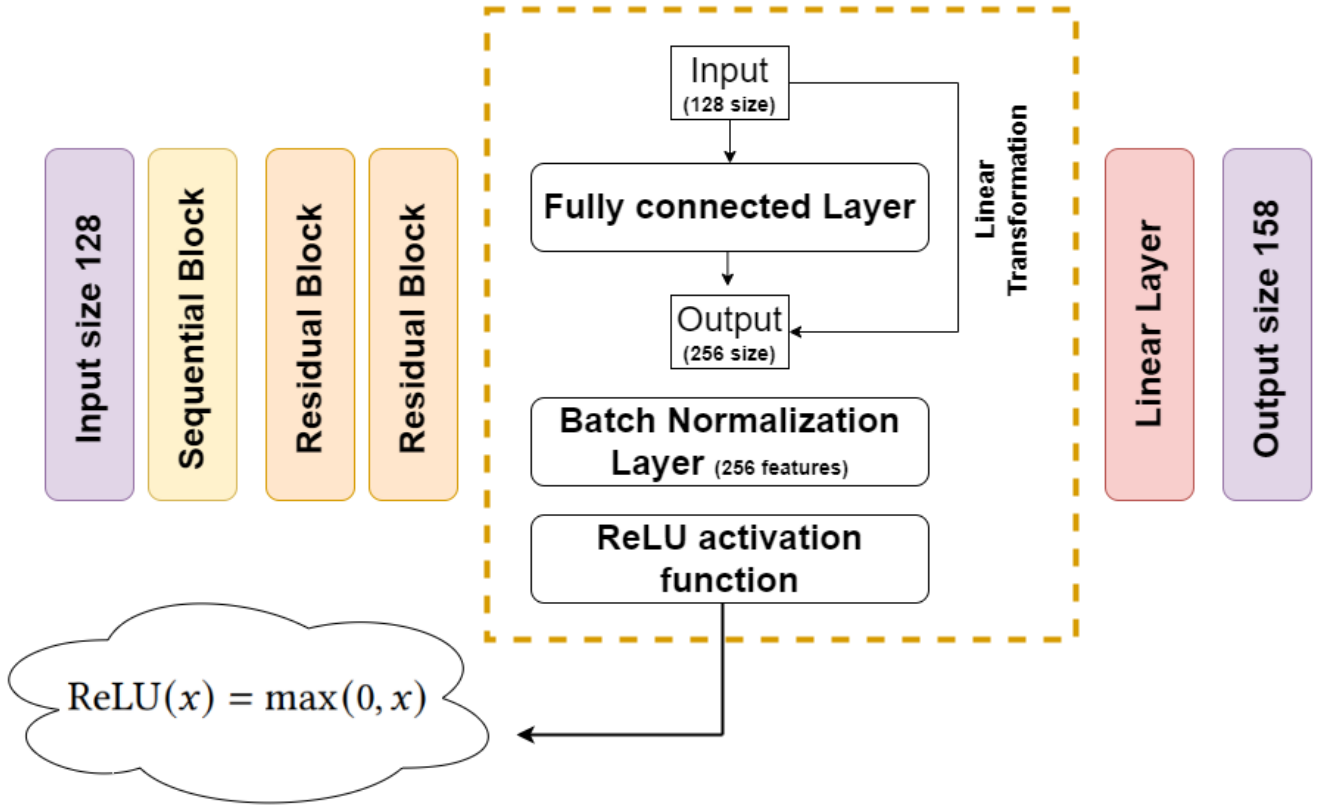
Conditional Tabular GAN (CTGAN) Attack

CTGAN Attacks



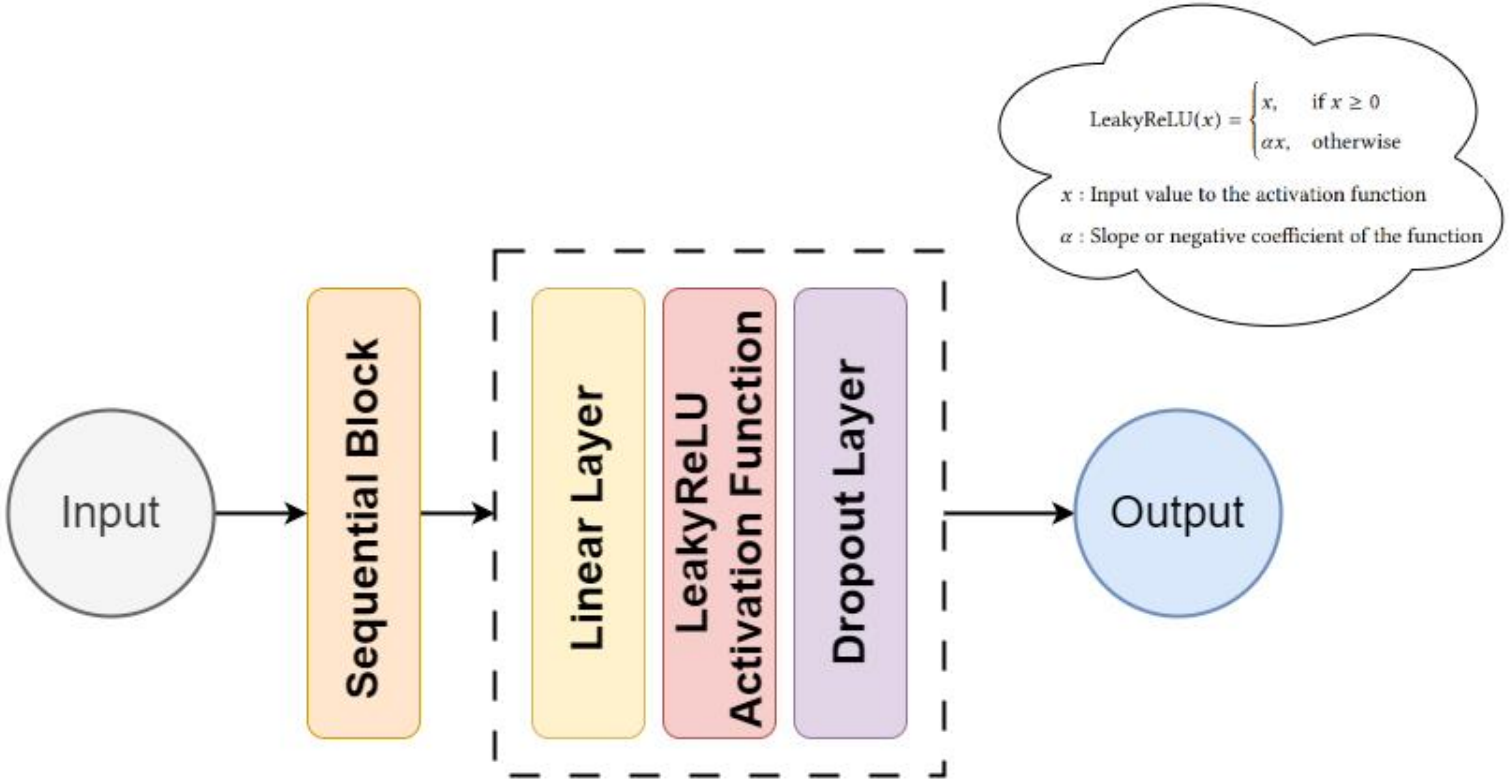
Used the FGSM evaded dataset to train the CTGAN model.

CTGAN Generator



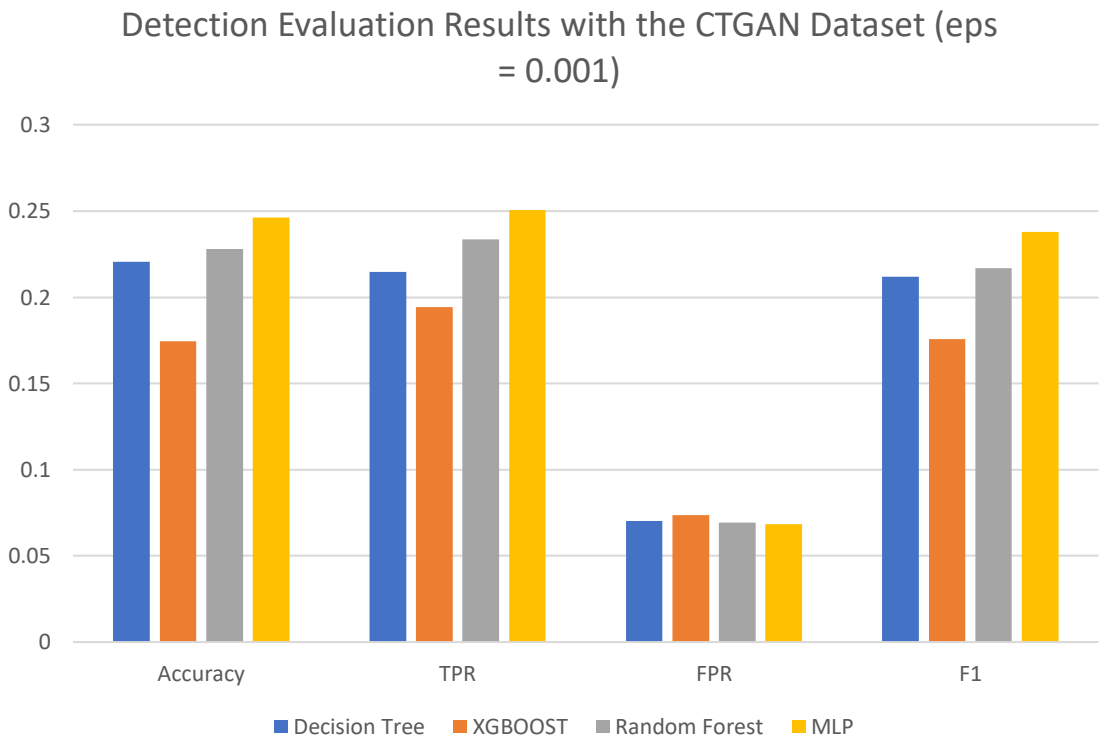
The generator model takes an input and applies two Residual blocks with batch normalisation and ReLU activation, and then feeds the output into a final Linear layer.

CTGAN Discriminator



The discriminator architecture consists of linear layers connected by LeakyReLU activations and dropout layers.

CTGAN Evaluation results



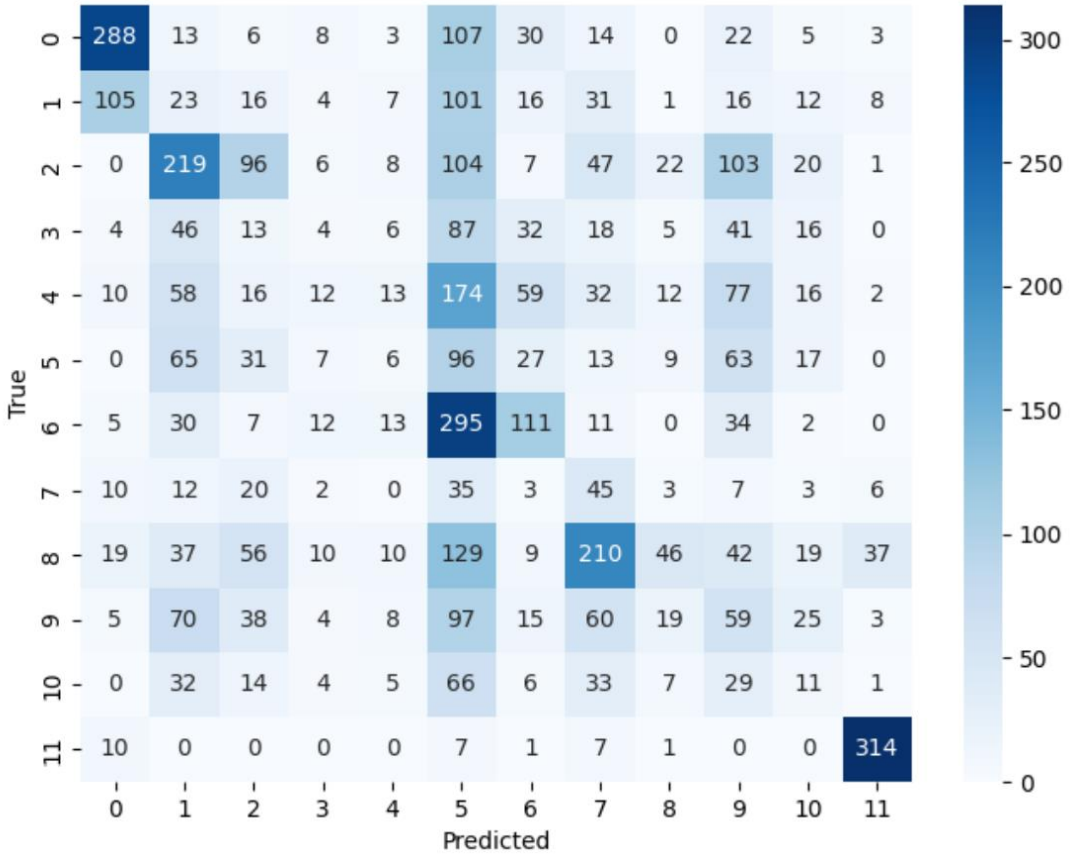
ML/DL Method	Accuracy	TPR	FPR	F1
Decision Tree	0.2206	0.2148	0.0702	0.2119
XGBOOST	0.1745	0.1942	0.0735	0.1756
Random Forest	0.2281	0.2337	0.0693	0.2169
MLP	0.2464	0.2505	0.0683	0.2378

MLP achieves the best performance with:

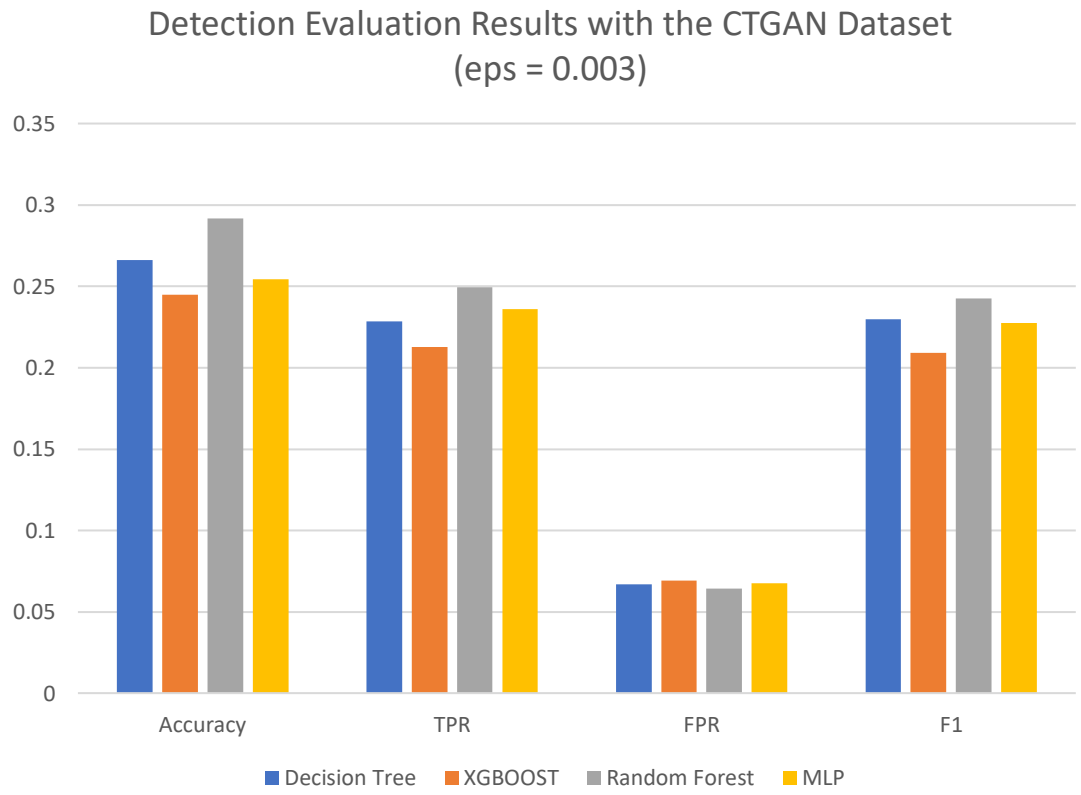
Accuracy = 0.2464,
TPR = 0.2361,
FPR = 0.0683, and
F1 = 0.2378

CTGAN Evaluation results

Confusion Matrix of Random Forest with the CTGAN Dataset (eps = 0.001)



CTGAN Evaluation results

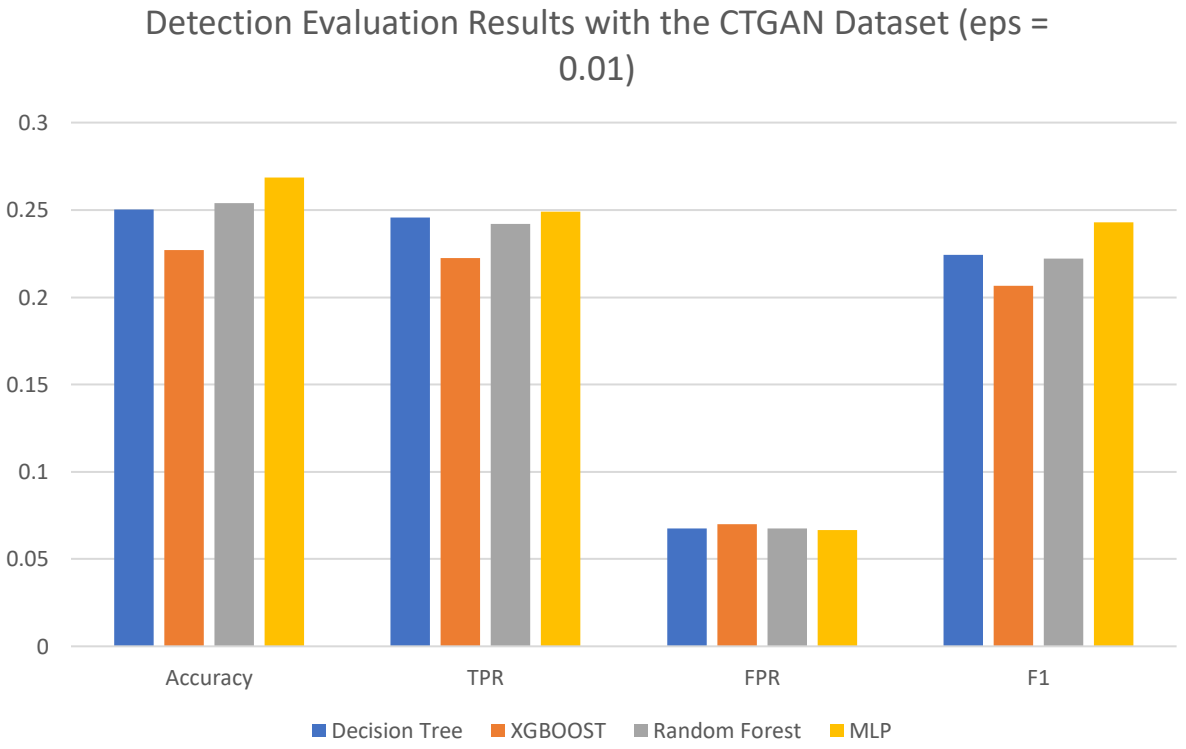


ML/DL Method	Accuracy	TPR	FPR	F1
Decision Tree	0.2660	0.2285	0.0671	0.2299
XGBOOST	0.2447	0.2127	0.0691	0.2093
Random Forest	0.2918	0.2495	0.0642	0.2425
MLP	0.2545	0.2361	0.0676	0.2275

Random Forest achieves the best performance with:

Accuracy = 0.2918,
TPR = 0.2495,
FPR = 0.0642, and
F1 = 0.2425

CTGAN Evaluation results



ML/DL Method	Accuracy	TPR	FPR	F1
Decision Tree	0.2502	0.2457	0.0676	0.2243
XGBOOST	0.2272	0.2225	0.0699	0.2066
Random Forest	0.2541	0.2422	0.0676	0.2223
MLP	0.2687	0.2490	0.0665	0.2429

MLP achieves the best performance with:

Accuracy = 0.2687,
TPR = 0.2490,
FPR = 0.0665, and
F1 = 0.2429

Conclusions & Future Work

Conclusion

An AI-powered IDPS was implemented for the IEC 60870-5-104 protocol, utilising four ML/DL methods: Decision Tree, Random Forest, XGBOOST and MLP.

The FGSM method was used in order to evaluate the resilience of the previous methods under the conditions of a typical adversarial attack.

Given the FGSM adversarial datasets, a CTGAN adversarial attack generator was implemented.

Future work

In general, the performance of the tested models (Decision Tree, XGBOOST, Random Forest, and MLP) is better on the FGSM adversarial datasets when compared to the CTGAN datasets. However, the difference in performance is less distinct for the MLP model, particularly at higher epsilon levels. This observation suggests that the MLP method might be more resilient to noise or better equipped to handle the specific characteristics or distribution of the CTGAN datasets.

Future work will investigate more complicated adversarial attacks and evaluate relevant countermeasures in order to strengthen and optimise the resilience of the AI security models.

MINDS

UNIVERSITY OF WESTERN MACEDONIA

Thank you for your attention!

ARES conference