

## Integrated Project Cyber security of energy systems for the digital-energy transition

*A method for smart grid intrusion detection through explainable deep learning*  
Giovanni Ciaramella - CNR-IIT

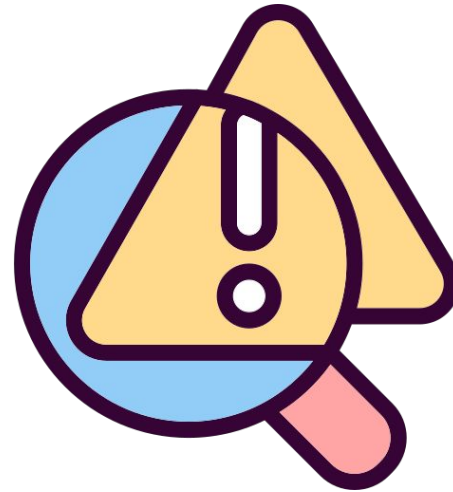


# 1

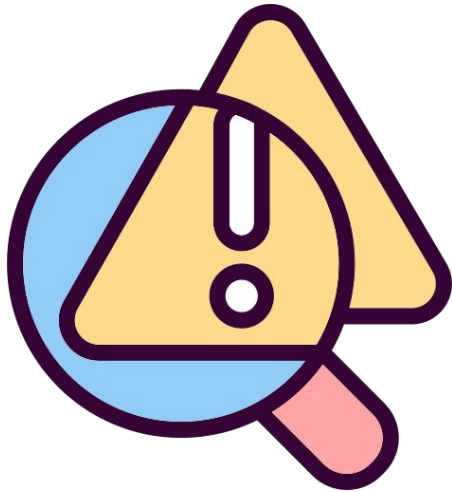
# INTRODUCTION



Cyberattacks



Main Problem



Main Problem



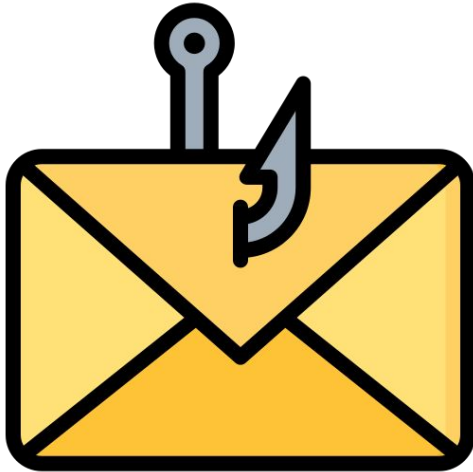
Human Factor

# 1.6%

Malicious Link Clicks by Employees  
Q2 2020

# 11.8%

Malicious Link Clicks by Employees  
2022



Phishing Attack



Malware



# 11s

Ransomware Attack

**2s**

Predicted Ransomware  
Attacks in 2031



The concept smart-grid was introduced in 2007 to solve one of the biggest problems of the new millennium: ***environmental problems.***





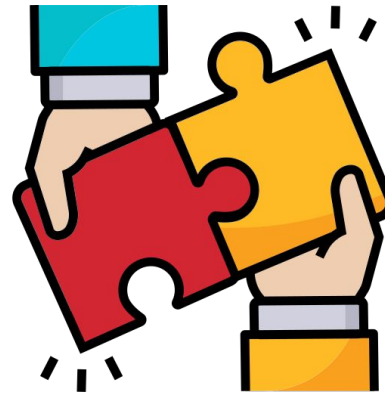
Denial of Service  
(DoS)



Distributed Denial of Service  
(DDoS)



Confidentiality



Integrity



Availability

In this paper we propose an intrusion detector based on Deep Learning (DL) to avoid the possible threat above mentioned.

For this purpose, we train and evaluate 10 different DL architectures.

We employed two datasets published on the web composed of PCAP (Packet Capture) converting them into images using a Python script written by the authors.

After training and testing, GradCAM++ and ScoreCAM were applied, and similarity between the generated images was assessed using the IF/IM-SSIM algorithm.

Journal of Computer Virology and Hacking Techniques (2025) 21:9  
<https://doi.org/10.1007/s11416-025-00549-1>

## RESEARCH



### A method for smart grid intrusion detection through explainable deep learning

Giovanni Ciarabella<sup>1,2</sup> · Fabio Martinelli<sup>3</sup> · Antonella Santone<sup>4</sup> · Francesco Mercaldo<sup>2,4</sup>

Received: 2 October 2024 / Accepted: 26 March 2025  
© The Author(s) 2025

#### Abstract

Over the years, cyber-attacks have increased drastically, and their execution changed with time. One of the targets of cyber criminals is trying to obtain sensitive information from mobile, cloud, or generally IoT devices. To avoid those risks, different countermeasures have been developed and implemented. For instance, the IEC 60870-5-104 protocol was developed to define the systems used for remote control in electrical engineering and power system automation applications. Starting from these considerations, in this paper, we propose an intrusion detector based on explainable Deep Learning (DL) that is able to detect possible attacks. In a nutshell, we consider several DL models, *i.e.*, AlexNet, DenseNet, EfficientNet, Inception, LeNet, MobileNet, ResNet50, Standard CNN, VGG16, and VGG19 to understand whether a network trace (stored in a PCAP file) is related to an attack. Moreover, to explain of the model attack prediction, we resort to two different Class Activation Mapping algorithms available in the literature: Grad-CAM++ and Score-CAM. As the last step, we also calculated the IF/IM-SSIM index to strengthen the robustness of the top-performing model and evaluate the similarity between the two CAM algorithms. Experimental results show the effectiveness of the proposed method, and we obtained an accuracy equal to 0.900 with the DenseNet. In conclusion, we applied the exact steps to a new dataset to confirm that the proposed methodology is scalable and applicable to other datasets and achieved promising results.

**Keywords** Smart grid · Deep learning · Intrusion · Security

#### 1 Introduction

During the last few years, the number of cyber-attacks has drastically increased. As a matter of fact, companies have become targets for cyber-criminals, and only a small number

of them are ready to combat them. One of the most important aspects of awareness is the human factor, which is a real danger for companies. As declared in a report published in 2023 by Infosecurity Magazine<sup>1</sup>, phishing encounters have increased every quarter since Q2 2020, and the number of employees who click on malicious links also rose from 1.6% in 2020 to 11.8% in 2022. One of the possible consequences of a phishing attack could be the introduction of ransomware in the companies' environment. With this type of malware, attackers get access to sensitive information to demand a ransom, and all data may be published if the latter is not delivered within a specific time frame. As reported by Cybercrime Magazine<sup>2</sup>, in the enterprise sector, one ransomware assault occurred every 11 seconds in 2021, and the frequency of these attacks on governments, corporations, individuals, and gadgets is predicted to increase over the next five years, reaching every two seconds in 2031. The spread of mobile, cloud, and

✉ Giovanni Ciarabella  
giovanni.ciarabella@itc.cnr.it;  
giovanni.ciarabella@imtlucca.it  
Fabio Martinelli  
fabio.martinelli@icar.cnr.it  
Antonella Santone  
antonella.santone@unimol.it  
Francesco Mercaldo  
francesco.mercaldo@unimol.it

<sup>1</sup> IMT School for Advanced Studies Lucca, Lucca, Italy

<sup>2</sup> Institute for Informatics and Telematics, National Research Council of Italy, Pisa, Italy

<sup>3</sup> Institute for High Performance Computing and Networking, Rende, Italy

<sup>4</sup> University of Molise, Campobasso, Italy

<sup>1</sup> <https://www.infosecurity-magazine.com/news/record-number-of-mobile-phishing/>

<sup>2</sup> <https://cybersecurityventures.com/global-ransomware-damage-costs-predicted-to-reach-20-billion-usd-by-2021/>

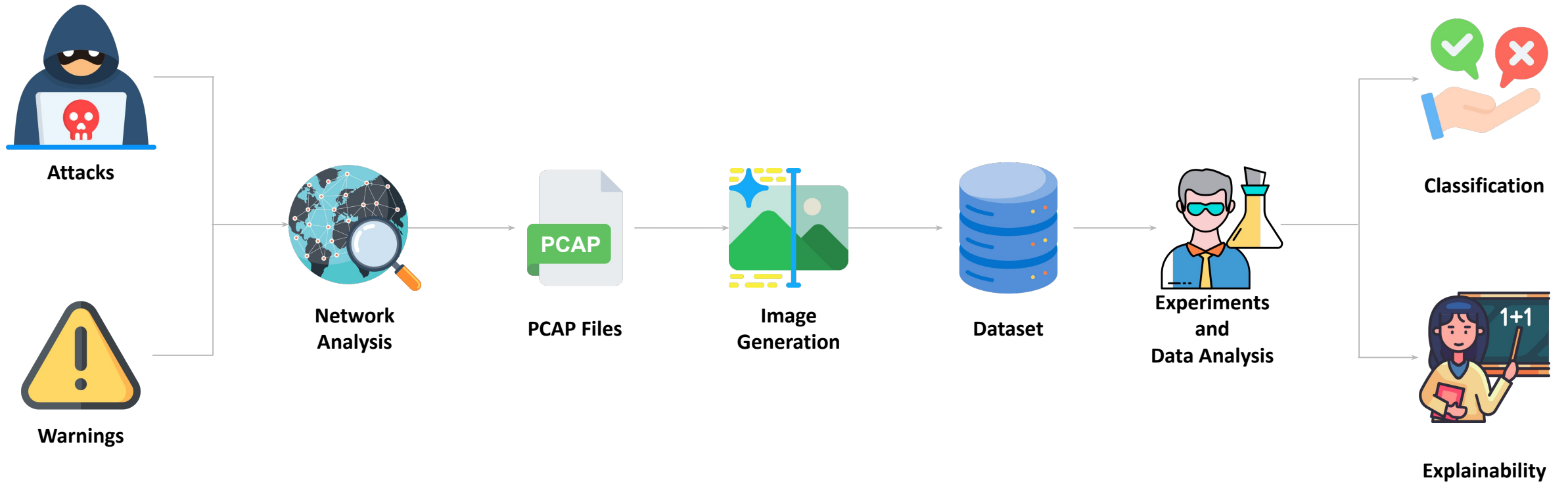
Published online: 05 May 2025

 Springer

# 2

# METHODOLOGY





## Dataset A

**Dataset Name:** IEC 60870-5-104 Intrusion Detection Dataset

**Brief Description:** The original dataset comprises several PCAP and CSV of three main labels: DoS attacks, MITM attacks, and a set of legitimate commands employed fraudulently.

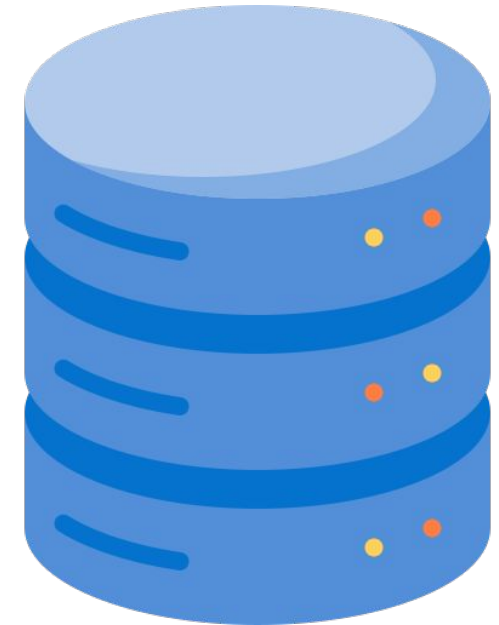
These attacks were simulated using IEC TestServer and validated with two actual Remote Terminal Unit (RTU) devices.



## Dataset B

**Dataset Name:** TII-SSRC-23

**Brief Description:** comprises several PCAP files obtained from eight traffic types (audio, background, text, video, bruteforce, DoS, information gathering, botnet) and 32 subtypes across both benign and malicious categories.

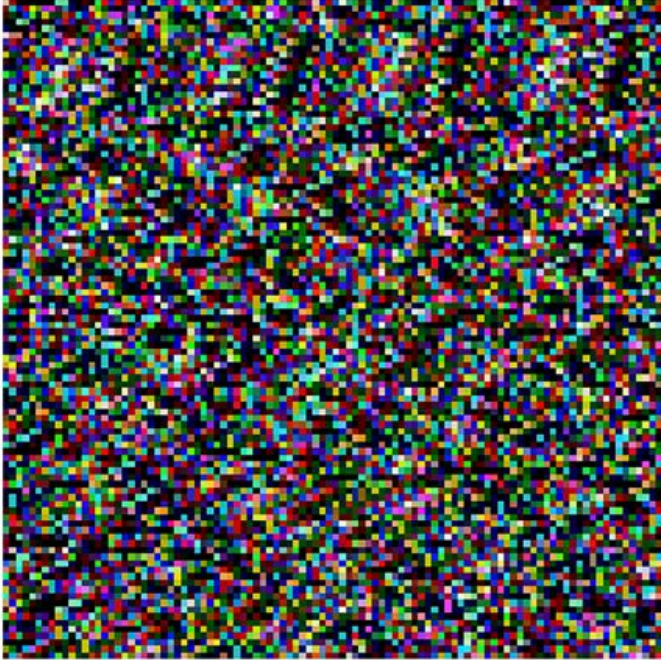


```

1 for each file in folder_files:
2
3     if file ends with ".pcap":
4         if file_size < image_size * image_size * 4:
5             pad data with zeros to reach image_size * image_size * 4
6
7         trim data to exactly image_size * image_size * 4 elements
8         reshape data to image_size x image_size x 4
9
10        height, width, channels = shape of data
11
12        new_height = height - (height % 2)
13        new_width  = width  - (width  % 2)
14
15        crop data to new_height x new_width x channels
16
17        half_height = new_height // 2
18        half_width  = new_width  // 2
19
20        part1 = data[:half_height, :half_width, :]
21        part2 = data[:half_height, half_width:, :]
22        part3 = data[half_height:, :half_width, :]
23        part4 = data[half_height:, half_width:, :]
24
25        create PIL image from each part using color_mode
26        save each image

```

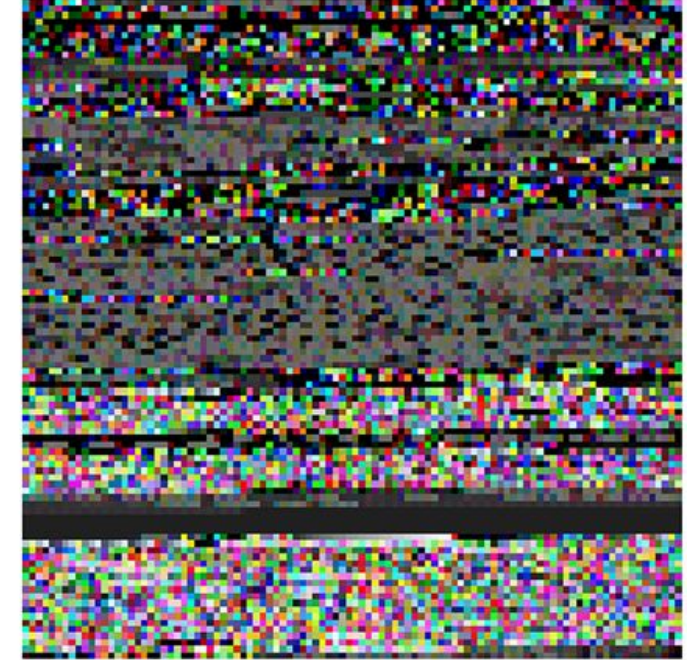




(a) DoS attack.



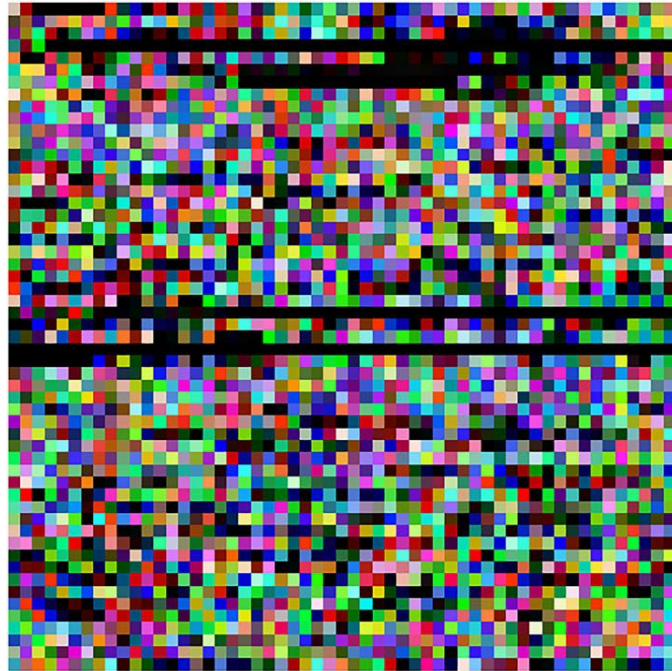
(b) MITM attack.



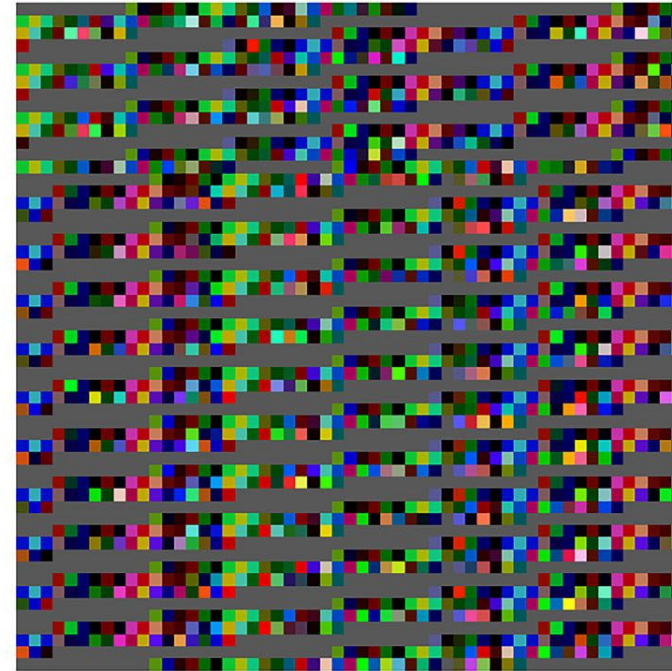
(c) Possible attack.

Samples of three different types of samples obtained after the conversion of PCAP files belonging to the “IEC 60870-5-104 Intrusion Detection Dataset”

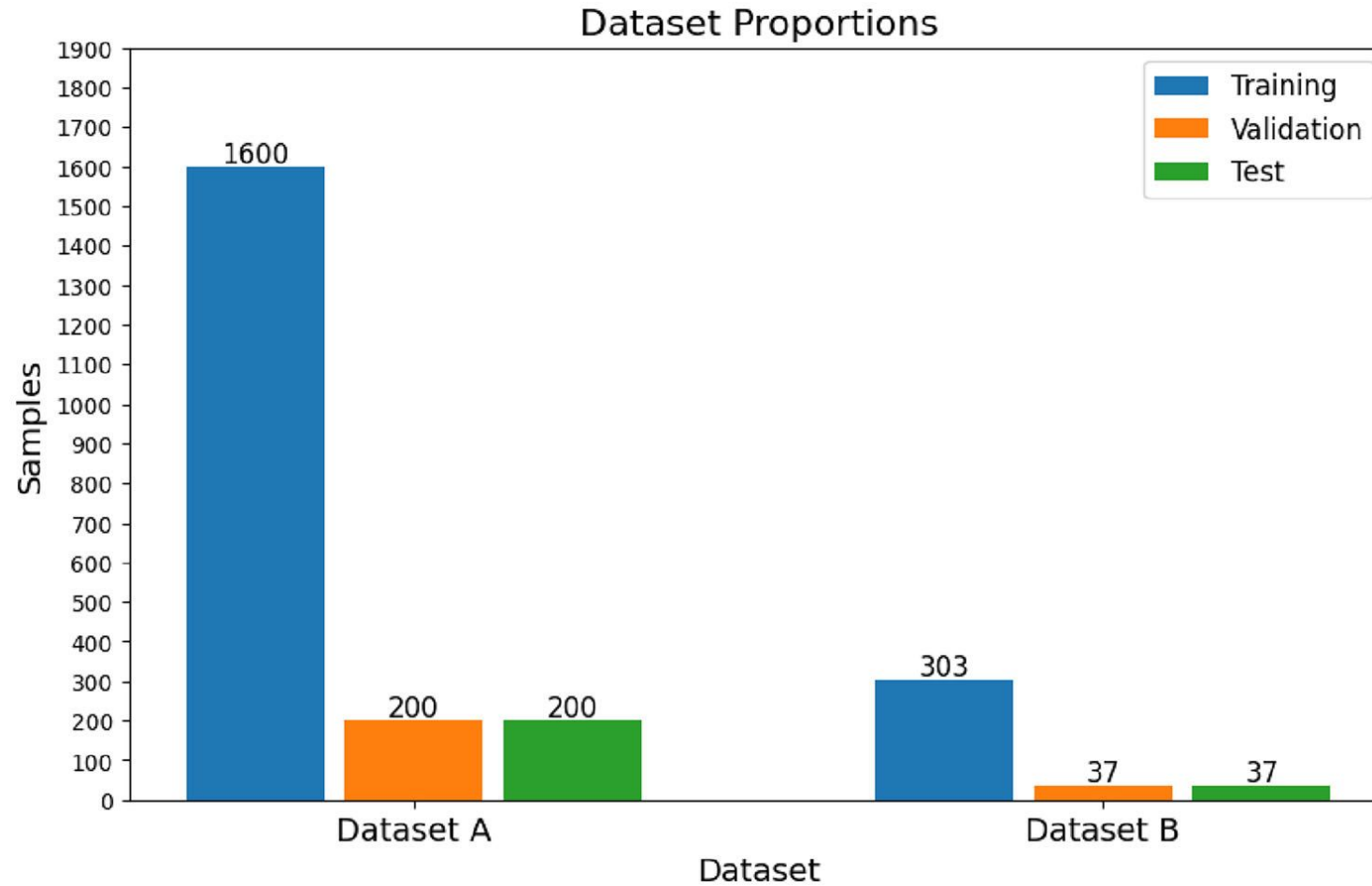




(a) Benign behaviour.



(b) Malicious behaviour.



Model	Epochs	Image Size	Batch Size	Learning Rate
AlexNet	35	110x3	32	1e-04
DenseNet	35	110x3	32	3e-04
EfficientNet	35	110x3	64	1e-04
Inception	35	110x3	64	1e-04
LeNet	20	110x3	32	1e-04
MobileNet	35	110x3	32	3e-04
ResNet50	30	110x3	32	3e-04
Standard CNN	40	110x3	64	1e-04
VGG16	40	110x3	64	1e-04
VGG19	40	110x3	32	1e-04



3

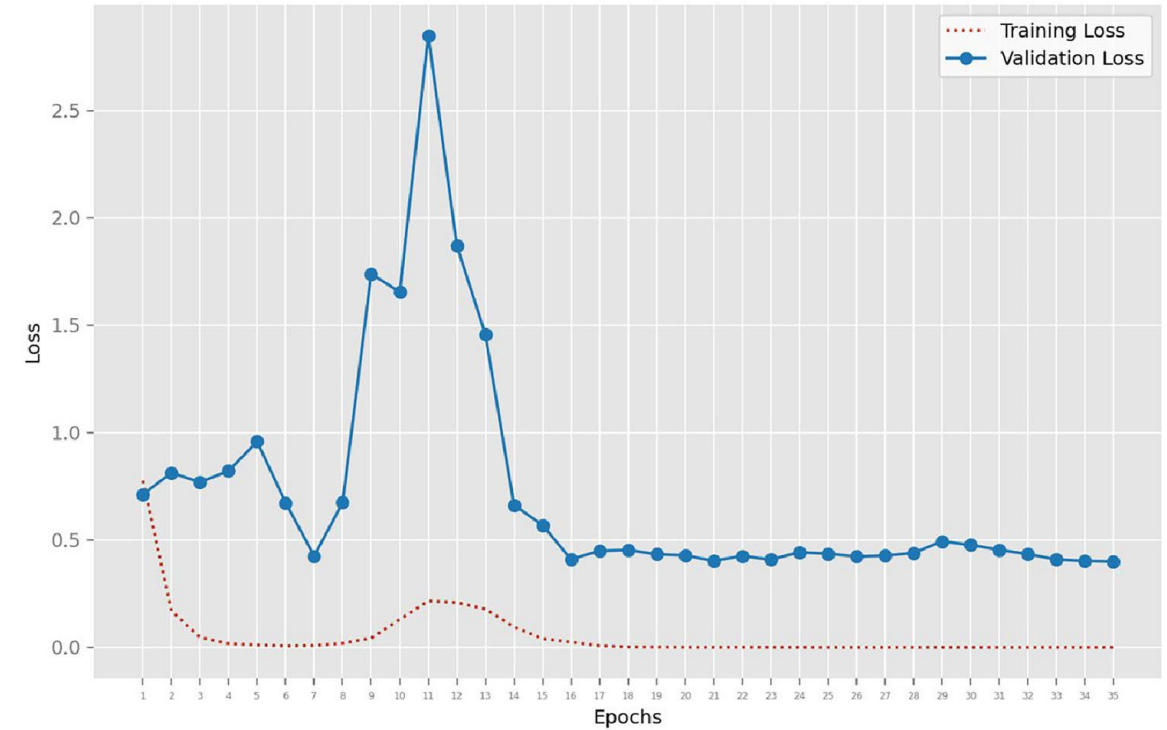
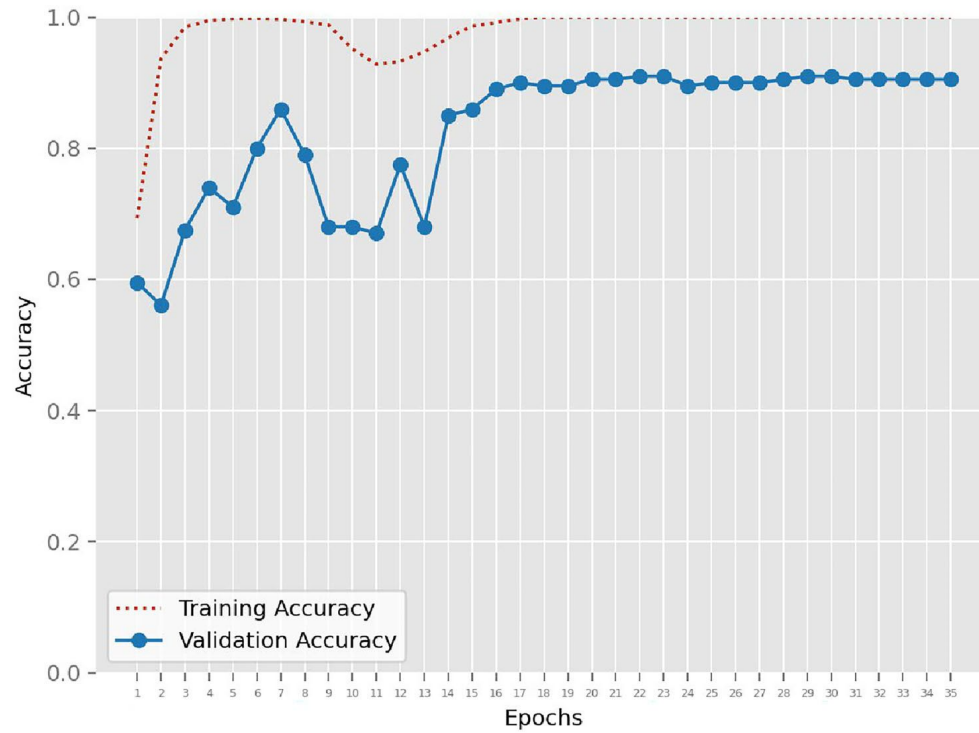
RESULTS

# 3.1

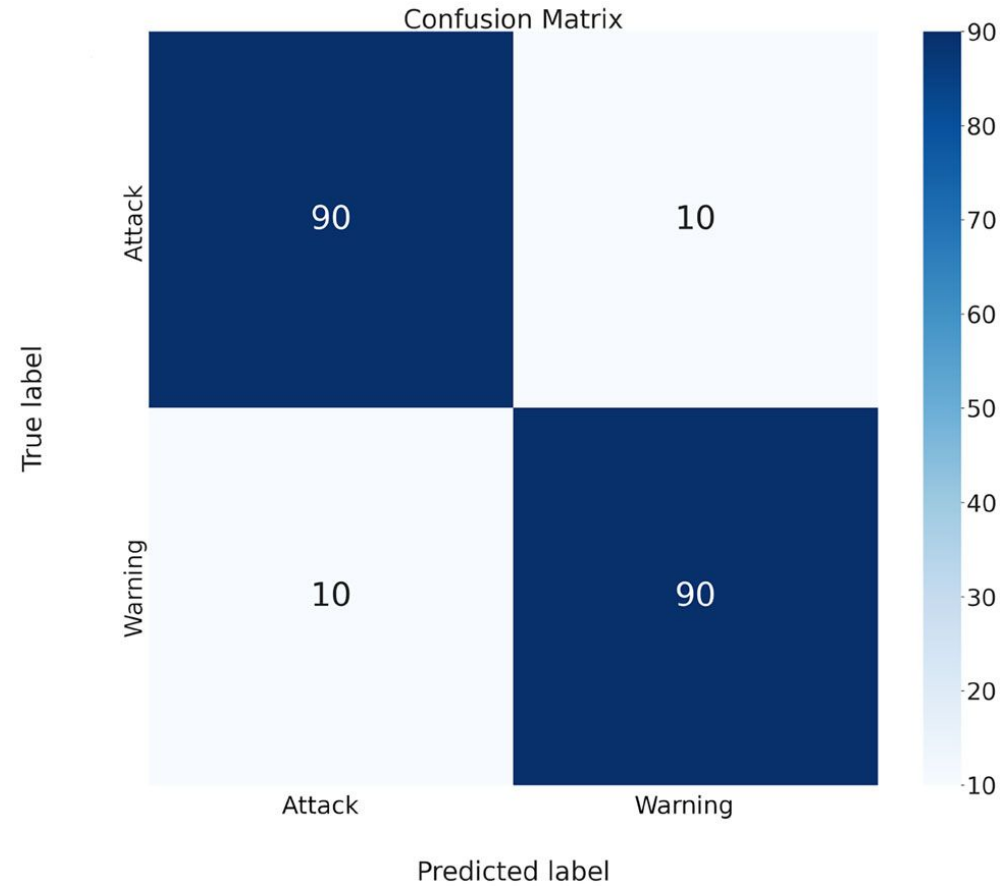
## DATASET A

Model	Loss	Accuracy	Precision	Recall	F-Measure	AUC
AlexNet	1.226	0.780	0.780	0.780	0.780	0.835
DenseNet	0.459	0.900	0.900	0.900	0.900	0.947
EfficientNet	0.960	0.635	0.635	0.635	0.635	0.689
Inception	0.568	0.870	0.870	0.870	0.870	0.916
LeNet	0.693	0.500	0.500	0.500	0.500	0.500
MobileNet	1.532	0.850	0.850	0.850	0.850	0.885
ResNet50	0.822	0.845	0.845	0.845	0.845	0.887
Standard CNN	0.575	0.780	0.780	0.780	0.780	0.867
VGG16	1.298	0.835	0.835	0.835	0.835	0.877
VGG19	0.693	0.500	0.500	0.500	0.500	0.500

Model	Loss	Accuracy	Precision	Recall	F-Measure	AUC
AlexNet	1.226	0.780	0.780	0.780	0.780	0.835
<b>DenseNet</b>	<b>0.459</b>	<b>0.900</b>	<b>0.900</b>	<b>0.900</b>	<b>0.900</b>	<b>0.947</b>
EfficientNet	0.960	0.635	0.635	0.635	0.635	0.689
Inception	0.568	0.870	0.870	0.870	0.870	0.916
LeNet	0.693	0.500	0.500	0.500	0.500	0.500
MobileNet	1.532	0.850	0.850	0.850	0.850	0.885
ResNet50	0.822	0.845	0.845	0.845	0.845	0.887
Standard CNN	0.575	0.780	0.780	0.780	0.780	0.867
VGG16	1.298	0.835	0.835	0.835	0.835	0.877
VGG19	0.693	0.500	0.500	0.500	0.500	0.500



Training and validation accuracy (left) and loss (right) trends across epochs for the DenseNet architecture.



Confusion matrix obtained after the testing phase of DenseNet.

0.9

Specificity

0.8

Matthews Correlation Coefficient

## Takeaway — Dataset A

The models achieved strong performance across the training, validation, and test phases. These results are further supported by additional evaluation metrics, including specificity and the Matthews Correlation Coefficient (MCC).

High **specificity** values indicate a strong ability to correctly reject negative instances

**MCC** scores suggest robust overall performance, reflecting a well-balanced prediction of both positive and negative classes.



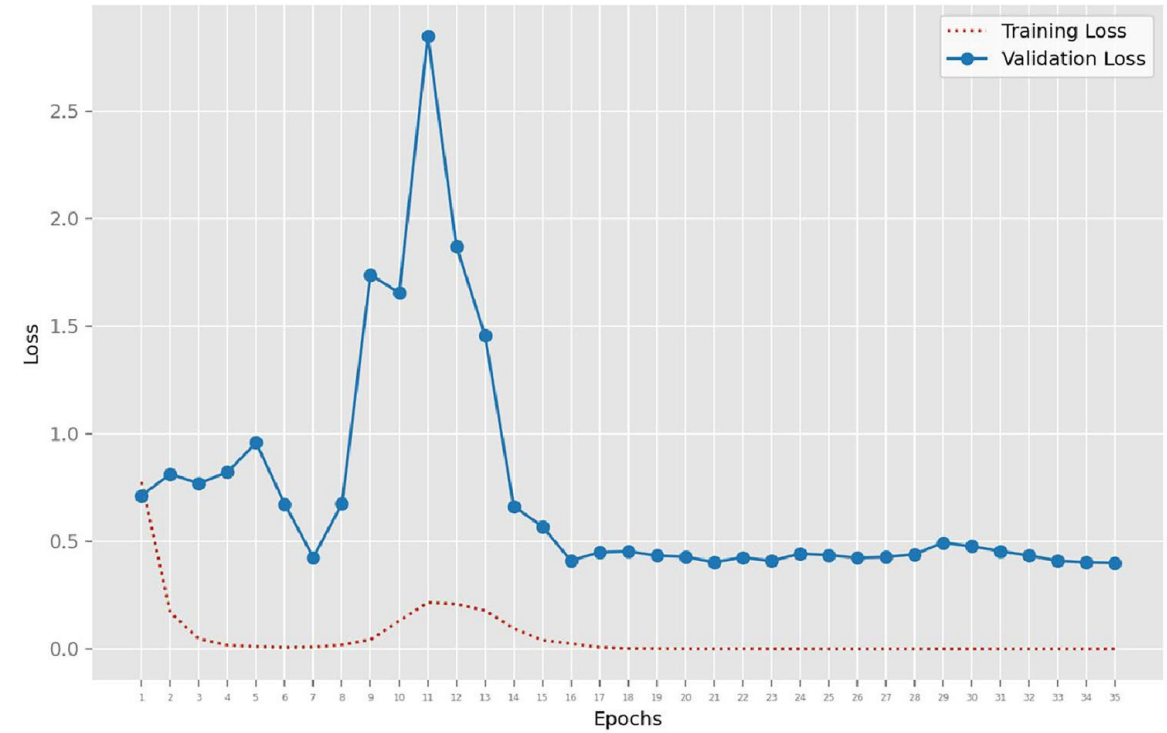
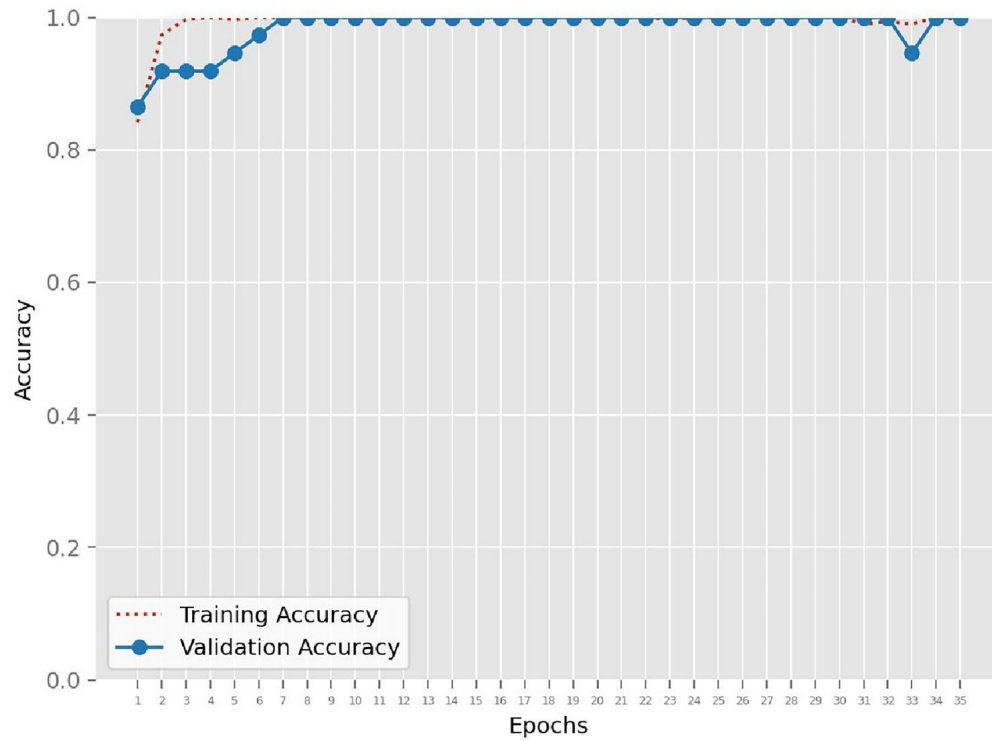


# 3.2

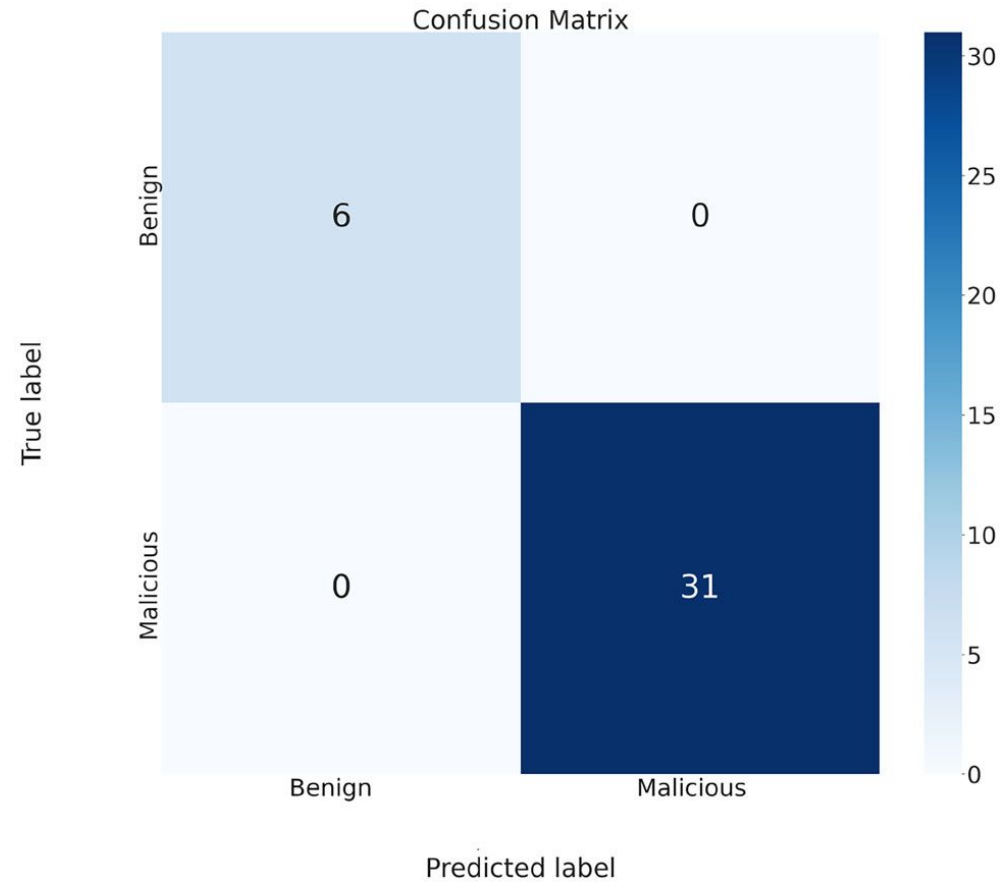
## DATASET B

Model	Loss	Accuracy	Precision	Recall	F-Measure	AUC
AlexNet	2.562	0.838	0.838	0.838	0.838	0.838
DenseNet	0	1	1	1	1	1
EfficientNet	2.532	0.162	0.162	0.162	0.162	0.27
Inception	0.606	0.946	0.946	0.946	0.946	0.993
LeNet	0.439	0.838	0.838	0.838	0.838	0.92
MobileNet	0.005	1	1	1	1	1
ResNet50	16.379	0.838	0.838	0.838	0.838	0.838
Standard CNN	0.053	0.973	0.973	0.973	0.973	0.999
VGG16	0	1	1	1	1	1
VGG19	0.447	0.838	0.838	0.838	0.838	0.838

Model	Loss	Accuracy	Precision	Recall	F-Measure	AUC
AlexNet	2.562	0.838	0.838	0.838	0.838	0.838
DenseNet	0	1	1	1	1	1
EfficientNet	2.532	0.162	0.162	0.162	0.162	0.27
Inception	0.606	0.946	0.946	0.946	0.946	0.993
LeNet	0.439	0.838	0.838	0.838	0.838	0.92
<b>MobileNet</b>	<b>0.005</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>1</b>
ResNet50	16.379	0.838	0.838	0.838	0.838	0.838
Standard CNN	0.053	0.973	0.973	0.973	0.973	0.999
VGG16	0	1	1	1	1	1
VGG19	0.447	0.838	0.838	0.838	0.838	0.838



Training and validation accuracy (left) and loss (right) trends across epochs for the MobileNet architecture.



Confusion matrix obtained after the testing phase of MobileNet.

**1**

Specificity

**1**

Matthews Correlation Coefficient

## Takeaway — Dataset A

Among the evaluated models, the MobileNet architecture achieved the best performance on Dataset B. Due to the limited size of Dataset B (377 samples), several architectures achieved perfect accuracy; however, most of these models exhibited clear signs of overfitting.

In contrast, MobileNet showed only mild overfitting during a small number of training epochs.

Additionally, both specificity and the Matthews Correlation Coefficient (MCC) reached a value of 1, indicating perfect and well-balanced classification performance.



# 3.3

## EXPLAINABILITY

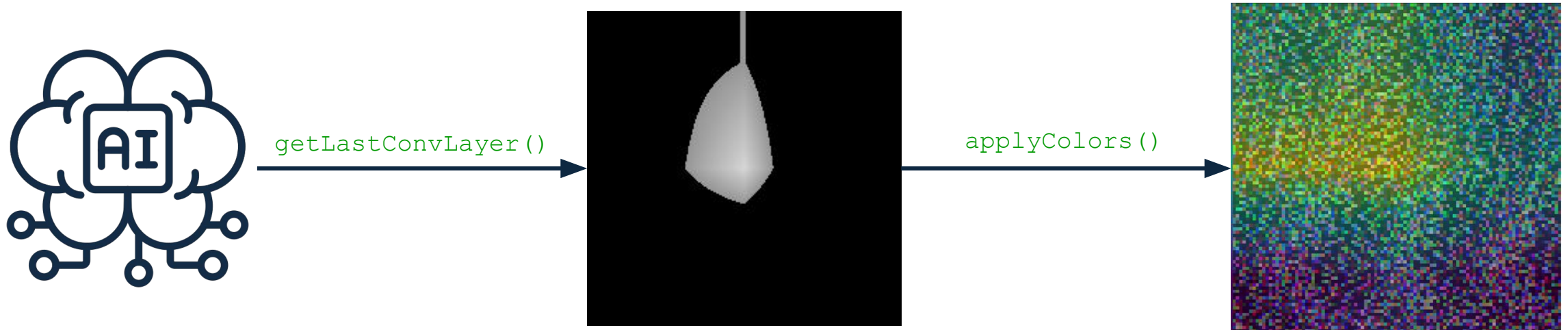


## Border definition:

Explainability is the set of techniques that make the decisions of artificial intelligence models understandable.

1. It helps explain why a system has produced a particular outcome.
2. It increases transparency, trust, and the ability to identify errors or biases.
3. It is especially important in critical domains such as healthcare, finance, and justice.







areas of disinterest to the model



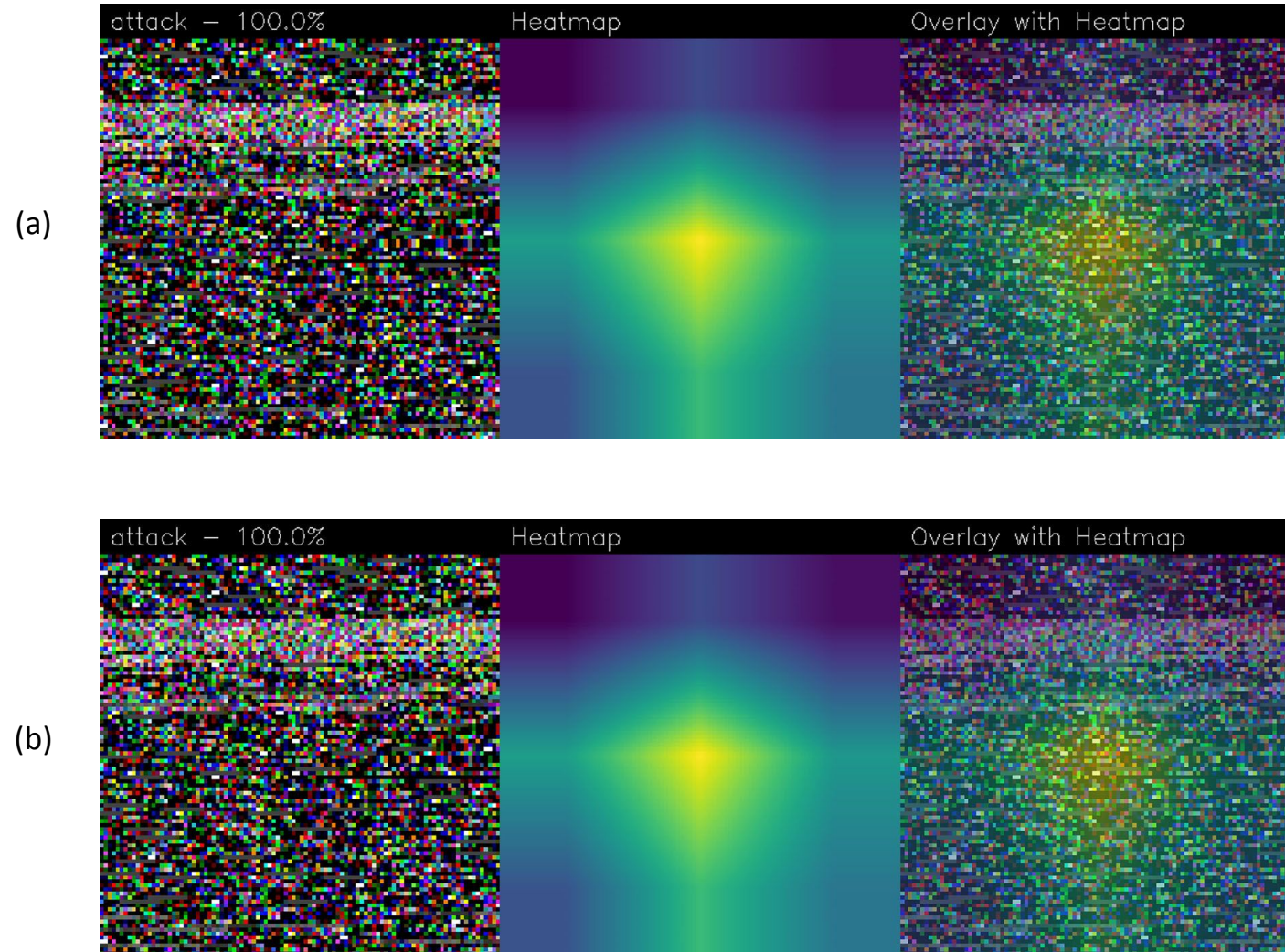
areas not very interesting for the model



areas of interest to the model

# 3.3.1

# XAI DATASET A



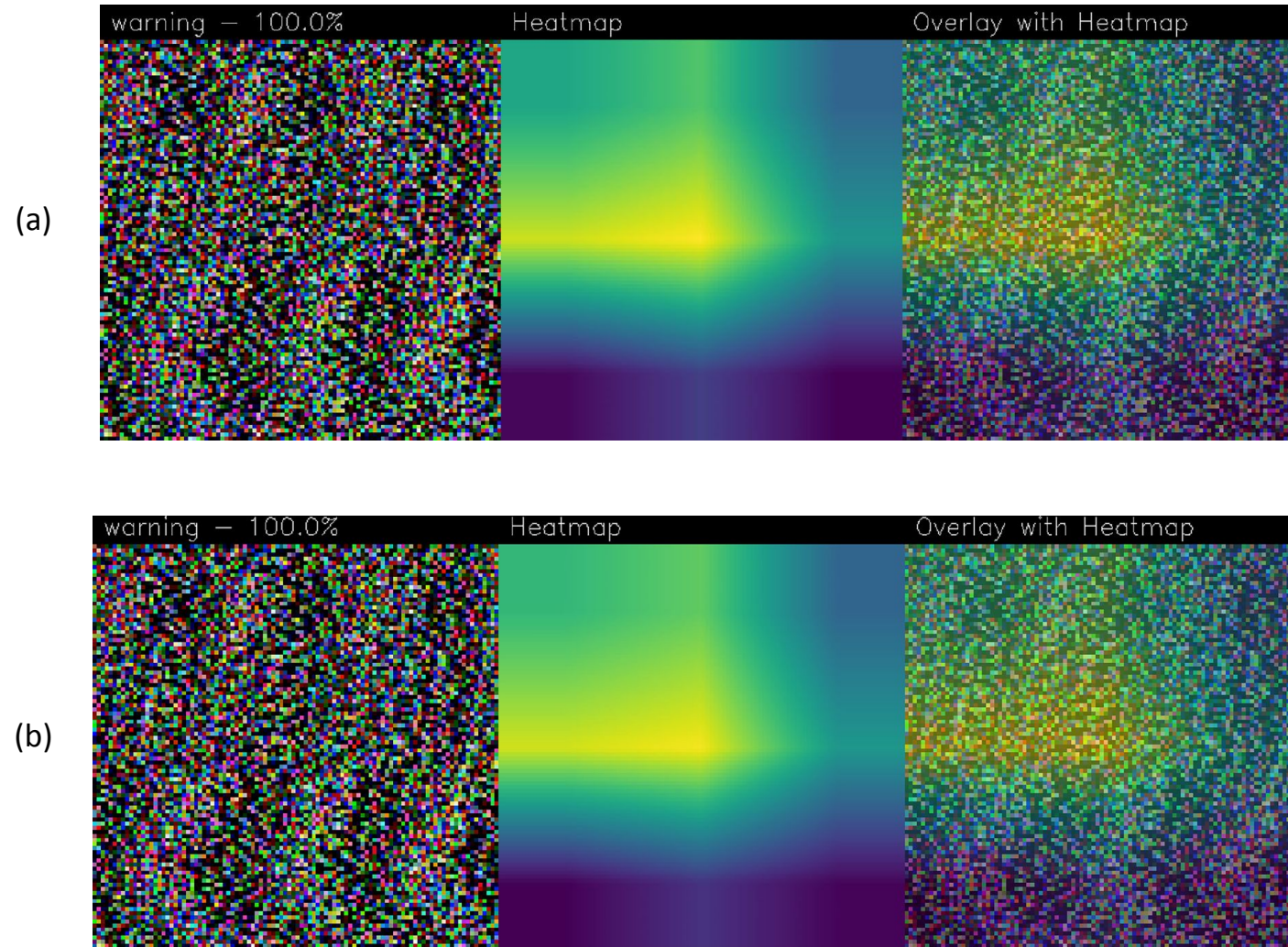
**0.816**

Score-CAM similarity index

**0.608**

Grad-CAM++ similarity index





**0.806**

Score-CAM similarity index

**0.552**

Grad-CAM++ similarity index



0.839

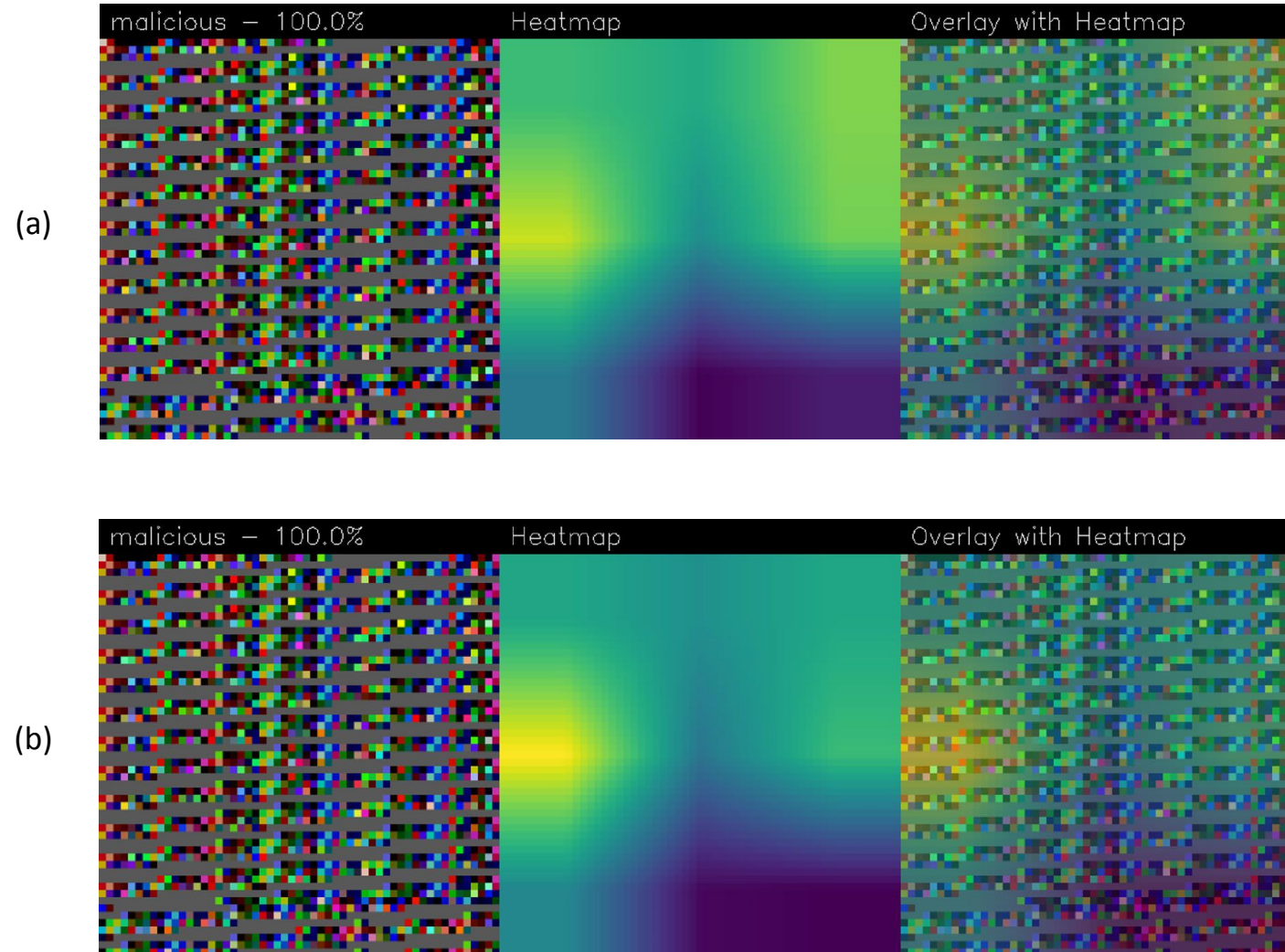
Attack class similarity index score

0.808

Warning class similarity index score

# 3.3.2

## XAI DATASET B

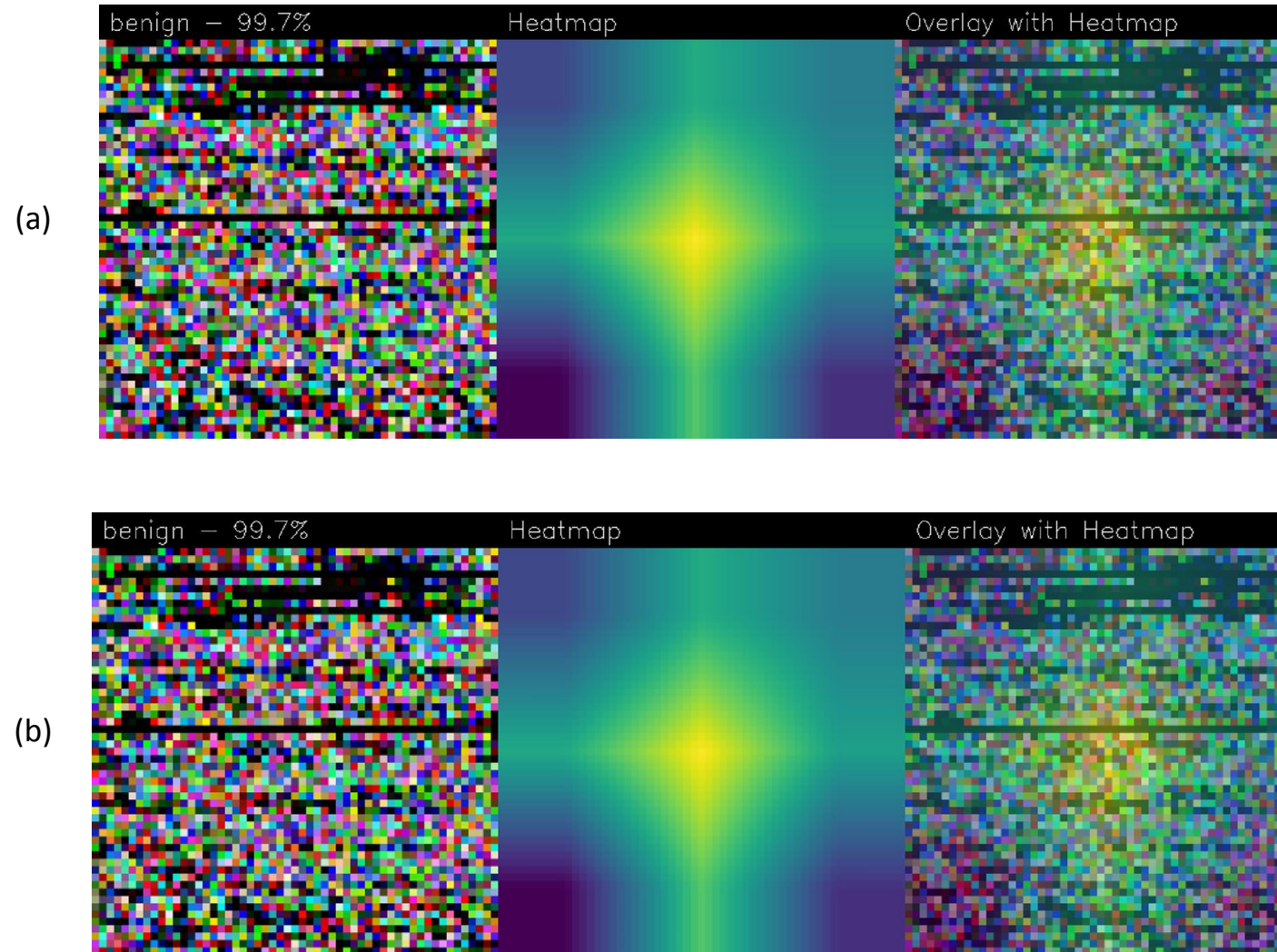


**0.365**

Score-CAM similarity index

**0.266**

Grad-CAM++ similarity index



Result of applying Grad-CAM++ (a) and Score-CAM (b) algorithm on “benign” samples (text)

**0.782**

Score-CAM similarity index

**0.457**

Grad-CAM++ similarity index

**0.503**

Malicious class similarity index score

**0.753**

Benign class similarity index score

# 3.4 | TIME PERFORMANCE ANALYSIS



$$TPA \text{ Image Generation} = \frac{tt}{nsc1} + \frac{tt}{nsc2}$$

$$TPA \text{ Training Phase} = \frac{tt}{ns}$$

$$TPA \text{ test, Grad-CAM++}, \text{ Score-CAM, IF / IM-SSIM} = \frac{tt}{nsc}$$

*tt* stands for total time employed for the execution in seconds;

*nsc1* the number of samples for the first class, and *nsc2* the number of samples for the second class of the dataset

Dataset	Image Generation (s)	Training (s)	Test (s)	Grad-CAM++ (s)	Score-CAM (s)	IF/IM-SSIM (s)	Total (s)
Dataset A	0.21	1.35	0.15	12.08	13.9	0.09	27.78
Dataset B	3.05	0.51	0.15	2.1	3.48	0.07	9.36

Average time (in seconds) for each phase of the proposed model, along with the total average time required to generate the results reported in this research, for both datasets.

## Takeaway — Time Performance Analysis

The time performance analysis (TPA) shows that Dataset A required more time during image generation and explainability phases due to its larger number of samples, while Dataset B exhibited lower execution times overall.

Training and testing phases were efficient for both datasets, with explainability methods (Grad-CAM++, Score-CAM, and IF/IM-SSIM) representing the most time-consuming steps, particularly for Dataset A.

Overall, the proposed model demonstrated scalable and manageable execution times across all phases



# 4

# CONCLUSION AND FUTURE WORK

1

Proposed a deep learning–based intrusion detection system capable of distinguishing network attacks from legitimate commands that may be exploited for malicious purposes.

2

Built and processed a PCAP-based dataset by converting network traffic into images and applying data augmentation to address the limited number of available samples.

3

Evaluated ten state-of-the-art deep learning architectures under different hyper-parameter settings, achieving the best performance with DenseNet, which reached an accuracy of 0.900.

1

Compare the results obtained using different learning approaches and neural network architectures to further assess performance and generalization capabilities.

2

Evaluate the robustness of the proposed method by leveraging Generative Adversarial Networks (GANs) to simulate adversarial and challenging scenarios.

3

Investigate the identification and classification of multiple threat types within smart grid environments through extended network-based analysis.

# LET'S KEEP IN TOUCH



# Thank you for the attention!

*Giovanni Ciaramella*