**UNIVERSITÀ DEGLI STUDI DI SALERNO**

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE ED
ELETTRICA E MATEMATICA APPLICATA

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA INFORMATICA


ARTIFICIAL INTELLIGENCE FOR CYBERSECURITY

# Security Evaluation of a Face Recognition System

**Gruppo 4**

| | |
|---|---|
| Nicola Lanzara | 0622702118 |
| Giulia Minichiello | 0622702127 |
| Simone Pacifico | 0622702115 |
| Lucia Senatore | 0622702089 |

**Anno Accademico 2023 – 2024**

# Sommario

# Overview

The goal of the project is to assess the robustness of a face recognition-based access control system against adversarial attacks and to explore the potential for implementing effective defences.

Face recognition systems are increasingly used for authentication and access control due to their ability to quickly and reliably identify individuals. However, they are not without vulnerabilities, especially concerning adversarial attacks, which exploit small perturbations, often imperceptible to the human eye, to deceive the facial recognition model. These attacks can lead to incorrect user identification and unauthorized access.

A critical aspect of adversarial attacks is their transferability, which refers to the ability of an attack designed for a specific artificial intelligence model to affect other models that have not been directly targeted. This implies that an adversary could deceive a facial recognition system used in real-world scenarios, even if the attack was initially developed for a different model.

During the project, a series of identities from the VGG-Face2 dataset were selected to create a representative test set, on which the accuracy of a face recognition model (NN1) was evaluated. Adversarial examples were then generated using the ART library to examine the effects of error-generic and error-specific attacks on NN1 and to measure performance degradation using Security Evaluation Curves. A second classifier (NN2), trained on the same dataset, was introduced to verify the transferability of the attacks. Finally, targeted defences against adversarial attacks were implemented and evaluated, exploring strategies such as adversarial sample detection and pre-processing techniques.

# Dataset Description

For the project, a sample of 100 identities was randomly selected from the training set of the VGG-Face2 dataset, available here: https://drive.google.com/file/d/1SXhc8m5PHxyM4lEWVEccSufIa8OiOjGW/view?usp=sharing.

The random selection was done to ensure that the identities are evenly represented and to minimize any bias in the selection process.

The test set was made up of a total of 1,000 images, divided equally among the selected identities. For each of the 100 identities, 10 images were chosen, ensuring enough data for a strong evaluation of the facial recognition system. The images were taken directly from the test set of the VGG-Face2 dataset, available here: https://drive.google.com/file/d/1K56kVYHHDfLA2Anm7ga0tQolMwlPk6R8/view?usp=sharing.

This choice helped to keep the quality and characteristics of the images consistent, avoiding the variations that could come from collecting images from other sources.

Using a well-defined and controlled test set is crucial for ensuring the reliability of the results during model evaluations. This approach ensures that the facial recognition system is tested on data that accurately reflects real-world use and the challenges that come with it.

# NN1 Evaluation

InceptionResnetV1 is a deep neural network model that combines two powerful and popular architectures in deep learning: Inception and ResNet. By combining these two architectures, InceptionResnetV1 benefits from the ability to capture complex features at different scales (thanks to Inception) and the ability to train effectively even with many layers (thanks to ResNet).

- **Inception**: The "Inception" part of the architecture uses multiple convolution filters of different sizes in parallel, then combines the results. This approach helps the model capture features at different scales, improving its ability to recognize complex patterns in images.
- **ResNet**: The "ResNet" part introduces residual blocks, which allow very deep networks to be built without problems like gradient degradation. Residual blocks use skip connections that bypass one or more layers, making it easier for the model to learn even with many layers.

The InceptionResnetV1 model is pre-trained on the VGGFace2 dataset, which contains over 3 million images of faces from more than 9,000 individuals. This training allows the model to learn rich and generalizable representations of human faces, making it very effective for facial recognition.

Thanks to its ability to extract detailed facial features, InceptionResnetV1 is often used in facial recognition applications such as:

- **Identity verification** (1:1 matching): Comparing two images to determine if they represent the same person.
- **Facial recognition** (1 matching): Identifying a person by comparing their image to a database of faces.
- **Facial embeddings**: Extracting feature vectors (embeddings) that represent a face in a latent space, which can be used for tasks like clustering or classification.

## Accuracy

The accuracy of the facial recognition network (NN1) evaluated on the test set was found to be 92.00799200%. While this result appears very positive, it should be interpreted with caution. The high accuracy is largely due to the fact that the data used for evaluation, specifically the images in the test set, are the same as those used during the training phase of the network.

When test data coincides with training data, the model has already "seen" that data and had the opportunity to adapt perfectly to it. This leads to an overestimated accuracy result, as the network is not being tested on its true ability to generalize to new, unseen data.

To improve generalization ability and make the evaluation more realistic, we tried using MTCNN (Multi-task Cascaded Convolutional Networks) for facial recognition. However, in this case, the accuracy obtained was lower than that of NN1. This suggests that, although MTCNN may offer a more robust and generalizable approach, the model was not as effective in recognizing faces in our specific dataset.

# Generation of Adversarial Examples on NN1

## FGSM

FGSM is a white-box attack, meaning the attacker knows everything about the neural network they're trying to trick. The idea is to use this knowledge to create a modified input that confuses the network by making it more likely to misclassify.

The parameter epsilon (ε) in FGSM controls how much the input is changed. A larger epsilon means bigger changes to the input, which can make the attack more effective but also easier to detect. A smaller epsilon means smaller changes, which are harder to notice but might not be strong enough to fool the network.

### Error Generic

### Error Specific

## BIM

### Error Generic

### Error Specific

## PGD

### Error Generic

### Error Specific

## Carlini Wagner

### Error Generic

### Error Specific

## DeepFool

DeepFool is an adversarial attack designed to deceive neural network models, particularly those employed in image classification tasks. Its objective is to identify the smallest possible perturbation that can be applied to an image to cause the model to alter its classification prediction. The algorithm operates iteratively, approximating the decision boundary of the classifier as if it were linear, and incrementally moving the perturbed image toward this boundary until a class change is achieved. This process minimizes the magnitude of the perturbation, often rendering it imperceptible to the human eye. DeepFool is recognized for its effectiveness and computational efficiency compared to other attack methods, highlighting the vulnerability of deep learning models to adversarial attacks.

### Error Generic

In our experiment, we set the `max_iter` parameter to 15, as this was deemed enough iterations to find an effective perturbation. This value was chosen based on experimental and practical considerations: a higher number of iterations might slightly increase the precision of the perturbation

but would come at a higher computational cost, while a lower number might not allow the algorithm to reach the classifier’s decision boundary.

Regarding `epsilon`, we tested various values to find the best trade-off between the magnitude of the perturbation and the ability to deceive the model. `Epsilon` defines the tolerance for the perturbation’s precision: lower values lead to smaller and less visible perturbations, which may not be sufficient to alter the classification, while higher values ensure that the classification is changed but with a more noticeable perturbation. This approach allowed us to determine the optimal `epsilon` value that balances the attack's effectiveness with its imperceptibility.

## NN2 Evaluation

## Transferability Analysis

## Adversarial Example Detector