Is my Data in your AI Model? Membership Inference Test with Application to Face Images

Daniel DeAlcala, Aythami Morales, Julian Fierrez, Gonzalo Mancera, Ruben Tolosana, Javier Ortega-Garcia Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Spain

{ daniel.dealcala, aythami.morales, julian.fierrez, gonzalo.mancera, ruben.tolosana, javier.ortega } @uam.es

Abstract—This article introduces the Membership Inference Test (MINT), a novel approach that aims to empirically assess if given data was used during the training of AI/ML models. Specifically, we propose two MINT architectures designed to learn the distinct activation patterns that emerge when an Audited Model is exposed to data used during its training process. These architectures are based on Multilayer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs). The experimental framework focuses on the challenging task of Face Recognition, considering three state-of-the-art Face Recognition systems. Experiments are carried out using six publicly available databases, comprising over 22 million face images in total. Different experimental scenarios are considered depending on the context of the AI model to test. Our proposed MINT approach achieves promising results, with up to 90% accuracy, indicating the potential to recognize if an AI model has been trained with specific data. The proposed MINT approach can serve to enforce privacy and fairness in several AI applications, e.g., revealing if sensitive or private data was used for training or tuning Large Language Models (LLMs).

Index Terms—Audit, Fairness, Reliability, Membership Inference, MIA, MINT, Face Recognition

I. INTRODUCTION

N June 2023, the European Parliament adopted its position with respect to Artificial Intelligence (AI) [1]. Concretely, they have imposed some obligations to all AI companies to guarantee the protection of the citizen rights. For example, this new regulation imposes the registration of AI/ML models in a European Union (EU) database and gives national authorities of EU countries the power to request access to both the trained AI models and the associated procedural details conducted in the development of those models. This new regulation is a game changer, imposing more transparency in the development of AI technologies in Europe. As a result, novel auditing tools based on the access to trained AI models must be developed to monitor AI technologies and their correct deployment in our society.

Nowadays, AI systems have become increasingly advanced and pervasive across different domains, collecting, analyzing, and processing vast amounts of data. These data often contain sensitive information [2] about individuals as well as copyright content [3], raising concerns about privacy [4]–[6] and

The authors are with the Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, 28049 Madrid, Spain

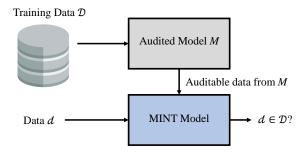


Fig. 1. The proposed Membership Inference Test (MINT) model is trained to predict if a given data (d) was used during the training process of an AI Model (M), trained with a database (D).

unauthorized access [7]–[9]. Consequently, it results critical the fast development of novel tools to clarify what data has been used to train AI models, preventing therefore the use of non-authorized data and increasing the transparency and explainability of the AI models [10]–[12]. These tools aim to prevent learning frameworks from hiding the training data within the model's parameters, thereby unveiling the opacity that they have traditionally relied upon. These considerations lead us to the main objective of the present study: checking if given data was used to train AI models [13], as summarized in Fig. 1.

This objective is related to the field of Membership Inference Attacks (MIAs), which refers to privacy attacks that target AI models trained on sensitive data [14], [15]. MIAs seeks to infer membership information about the training dataset used by exploiting the behavior and responses of the model. Also, MIAs leverage the potential memorization of the training data used by the AI models, especially when the models overfit [16], [17] or when the training data contain unique or distinctive patterns [14]. However, the performance of MIAs has been questioned in the last years due to the inherent complexity of the task [18], [19].

While MIAs are traditionally considered as attacks to quantify the potential privacy leakage of AI models, in the present article we introduce a novel perspective in line with the new AI regulation [1]. Concretely, we consider it as an auditing tool for detecting the potential use of unauthorized data for training AI models, a concept we term as Membership

Inference Test (MINT). The aim of MIAs and MINT is similar: to determine whether a given data sample was used during the training of a target AI model. However, although MIAs and our proposed MINT share the same objective, they diverge in their environmental conditions, leading to different methodological approaches. In MIAs, the model is trained assuming the role of an attacker who does not have access to the original model, but to a replica as similar as possible to it (shadow model) [14]. In MINT, we assume the role of an auditor who has direct access to the original model or partial information of it (e.g., intermediate activations). This is aligned with the new AI regulation [1], which imposes the registration of the trained AI models in an EU database. As a result, this new paradigm moves from the attacker to the auditor point of view, allowing the development of MINT models under different environmental conditions.

In the present study we concentrate on the challenging task of Face Recognition, due to the sensitive nature of face biometrics data [20]-[22] according to the recent General Data Protection Regulation (GDPR) promoted by EU [23]. Although MIAs and MINT differ due to their distinct environmental conditions, they also share significant similarities, so methods in one domain can inspire advancements in the other as well. To the best of our knowledge, there is no literature on the application of MIAs to Face Recognition. This gap in the state of the art is particularly crucial as MIAs frequently deal with classification tasks, using the network's output vector to determine membership in the training data, like a form of 'probability' of belonging to each class [16], [19], [24]-[26]. However, due to the characteristics of Face Recognition systems, focused on recognition rather than classification, such vector usually does not exist. This multiclass probability vector is traditionally replaced by an embedding feature vector representing individual characteristics in a learned space. Consequently, the methods and results diverge even more from what we typically see in state-of-the-art MIAs studies [19].

The main contributions can be summarized as follows:

- We introduce MINT as a novel perspective to detect if given data was used during the training process of AI models from an AI auditor standpoint.
- We propose two MINT architectures: i) the Vanilla MINT model based on a Multilayer Perceptron (MLP) network, trained by applying max pooling to the activation maps of both training and non-training data samples, and ii) the Convolutional Neural Network (CNN) MINT model, trained using the entire activation blocks of both training and non-training data samples.
- We conduct an extensive evaluation of the two proposed MINT architectures considering three state-of-the-art Face Recognition systems and six publicly available databases, comprising over 22 million face images. Different experimental scenarios are considered depending on the context available of the Face Recognition systems. The proposed MINT approaches achieve promising results, with up to 90% accuracy detecting if data was used for training, reducing the error rates by over 55% compared to other state-of-the-art methods.

The article is structured as follows. Sec. II provides a revision of the state of the art in the scope of the article. Sec. III describes the main concepts of the proposed MINT, and the specific details for the task considered, Face Recognition. The databases, Face Recognition systems, and experimental protocol are described in Sec. IV. Results are discussed in Sec. V, whereas conclusions and future work are finally drawn in Sec. VII.

II. RELATED WORKS

First, it is important to remark that, to the best of our knowledge, the specific problem addressed in the present article has not been directly tackled in the current literature. Nevertheless, there are studies and research directions that touch upon similar topics or are grounded in analogous principles, which we will briefly discuss in this section. Our study closely relates to three fundamental research directions: "Membership Inference Attacks", "Reconstructing Training Data from Neural Networks", and "Data Tracing". In the following, we provide concise explanations of these three primary research areas, with special emphasis on MIAs due to its significant similarity to our own research.

A. Membership Inference Attacks

The capability to detect the data used for training AI models presents a security concern that can potentially unveil sensitive [2] or private information [27], encompassing areas such as shopping preferences, health records, and pictures, among many others. TThis capability has been explored in the literature mainly under the name of Membership Inference Attacks (MIAs). These attacks aim to extract such sensitive information through an adversarial approach, without access to the model or its training data, as providers are hesitant to disclose such data. The pioneering work by Shokri et al. [14] laid the groundwork for this line of research. In their approach, they used "shadow models" that emulate the functionality of the original model, despite having no direct access to it. To construct these so-called "shadow models", it is essential to have knowledge about the architecture of the original model they are based on, along with a subset of samples and statistics derived from the original dataset used for training. As a result, they crafted "shadow training sets" mirroring the original dataset, subsequently training shadow models to mimic the original model. This approach offers complete control over the training and non-training data of these models. A binary classifier was trained to distinguish between the training and non-training data used in the shadow models. Images were passed through the network, and output embeddings from all shadow models were used to train a simple binary classifier. Better results were observed with an increasing number of shadow models.

Other approaches have emerged from the original publication [14], following the same principle, but instead of training a binary classifier, relying on the value of specific metrics and thresholds. For example, the authors of [28] analyzed whether the loss value of input data was above a threshold or not, or [16], [29], which used the prediction value.

An alternative method was presented by Nasr *et al.* [30]. Their work introduced the Black-box and White-box terminology into this context. Previous efforts primarily relied on the output of shadow models for classification, essentially the output embedding (Black-box). However, Nasr *et al.* proposed a White-box framework, granting them access to the activations, loss, and gradients. The authors demonstrated that access to the White-box information offers limited utility in distinguishing whether a sample was used for training or not. The best results were achieved using gradients [30] where as activations did not yield promising results [18], [31]. In general, in the White-box context, the number of studies in the literature is limited, and with no significant results.

More recently, the authors of [18] delved into the challenges of MIAs and the inherent complexity of the task. They concluded that the actual performance in the task is notably worse than the results published in the state of the art. This is largely due to the insufficient evaluation methodology. Furthermore, they explored previously untapped White-box information involving gradients, although they were not able to significantly improve the Black-box framework performance.

Finally, it is noteworthy to mention a series of studies that may look similar [32]–[34]. These studies aim to detect *users* who have been utilized in training a Face Recognition system through distances between users in the embedding space [32], [34] or via guided captioning [33]. While these studies are intriguing, they diverge significantly in terms of objectives and methodologies from both our proposed method and the broader MIAs and MINT research lines.

B. Reconstructing Training Data from AI Models

The goal of MIAs is to figure out the data used to train the AI model with some prior knowledge of the data. On the contrary, the research line in the present section is purely based on the network itself, without any prior knowledge of the data. Initial methods tried to find inputs that made intermediate neurons highly active. However, this often resulted in noisy data [35]. To tackle the noisy data problem, some approaches introduced prior knowledge or image generators [36] [37]. The drawback is that the image generators can lead to adversarial examples [38], [39]. The recent work by Haim et al. [40] represents a cutting-edge approach, showing very promising outcomes. The techniques described there can generate images resembling those used for training, offering insights into the utilized images, but they can not guarantee if specific images were part of the training set or not. It is worth noting other similar approaches like model-inversion, which aims to find representatives of each output class [41], [42].

C. Data Tracing

Methods described in previous sections can be seen as "passive methods" that rely on an already trained model without interfering with it. In contrast, the Data Tracing research direction falls under the category of "active methods", where the training data is directly altered to influence the model [43]. Data tracing involves changing the training data to identify if it was used during training. Several studies have

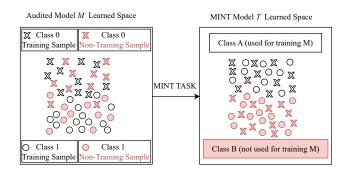


Fig. 2. The MINT task is represented graphically. On the left, we show the Audited Model Learned Space, illustrating two classes (0 and 1) each with samples used and not used in the Audited Model training. On the right, we present the feature space we aim to learn with our proposed MINT Model, where previous embeddings represented in the left plot will be projected so they become easily separable for the new binary classification task: used or not used (A or B) in the training of the Audited Model.

tried to tackled this challenge, demonstrating that with smart changes in the training data, it is feasible to detect its presence during inference [44], [45]. For instance, Radioactive Data [46] achieved this by inserting "radioactive marks" into the training data to detect if a database was used to train a model.

III. MEMBERSHIP INFERENCE TEST (MINT): PROPOSED APPROACH

MINT is a Membership Inference Test for a neural network's training data that aims to comply with current legislation [1], hence, assuming a certain level of collaboration from the model developer. The ultimate goal is to develop techniques for all data types (text, images, audio, etc.), taking into account different amounts of information provided by the model developer about the model training and operation. This spans from scenarios where no data information is provided (unsupervised) or very little (semi-supervised), up to when sufficient information about the training data is granted (supervised). Simultaneously, it considers whether you have access to the entire model (white-box) or merely to the model's output (black-box). In the present study, we analyze a broad range of scenarios, varying the quantity of data available for training the MINT model and contemplating both white-box and black-box settings.

Detecting the data used to train a model is a complex task in some sense opposite to the primary goal of common machine learning (ML) problems: generalize well during inference to samples unseen during learning, i.e., behaving similarly for both training and non-training samples. Consequently, in standard ML we expect samples of the same class to be proximate in the model representation space, regardless of their presence in the training set (Fig. 2). MIAs and MINT on the other hand, seek to distinguish samples (used or not for training the Audited Model M) that the model M intends to represent as close as possible. Hence, much of the existing MIA literature revolves around overfitted networks that do not generalize well [14], [19], [47]–[50].

In MINT, we assume the role of model auditors with access to the original model, or at least to partial information of it.

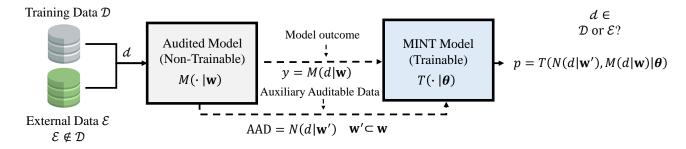


Fig. 3. The Membership Inference Test (MINT) Model (T) is trained to predict if a specific data sample (d) was used during the training process of an Audited AI/ML Model (M), which was previously trained with a database (\mathcal{D}) . The input of the MINT Model is AAD (e.g., activations maps for data samples d) and/or the model outcome obtained from M.

This is a consistent scenario either due to the legal considerations outlined in the present study (i.e., granting national authorities the possibility to request the model [1]), or as a result of developer transparency initiatives aimed at end-users. This is an important distinction from related works around MIAs, as here we do not need to train "shadow models" to simulate the original model's behavior. The "shadow models", while a useful approximation, operate in a space different from the original (Fig. 2 left), potentially imposing limitations on the achievable performance [18]. Instead, with access to the original model, we can directly apply our proposed approach, which may yield different detection results and methods for similar tasks.

A. Problem Statement and Terminology

Let us consider a Training Dataset (\mathcal{D}) , an External Dataset (\mathcal{E}) and a collection of samples d $(d \in \mathcal{D} \cup \mathcal{E})$. We assume a learned model (M) that is trained for a specific task (e.g., text generation, Face Recognition, etc.) using the dataset \mathcal{D} . For any input data record (d), the model (M) generates an outcome (y) based on d and a set of parameters (\mathbf{w}) learned during the training process, i.e., $y = M(d|\mathbf{w})$.

We hypothesize that an authorized auditor possesses access to model M, enabling him the acquisition of information regarding how M processes data d. Alternatively, the auditor may possess information detailing how M has processed data d, even in the absence of direct model access. This information comprises the generated Model Outcome $y = M(d|\mathbf{w})$ and if possible also some intermediate results (e.g., activation maps of specific layers in a neural network). These intermediate outcomes $N(d|\mathbf{w}')$ provide insights about a subset of parameters $\mathbf{w}' \subset \mathbf{w}$. We define these intermediate outcomes as Auxiliary Auditable Data (AAD). The auditor does not need the full description of the model M or the values of the complete set of parameters \mathbf{w} to obtain the AAD.

The aim of the proposed MINT is to determine if given data d was used to train the model M. To this end, an authorized auditor employs a collection of AAD and/or Model Outcomes g to train a MINT Model $T(\cdot|\theta)$ able to predict if a data sample g belongs to the Training Data g or External Data g (g g g). The proposed MINT models exploit the memorization capacity of machine learning processes. The key elements of MINT are defined bellow:

- Audited Model M: a learned model defined by an architecture and a set of parameters w.
- Training Data \mathcal{D} : collection of data used to train M.
- External Data \mathcal{E} : any data out of the collection (\mathcal{D}) .
- Model Outcome $y = M(d|\mathbf{w})$: final outcome of M that results from processing an input data d using the set of parameters \mathbf{w} .
- Auxiliary Auditable Data AAD = N(d|w'): intermediate outcomes of M that result from processing an input data d using a subset w' of the parameters w. The model outcome can be seen as the case where w' = w in which case N(d|w') = M(d|w).
- MINT Model T: a model defined by an architecture and a set of parameters θ trained using AAD $N(d|\mathbf{w}')$ and/or Model Outcomes $M(d|\mathbf{w})$ obtained from the two subsets of samples \mathcal{D} and \mathcal{E} .

To provide a better comprehension, we include in Fig. 3 the complete diagram of the system's operation.

B. MINT: Application to Face Recognition Models

In the previous Sec. III-A, we have introduced the general concepts of our MINT proposal, which can be applied to any type of Audited Model M. Next, we detail the various components of the proposal for the application of MINT to the challenging task of Face Recognition.

The data considered in the experimental framework are face images. On one side, we have the Training Data \mathcal{D} , and on the other side, the External Data \mathcal{E} .

- \mathcal{D} is the Face Recognition database considered for training the Audited Model M.
- E consists of images from external Face Recognition databases. We could add images unrelated to faces in E. However, we included images with similar characteristics to D in order to prevent the network from learning to detect database-dependent features of face images instead of focusing in the main task: database-independent detection of used or not for training. It is crucial that these images do not overlap with those in D, i.e., E ≠ D.

The audited model M is a Face Recognition model. These models are trained in such a way that they learn a feature space where faces of the same person are located close together, while faces of different individuals are separated. The output

of this model is not a classification vector but rather a feature embedding of the input face, within the learned feature space.

The AAD $N(d|\mathbf{w}')$ consists of intermediate outputs of the network using a subset \mathbf{w}' of the parameters \mathbf{w} . In this case, the Audited Model M is a Face Recognition model, which primarily consists of convolutional layers. These intermediate outputs, as an application example, comprise the activation maps generated when a face image is propagated through the network during the feedforward process [51], [52]. The Model Outcome (y) is the output of the entire Face Recognition model when a face is input. As we previously explained, this will be an embedding representing that face in the multidimensional space learned by the network.

C. Proposed MINT Models

The architecture of the MINT Model T can vary significantly based on the type of information used. First, let's focus on the format of the AAD and the Model Outcome. In the case study around face biometrics developed in the present paper, the format of activations as they pass through convolutional layers has a resolution of $H \times W$, which is influenced by padding and filter size of the convolutional layers. It also has a number of channels C, corresponding to the number of filters in that layer, which generally increases with deeper layers. On the other hand, the Model Outcome is simply a vector of size L, where L is the number of dimensions/features in the learned multidimensional space. We present two MINT Models:

- Vanilla MINT Model: We propose a MLP network to capture the training patterns within the vectorized information of the audited model (M). When it comes to the output embedding, it is already a vector of size L that can be analyzed by the MLP. For intermediate activations to be analyzed (convolutional layers), they need to be transformed into a vector. Based on existing literature [35], [53], data used for training maximizes the activations of intermediate neurons. Therefore, it seems reasonable to use the maximum value for each channel. obtaining C values for each group of activations. We have also explored other options beyond channel maximum were explored (e.g., mean value), but they yielded inferior results and thus were discarded. This way, we transform the activations from $H \times W \times C$ into a vector of size C, which can be analyzed by a MLP. If the owner of the model provides access to this information, activations and embeddings can be analyzed together or separately, depending on what yields better results. More details regarding the specific architecture of the Vanilla MINT Model are provided in Fig. 4 (a).
- Convolutional Neural Network MINT Model: We propose a CNN to capture the training patterns in the activation maps obtained from the convolutional layers of the audited model (M). Taking the maximum activations of each channel for MINT as done in the previous Vanilla MINT Model involves discarding a significant amount of information. Analyzing the full activations with a CNN is a more information-rich approach, which can potentially lead to better results. In Fig. 4 (b), we provide the implementation details of the proposed CNN MINT Model.

Using the AAD with dimensions $H \times W \times C$ extracted from model M, we train a CNN with an architecture specifically designed for this input format. The utilization of the Model Outcome, which comprises vectorial information, doesn't directly fit within the MINT CNN architecture, hence it is not used as input for T in this case, see Fig. 4 (b).

IV. EXPERIMENTAL FRAMEWORK

We first introduce the Face Recognition models and databases used in the present study. Then, we present the details regarding the experimental protocol.

A. Face Recognition Data and Models

We consider three popular Face Recognition models from the InsigthFace project [54]. The models used (M in Fig. 3) are:

- A ResNet-100 network [55], trained on the MS1Mv3 database [56] (D in Fig. 3) with ArcFace loss function [57]. The MS1Mv3 database comprises 5.2M images from 91K identities.
- 2) A ResNet-100 network, trained on the Glint360K database [58] (D) with CosFace loss function [59]. The Glint360K database comprises 17M images from 360K identities.
- A partial fully-connected ResNet-100 network [60], trained on the Glint360K database (D) with CosFace loss function.

As external data (\mathcal{E}) to train and test the MINT model T we use the following datasets: The IJB-C [61] (3.5K identities and 138K face images), FDDB [62] (5.2K face images), GANDiffFace [63], [64] (10K identities and 500K face images), and Adience [65] (2.2K identities 26.5K face images).

B. Experimental Protocol

Below, we elaborate on the diverse factors considered in this study to establish an equitable experimental protocol:

- 1) No identity overlap between Train and Evaluation, ensuring that performance is not influenced by identity recognition rather than the primary task of recognizing training and non-training data.
- 2) 4 different databases as External Data \mathcal{E} . Three of them (IJB-C, FDDB, GANDiffFace) are used to extract samples for training the MINT Model, while the fourth one (Adience) is used for evaluation. We will refer to this choice of databases as the "Baseline Case". Having either IJB-C or GANDiffFace, or both, in the training set is necessary to achieve balanced training (enough \mathcal{E} samples). If we include only GANDiffFace from these two databases in the training set, the network might learn to recognize specific patterns from synthetic images instead of focusing on the MINT task. Therefore, we always include IJB-C in the training set. However, the remaining three databases are interchangeable, allowing evaluation with GANDiffFace or FDDB and utilization

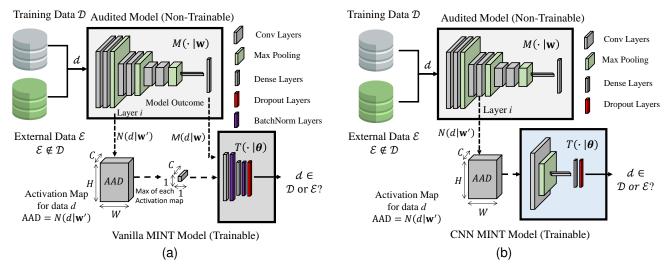


Fig. 4. Learning framework of the Vanilla MINT Model (a) and the CNN MINT Model (b) trained with the AAD obtained from the Convolutional Layer *i* and/or the model outcome if possible.

- of Adience in training. We refer to these two cases as "Case A" and "Case B", respectively. We present some results interchanging this evaluation database in our experiments. The Baseline case provides the most robust evaluation, as Case A involves evaluation with synthetic images, which may raise doubts due to the nature of these images, and Case B involves evaluation with only 5.2K images. Therefore, the majority of results will be presented under the "Baseline Case".
- 3) In the evaluation, there is an equal number of samples for each class (the same from \mathcal{D} and \mathcal{E}) to ensure balanced evaluation results. For example, as the Adience database E used for evaluation in "Baseline Case" has 26.5K images, we took the same number of images from D, leaving the rest of the D database available for training.
- 1) Data Pre-processing: We first apply the same face detector to the images of all databases. The face detector used can be found in the InsightFace project [54], and it is the SCRFD Detector [66]. This face detector is the same one used for the training of the Face Recognition models. With this face detector, we extract faces from the images, crop, align them, and then resize them to 112×112 . This ensures that all the images consist of a face in the center of the image, with the same orientation and size. This way, we establish similar conditions for all the images thereby promoting focus on the MINT task.
- 2) Auxiliary Auditable Data and Model Outcome: Once the data is preprocessed, we proceed with the feedforward pass through the Face Recognition network. We feed the face images from datasets \mathcal{D} and \mathcal{E} into the network M and extract the AAD and the Model Outcome. For the Model Outcome, we simply retrieve the output feature embedding. The AAD is derived from intermediate outputs of the model. Therefore, we must determine the depth within the model to obtain this AAD. The three Face Recognition models are based on the ResNet100 architecture as proposed by Han $et\ al.\ [55]$. This

architecture comprises four primary layers, each consisting of the designated Building Blocks. In this study, we collect the output of the final convolutional layer in the last Building Block of each of the four primary layers, resulting in four distinct AAD groups.

3) MINT Models Training: With the previously obtained AAD and Model Outcome, we propose two MINT models. Depending on how we handle the acquired information, we propose two architectures (Sec. III-C): a Vanilla MLP model and a CNN model. The CNN utilizes the activations of the convolutional blocks of M as-is, while the Vanilla model, as previously explained, applies max pooling per channel to convert the activations of the convolutional blocks into a vector that can be processed by the MLP.

The specific architecture of the Vanilla model consists of a dense layer with 128 neurons, ReLU activation, and an L1 regularizer with a value of 0.1. This is followed by a batch normalization layer, a dropout layer with a value of 0.5, and finally, the output layer with 1 neuron and sigmoid activation. The training algorithm consists of Adam optimizer with default settings and lr=0.001, training for 20 epochs with batches of 128 samples, each with an equal number of samples per class. Binary cross-entropy is employed as the loss function.

The CNN architecture consists of a convolutional layer with 64 filters, each of dimension C with a kernel = 5, and ReLU activation. The architecture is completed by a maxpooling layer and, finally, a fully connected layer with C neurons, ReLU activation, and a dropout of 0.5 followed by the output layer with 1 neuron and sigmoid activation. The training algorithm consists again of Adam optimizer with default settings and lr=0.001, training for 30 epochs with batches of 128 samples, each with an equal number of samples per class. Binary cross-entropy as the loss function.

Both architectures are chosen after extensive experimentation. It was found that the architectures did not need to be overly complex, as performance did not improve significantly with more complex architectures. In the experimental section (Section V-A), an experiment has been included to demonstrate the model's performance versus the network complexity.

V. EXPERIMENTS AND RESULTS

This section presents the results achieved by our proposed MINT approach. As it involves binary classification between \mathcal{D} and \mathcal{E} , and considers a balanced evaluation, the baseline accuracy (i.e., random guess) stands at 50%. Additionally, we remind that all results are derived from samples in \mathcal{E} that do not belong to the databases used in training the MINT Model, as elaborated in Sec. IV-B.

The nomenclature used in this section is as follows::

- FR Model (M in Fig. 3): These are the Face Recognition models presented in Sec. IV-A. The number corresponds to the one specified in that section.
- *Model Outcome:* Refers to the Output Embedding of the Face Recognition model in Sec. III-B.
- Conv Layer X: Represents the AAD obtained from each of the four primary layers, as explained in Sec. IV-B2
- *Combination:* Indicates the concatenation of the AAD obtained from the four primary layers.
- Training Data: Represents three possible scenarios based on the amount of data available for training our MINT Model T. The high scenario (100K) assumes a significant amount of data, with 50K data \mathcal{D} and 50K external data \mathcal{E} . In the medium scenario (50K), there are 25K data samples each for \mathcal{D} and \mathcal{E} . The low scenario (1K) considers only 500 data samples for both \mathcal{D} and \mathcal{E} . To put these numbers in context, 50K data samples from \mathcal{D} represent only 1% of the MS1Mv3 database used for training and 0.3% of Glint360k.

A. Results

Table I shows the classification accuracy results for the 3 Face Recognition models using different information to train the Vanilla MINT Model T (AAD and Model Outcomes). All of this is done under a scenario of high data available for training. In these results, we can observe a superiority in accuracy when combining all the AAD information (Combination). In terms of the AAD it is important to highlight that in FR Model 1, Conv Layer #2 yields the best results after Conv Layer #1, while in FR Model 2 and 3, Conv Layer #4 follows with better accuracy. This might be attributed to the fact that ResNet architectures are based on "Skip Connections", allowing information to traverse through the network from first layers to the last ones [67].

The results in this section pertain to the Baseline Case presented in Sec. IV-B, where the databases IJB-C, FDDB, and GANDiffFace are used in training the MINT Models, while Adience is used for evaluation. This is the case that provides a more robust evaluation (Sec. IV-B), and therefore the one mostly used throughout the experimental section. However, to demonstrate the robustness of the results, Table II showcases the same results as Table I with exchanged training and evaluation databases for the other two scenarios (see Sec. IV-B):

TABLE I

Classification accuracy for various AAD and Model Outcomes using the Vanilla MINT Model. MINT model trained with 100K samples (1% of the total Face Recognition Model training set for FR Model 1 and 0.3% for FR Models 2 and 3).

Auditable Data	FR MODEL 1	FR MODEL 2	FR MODEL 3
Conv Layer #1	0.77	0.80	0.78
Conv Layer #2	0.74	0.68	0.68
Conv Layer #3	0.68	0.59	0.62
Conv Layer #4	0.69	0.76	0.75
Model Outcome	0.75	0.78	0.76
Combination	0.84	0.84	0.81

TABLE II

Classification accuracy for various AAD and Model Outcomes using the Vanilla MINT Model. MINT model trained with 100K samples (1% of the total Face Recognition Model training set for FR Model 1 and 0.3% for FR Models 2 and 3). We display the results for Cases A and B (Case A/Case B in Table).

Auditable Data	FR MODEL 1	FR MODEL 2	FR MODEL 3
Conv Layer #1	0.84/0.78	0.83/0.75	0.86/0.75
Conv Layer #2	0.68/0.68	0.65/0.67	0.63/0.66
Conv Layer #3	0.54/0.53	0.51/0.58	0.53/0.59
Conv Layer #4	0.53/0.55	0.77/0.75	0.71/0.72
Model Outcome	0.73/0.64	0.86/0.79	0.84/0.74
Combination	0.88/0.78	0.87/0.82	0.90/0.80

- Case A: IJB-C, FDDB, and Adience are used for training, with GANDiffFace for evaluation.
- Case B: IJB-C, GANDiffFace, and Adience are used for training, with FDDB for evaluation.

As shown in Table II, the outcomes are slightly superior for Case A and slightly inferior for Case B compared to the Baseline Case, yet the conclusions remain unchanged (results are presented as Case A / Case B).

In Table III, we can observe the classification accuracy when we vary the data available to train the Vanilla MINT model for the best obtained Auditable Data (combination of all convolutional layers). Naturally, accuracy decreases when we reduce the number of training samples, demonstrating that the patterns to be learned are not trivial. A noteworthy result is that in the 50K scenario, accuracy decreases very slightly with respect to 100K. This result suggests that an intermediate number of training samples may be enough to achieve acceptable results. Another interesting result is what occurs in the 1K scenario. For FR Model 1, performance decreases by 18%. However, for FR Model 2 and FR Model 3, it decreases by only 8% and 6%, respectively. From these results, we can infer that depending on the model M (its training, architecture, loss function, etc.), it may be possible to develop our MINT Model T with more or less training data.

In Table IV, we can observe the different classification accuracy results for the 3 Face Recognition Models using the different information available to train the CNN MINT Model T under the high data scenario. Here, the CNN is trained using the full activation blocks of each convolutional layer. In this case, just like before, the best results are achieved with Conv Layer #1 (i.e., the layer closest to the image domain), progressively decreasing as the layer gets closer to the end of the network (i.e., the layer closest to the output domain). It is

TABLE III

CLASSIFICATION ACCURACY FOR THE VANILLA MINT MODEL (COMBINATION AS AUDITABLE DATA) IN THE VARIOUS DATA PROVIDED SCENARIOS DISCUSSED FOR THE THREE FACE RECOGNITION (FR) MODELS.

Training Data	FR Model 1	FR Model 2	FR Model 3
100K	0.84	0.84	0.81
50K	0.84	0.82	0.79
1K	0.66	0.76	0.75

TABLE IV

Classification accuracy for various AAD using the CNN MINT model. MINT model trained with 100K samples (1% of the total Face Recognition Model training set for FR model 1 and 0.3% for FR models 2 and 3).

Auditable Data	FR Model 1	FR Model 2	FR Model 3
Conv Layer #1	0.86	0.89	0.90
Conv Layer #2	0.85	0.86	0.86
Conv Layer #3	0.83	0.75	0.76
Conv Layer #4	0.73	0.74	0.73

important to note that, due to the different sizes of activation maps in the different layers ($H \times W$ depends on the depth at which you take the activation map), combining multiple layers to form the AAD in this CNN MINT would require a specific architecture or preprocessing, which is out of the scope of this paper. The Model Outcome is also not used, as explained in Sec. III-C, because, as we discussed there, it is vectorial information that does doesn't fit in a CNN architecture in a straightforward way.

In Table V, we observe the impact of varying the number of available data used to train the CNN MINT Model T for the best considered Auditable Data (Conv Layer #1). We see that for FR Model 1, a scenario with limited data for training the MINT Model negatively affect our results. However, for the other two FR Models, it has almost no effect on accuracy, achieving consistently high accuracy despite decreasing significantly the amount of training data. This indicates that for this MINT architecture depending on the model being audited, few samples (500 from \mathcal{D}) are enough to achieve good results. This result is very interesting as it implies that the auditor would require little cooperation from the developer.

To conclude we compare our proposed MINT model to the state of the art, in Table VI. As mentioned earlier, to the best of our knowledge, no state-of-the-art studies address the problem of detecting if given data was used for training a Face Recognition model. Therefore, making direct comparisons with other studies is not possible. Nevertheless, most studies around MIAs often use the classification output vector of the

TABLE V
CLASSIFICATION ACCURACY FOR THE CNN MINT MODEL (CONV LAYER
#1 AS AUDITABLE DATA) IN THE VARIOUS DATA PROVIDED SCENARIOS
DISCUSSED FOR THE THREE FACE RECOGNITION (FR) MODELS.

Training Data	FR Model 1	FR Model 2	FR Model 3
100K	0.86	0.89	0.90
50K	0.85	0.89	0.87
1K	0.73	0.87	0.87

TABLE VI

CLASSIFICATION ACCURACY COMPARISON BETWEEN MIA AND THE TWO MINT APPROACHES PROPOSED IN THIS WORK. *WE HAVE ADAPTED THE MIA APPROACH USED IN [18] USING THE FACE RECOGNITION MODEL OUTCOME. NOTE THAT WE USED THE ORIGINAL MODEL INSTEAD OF SHADOW MODELS. **WE USED THE VANILLA MINT MODEL BASED ON THE COMBINATION OF ALL CONVOLUTIONAL LAYERS. ***WE USED THE CNN MINT MODEL TRAINED WITH THE CONV LAYER #1.

Method	FR Model 1	FR Model 2	FR Model 3
MIA [18]*	0.75	0.78	0.76
MINT Vanilla**	0.84	0.84	0.81
MINT CNN***	0.86	0.89	0.90

network [14], [16], [28], [29], or the output from the layers just before this output [18], [30]. These approaches are somewhat analogous to our Model Outcome (the Model Outcome of a Face Recognition Model is similar to the layers just before the output vector in a classification model), and thus, we use them as a reference for interpreting the state of the art, despite the environmental differences between the works around MIAs and our proposed MINT.

Subsequently, we compare in Table VI the best published MIA results with the best results achieved in our present study. As observed, the methods proposed in the present study significantly outperform what could be considered the state of the art in Face Recognition MIAs, surpassing it for all three FR Models by more than 10%. Additionally, it is important to note that the results from the CNN MINT Model suggest even higher performance may be obtained if we combine information from all convolutional layers with specific architectures. Finally, Fig. 5 provides a comparison of the ROC curves for the three methods outlined in the table VI, along with the results for the Vanilla network with information from individual Conv Layers. An important leap is evident between MIA (purple curve) and our best result using the CNN MINT Model (black curve), which we have also observed in Table VI. One of the main criticisms discussed in [18] and [15] is that the evaluation protocol and metrics commonly used in MIA approaches are not adequate, and the results presented in previous studies suffer from a high False Positive Rate. In Fig. 5, we can see that this is not the case in our study.

As explained in Sec. IV-B3, an in-depth heuristic analysis (not fully reported here) has been conducted varying the complexity of the presented MINT Model architectures. From that study we can conclude that increasing the complexity of the architecture does not easily improve the performance. In our experiments, more complex architectures led to overfitting and a loss of performance in evaluation. To demonstrate this, Table VII presents results by varying the number of parameters of the MINT CNN architecture. We provide results with the original model as explained in Sec. IV-B3, another architecture with approximately three times the number of parameters, and the last one with ten times the number of parameters. We also include an experiment with a model three times simpler to demonstrate that the CNN requires enough complexity.

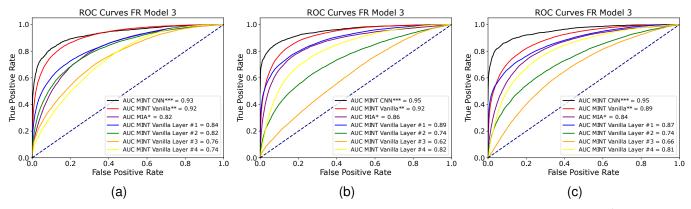


Fig. 5. ROC curves obtained of the different MINT approaches for the FR Model 1 (left), FR Model 2 (center), and FR Model 3 (right). *We have adapted the MIA approach used in [18] using the Face Recognition Model outcome. Note that we used the original model instead of shadow models. **We used the Vanilla MINT model based on the combination of all convolutional layers. ***We used the CNN MINT Model trained with the Conv Layer #1.

TABLE VII

CLASSIFICATION ACCURACY COMPARISON WHILE VARYING THE COMPLEXITY OF THE MINT CNN NETWORK USED. * CNN MINT ARCHITECTURE WITH $\div 3$ PARAMETERS. ** ORIGINAL MINT CNN ARCHITECTURE PRESENTED IN IV-B3. *** CNN MINT ARCHITECTURE WITH $\times 3$ PARAMETERS. **** CNN MINT ARCHITECTURE WITH $\times 10$ PARAMETERS.

Method	# Param.	FR Model 1	FR Model 2	FR Model 3
MINT CNN*	580	0.79	0.81	0.82
MINT CNN**	1.7K	0.86	0.89	0.90
MINT CNN***	5 K	0.86	0.88	0.90
MINT CNN****	18K	0.85	0.88	0.88

VI. DISCUSSION

It is important to contextualize the results achieved in the present study. While in Sec. V-A we provide a comparison with the state of the art in MIAs, this is for a simple interpretation as there is no specific work in the state of the art that aims at detecting if given data was used for training on the task of Face Recognition. In the state of the art of MIAs, results are available for classification tasks with very simple databases such as CIFAR-10 [68], where accuracy reaches up to 0.631 [69]. In the present study, we present results for the Face Recognition task. The complexity of this task is much higher, and with more relevant real-world application. Furthermore, the results achieved with the proposed MINT approach reach up to 90%, significantly surpassing those for CIFAR-10.

On the other hand, it is important to emphasize the context of our study and the results achieved. As explained in Sec. III, existing MIAs and our proposed MINT operate under different environmental conditions. These differences arise from the distinct application scenarios of both technologies. While MIA is designed as an attack, MINT is conceived as an auditing tool motivated on the recent legislation and aims for user transparency. As a result, MIAs require the creation of shadow models to simulate the behavior of the original model to attack, whereas in MINT we have direct access to the original model. This leads to divergent outcomes and the potential utilization of information in MINT that is not particularly useful in

MIA. For instance, the use of intermediate activations seems not to yield substantial improvements in MIAs [18], [30], [31]. Nevertheless, as demonstrated in our study, appropriately leveraging this information significantly enhances performance in MINT. This implies that ineffective methods in MIA do not necessarily perform poorly in MINT, suggesting opportunities for technology-specific advancements. As an example, exploring the use of gradients could be a promising avenue for future research in MINT.

VII. CONCLUSIONS

This paper has presented the Membership Inference Test (MINT), a novel approach to empirically assess whether specific data was used in the training of Artificial Intelligence (AI) models from an auditor's standpoint. Our research introduces two innovative architectures for modeling activation patterns: one based on Multilayer Perceptron (MLP) networks and the other on Convolutional Neural Networks (CNNs). These models have been rigorously evaluated in the challenging context of Face Recognition, using three state-of-the-art Face Recognition models and six publicly available databases comprising over 22 million face images.

Our findings demonstrate that our proposed MINT approach can achieve up to 90% accuracy in identifying whether a particular AI model has been trained with specific data. This is a significant advancement, especially considering the increasing demands for transparency and accountability in the use of AI technologies.

This work opens numerous avenues for future research. A direct path is to enhance the presented work by employing a different architecture for the MINT Model, such as using Visual Transformers or incorporating additional information like the gradients of the data. The study demonstrates the effectiveness of the MINT Model even with only 1K training samples; a subsequent step would be to explore unsupervised training and identify patterns in the activations in an unsupervised manner, eliminating the need for developer involvement. Another future direction could involve training the MINT Model alongside the Audited Model with a joint loss function to optimize performance for both tasks simultaneously.

This would necessitate active collaboration from the model developer, leading to what we might call Active Membership Inference Testing, as opposed to the Passive Membership Inference Testing presented in this work. Additional lines of research include applying this technology to other applications, such as image generation models [70], [71], or to different types of data, such as text [72], and studying MINT on widely available LLMs [73].

ACKNOWLEDGEMENT

This study has been supported by projects BBforTAI (PID2021-127641OB-I00 MICINN/FEDER), HumanCAIC (TED2021-131787B-I00 MICINN) and Cátedra ENIA UAM-VERIDAS en IA Responsable (NextGenerationEU PRTR TSI-100927-2023-2). The work of D. deAlcala is supported by a FPU Fellowship (FPU21/05785) from the Spanish MIU. A. Morales is supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with Universidad Autónoma de Madrid in the line of Excellence for the University Teaching Staff in the context of the V PRICIT (Regional Programme of Research and Technological Innovation). The work has been conducted within the ELLIS Unit Madrid.

REFERENCES

- T. Madiega, "Artificial intelligence act," European Parliament: European Parliamentary Research Service, vol. PE 698.792 (Updated June 2023), 2021.
- [2] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana, "SensitiveNets: Learning agnostic representations with application to Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2158–2164, June 2021.
- [3] S. Yu, F. Carroll, and B. L. Bentley, "Insights into privacy protection research in AI," *IEEE Access*, vol. 12, pp. 41704–41726, 2024.
- [4] Y. Wang, J. Wan, J. Guo, Y.-M. Cheung, and P. C. Yuen, "Inference-Based similarity search in randomized montgomery domains for privacy-preserving biometric identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 7, pp. 1611–1624, 2018.
- [5] A. Hassanpour, M. Moradikia, B. Yang, A. Abdelhadi, C. Busch, and J. Fierrez, "Differential privacy preservation in Robust Continual Learning," *IEEE Access*, vol. 10, February 2022.
- [6] M. Gomez-Barrero, J. Galbally, A. Morales, and J. Fierrez, "Privacy-Preserving comparison of Variable-Length Data with application to Biometric Template Protection," *IEEE Access*, vol. 5, pp. 8606–8619, June 2017.
- [7] K. Manheim and L. Kaplan, "Artificial intelligence: Risks to privacy and democracy," Yale JL & Tech., vol. 21, p. 106, 2019.
- [8] A. Peña, I. Serna, A. Morales, J. Fierrez, A. Ortega, A. Herrarte, M. Alcantara, and J. Ortega-Garcia, "Human-centric multimodal Machine Learning: Recent advances and testbed on AI-based recruitment," SN Computer Science, vol. 4, no. 5, p. 434, 2023.
- [9] P. Delgado-Santos, G. Stragapede, R. Tolosana, R. Guest, F. Deravi, and R. Vera-Rodriguez, "A survey of privacy vulnerabilities of mobile device sensors," ACM Computing Surveys, vol. 54, no. 11s, pp. 1–30, 2022.
- [10] Y. Rong, T. Leemann, T.-T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, and E. Kasneci, "Towards Human-Centered explainable AI: A survey of user studies for model explanations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2104–2122, 2024.
- [11] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.
- [12] J. Tello, M. de la Cruz, T. Ribeiro, J. Fierrez, A. Morales, R. Tolosana, C. L. Alonso, and A. Ortega, "Symbolic AI (LFIT) for XAI to handle biases," in *Proceedings of the European Conference on Artificial Intelligence Workshops*, vol. 3523, October 2023.

- [13] M. Jegorova, C. Kaul, C. Mayor, A. Q. O'Neil, A. Weir, R. Murray-Smith, and S. A. Tsaftaris, "Survey: Leakage and privacy at inference time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 9090–9108, 2023.
- [14] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks against Machine Learning models," in *Proceedings* of IEEE Symposium on Security and Privacy, 2017, pp. 3–18.
- [15] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership Inference Attacks from first principles," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2022, pp. 1897–1914.
- [16] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-Leaks: Model and data independent Membership Inference Attacks and defenses on Machine Learning models," in *Proceedings of the Annual Network and Distributed System Security Symposium*, 2018.
- [17] K. Leino and M. Fredrikson, "Stolen memories: Leveraging model memorization for calibrated White-Box membership inference," in *Pro*ceedings of the USENIX Security Symposium, 2020, pp. 1605–1622.
- [18] S. Rezaei and X. Liu, "On the difficulty of Membership Inference Attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7892–7900.
- [19] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership Inference Attacks on Machine Learning: A survey," ACM Computing Surveys, vol. 54, no. 11s, pp. 1–37, 2022.
- [20] Y. Lyu, Y. Jiang, Z. He, B. Peng, Y. Liu, and J. Dong, "3D-Aware adversarial makeup generation for facial privacy protection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13438–13453, 2023.
- [21] P. Melzi, H. O. Shahreza, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, J. Fierrez, S. Marcel, and C. Busch, "Multi-IVE: Privacy enhancement of multiple Soft-Biometrics in Face Embeddings," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, January 2023, pp. 323–331.
- [22] I. Serna, A. Morales, J. Fierrez, and N. Obradovich, "Sensitive Loss: Improving accuracy and fairness of Face Representations with Discrimination-Aware Deep Learning," *Artificial Intelligence*, vol. 305, p. 103682, April 2022.
- [23] European Parliament and the Council of the European Union. (2016) General Data Protection Regulation GDPR. EU 2016/679. [Online]. Available: https://gdpr-info.eu/
- [24] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only Membership Inference Attacks," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 1964–1974.
- [25] A. Pyrgelis, C. Troncoso, and E. De Cristofaro, "Knock knock, who's there? Membership inference on aggregate location data," in *Proceedings* of the Network and Distributed Systems Security Symposium, 2018.
- [26] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Demystifying Membership Inference Attacks in Machine Learning as a service," *IEEE Transactions on Services Computing*, vol. 14, no. 6, pp. 2073–2089, 2019.
- [27] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, and C. Busch, "An overview of privacy-enhancing technologies in biometric recognition," ACM Computing Surveys, 2024.
- [28] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in Machine Learning: Analyzing the connection to overfitting," in *Proceedings of the IEEE Computer Security Foundations Symposium*, 2018, pp. 268–282.
- [29] L. Song and P. Mittal, "Systematic evaluation of privacy risks of Machine Learning models," in *Proceedings of the USENIX Security Symposium*, 2021, pp. 2615–2632.
- [30] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of Deep Learning: Passive and active white-box inference attacks against Centralized and Federated Learning," in *Proceedings of* the IEEE Symposium on Security and Privacy, 2019, pp. 739–753.
- [31] A.-M. Cretu, D. Jones, Y.-A. de Montjoye, and S. Tople, "Re-aligning shadow models can improve white-box Membership Inference Attacks," arXiv preprint arXiv:2306.05093, 2023.
- [32] M. Chen, Z. Zhang, T. Wang, M. Backes, and Y. Zhang, "FACE-AUDITOR: Data auditing in Facial Recognition systems," in *Proceedings of the USENIX Security Symposium*, 2023, pp. 7195–7212.
- [33] D. Hintersdorf, L. Struppek, M. Brack, F. Friedrich, P. Schramowski, and K. Kersting, "Does CLIP know my face?" arXiv preprint arXiv:2209.07341, 2022.
- [34] G. Li, S. Rezaei, and X. Liu, "User-level Membership Inference Attack against metric embedding learning," in *Proceedings of the International Conference on Learning Representations*, 2022.

- [35] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.
- [36] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE/CVF Conference* on Computer Vision and Pattern Recognition, 2015, pp. 5188–5196.
- [37] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," in *Proceedings of the International Coference on Machine Learning*, 2015.
- [38] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," stat, vol. 1050, p. 20, 2015.
- [39] M. Ghafourian, J. Fierrez, L. F. Gomez, R. Vera-Rodriguez, A. Morales, Z. Rezgui, and R. Veldhuis, "Toward Face Biometric De-identification using Adversarial Examples," in *IEEE Conference on Computers, Soft*ware, and Applications (COMPSAC), June 2023, pp. 723–728, also presented at AAAI Workshop on Artificial Intelligence for Cyber Security, AICS 2023.
- [40] N. Haim, G. Vardi, G. Yehudai, O. Shamir, and M. Irani, "Reconstructing training data from trained neural networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22911–22924, 2022.
- [41] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1322–1333.
- [42] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, "Dreaming to distill: Data-free knowledge transfer via deepinversion," in *Proceedings of the IEEE/CVF Conference* on Computer Vision and Pattern Recognition, 2020, pp. 8715–8724.
- [43] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Madry, B. Li, and T. Goldstein, "Dataset security for Machine Learning: Data poisoning, backdoor attacks, and defenses," *IEEE Transactions* on Pattern Analysis and Machine Intelligence, vol. 45, no. 2, pp. 1563– 1580, 2023.
- [44] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against Support Vector Machines," in *Proceedings of the International Coference on Machine Learning*, 2012, pp. 1467–1474.
- [45] J. Steinhardt, P. W. W. Koh, and P. S. Liang, "Certified defenses for data poisoning attacks," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [46] A. Sablayrolles, M. Douze, C. Schmid, and H. Jegou, "Radioactive data: tracing through training," in *Proceedings of the International Conference* on *Machine Learning*, vol. 119, 13–18 Jul 2020, pp. 8326–8335.
- [47] S. M. Tonni, D. Vatsalan, F. Farokhi, D. Kaafar, Z. Lu, and G. Tangari, "Data and model dependencies of Membership Inference Attack," arXiv preprint arXiv:2002.06856, 2020.
- [48] P. Irolla and G. Châtel, "Demystifying the Membership Inference Attack," in *Proceedings of the IEEE CMI Conference on Cybersecurity* and Privacy, 2019, pp. 1–7.
- [49] H. Hu, Z. Salcic, G. Dobbie, Y. Chen, and X. Zhang, "EAR: An enhanced adversarial regularization approach against Membership Inference Attacks," in *Proceedings of the IEEE International Joint Con*ference on Neural Networks, 2021, pp. 1–8.
- [50] Y. Kaya and T. Dumitras, "When does data augmentation help with Membership Inference Attacks?" in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 5345–5355.
- [51] I. Serna, A. Peña, A. Morales, and J. Fierrez, "InsideBias: Measuring Bias in deep Networks and application to Face Gender Biometrics," in IAPR International Conference on Pattern Recognition, January 2021, pp. 3720–3727.
- [52] I. Serna, D. DeAlcala, A. Morales, J. Fierrez, and J. Ortega-Garcia, "IFBiD: Inference-Free Bias Detection," in AAAI Workshop on Artificial Intelligence Safety, ser. CEUR-WS, vol. 3087, February 2022.
- [53] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter, "Zoom in: An introduction to circuits," *Distill*, vol. 5, no. 3, pp. e00 024–001, 2020.
- [54] InsightFace Team. (2023) InsightFace Models. [Online]. Available: https://insightface.ai/
- [55] D. Han, J. Kim, and J. Kim, "Deep pyramidal residual networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 5927–5935.
- [56] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for Large-Scale Face Recognition," in *Proceedings of* the European Conference on Computer Vision, 2016, pp. 87–102.
- [57] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for Deep Face recognition," in *Proceedings of the IEEE/CVF* Conference on Computer Vision and Pattern Recognition, 2019.

- [58] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang et al., "Partial FC: Training 10 million identities on a single machine," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1445–1449.
- [59] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.
- [60] X. An, J. Deng, J. Guo, Z. Feng, X. Zhu, Y. Jing, and L. Tongliang, "Killing two birds with one stone: Efficient and robust training of Face Recognition CNNs by partial FC," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [61] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother, "IARPA Janus Benchmark C: Face dataset and protocol," in *Proceedings of the International Conference on Biometrics*, 2018, pp. 158–165.
- [62] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," UMass Amherst technical report, Tech. Rep., 2010.
- [63] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, D. Lawatsch, F. Domin, and M. Schaubert, "GANDiffFace: Controllable generation of synthetic datasets for Face Recognition with realistic variations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2023.
- [64] P. Melzi, R. Tolosana, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, I. DeAndres-Tame, A. Morales, J. Fierrez, J. Ortega-Garcia et al., "FRCSyn-onGoing: Benchmarking and comprehensive evaluation of real and synthetic data to improve face recognition systems," *Information Fusion*, vol. 107, p. 102322, 2024.
- [65] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.
- [66] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, "Sample and computation redistribution for efficient face detection," in *Proceedings of the International Conference on Learning Representations*, 2021.
- [67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [68] A. Krizhevsky, G. Hinton et al., "Learning multiple layers of features from tiny images," *Toronto, ON, Canada*, 2009.
- [69] L. Watson, C. Guo, G. Cormode, and A. Sablayrolles, "On the importance of difficulty calibration in Membership Inference Attacks," Proceedings of the International Conference on Learning Representations, 2022.
- [70] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proenca, and J. Fierrez, GAN Fingerprints in Face image synthesis. Springer, April 2022, ch. GAN Fingerprints in Face Image Synthesis, pp. 175–204.
- [71] A. Hassanpour, F. Jamalbafrani, B. Yang, K. Raja, R. Veldhuis, and J. Fierrez, "E2F-Net: Eyes-to-face inpainting via StyleGAN latent space," *Pattern Recognition*, vol. 152, p. 110442, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320324001936
- [72] A. Peña, A. Morales, J. Fierrez, J. Ortega-Garcia, I. Puente, J. Cordova, and G. Cordova, "Continuous document layout analysis: Human-in-the-loop AI-based data curation, database, and evaluation in the domain of public affairs," *Information Fusion*, vol. 108, p. 102398, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253524001763
- [73] I. DeAndres-Tame, R. Tolosana, R. Vera-Rodriguez, A. Morales, J. Fier-rez, and J. Ortega-Garcia, "How good is ChatGPT at face biometrics? A first look into recognition, soft biometrics, and explainability," *IEEE Access*, vol. 12, pp. 34390–34401, 2024.



Daniel DeAlcala , a research professional, obtained his B.Sc. in Telecommunications Engineering from Universidad Autonoma de Madrid (UAM) in 2020, graduating with outstanding academic achievements. Subsequently, he pursued a Master's degree in Deep Learning, which he successfully completed in 2021. In 2022, Daniel embarked on his Ph.D. journey by joining the esteemed Biometrics and Data Pattern Analytics Laboratory (BiDA Lab) at UAM. His research primarily centers around Fair and Transparent AI and innovative architectural developments.

Daniel has presented his work at prestigious conferences, including CVPR, and continues to make significant contributions to the field of AI.



Aythami Morales received his M.Sc. degree in Electrical Engineering in 2006 from Universidad de Las Palmas de Gran Canaria. He received his Ph.D. degree in Artificial Intelligence from La Universidad de Las Palmas de Gran Canaria in 2011. He performs his research works in the BiDA Lab – Biometric and Data Pattern Analytics Laboratory at Universidad Autónoma de Madrid, where he is currently an Associate Professor (CAM Lecturer Excellence Program). He is a member of the ELLIS Society (European Laboratory for Learning and Intelligent

Systems). He has performed research stays at the Biometric Research Laboratory at Michigan State University, the Biometric Research Center at Hong Kong Polytechnic University, the Biometric System Laboratory at the University of Bologna, and Schepens Eye Research Institute (Harvard Medical School). He is the author of more than 100 scientific articles published in international journals and conferences and 2 patents.



Ruben Tolosana received the M.Sc. degree in Telecommunication Engineering, and the Ph.D. degree in Computer and Telecommunication Engineering, from Universidad Autonoma de Madrid, in 2014 and 2019, respectively. In 2014, he joined the Biometrics and Data Pattern Analytics – BiDA Lab at the Universidad Autonoma de Madrid, where he is currently an Assistant Professor. He is a member of the ELLIS Society, the Technical Area Committee of EURASIP, and the Editorial Board of the IEEE Biometrics Council Newsletter. His research interests

are mainly focused on signal and image processing, pattern recognition, and machine learning, particularly in the areas of DeepFakes, Human-Computer Interaction, Biometrics, and Health. He is the author of more than 80 scientific articles published in international journals and conferences. He has served as General Chair and Program Chair (AVSS 2022), and Area Chair (IJCB 2023, ICPR 2022) in top conferences. Dr. Tolosana has also received several awards such as the European Biometrics Industry Award (2018) from the European Association for Biometrics (EAB) and the Best Ph.D. Thesis Award in 2019-2022 from the Spanish Association for Pattern Recognition and Image Analysis (AERFAI).



Gonzalo Mancera earned his B.Sc. in Telecommunications Engineering from Universidad Autónoma de Madrid (UAM) in 2021. Following this, he pursued a Master's degree in Deep Learning fir Audio and Video, which he completed in 2022. In 2023, Gonzalo commenced his Ph.D at the renowned Biometrics and Data Pattern Analytics Laboratory (BiDA Lab) at UAM. His research initially focused on Fair and Transparent AI and pioneering architectural advancements, and he has since transitioned to working in Natural Language Processing (NLP).



Julian Fierrez received the M.Sc. and the Ph.D. degrees in telecommunications engineering from Universidad Politecnica de Madrid, Spain, in 2001 and 2006, respectively. Since 2002 he has been affiliated as a PhD candidate with the Universidad Politecnica de Madrid, and since 2004 at Universidad Autonoma de Madrid, where he is currently a Full Professor since 2022. From 2007 to 2009 he was a visiting researcher at Michigan State University in the USA under a Marie Curie fellowship. Since 2016 he has been Associate Editor for Elsevier's Information

Fusion and IEEE Trans. on Information Forensics and Security, and since 2018 also for IEEE Trans. on Image Processing. He has been the General Chair of the IAPR Iberoamerican Congress on Pattern Recognition (CIARP 2018) and the Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2019). He is also the recipient of a number of world-class research distinctions, including: the EBF European Biometric Industry Award 2006, EURASIP Best PhD Award 2012, Medal in the Young Researcher Awards 2015 by the Spanish Royal Academy of Engineering, and the Miguel Catalan Award. He is an ELLIS Member since 2020



Javier Ortega-Garcia received the M.Sc. degree in electrical engineering and the Ph.D. degree (cum laude) in electrical engineering from Universidad Politecnica de Madrid, Spain, in 1989 and 1996, respectively. He is currently a Full Professor at the Signal Processing Chair at Universidad Autonoma de Madrid - Spain, where he holds courses on biometric recognition and digital signal processing. He is a founder and Director of the BiDA-Lab, Biometrics and Data Pattern Analytics Group. He has authored over 300 international contributions,

including book chapters, refereed journals, and conference papers. His research interests are focused on biometric pattern recognition (on-line signature verification, speaker recognition, human-device interaction) for security, ehealth, and user profiling applications. He chaired Odyssey-04, The Speaker Recognition Workshop, ICB-2013, the 6th IAPR International Conference on Biometrics, and ICCST2017, the 51st IEEE International Carnahan Conference on Security Technology.