IEEE Open Journal of the Computer Society

Received 17 March 2025; revised 6 May 2025; accepted 17 May 2025. Date of publication 21 May 2025; date of current version 6 June 2025. The review of this article was arranged by Associate Editor Xiang Sun.

Digital Object Identifier 10.1109/OJCS.2025.3572244

Membership Inference Attacks and Differential Privacy: A Study Within the Context of Generative Models

BORJA ARROYO GALENDE [1], PATRICIA A. APELLÁNIZ [1], JUAN PARRAS [1], SANTIAGO ZAZO [1], AND SILVIA URIBE [2]

[1]Information Processing and Telecommunications Center, Escuela Tecnica Superior de Ingeniería de Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain

[2]Escuela Técnica Superior de Ingeniería de Sistemas Informáticos, Universidad Politécnica de Madrid, 28040 Madrid, Spain

CORRESPONDING AUTHOR: BORJA ARROYO GALENDE (e-mail: borja.arroyog@upm.es).

This work was supported by the Synthetic Generation of Hematological Data over Federated Computing Frameworks (SYNTHEMA) Project from European Union's Horizon Europe under Grant 101095530.

ABSTRACT Membership attacks pose a major issue in terms of secure machine learning, especially in cases in which real data are sensitive. Models tend to be overconfident in predicting labels from the training set. Nevertheless, its application has traditionally been limited to supervised models, while in the case of generative models we have found that there is a lack of theoretical foundations to bring this concept into the scene. Hence, this article provides the theoretical background in the context of membership inference attacks and their relationship to generative models, including the derivation of an evaluation metric. In addition, the link between these types of attack and differential privacy is shown to be a particular case. Lastly, we empirically show through simulations the intuition and application of the concepts derived.

INDEX TERMS Generative AI, computer security, machine learning, private machine learning, differential privacy.

I. INTRODUCTION

Synthetic data generation is a topic that is attracting more and more attention. Generative models have already shown great applicability in fields where data is abundant, such as language and vision [22], [30]. Furthermore, this trend has successfully reached fields where data have confidential or sensitive restrictions, in which synthetic data provide a valuable alternative for downstream applications [9], [14], and even for data anonymisation [17].

However, the scarcity of data in some domains poses several challenges in ensuring the quality and confidentiality of synthetic samples [11]. Therefore, synthetic data should be evaluated in terms of fidelity, privacy, and utility [19]. This work focusses on the security of synthetic data. The broader concept of security in machine learning, based on the taxonomy proposed in [4], involves three different types of attack:

1) Attacks against integrity. The system malfunctions without noticing.
2) Attacks against a system's availability. The system does not respond to requests from legitimate users.
3) Attacks against privacy and confidentiality. The attacker's objective is to obtain private information about the system, its users, or data.

Based on this classification, the context for our work arises around the latter, where privacy should be preserved while guaranteeing the utility of synthetic data for downstream tasks. From the privacy and confidentiality perspective (which we will shorten as privacy from now on), the survey [23] provides a comprehensive study. They define a threat model that involves several actors and assets, including information about the model and training data. Therefore, attacks can be considered as black box, where the attacker has no knowledge of the assets; and white box, where the adversary has complete access to the victim's model parameters or learning updates [12]. Between these two, there is a wide range of possibilities where the attacker does not have full access to the assets, but rather knows partial information.

In addition to classification in terms of adversary knowledge, there are alternative typologies that focus on the target of the attacker. Common attacks are membership inference attacks [24] and reconstruction attacks [25], which are related to the joint distribution of training data. This work focusses on the former, membership inference attacks (MIA), which attempt to determine whether an input sample was used to train a model. This type of attack can also be viewed from the perspective of data audit to determine whether a sample has been used without authorisation [20].

MIA has been used to disclose information for a wide variety of machine learning models, including, but not limited to, traditional machine learning or deep learning models used for classification and regression [16]. However, this work focusses on its application to generative models, in which the model can be seen as the probability distribution of the real data. The need for a privacy analysis in generative models is currently a topic of importance due to the abuse of existing online data [10], used to train, among others, large language models[1]. Privacy mechanisms could play a vital role in future policies on massive models. In that regard, we propose a unified theoretical framework to express MIA, which is particularised for several well-established methods.

II. OBJECTIVES AND CONTRIBUTIONS

The goal of this article is to present a unified theoretical framework, derived from Bayesian analysis, that encompasses the definition of MIA and its relationship with several state-of-the-art approaches. More specifically, this work presents the following.

- A Bayesian derivation for an expression to evaluate MIA risk. Note that this work extends the proof from [6] for black-box attacks. This derivation includes the actual metric we propose and the theoretical framework behind the metric, which is missing in the state-of-the-art.
- The relationship between the metric proposed with several particular cases from the state of the art whose aim is to estimate MIA as well. For that, we demonstrate how our metric provides a general common background from which the rest of metrics can be derived.
- A novel definition of differential privacy (DP) in the context of generative models and its relationship to MIA. In contrast to other methods with no theoretical guarantees, we propose a theoretical framework for DP that is directly linked to synthetic data. Differentially private stochastic gradient descent is a well-known example in which the DP framework is applied to the neural network training algorithm but with no guarantee of DP constraints on the output.
- A detailed study of the implications of the assets including their impact on the metric by running simulations. In this article, the term asset includes the prior information, the reference set, the models, etc.

[1]As an example, Meta has been sued for infringing on property rights for downloading books with copyrights to train their models: https://www.reuters.com/technology/artificial-intelligence/french-publishers-authors-file-lawsuit-against-meta-ai-case-2025-03-12

TABLE 1. Summary of Mathematical Notation

Symbol	Explanation
$I(\cdot)$	Inference problem representing the MIA risk
$p(\cdot)$	Probability distribution
$\sigma(\cdot)$	Sigmoid transformation
θ	Parameters of a (generative) model
θ_v	Parameters of the victim model threatened by an attacker
θ_r	Parameters of a reference model with similar properties. to the victim model trained on a reference set
x_q	Query sample as stated in the literature of MIA
m_q	Membership boolean of the query sample
S	Set with all data samples including the training and reference subsets
S_t	Train set used to train the generative model
S_{-q}	Train set excluding a query sample
$f(\cdot)$	Monotonic transformation
$d(\cdot,\cdot)$	General expression for a distance metric
χ	Magnitude of the perturbation in DPSGD
ω	Letter used to establish an upper bound in the metric

III. MATHEMATICAL CONCEPTS AND NOTATION

This section provides an overview of all the mathematical concepts included within the text. Table 1 collects these concepts and provides a brief explanation of each.

Throughout this text, two key mathematical objects appear frequently: the training set and the reference set. Both are made up of real samples and are mutually exclusive subsets of S. These sets play a crucial role in the contexts of MIA and DP, as detailed below.

1) MIA context:
   - The victim model, denoted by its parameters θ_v, is trained using the training set.
   - The reference model, denoted by its parameters θ_r, is trained using the reference set.
   - The reference model provides a baseline for the attacker, allowing density comparisons between synthetic and real data. High-density regions in the synthetic domain must be calibrated, as they may overlap with high-density regions in real data. Thus, if high-density regions coincide in both domains, there is no reason to suspect information leakage.
2) DP context:
   - DP builds on the MIA framework by refining the definition of the reference set.
   - Specifically, the reference model θ_r is redefined as $\theta_{v'}$, the model trained on S_{-q} (i.e., the training set excluding a query sample x_q).
   - This formulation expresses DP as a special case of MIA, where the reference set is a neighbouring dataset, i.e. two datasets are neighbouring if they differ in a single sample. In this case, the difference is established so that the reference set omits one sample from the training set, that is, the query.

IV. METHODOLOGY

A. PROBLEM STATEMENT AND ANALYSIS

MIA can be defined as the inference problem of estimating the boolean membership of a given sample, usually denoted as a query, given the knowledge an attacker has of the environment. This inference problem (I) is stated in (1), where $\mathcal{M}$ represents the knowledge that the attacker has about the victim model, and S_t refers to the data set used to train the model. The term knowledge is expressed through the parameters of the model constructed by the attacker, θ_v. In that equation, the subindex q denotes the query; hence x_q is the sample and $m_q \in \{0, 1\}$ is the membership boolean.

$$I(\theta_v, x_q) = p(x_q \in S_t|\mathcal{M}) = p(m_q = 1|\theta_v, x_q) \tag{1}$$

Let us define the following sets that will be used subsequently, being $S_{-q} \subset S_t \subset S$.

$$S = (x_i, m_i)_{i=1}^{N}$$
$$S_t = \{x_i \in S | m_i = 1\}$$
$$S_{-q} = \{s_i \in S_t | i \neq q\}$$

These sets denote, respectively, all real data (where N denotes the total number of samples), the real data used for training (i.e., those with $m_i = 1$), and the training set used for training except for a single query sample (i.e., S_t except the query x_q). Hence, the membership probability can be expressed as

$$\begin{aligned} p(m_q = 1|\theta_v, x_q) &= \int p(m_q = 1, S_{-q}|\theta_v, x_q) dS_{-q} \\ &= \int p(m_q = 1|\theta_v, x_q, S_{-q}) p(S_{-q}) dS_{-q} \end{aligned} \tag{2}$$

This expression can be manipulated using Bayes' theorem (see Appendix A) to obtain the canonical equation for the MIA problem (3), in which σ is the sigmoid function.

$$\begin{aligned} &p(m_q = 1|\theta_v, x_q) \\ &= \int \sigma\left(\ln \frac{p(\theta_v|m_q = 1, x_q, S_{-q}) p(m_q = 1)}{p(\theta_v|m_q = 0, x_q, S_{-q}) p(m_q = 0)}\right) \\ &\quad p(S_{-q}) dS_{-q} \end{aligned} \tag{3}$$

Let us assume that the posterior distribution of θ_v is proportional to the exponential function, while the prior over θ_v is uniform in its domain, then we can derive (4) (see Appendix B).

$$\begin{aligned} &p(m_q = 1|\theta_v, x_q) \\ &= \int \sigma\Bigg[\left(-d(x_q, \theta_v)\right) \\ &\quad - \ln \int \exp\left(-d(x_q, \phi_v)\right) p(\phi_v|S_{-q}) d\phi_v \\ &\quad + \ln \frac{p(m_q = 1)}{p(m_q = 0)}\Bigg] p(S_{-q}) dS_{-q} \end{aligned} \tag{4}$$

Given a well-known approximation for the expression within the integral (delta approximation), we can further simplify the equation to obtain the desired metric for MIA (5) (see Appendix C for details).

$$I(\theta_v, x_q) \approx \sigma\left[\ln \frac{p(x_q|\theta_v)}{p(x_q|\theta_r)} + \ln \frac{p(m_q = 1)}{p(m_q = 0)}\right] \tag{5}$$

In this equation, $p(x_q|\theta)$ represents the likelihood of x_q with respect to the victim model θ_v or with respect to the reference model θ_r. As we treat estimates as actual probabilities, the final expression must be transformed into the interval [0, 1], so that both the estimate and the error are bounded.

Furthermore, the transformation $\sigma[\ln(\cdot)]$ can be generalised to other monotonic transformations f. However, since $I(\theta_v, x_q)$ is an actual probability, f must satisfy that it is bounded in the interval [0, 1]. Lastly, note that the last term of the final expression can be cancelled out in the case of an uninformative uniform prior, that is,

$$f\left(\frac{p(x_q|\theta_v)}{p(x_q|\theta_r)} \cancelto{1}{\frac{p(m_q = 1)}{p(m_q = 0)}}\right) = f\left(\frac{p(x_q|\theta_v)}{p(x_q|\theta_r)}\right)$$

B. EXISTENT APPROACHES AS PARTICULARISATONS

In the state of the art, there are several approaches to estimate MIA that are particular cases of our metric. This section links our general framework to each of them.

An initial exploration of the topic is given in [15], where they approximated MIA with the expression

$$f\left(\frac{p(x_q|\theta_v)}{p(x_q|\theta_r)}\right) \approx \frac{p(x_q|\theta_v)}{1} \approx \frac{1}{Nh} \sum_{i=1}^{n} K\left(\frac{x_q - g_i}{h}\right)$$

where they denoted g_i as each synthetic sample obtained from the generative model. Note that this approach employs a kernel (K) to build a density estimate (kernel density estimate) to compute the density for the victim model. In addition, they ignored the effect of the denominator, as the density for the reference model was not treated in their analysis. In contrast, we take into account the effect of the denominator through a reference set.

Moreover, in [13], the approximation was driven by a classifier/discriminator such that

$$f\left(\frac{p(x_q|\theta_v)}{p(x_q|\theta_r)}\right) \approx \frac{p(x_q|\theta_v)}{p(x_q|\theta_r)} = r(x_q) \approx r_\psi(x_q)$$

where $r(x)$ is the density ratio evaluated at x_q, which is approximated with the parameters ψ as can be seen in [26]. In this way, they parameterized a complex density by relying on a neural network to approximate it. In contrast to our metric, they did not provide an actual probability, which limits the expressiveness of their metric.

Furthermore, [6] and [7] expressed the dependency with respect to the parameters via the data itself, that is,

$$f\left(\frac{p(x_q|\theta_v)}{p(x_q|\theta_r)}\right) \approx -\ln\left(\frac{\exp\left(-d(x_q, \theta_v)\right)}{\exp\left(-d(x_q, \theta_r)\right)}\right)$$

$$
\begin{aligned}
&= d(x_q, \theta_v) - d(x_q, \theta_r) \\
&\approx \min_{x_v}\{d(x_q, x_v) | x_v \in S_v\} \\
&\quad - \min_{x_r}\{d(x_q, x_r) | x_r \in S_r\}
\end{aligned}
$$

where S_v and S_r denote, respectively, a data set produced with the victim model and a data set produced by a reference model, and where we assume that the normalisation constant ratio is approximately one. In this case, they proposed a quantile to summarise the whole distribution, which may limit the expressiveness of their analysis.

Lastly, [27] expressed the inference problem as [13], that is, with the ratio of densities. The main difference between both is that in this case they estimated this ratio by first estimating both densities either through a kernel density estimate or via normalising flows.

Hence, the approaches [27] and [13] are similar to ours. However, several key elements are missing in the previous work:

- A unified theoretical approach derived through Bayesian analysis to obtain the metric.
- A probabilistic perspective such that f is well known. Indeed, within our Bayesian framework, we propose the sigmoid function.
- Related to the previous point, the effect of the prior is not considered.

C. RELATION TO DIFFERENTIAL PRIVACY

DP, as established in [8], is a framework originally developed in the context of statistical databases to quantify and control the loss of privacy resulting from the release of data derived from such databases. In a statistical database, users query aggregated information rather than access individual records directly.

The concept of ϵ-Differential Privacy (ϵ-DP) formalises this framework by introducing a parameter ϵ (epsilon) to quantify privacy loss. A mechanism satisfies ϵ-DP if the inclusion or exclusion of a single database entry causes only a slight change in the probability distribution of the output. Specifically, for all possible pairs of datasets differing in one entry, and for all possible outputs, the probability that the mechanism produces a given output differs by at most a factor of e^{ϵ}.

This ensures that any single individual's data point has a minimal impact on the overall output, preventing significant information leakage even if an attacker gains access to the results of differentially private queries. Originally coined for databases, ϵ-DP has evolved to encompass machine learning models, providing a well-known framework for privacy evaluation in various data-driven fields [18].

The mathematical definition of ϵ-DP [8] is

$$
\frac{p(A(D_1) \in Q)}{p(A(D_2) \in Q)} \leq e^{\epsilon}
$$

where A denotes a randomised algorithm applied on a data set D_n, D_1 is a neighbour data set with respect to D_2, and Q is the image of A. In the context of generative models, we can use the following properties to propose a new definition.

- A generative model $\mathcal{M}$ is a randomised algorithm as its behaviour depends on a random variable commonly expressed as Z, as in the case of the latent space of a variational autoencoder, or the input noise for a generative adversarial network.
- A different $\mathcal{M}$ is learnt from each data set, so we know that $\mathcal{M}(D_1) \neq \mathcal{M}(D_2)$ if these two sets are different. We can define $D_1 = S_t$ and $D_2 = S_{-q}$ as these two data sets are neighbouring.
- We are no longer interested in the original definition of membership with respect to the image Q, but in membership with respect to the training set.

Let us now try to find a relationship between the proposed metric for MIA and ϵ-DP. For that, we can start by setting an upper threshold in the MIA metric $I(\theta_v, x_q) \leq \omega$, where ω is a probability; then we rearrange the terms so that we can derive an expression for ϵ, which is the main driver for DP.

$$
\begin{aligned}
I(\theta_v, x_q) &= \sigma\left[\ln \frac{p(x_q|\theta_v)p(m_q=1)}{p(x_q|\theta_r)p(m_q=0)}\right] \leq \omega \\
\ln \frac{p(x_q|\theta_v)p(m_q=1)}{p(x_q|\theta_r)p(m_q=0)} &\leq \sigma^{-1}(\omega) = \ln \frac{\omega}{1-\omega} \\
\frac{p(x_q|\theta_v)p(m_q=1)}{p(x_q|\theta_r)p(m_q=0)} &\leq \frac{\omega}{1-\omega} = \exp\left(\ln \frac{\omega}{1-\omega}\right)
\end{aligned}
$$

We now omit the influence of the prior ratio and assume that $\theta_r = \theta'_v$, meaning that the reference model has been trained with S_{-q}. Hence, there is a strong relationship with respect to MIA analysis, but in this case the reference model learns from the training set except for the query sample S_{-q}. Thus, we have

$$
\frac{p(x_q|\theta_v)}{p(x_q|\theta'_v)} \leq \exp\left(\ln \frac{\omega}{1-\omega}\right) = e^{\epsilon} \tag{6}
$$

Therefore, we can derive an expression for ϵ in the context of DP based on the concept of MIA given two neighbouring data sets for the victim and reference sets.

$$
\epsilon = \ln \frac{\omega}{1-\omega}
$$

This relationship shows how DP can be considered as a particular case of MIA in the case of a generative model. Hence, its estimation is related to the estimation of the bound in terms of MIA.

V. RESULTS

This section provides insights and evidence from the various perspectives of theoretical analysis by examining several simulations. First, we examine the modelling problem consistent of a prior and posterior Gaussian mixture, being the trained model subject to overfitting up to some degree; second, we demonstrate that any prior belief could boost the attacker's performance; third, we show the relationship between overfitting and DP through particularising MIA; fourth,

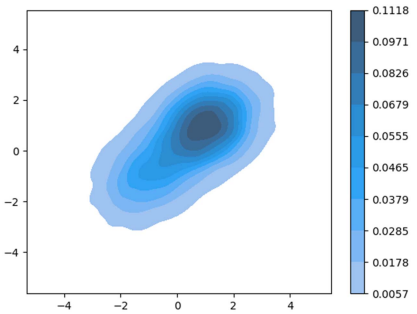


FIGURE 1. Kernel density of the proposed Gaussian mixture used as example.

we hypothesise about a more likely situation in which real data is not achievable but the attacker possesses a synthetic data set similar to the real data; lastly, we show a real-world application through image generation.

A. MODELING PROBLEM

Consider the mixture of two Gaussian distributions, with the components' distribution being $\{\pi_1 = 0.7, \tilde{}\pi_2 = 0.3\}$, means at $\mu_1 = [1, 1]$ and $\mu_2 = [-1, -1]$, and both having the same covariance matrix given by

$$\Sigma = \begin{pmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{pmatrix}$$

The density plot of this distribution is shown in Fig. 1, where the colour indicates the density approximated through kernel density estimation at that coordinates.

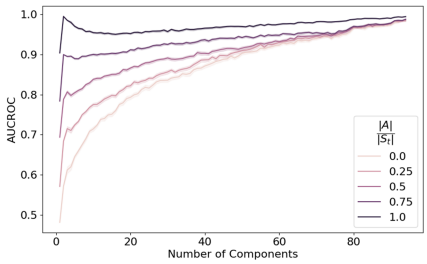
We approximate this distribution by fitting another Gaussian mixture via the expectation maximisation (EM) algorithm. The degree of overfitting can be controlled through the number of components of the learnt distribution, which is a hyperparameter of EM. The learnt distribution serves to demonstrate the effect of overfitting.

B. EFFECT OF THE PRIOR IN THE METRIC

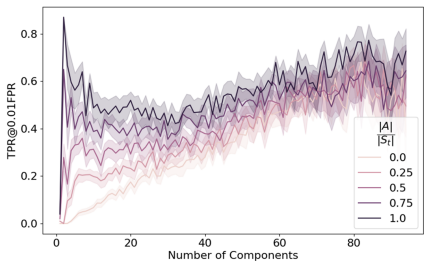
The final expression after the derivation, (5), expresses the inference problem as a product between the likelihood ratio and the prior ratio. Prior belief is an important topic in Bayesian analysis, but it is normally problem-dependent and poses several challenges, the most important being the domain knowledge of the data and the physical processes involved in the context of those data.

A situation in which the priors can make a difference occurs when an attacker possesses a data set that could have been used to train a model and knows that a subset of it was likely used during the training phase. Hence, the relationship between the sets can be expressed as $A \subseteq S_t \subset S$, where A is the subset that was likely used for training.

As the size of A increases, the attacker knows more about S_t, which, at the limit $A = S_t$, produces an almost perfect membership attack ("perfect" depends on the strength of the belief: the prior).



(a) AUCROC for a privacy attack where the different curves denote the fraction of samples likely used to train the generative model.



(b) TPR@0.1FPR for a privacy attack where the different curves denote the fraction of samples likely used to train the generative model.

FIGURE 2. Results for privacy attacks based on our metric. Each subfigure illustrates different aspects of the attack's performance. These subfigures demonstrate that prior belief can determine the effectiveness of an attack even in non-overfitting cases.

One widely used metric in the context of MIA is the area under the curve ROC (AUCROC), which fits in this problem because it resembles a traditional classification task. Another widely used metric with potential benefits is the true positive rate at a low false positive rate (TPR@0.1FPR) [5]. Fig. 2 shows how TPR and AUCROC vary depending on the number of training samples in which the attacker has a prior belief, expressed as a ratio over the total number of samples, with respect to the cardinality of all training samples ($\frac{|A|}{|S_t|}$). The aforementioned ratio has an impact on the MIA estimates, where MIA's success becomes more likely as this ratio gets closer to 1.0. In contrast, the curve with a null ratio (0.0) denotes the situation in which the prior has no effect and the attack is less effective. Note that in the case of TPR, the curve exhibits a peak in a model with low overfitting (small number of components). This could lead to potential leakages in cases where an attacker takes advantage of prior information.

C. EPSILON AS AN UPPER THRESHOLD

The relationship between overfitting and DP can be studied within the context of our problem to show how loss of privacy, measured through ϵ, increases as the number of components increases.

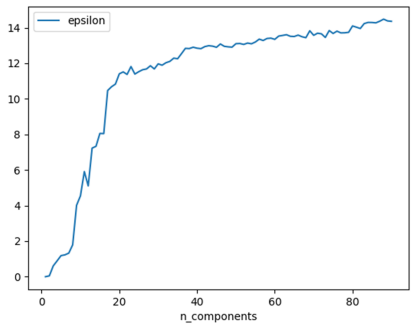


FIGURE 3. Epsilon estimates based on the definition from (6) where the value is given by the upper threshold from all different selections of x_q from S_t. The number of components, which is an indicator for overfitting, determines privacy risk through epsilon. Note that higher epsilons denote less privacy.

TABLE 2. ϵ Estimated by Adding a Gaussian Noise to the Training Data With the Standard Deviation Presented in the First Row in a Setting of 100 Training Samples and 60 Components

scale	0.01	0.1	0.5	1	2	5	10
ϵ	8.31	3.33	2.92	2.79	2.67	1.26	0.15

The value of ϵ, interpreted as an upper bound, can be estimated by iterating over the samples of a dataset and computing the ratio defined in (6). In this ratio, the numerator corresponds to the density obtained from a model trained with a data set that includes x_q (that is, S_t), while the denominator represents the density of a model trained without x_q (that is, S_{-q}).

The dependence of ϵ on the degree of overfitting is direct. Fig. 3 shows how ϵ increases with the number of components of the Gaussian mixture of the problem.

Based on previous work [21] along with ours, the concrete epsilon values may differ, but the trend in Fig. 3 is as expected, with a higher epsilon, that is, less privacy, as the overfitting increases due to the growing number of components.

D. EFFECT OF NOISE ON DP

Current research trend focusses on adding a perturbation step within the training process of generative models [2], [21], [28]. Overall, this procedure adds noise that frequently comes from a Laplacian or a Gaussian distribution to the learnt parameters (or gradients).

To mimic this behaviour, the EM algorithm can be perturbed by adding Gaussian noise to the training data. So, rather than introducing a bias over the parameters directly, this is indirectly done as these parameters will not model the original data but a modified version. In this way, we can observe how DP evolves depending on the magnitude of the perturbation, as presented in Table 2, where higher magnitudes, determined by the scale, imply lower epsilon values. Thus, increasing the overfitting produces an increase in ϵ.

E. REFERENCE MODEL VS REFERENCE DATA

The expression derived to estimate the probability of MIA for a point involves the ratio of the probability of the query given the parameters of the victim model with respect to the probability of the query given the parameters of the reference model. Previous works have paid little attention to how the latter term can affect the estimation as they have assumed that there exists a reference set to which the attacker has access.

However, this assumption may not hold, or at least does not seem to be realistic given that an attacker may not easily obtain real data. There are several reasons that support this statement.

- Data are scarce when it comes to sensitive information. Note that privacy leakages are closely related to the cardinality of the training data set, so it would not be unusual that a single dataset exists or at least a single dataset is possessed by the attacker.
- Distribution shifting could easily lead to errors. Generative models learn the correlations driving some phenomenon. The variability of data based on their location or other circumstances could potentially lead to different correlations, which would make the analysis unsolvable.
- Raw or pseudonimised data seem to be hardly exposed, while synthetic or anonymous data seem more likely to be disclosed.

Therefore, for all the reasons exposed, the MIA analysis may be affected by the definition of S_r and $p(x_q|\theta_r)$. In general, previous work has considered two cases (see Section III.IV-B), the first being the assumption that $p(x_q|\theta_r)$ can be ignored, while the second being the assumption that $p(x_q|\theta_r)$ indeed emanates from the exact same distribution as S_t and not from a learnt probability $\hat{p}(X)$.

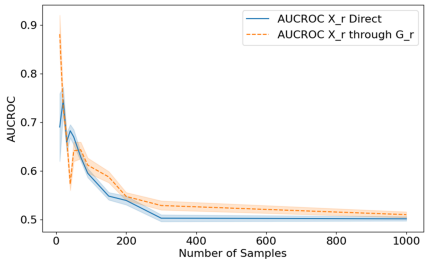
This section proposes an alternative approach, where we assume that $p(x_q|\theta_r)$ should be treated as another generative model, which is very similar to the victim model in architecture, but sees different data. Fig. 4 shows the results when a "shadow" generative model with the exact same architecture and training algorithm is considered. This result indicates a performance improvement and stabilisation in terms of AUCROC. Additionally, the TPR suggests the presence of a region where overfitting could lead to membership leakage.

F. CASE STUDY: FASHION-MNIST

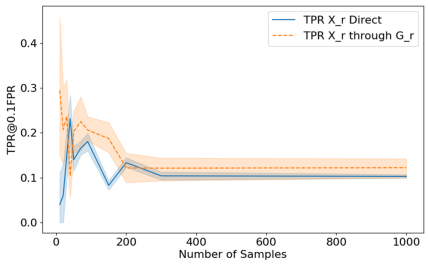
Once intuition is clear through white-box simulations, a more realistic scenario is presented in this section to illustrate the applicability of this framework in the field of image generation. Refer to the Github repository in https://github.com/BorjaArroyo/mia-theoretical-foundations to check the actual implementation.

The experimental setup is based on a previously published generative model [3], evaluated under varying levels of privacy—quantified by the parameter χ[2]—in the context of differentially private stochastic gradient descent

[2]We use χ to avoid confusion with σ, which denotes the sigmoid transformation.



(a) Variation of the AUCROC.



(b) Variation of the TPR@0.1FPR.

FIGURE 4. Variation of classification scores based on our metric due to the cardinality of the training set for the cases with and without a shadow model. Each subfigure illustrates different aspects of the attack's performance. These results indicate how a synthetic reference set is able to provide even better results compared to a real reference set.

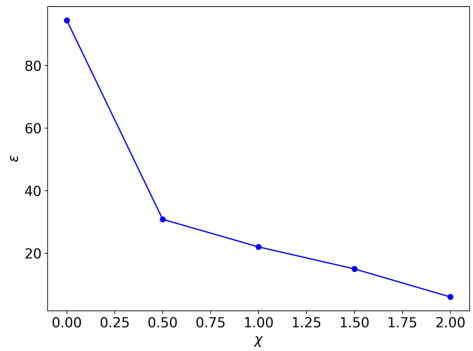


FIGURE 5. Lineplot of the ϵ values estimated through the approach described in the text. Overall, the trend demonstrates a decrease in ϵ, that is, increasing the privacy guarantees, which can be induced by increasing the perturbation in DPSGD. These results align with previous results.

(DPSGD) [1]. The model is trained on the Fashion-MNIST data set [29], with the goal of analysing how different values of χ affect the proposed metric. It is important to note that our aim is to demonstrate the application of our framework to an existing approach, specifically the one available at https://github.com/alexbie98/dpgan-revisit.

The way in which privacy risk concerns are analysed emanates from the causal relationship between overfitting and DP. As is subsequently mentioned, the generative model sees just a single example from one of the classes of Fashion-MNIST. The model tends to generate samples very close to that unique sample when conditional sampling is applied on its label. Therefore, we suppose that the most extreme case in which the density ratio may increase is around that specific sample. Hence, we pick that unique sample as the query, and we compute the density ratio at that point. We then assume that this ratio is a tight estimate of ϵ as it reflects a suitable upper bound in terms of density differences.

The simulation process, due to the slow convergence of the GAN models, is pragmatic in several ways.

- Gradient noise, χ, is defined as the set $\{0.0, 0.5, 1.0, 1.5, 2.0\}$.
- The query always belongs to the first class of the fashion-MNIST data set (T-shirt/top).
- We consider a victim model trained on samples from all classes except the query class but including the query sample.
- We consider a reference model trained on samples from all classes but the query class.
- We choose a sufficiently small number of samples per class: $N_i = 50$. Hence, the total number of samples for the victim model is $\sum_{i=1}^{9} N_i + 1 = 451$. Therefore, the total number of samples for the reference model is 450.
- Density estimation is performed via cascading principal component analysis and kernel density estimation.
- Several seeds are run to compute an upper threshold of the metric. We select the maximum value of the ratios obtained in the process.

Once the modelling problem and its context are understood, the stages to estimate ϵ for a specific χ and seed are as follows.

1) The query is selected as the first sample within the first class. Neighbouring data sets are created by following the criteria described above. Samples from a class are included in order until the $N_i = 50$ threshold is reached.
2) The victim and reference models are trained on their respective neighbouring dataset. The training process follows the ideas considered in [3] in which a scheduler decides at each iteration what model component (generator or discriminator) is updated. In contrast to the repository referenced, we use a conditional GAN to generate samples of the desired class.
3) For each model, a sufficiently large amount of synthetic samples is generated with equal representation of each class, including the query class.
4) A density estimation model is trained for each of these sets of synthetic samples.
5) Density estimation is performed on the query to evaluate its likelihood given each synthetic set.
6) The ratio is estimated.

Fig. 5 shows the results obtained through this process in terms of the upper threshold estimation with respect to various χ values. The trend of this curve is concordant with the trend denoted for the Gaussian mixture simulation. Note that the epsilon rapidly decreases as the perturbation increases.

In addition to these results, we provide some sample images to demonstrate the effect of the magnitude of DPSGD on the

FIGURE 6. Table with figures generated from models trained with different values of χ. From left to right: real images, χ = [0.0, 0.5, 1.0, 1.5, 2.0, 5.0, 10.0]. Synthetic images are selected to be the closest to the real case.

resulting synthetic images in Fig. 6. The trend from left to right shows a clear impoverishment in terms of image quality in such a way that the sharpness and edges of the samples worsen, which aligns with the trend in terms of DP described in the previous figure.

VI. DISCUSSION

Several key insights are provided through the results that denote the importance of our work.

- The prior has a great influence on the behaviour of the metric. Through the simulations, we have illustrated how the attacker's belief could boost MIA effectiveness largely to reach an AUCROC of almost a perfect classifier (see the uppermost curve in Fig. 2). This scenario poses a challenge that could lead to data leakage in any case, so it is of utmost importance not to allow this belief to exist. Moreover, given the results shown in Fig. 2 with regards to the TPR, it seems that even with scarce overfitting the risk of leakage rises.
- Overfitting has a high correlation with data leakage. By varying the degree of overfitting through the number of components in the Gaussian mixture, we have found a link between the degree of overfitting and the value of ϵ (see Fig. 3). The trend shown, which denotes a direct relationship between these two factors, shows how the traditional measure of privacy (ϵ) increases exponentially up to a certain threshold (in Fig. 3 roughly 20 components), while after this threshold it exhibits a much lower slope. This result offers two related insights: (1) the complexity of a model, which is one of the main promoters of overfitting, could lead to a greater prevalence of information leakage; and (2) the influence of overfitting has a large impact, especially in the early stages (exponential behaviour), on the risk of MIA.
- Additive noise decreases ϵ. Although overfitting promotes information leakage, we can regularise the degree of overfitting by adding noise to the training process (see Table 2). In our experiment with 100 training samples and 60 components in the Gaussian mixture, we show how sensitive ϵ is to additive noise, reducing its value from roughly 12 (see Fig. 3), to 8.31, or even 3.33, with a tiny perturbation over the training data (scales 0.01 and 0.1 respectively). Although we have added noise directly to the input data, there are other approaches that directly perturb the parameters during the training process. This is the case of one of the most established techniques, DP-SGD (differentially private stochastic gradient descent) [1]. Any of these techniques are mechanisms in the context of DP, but still pose privacy risks: the output of the generative model that is trained using DP does not guarantee that it is inmune to attacks. Practical risks such as MIA can still arise when ϵ is large enough or model overfitting occurs, even under DP training Therefore, we should consider additive noise as one of the regularisation techniques that can be used in our toolkit to prevent data leakage. Hence, the selection of a regularisation technique along with its own parameters is one of the hyperparameters to decide in order to maximise fidelity and utility while preserving privacy guarantees.
- A reference model is enough to disclose information. One of the main limitations of our method is that it depends on a reference data set which, for several reasons, could be unattainable. Nevertheless, we have shown how the performance in terms of AUCROC and TPR is risky in cases in which the attacker can only rely on a reference generative model. Of course, the success of this attack is influenced by the similarity of the architecture and training process of the victim and reference models, and the similarity of the victim data with respect to the reference data used to train that model. In the case of high similarities, Fig. 4 illustrates the AUCROC and TPR given this two settings, starting from a performant attack (in terms of AUCROC) with a real set used as reference, compared to the case in which a synthetic set is used as reference. Therefore, MIA could offer a superior performance in a setting in which the attacker has a reference model.
- Epsilon estimation is coherent in the image domain. Training a conditional GAN via DPSGD with adaptative scheduling results in good quality images with better privacy guarantees, as denoted in Fig. 5.

VII. CONCLUSION AND FUTURE LINES

The notion of MIA seems to be extensible to generative models, where its definition emanates from a Bayesian treatment of the problem. We have found that

- Prior belief can lead to significant improvements on the attacker's side.

- The quotient of probabilities seems to be the most valid magnitude to quantify the MIA risk.

Furthermore, we have found that there is a relationship between MIA and ϵ-DP based on the definition of the latter in the context of generative models. More importantly, ϵ-DP, in generative models, is a particular case of MIA where the reference set is a neighbour data set of the original data set.

Moreover, besides all the theoretical formulations of the problem, we have demonstrated that the intuition behind these ideas applies to reality with several simulation examples in which we have used a Gaussian mixture model to evaluate our derivations.

Our work could be extended in various pathways, starting from the consolidation of our approach regarding differential privacy in generative models, a comparison in terms of the effectiveness of differential privacy and other regularisation techniques with respect to ϵ, and the application of our metric to generative models in general.

APPENDIX A
CANONICAL EQUATION FOR MIA

From Bayes theorem

$$
\begin{aligned}
p(m_q = 1|\theta_v, x_q, S_{-q}) &= \frac{p(\theta_v|m_q = 1, x_q, S_{-q})p(m_q = 1)}{p(\theta_v|x_q, S_{-q})} \\
&= \frac{p(\theta_v|m_q = 1, x_q, S_{-q})p(m_q = 1)}{\sum_{y\in\{0,1\}} p(\theta_v|m_q = y, x_q, S_{-q})p(m_q = y)} \\
&= \frac{1}{1 + \frac{p(\theta_v|m_q=0,x_q,S_{-q})p(m_q=0)}{p(\theta_v|m_q=1,x_q,S_{-q})p(m_q=1)}} \\
&= \frac{1}{1 + \exp\left(\ln \frac{p(\theta_v|m_q=1,x_q,S_{-q})p(m_q=1)}{p(\theta_v|m_q=0,x_q,S_{-q})p(m_q=0)}\right)} \\
&= \sigma\left(\ln \frac{p(\theta_v|m_q = 1, x_q, S_{-q})p(m_q = 1)}{p(\theta_v|m_q = 0, x_q, S_{-q})p(m_q = 0)}\right)
\end{aligned}
$$

Therefore, substituting in the original expression, we have

$$
\begin{aligned}
&p(m_q = 1|\theta_v, x_q) \\
&= \int \sigma\left(\ln \frac{p(\theta_v|m_q = 1, x_q, S_{-q})p(m_q = 1)}{p(\theta_v|m_q = 0, x_q, S_{-q})p(m_q = 0)}\right) \\
&\quad p(S_{-q})dS_{-q}
\end{aligned}
$$

APPENDIX B
DEFINING THE PRIOR AND POSTERIOR

Let us assume that the posterior distribution of θ_v is proportional to the exponential function, while the prior over θ_v is uniform in its domain; then we have

$$
\begin{aligned}
p(\theta|S) &\propto \prod_i p(x_i|m_i = 1, \theta_v)p(\theta_v) \\
&\propto \prod_i \exp\left(-d(x_i, \theta_v)m_i\right) \\
&= \exp\left(-\sum_i d(x_i, \theta_v)m_i\right)
\end{aligned}
$$

The expression $d(\cdot, \cdot)$ will be defined as a proper metric or distance later. According to the last assumption, we have

$$
p(\theta_v|m_q = 1, x_q, S_{-q}) = \frac{\exp\left(-\sum_i d(x_i, \theta_v)m_i\right)}{\int \exp\left(-\sum_i d(x_i, \phi_v)m_i\right) d\phi_v}
$$

$$
p(\theta_v|m_q = 0, x_q, S_{-q}) = \frac{\exp\left(-\sum_{i\neq q} d(x_i, \theta_v)m_i\right)}{\int \exp\left(-\sum_{i\neq q} d(x_i, \omega_v)m_i\right) d\omega_v}
$$

Substituting these expressions into the odd ratio of the original expression we have

$$
\begin{aligned}
&\frac{p(\theta_v|m_q = 1, x_q, S_{-q})}{p(\theta_v|m_q = 0, x_q, S_{-q})} \\
&= \frac{\exp\left(-\sum_i d(x_i, \theta_v)m_i\right)}{\int \exp\left(-\sum_i d(x_i, \phi_v)m_i\right) d\phi_v} \Bigg/ \frac{\exp\left(-\sum_{i\neq q} d(x_i, \theta_v)m_i\right)}{\int \exp\left(-\sum_{i\neq q} d(x_i, \omega_v)m_i\right) d\omega_v} \\
&= \frac{\exp\left(-d(x_q, \theta_v)\right)}{\int \exp\left(-\sum_i d(x_i, \phi_v)m_i\right) d\phi_v} \Bigg/ \frac{}{\int \exp\left(-\sum_{i\neq q} d(x_i, \omega_v)m_i\right) d\omega_v}
\end{aligned}
$$

$$= \frac{\exp\left(-d(x_q,\theta_v)\right)}{\int \exp\left(-d(x_q,\phi_v)\right) \dfrac{\exp\left(-\sum_{i\neq q} d(x_i,\phi_v)m_i\right)}{\int \exp\left(-\sum_{i\neq q} d(x_i,\omega_v)m_i\right) d\omega_v} d\phi_v}$$

The quotient in the denominator can be simplified as

$$\frac{\exp\left(-\sum_{i\neq q} d(x_i,\phi_v)m_i\right)}{\int \exp\left(-\sum_{i\neq q} d(x_i,\omega_v)m_i\right) d\omega_v} = p(\phi_v|S_{-q})$$

Therefore, the above expression can be rewritten as

$$\frac{p(\theta_v|m_q=1,x_q,S_{-q})}{p(\theta_v|m_q=0,x_q,S_{-q})} = \frac{\exp\left(-d(x_q,\theta_v)\right)}{\int \exp\left(-d(x_q,\phi_v)\right) p(\phi_v|S_{-q}) d\phi_v}$$

Introducing this result into the original expression, we have

$$\begin{aligned} & p(m_q=1|\theta_v,x_q) \\ &= \int \sigma\left[\ln \frac{\exp\left(-d(x_q,\theta_v)\right)}{\int \exp\left(-d(x_q,\phi_v)\right) p(\phi_v|S_{-q}) d\phi_v} + \ln\frac{p(m_q=1)}{p(m_q=0)}\right] p(S_{-q}) dS_{-q} \\ &= \int \sigma\Bigg[\left(-d(x_q,\theta_v)\right) - \ln \int \exp\left(-d(x_q,\phi_v)\right) p(\phi_v|S_{-q}) d\phi_v + \ln\frac{p(m_q=1)}{p(m_q=0)}\Bigg] p(S_{-q}) dS_{-q} \end{aligned}$$

APPENDIX C
DERIVATION OF THE METRIC

We will focus on the second term within the sigmoid transformation, as its integral is intractable. As S_{-q} remains unknown, we alternatively consider a reference model trained over a reference set S_r such that $(x_q, m_q) \notin S_r$ by definition. In addition, assume that

$$p(\phi_v|S_{-q}) \approx p(\phi_v|S_r) \approx \delta(\phi_v - \theta_r)$$

because the set S_r has a sufficient amount of samples to consider that the posterior is very narrow around the parameters of the model trained with S_r (θ_r). Under these simplifications, the log of the integral can be rewritten as

$$\begin{aligned} & -\ln \int \exp\left(-d(x_q,\phi_v)\right) p(\phi_v|S_{-q}) d\phi_v \\ &\approx -\ln\left(\exp\left(-d(x_q,\theta_r)\right)\right) \\ &= d(x_q,\theta_r) \end{aligned}$$

Therefore, we have the membership probability as

$$p(m_q=1|\theta_v,x_q) \approx \int \sigma\left[d(x_q,\theta_r) - d(x_q,\theta_v) + \ln\frac{p(m_q=1)}{p(m_q=0)}\right] p(S_{-q}) dS_{-q}$$

where the dependence with respect to S_{-q} does not anymore exist, obtaining

$$\sigma\left(d(x_q,\theta_r) - d(x_q,\theta_v) + \ln\frac{p(m_q=1)}{p(m_q=0)}\right)$$

Thus, we can reformulate the inference problem as

$$\begin{aligned} & I(\theta_v,x_q) \\ &= p(m_q=1|\theta_v,x_q) \\ &= \sigma\left[d(x_q,\theta_r) - d(x_q,\theta_v) + \ln\frac{p(m_q=1)}{p(m_q=0)}\right] \\ &= \sigma\Bigg[-\ln\left(\exp\left(-d(x_q,\theta_r)\right)\right) + \ln\left(\exp\left(-d(x_q,\theta_v)\right)\right) + \ln\frac{p(m_q=1)}{p(m_q=0)}\Bigg] \\ &= \sigma\left[\ln\frac{\exp\left(-d(x_q,\theta_v)\right)}{\exp\left(-d(x_q,\theta_r)\right)} + \ln\frac{p(m_q=1)}{p(m_q=0)}\right] \\ &\approx \sigma\left[\ln\frac{p(x_q|\theta_v)}{p(x_q|\theta_r)} + \ln\frac{p(m_q=1)}{p(m_q=0)}\right] \\ &\approx f\left(\frac{p(x_q|\theta_v)p(m_q=1)}{p(x_q|\theta_r)p(m_q=0)}\right) \end{aligned}$$

ACKNOWLEDGMENT

Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

[1] M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 308–318, *arXiv:1607.00*.

[2] M. Azadmanesh, B. S. Ghahfarokhi, and M. A. Talouki, "ADAM-DPGAN: A differential private mechanism for generative adversarial network," *Appl. Intell.*, vol. 53, no. 9, pp. 11142–11161, May 2023.

[3] A. Bie, G. Kamath, and G. Zhang, "Private GANs, revisited," Oct. 2023, *arXiv:2302.02936*.

[4] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognit.*, vol. 84, pp. 317–331, Dec. 2018, *arXiv:1712.03141*.

[5] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership inference attacks from first principles," in *Proc. 2022 IEEE Symp. Secur. Privacy*, 2022, pp. 1897–1914.

[6] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-Leaks: A taxonomy of membership inference attacks against generative models," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2020, pp. 343–362, *arXiv:1909.03935*.

[7] Y. Du and N. Li, "Systematic assessment of tabular data synthesis algorithms," Feb. 2024.

[8] C. Dwork, "Differential Privacy," *Int. Colloq. Automata, Languages Program.* Berlin, Heidelberg: Springer, pp. 1–12, 2006.

[9] J. Fonseca and F. Bacao, "Tabular and latent space synthetic data generation: A literature review," *J. Big Data*, vol. 10, no. 1, Jul. 2023, Art. no. 115.

[10] G. Franceschelli and M. Musolesi, "Copyright in generative deep learning," *Data Policy*, vol. 4, Jan. 2022, Art. no. e17.

[11] G. Ganev, K. Xu, and E. De Cristofaro, "Understanding how differentially private generative models spend their privacy budget, May 2023, *arXiv:2305.10994*.

[12] T. Ha, T. K. Dang, and N. Nguyen-Tan, "Comprehensive analysis of privacy in black-box and white-box inference attacks against generative adversarial network," in *Proc. Future Data Secur. Eng., 8th Int. Conf.*, Cham, Springer Int. Publishing, 2021, pp. 323–337.

[13] J. Hayes, L. Melis, G. Danezis, and E. L. De Cristofaro, "Membership Inference Attacks Against Generative Models," Aug. 2018, *arXiv:1705.07663*.

[14] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic data generation for tabular health records: A systematic review," *Neurocomputing*, vol. 493, pp. 28–45, Jul. 2022.

[15] B. Hilprecht, M. Härterich, and D. Bernau, "Monte Carlo and reconstruction membership inference attacks against generative models," in *Proc. Privacy Enhancing Technol.*, 2019, pp. 232–249.

[16] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership Inference Attacks on Machine Learning: A Survey," *ACM Comput. Surveys*, vol. 54, no. 11, pp. 1–37, Sep. 2022.

[17] H. Hukkelås, R. Mester, and F. Lindseth, "DeepPrivacy: A generative adversarial network for face anonymization," in *Proc. Adv. Vis. Comput.*, 2019, pp. 565–578.

[18] Z. Ji, Z. C. Lipton, and C. Elkan, "Differential privacy and machine learning: A survey and review," Dec. 2014, *arXiv:1412.7584*.

[19] O. Mendelevitch and M. D. Lesh, "Fidelity and privacy of synthetic medical data," Jun. 2021, *arXiv:2101.08658*.

[20] Y. Miao et al., "No-label user-level membership inference for ASR model auditing," in *Proc. Eur. Symp. Res. Comput. Secur.*, 2022, pp. 610–628.

[21] J. Near, "Differential privacy: Future work & open challenges," *NIST*, Jan. 2022.

[22] G. Raut and A. Singh, "Generative AI in vision: A survey on models, metrics and applications," Feb. 2024, *arXiv:2402.16369*.

[23] M. Rigaki and S. Garcia, "A survey of privacy attacks in machine learning," *ACM Comput. Surveys*, vol. 56, no. 4, pp. 1–34, Nov. 2023.

[24] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 3–18.

[25] P. Stock, I. Shilov, I. Mironov, and A. Sablayrolles, "Defending against reconstruction attacks with Rényi differential privacy," Feb. 2022, *arXiv:2202.07623*.

[26] L. C. Tiao, A. Klein, M. W. Seeger, E. V. Bonilla, C. Archambeau, and F. Ramos, "BORE: Bayesian optimization by density-ratio estimation," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, Feb. 2021, pp. 10289–10300. [Online]. Available: https://proceedings.mlr.press/v139/tiao21a.html

[27] B. van Breugel, H. Sun, Z. Qian, and M. van der Schaar, "Membership inference attacks against synthetic data through overfitting detection," Feb. 2023, *arXiv:2302.12580*.

[28] B. Weggenmann, V. Rublack, M. Andrejczuk, J. Mattern, and F. Kerschbaum, "DP-VAE: Human-readable text anonymization for online reviews with differentially private variational autoencoders," in *Proc. ACM Web Conf.*, WWW '22, 2022, pp. 721–731.

[29] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," Sep. 2017, *arXiv:1708.07747*.

[30] W. X. Zhao et al., "A survey of large language models," Nov. 2023, *arXiv:2303.18223*.

BORJA ARROYO GALENDE received the degree in natural environment engineering from ETSIMFMN, Universidad Politica de Madrid, Madrid, Spain, the degree in computer engineering, and the master's degree in engineering and data science from ETSI Informática, Universidad Nacional de Educación a Distancia (UNED), Madrid. He is currently developing his thesis in the context of machine learning in federated environments. He is also a Software Developer/Architect with GATV.

PATRICIA A. APELLÁNIZ received the bachelor's degree in telecommunication engineering from the Universidad Autónoma de Madrid, Madrid, Spain, in 2018, and the master's degree in telecommunication engineering from the Universidad Politécnica de Madrid (UPM), Madrid, where she is currently working toward the Ph.D. degree in deep learning algorithms for medical applications. She is also a Researcher with UPM. Her research focuses on audio and image signal processing to different deep learning applications.

JUAN PARRAS received the B.S. degree in telecommunications engineering from the Universidad de Jaén, Jaén, Spain, in 2014, and the M.Sc. and Ph.D. degrees in telecommunications engineering from the Universidad Politécnica de Madrid (UPM), in 2016 and 2020, respectively. He is currently an Assistant Professor with UPM. His research interests include deep generative models, deep reinforcement learning, game theory, and optimization with health and communications applications.

SANTIAGO ZAZO is currently a Dr. Engineer with the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 1995. In 1998, he joined UPM, where he is currently a Full Professor in signal theory and communications. He is the author or co-author of more than 40 journal articles and about 200 conference papers. His main research interest includes signal processing, deep learning with medical applications, distributed optimization, game theory, and reinforcement learning.

SILVIA URIBE received the Telecom Engineering degree (Hons.), the master's degree in communications technologies and systems, the master's degree in telecommunication management, and the Ph.D. degree (cum laude) from the Universidad Politécnica de Madrid, Madrid, Spain, in 2008, 2010, 2013, and 2016, respectively. She has been a Member of the Visual Telecommunication Application Group since 2006. Her research interests include interactivity and content personalization technologies.