Trading-Off Privacy, Utility, and Explainability in Deep Learning-Based Image Data Analysis

Wisam Abbasi, Paolo Mori, and Andrea Saracino

***Abstract*—This paper proposes a novel approach for multi-party collaborative data analysis problems, where analysis accuracy and divergence are required, as well as both privacy of shared data and explainability of results. The proposed approach aims at trading-off data privacy, decision explainability, and data utility by analytically relating these three measures, evaluating how they impact each other, and proposing a methodology to find the best possible *trade-off* among them. In particular, given a set of requirements from the participants for a collaborative analysis problem, we propose a method to properly tune the parameters of privacy-preserving mechanisms and explainability techniques to be adopted by all participants, obtaining the best *trade-off*. The paper is focused on deep learning-based image data analysis problems, though the approach can be generalized to other data types. The (ϵ, δ)-*Differential Privacy* and the *Autoencoders* privacy-preserving techniques have been adopted to preserve data privacy, while the *SmoothGrad* mechanism has been used to provide decision explainability. The proposed methodology has been validated with a set of experiments on three multi-class deep learning classifiers and three well-known image datasets, MNIST, FER, and CIFAR-10.**

***Index Terms*—Data privacy, explainable AI, privacy-preserving data analysis, trustworthy AI.**

I. Introduction

HIGH data availability and increasing computational power have increased in the last years the applications of Artificial Intelligence (AI), where novel systems are being implemented to automatically analyze and correlate vast amounts of available data. These AI-based systems can produce accurate results, that are used to drive critical or strategic decisions. This can be game-changing for critical applications such as CyberThreat Information (CTI) analysis [1]. AI-based analysis systems entail some concerns related to data privacy and also to the transparency of the analysis process [2], [3], and such concerns are relevant particularly when data analysis is performed collaboratively [4]. In particular, data can bring privacy sensitive information, and their exposure to the other participants of a collaborative data analysis system might tamper the reputation of a party or disclose personal, sensitive, or critical data. To address this issue, Privacy-Preserving Machine Learning (PPML) [5] techniques are used to protect such data from being disclosed, while still allowing the execution of meaningful data analysis. Performing data analysis involves other concerns beyond *privacy*. In particular, recently, there has been a growing need for *decision explainability*. In fact, by following recent directives, such as the European Artificial Intelligence Act, understanding decision-making criteria is important for both technical and ethical reasons. As privacy requirements and their enforcement through PPML normally affect the effectiveness (*accuracy*) of a decision system, considering explainability adds another dimension, which might be in contrast with both the accuracy of the decision and ensured privacy. Still, all these elements are of crucial importance as they pose the basis for the Trustworthy AI paradigm. To the best of our knowledge, up to now, these three concepts have been addressed as standalone, by focusing either on Explainable/ Explicable AI (XAI) concepts [6], [7], or on trading-off privacy with data utility (accuracy) [4], [8], [9], [10], [11], [12], [13].

This work proposes a methodology and logical framework for collaborative data analysts, to shape and define the optimal *trade-off* between *data utility*, *privacy*, and *explainability*, applied to a deep learning-based image data analysis system. In particular, we selected techniques for preserving data privacy and ensuring model explainability, which can be tuned through input parameters, and we applied them to a data analysis problem involving image data (pictures or video streams). In this context, the proposed approach defines a measure for evaluating the *trade-off* among the data *Privacy Gain*, *Data Utility Loss*, and model *Explainability Gain* obtained by applying such techniques. This measure is meant to allow the tuning of the techniques and parameters for preserving data privacy and ensuring model explainability for a specific data analysis problem, to define the configuration that maximizes the *Privacy Gain* and the *model explainability* while minimizing the *Data Utility Loss*. The methodology defines, thus, a general *trade-off* measure and exploits novelly designed tridimensional *compatibility matrices* to find the configuration yielding the best *trade-off* value. Furthermore, given a set of requirements on minimum *Privacy Degree* and explainability level, which are provided by the participants to collaborative analysis, the proposed framework will calculate the optimal solution satisfying these requirements.

As a reference example, we use a facial expression recognition model, which classifies individual faces into different emotion categories, such as happy, sad, or angry, based on the

Manuscript received 5 May 2023; revised 23 April 2024; accepted 6 May 2024. Date of publication 14 May 2024; date of current version 16 January 2025. This work was supported by H2020 SIFIS-Home action and PRIN projects ASCOT-SCE and ASSISTANTS funded through the Next Generation EU program. *(Corresponding author: Andrea Saracino.)*

Wisam Abbasi is with the Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, 56124 Pisa, Italy, and also with the Department of Computer Science, The University of Pisa, 56124 Pisa, Italy (e-mail: wesam.alabbasi@iit.cnr.it).

Paolo Mori is with the Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, 56124 Pisa, Italy (e-mail: paolo.mori@iit.cnr.it).

Andrea Saracino is with the Department of Excellence in Robotics, AI, Scuola Superiore Sant'Anna, 56127 Pisa, Italy, and also with the Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, 56124 Pisa, Italy (e-mail: andrea.saracino@santannapisa.it).

Digital Object Identifier 10.1109/TDSC.2024.3400608

expression detected in an image or a video frame. Both *privacy* and *explainability* are extremely relevant in this problem, as privacy of classified faces might be a requirement, and it is interesting to understand which physical features are relevant for the classification decision. We will refer in the following to the Facial Expression Recognition (FER) dataset.[1] For the sake of validation and to demonstrate the capability of generalization of the proposed methodology, we also apply it to the classification of handwritten digits, using the MNIST dataset,[2] and to object recognition, using the CIFAR-10 dataset.[3] Privacy preservation and explainability are handled by applications of *Autoencoders*, *Differential Privacy*, and the *SmoothGrad* techniques.

This paper provides the following contributions:

- we propose a novel framework for collaborative data analysis that balances *data privacy*, *data utility loss*, and *model explainability* in deep learning classification of image-based datasets;
- we propose a methodology and a measure to compute the optimal *trade-off* between *privacy*, *data utility*, and *explainability* aimed at evaluating the impact of the application of privacy-preserving and explainability techniques to the data analysis process;
- we define the concept of the *tri-dimensional compatibility matrix*, extending the concept defined in [4] as a tool to find the optimal *trade-off* among *data utility*, *privacy*, and *explainability* for defined constraints;
- we implemented an instance of the classification problem of handwritten digit classification, facial expression recognition, and object recognition by using *Autoencoders*, *Differential Privacy*, and the *SmoothGrad* techniques, and we measured *Privacy Gain*, *Explainability Gain*, classification *Utility Loss*, and the *trade-off* among them;
- we validate the proposed methodology through three use cases based on image datasets, by performing a set of experiments to measure the capability of reaching the best *trade-off*, where *privacy* and *explainability* are maximized, still with a limited impact on *data utility*. Specifically, we show and discuss how it is possible to tune the *Privacy Degree* and the *Explainability Degree* parameters to maximize at the same time *privacy* and *explainability*.

The experiments conducted in this paper show that the adoption of the *Autoencoders*, the *Differential Privacy*, and the *SmoothGrad* techniques for preserving *privacy* and increasing *explainability* always positively affected the *trade-off score*. In particular, the results we obtained shed light on the relationship between privacy, explainability, and data utility. As a matter of fact, according to our measures, a Utility Loss has been observed when we adopted the two privacy preserving techniques previously mentioned, but the extent of this loss remains relatively low and stable even increasing Privacy Degree parameters. Moreover, the Utility Loss resulting from adopting privacy mechanisms is notably more pronounced in facial expression recognition due to the sensitivity associated with identifying facial attributes. Achieving an optimal *trade-off* score involves balancing privacy and explainability degrees, considering the specific requirements of each use case.

[1]https://www.kaggle.com/msambare/fer2013.
[2]https://huggingface.co/datasets/mnist
[3]http://www.cs.toronto.edu/ kriz/cifar.html

While the overall contribution of this work is focused on image-based data analysis, the underlying concept promise for potential application to tabular data, through specific adaptation. For an exploration of the application of this methodology to tabular data, please refer to [14].

The rest of the paper is structured as follows. In Section II, background about privacy-preserving and explainable AI is reported. Section III describes the reference scenario of this paper and the problem we are addressing. Section IV formally presents the measures used in this work for privacy, explainability, and data utility. Section V describes the proposed *trade-off* formulation and the methodology to compute the optimal *trade-off*. Section VI presents the experiments conducted. Section VII presents and compares related work, while Section VIII concludes by proposing some future work directions.

II. Background

This section discusses privacy-preserving and XAI techniques.

A. Privacy-Preserving Techniques

Data privacy-preserving techniques are aimed at protecting dataset attributes that are considered sensitive or may lead to the re-identification of the person the data refers to. There are two main approaches for data privacy protection: the first one is related to hiding sensitive attributes to prevent person re-identification, and the second is about delivering models that perform responsible data analysis by learning general patterns instead of memorizing specific sensitive attributes and data instances. To protect data privacy when being shared or analyzed by third parties, *anonymization-based* techniques are exploited. The most popular ones are *k-anonymity* [8], *l-diversity* with all its variants [9], *t-closeness* [10], [15], (ϵ, δ)-Differential Privacy [11] and *Autoencoders* [16]. In this work, we use the last two techniques, which are powerful and effective, especially when working with image datasets.

1) (ϵ, δ)-Differential Privacy*:* (ϵ, δ)*-Differential Privacy* (DP) is considered one of the most powerful privacy-preserving techniques. DP technique differs from the traditional methods in the mechanism it uses to add noise to the data either before or during analysis, which makes it invulnerable to re-identification or data reconstruction attacks. More in details, traditional methods add noise to individual records, while the DP technique adds Laplace or Gaussian distribution noise during the analysis phase and to the learning model used. When the DP technique is applied to two neighboring datasets (i.e., datasets differing by one data instance), the outputs of the same data analysis on the two datasets are indistinguishable, thus don't disclose whether the given data instance was included in the original dataset or not. The degree to which these outcomes are indistinguishable depends on the values of the privacy budget parameter and the sensitivity parameter, which measures the algorithm sensitivity to the insertion or removal of an individual item from the dataset [11]. (ϵ, δ)*-Differential Privacy* equation is presented in (1) [11].

$$Pr[M(D_1) \in S] \leq Pr[M(D_2) \in S] \times exp(\epsilon) + \delta \quad (1)$$

In (1), Pr is the probability of the event, ϵ is the privacy budget that is used in our approach, δ is the failure probability, M is the randomized algorithm which provides (ϵ, δ)-*Differential Privacy*, and D_1 and D_2 are two datasets that differ in at least one data instance, and $S \subseteq Range(M)$.

For privacy analysis we use the *moments accountant* privacy budget tracking method [17], which uses a Differential Private Stochastic Gradient Descent (DP-SGD) algorithm with an additive Sampled Gaussian Mechanism (SGM). SGM is an *additive Gaussian noise* [18] and *Sampling* [19] mechanism used in *Differential Privacy* as defined in (2) for a real-valued function f mapping subsets of D to $\mathbb{R}^d$:

$$M(D) \triangleq f(D) + N(0, S_f \sigma^2) \tag{2}$$

where D is the dataset and a subset of its elements are sampled randomly and independently from each other with sampling rate $0 < q \leq 1$ to be used by the algorithm f. $N(0, \sigma^2)$ is the Gaussian distribution of the noise added with a mean equal to 0, and σ is the noise added with $S_f \sigma^2$ standard deviation of the noise bounded to ℓ_2 sensitivity. The moments accountant method exploits the *Composability* and *Group privacy* attributes of DP for several applications of the Gaussian mechanism on random samples of the dataset and accumulates the overall privacy budget for these executions using the *privacy accountant* concept introduced in [20] by implementing the *accountant* procedure at each execution of the Gaussian mechanism on a sample data according to the sampling ratio and performs privacy budget accumulation at the end of all procedure executions.

Using moments accountant, the accounting procedure allows proving that an algorithm is (ϵ, δ)-differentially private for appropriately selected configurations of the parameters for any $\epsilon < c_1 q^2 \, T$ and for any $\delta > 0$ if the noise multiplier σ was defined to be as in (3) proposed in [17]:

$$\sigma \geq c_2 \frac{q\sqrt{T log(1/\delta)}}{\epsilon} \tag{3}$$

where c_1 and c_2 are constants so that given the sampling probability $q = L/n$, L is the sampling ratio of each Lot, n is the size of the dataset, and T is the number of training steps and $T = \frac{E}{q}$ and E is the number of Epochs. The relationship between the noise multiplier and the privacy budget ϵ is negative, which implies better privacy protection when increasing the value of the noise multiplier. TensorFlow Privacy provides an implementation of the DP privacy accountant method for SGM[4] and a documentation[5] for the privacy framework.

There are three ways to add noise to machine learning models to satisfy the *Differential Privacy* property, the first one is to add the noise to the objective function, the second is to add noise to the gradients at each iteration of the training phase, and the third is to add noise to the output of the training phase [21].

2) Autoencoders: This technique is a type of neural network that is trained to compress data and represent their most important features as a latent representation code. The technique involves two components: an encoder and a decoder. The encoder takes an input dataset and encodes it in a latent representation (code), to get its important features represented as a code. Then, the decoder takes the latent representation and decodes it to get the reconstructed image. The *Utility Loss* in *Autoencoders* is represented as the distance between the original image and the reconstructed one. The anonymization degree is controlled by the code size, i.e., the latent representation: the lower the code size, the higher anonymization we get. The underlying mechanism can be considered similar to the compression one [22], [23]. As illustrated in Fig. 1, by compressing the original input data into a latent representation code (the compression structure), the reconstructed data loses many of its features but keeps the ones that most affect the analysis function decision.

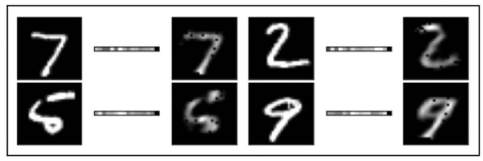


Fig. 1. Autoencoder effect on the MNIST dataset.

Autoencoders are used as a privacy-preserving technique themselves [24], [25], but they may be fine-tuned or combined with other privacy-preserving methods [22], [26].

B. Explainable Artificial Intelligence

Explainable/Explicable Artificial Intelligence (XAI) is a recent concept that has been proposed to enforce safety, fairness, trust, and transparency in AI models. XAI provides insights into how the predictive model works, the correlations among data sources and features. Hence, XAI can be exploited for discrimination avoidance and for granting users the right to get an explanation of why certain decisions have been made by AI models. On the other hand, XAI represents a threat to data privacy and allows exploration and exploitation of the model by possible malicious users. In this paper, we use Saliency Maps enhanced with the *SmoothGrad* technique for addressing explainability.

1) Saliency Maps: The Saliency Maps [27] is a gradient-based method to explain deep neural networks. The maps generated using this technique make the pixels of the input image that have the highest gradient, i.e., the most influence on the classification of the image, more visible in the image. These gradients are computed using two alternative methods which differ in the scope (local vs global):

Image-specific class saliency visualization method (local scope): an approximation class score is being calculated in this technique using (4) where $S_l(d)$ is the class scoring function, d represents the input image, l is the label class, w_l is the weight vector, and b_l is the model bias.

$$S_l(d) = w_l^T d + b_l \tag{4}$$

In the case of a Convolutional Neural Network (CNN) model, the equation is updated to be as in (5a), where w is the derivative of S_l concerning the image d at the point d_0 computed using (5b):

$$S_l(d) \approx w^T d + b \tag{5a}$$

$$w = \frac{\partial S_l}{\partial d}\bigg|_{d_0} \tag{5b}$$

Class model visualizations (global scope): generate numerically computed images representing the appearance of classes

[4]https://github.com/tensorflow/privacy

[5]https://tensorflow.org/responsible_ai/privacy/guide/measure_privacy

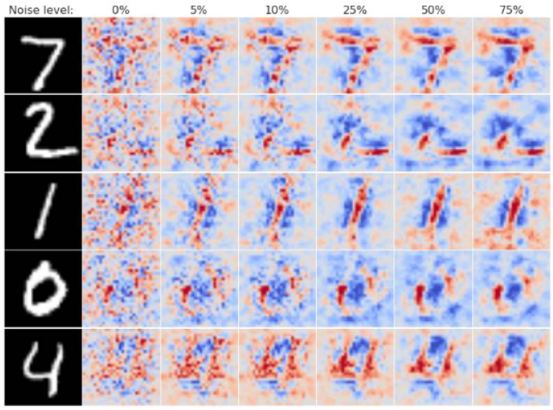


Fig. 2. Effect of Gaussian noise [28].

for a given CNN model and a learned classification by this model. It uses a regularisation parameter Ω to generate an L_2-regularised image for a specific class l with respect to a specific ConvNet layer for an image d, given a high score value for the class scoring function S_c by (6):

$$\arg\max_d S_l(d) - \Omega \|d\|_2^2 \tag{6}$$

2) SmoothGrad: This technique smooths the saliency maps produced by gradient-based models through noise elimination by adding *Gaussian noise*, which is the standard deviation of the Gaussian perturbations. The perturbation is responsible for saliency maps sharpening to generate better and more accurate explanations of the model predictions. The explanation quality can be controlled by regulating the introduced Gaussian noise, as illustrated in Fig. 2 [28], unlike other explainability mechanisms. The choice of the *SmoothGrad* method as an explainability mechanism in this paper is motivated by its unique ability to allow to control the explanation quality through the manipulation of the explainability noise parameter. Indeed, this distinctive feature makes *SmoothGrad* particularly suitable for our goal of dynamically investigating the trade-off between *privacy*, *explainability*, and *data utility* varying the explainability quality.

The higher the value of Gaussian noise, the better the explanation. According to [28], SmoothGrad method visually sharpen gradient-based sensitivity maps based on the Gaussian Noise parameter.

III. Reference Scenario and Problem Statement

The reference scenario we are considering is made of several stakeholders who produce image data, i.e., faces taken from surveillance cameras, and they are looking to perform collaborative analysis to increase the accuracy of a facial expression detector. To this aim, such stakeholders share their data with a centralized *honest-but-curious* server that will perform the analysis (as shown in Fig. 3). Anonymization might be required before sharing, through the techniques described in Section II-A. Fig. 3 shows an example where the stakeholders on the left adopt the *Differential Privacy* technique for model anonymization, and they share the predictions obtained using these differential private models and/or locally trained model weights with the central server, while the stakeholders on the top right adopt the

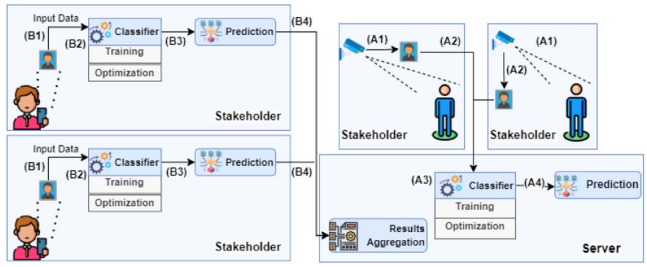


Fig. 3. Privacy-preserving image recognition scenario.

Autoencoders technique for data anonymization, and they share the anonymized datasets with the central server.

Explainability is another key requirement for the image recognition model, as it is imperative for stakeholders to comprehend the criteria underlying decision-making processes. For instance, a stakeholder should have the capability to understand why a given image submitted by his smart device has been classified as a "happy" face . Furthermore, the adoption of privacy and explainability techniques might impact the overall *Data Utility*, and they may mutually affect each other.

The methodology we propose in this paper addresses the problem of properly configuring the techniques to preserve data privacy and to provide explainability in order to obtain the best *trade-off* among *Privacy Gain*, *Data Utility Loss*, and *Explainability Gain* (which are formally defined in the next section) related to the process for the classification analysis of image datasets. Moreover, the problem could also include constraints on the values of the parameter passed to the privacy and explainability techniques, i.e., the *Privacy Degree* and the *Explainability Degree*, as requirements imposed by the stakeholders for sharing their data. For instance, a stakeholder could authorize the usage of their dataset only if the *Autoencoders* technique is applied and the code size, i.e., a *Privacy Degree*, is less than 55% of the image size. Although in this paper we take into account only the techniques shown in Section II, the methodology could be easily extended to take into account further techniques. This can be accomplished by seamlessly incorporating alternative privacy and explainability mechanisms, provided they meet specific criteria: i) both the levels of privacy and explainability should be configurable acting on proper input parameters; ii) a proper metric for assessing the *Privacy Gain* and a proper metric for assessing the *Explainability Gain* resulting from the application of such alternative privacy and explainability techniques should be available; and iii) a proper metric for measuring the associated *Data Utility Loss* should be available as well. For instance, our study employs two distinct *Data Utility Loss* evaluation metrics that are tailored specifically to the characteristics of the two privacy mechanisms we utilized.

IV. Concepts and Model

This section formally discusses the three key measures on which the proposed approach is based, namely: the *Privacy Gain*, the *Data Utility Loss*, and the *Explainability Gain*.

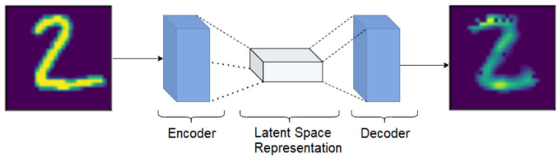


Fig. 4. *Autoencoders* applied to "2" of the MNIST dataset.

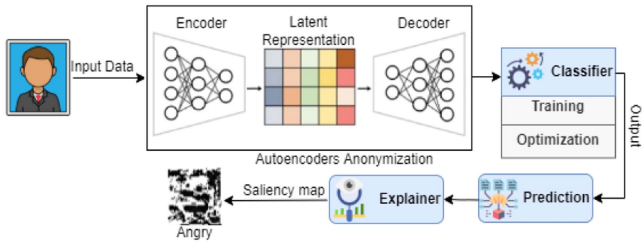


Fig. 5. *Autoencoders* for dataset anonymization with multi-class classifier architecture.

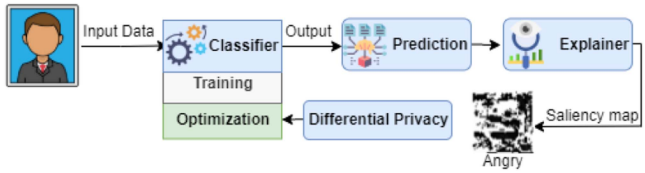


Fig. 6. DP with multi-class classifier architecture.

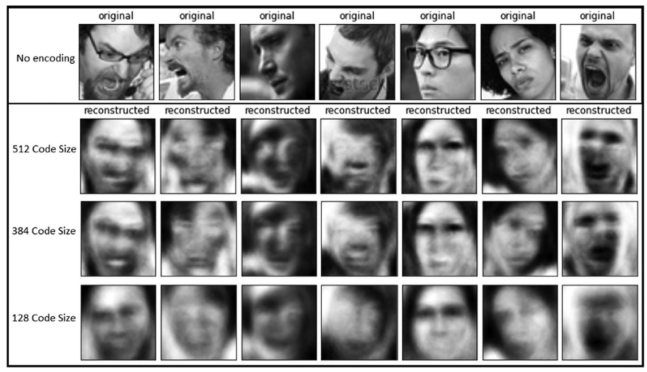


Fig. 7. *Autoencoders* privacy-preserving with different code (latent space representation) sizes on FER Dataset.

A. Anonymization and Privacy Gain

Autoencoders technique is used to disturb the original images before being fed to the model, i.e., before being sent to the server, while the *Differential Privacy* technique in this proposal is used during the analysis phase, i.e., on the stakeholder side to restrict model memorization of specific attributes and dataset entries and to let the model only learn general patterns about the dataset. Cryptographic methods are another alternative that could be used to preserve privacy in image analysis and ensure that data remain confidential and can be accessed only by authorized parties. For instance, Fully Homomorphic Encryption (FHE) has been used in [29] to protect privacy in neural network models used to deliver AI as a service on the cloud. Following such an approach, FHE-encoded data are analyzed by neural networks to produce encrypted results within an acceptable time. In addition, the authors of [30] proposed a privacy-preserving model for healthcare services in IoT environments, leveraging secure multiparty computing (SMPC) and federated learning. Moreover, in [31], SMPC has been used as a privacy mechanism with human activity recognition models, yielding the same accuracy gains as with no privacy mechanism applied. However, cryptographic methods impact the explanations generated by explainability mechanisms. For instance, the application of Homomorphic Encryption (HE) shows a noticeable impact on the highlighted regions of GradCAM attention maps in domains like Face Recognition [32] and Medical Image Analysis [33]. In contrast, our approach integrates flexible privacy protection methods with lower overhead and controllable impact on model explainability.

Referring back to the facial expression recognition example, applying the previous two privacy-preserving mechanisms on the face images would alter either the original images to produce anonymized faces that are unrecognizable, or the gradients computed during the data processing phase. However, in both cases, this anonymization could affect the accuracy of the facial expression recognition model causing some facial expressions to be not correctly recognized anymore. Moreover, the explainability of the decisions in terms of saliency maps could be affected since the dataset images are less visible.

Autoencoder deep neural networks, first encodes the original image into a latent representation and then decodes the latent representation to reconstruct the anonymized image. This phase is executed on the smart device which captures the image. The anonymized image is then fed to the classifier, which runs on the server. Fig. 4 shows an example of the application of such a technique to an MINST dataset element. The integration of the *Autoencoders* technique within the image classification system shown in Fig. 3 is illustrated in Fig. 5.

Differential Privacy, instead of operating on the dataset images, it adds noise to the gradients during the training process using a differential private optimizer. Hence, the classification model memorization of data instances is limited, as the used model only learns general patterns of the dataset. In fact, the trained differential private model can't distinguish whether specific dataset instances were used for training, thus protecting against reconstruction and membership inference attacks [34], [35], [36]. The integration of a *DP* mechanism within our image classification system is depicted in Fig. 6.

Both the previous anonymization techniques accept an input parameter, the *Privacy Degree*. In particular, when adopting the *Differential Privacy* technique, such a parameter is represented by the noise introduced in the model, as explained in Section II-A. Instead, when adopting the *Autoencoders* technique, such a parameter is represented by the size of the code, as explained in Section II-A as well. Figs. 7 and 8 illustrate, respectively, two examples of the application of the *Autoencoders* technique on the FER dataset and on the CIFAR-10 dataset with different code sizes.

Applying such techniques in the image classification process results in a *Privacy Gain*, i.e., the modifications made by these techniques introduce an *uncertainty degree* in the resulting model. This uncertainty is due to the lack of knowledge of all dataset attributes caused by the application of the anonymization

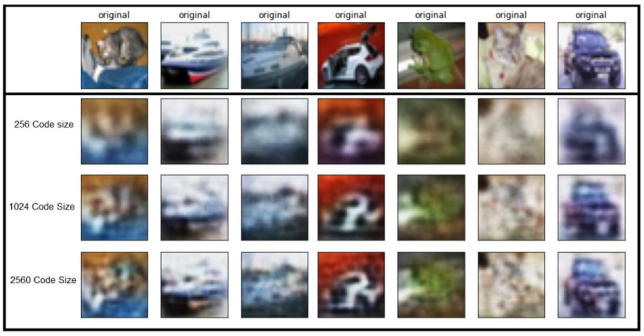


Fig. 8. Autoencoders privacy-preserving with different code (latent space representation) sizes on CIFAR-10 Dataset.

techniques, and it reduces the confidence in the predictions performed using this model [37].

The *Privacy Gain* obtained by applying the *Differential Privacy* technique to the dataset D and the classifier λ, denoted by $PG_{DP}(D, \lambda)$, is measured according to the returned privacy budget, i.e., the ϵ value we got from (1). The lower the ϵ value, the higher the *Privacy Gain* returned, as shown in (7a).

When the *Autoencoders* method is applied, the *Privacy Gain* is measured as the percentage of the code size C compared to the original image size I, as expressed in (7b). Considering that the code size C is always not greater than the size of the images in the dataset D (denoted by $I(D)$), we observe that $PG_{Auto}(C, D)$ is normalized in the interval (0,1].

$$PG_{DP}(D, \lambda) = \frac{1}{\epsilon(D, \lambda)} \tag{7a}$$

$$PG_{Auto}(C, D) = 1 - \frac{C}{I(D)} \tag{7b}$$

B. Classifier and Data Utility Loss

The enforcement of the privacy-preserving techniques discussed in Section IV-A may reduce data utility and affect model performance. *Data Utility Loss* computation depends on the privacy technique used. When adopting the *Autoencoders* technique, the *Utility Loss*, U_{Auto}, can be measured as shown in (8) based on the divergence concept, which represents the change between two distributions (or two datasets in this case):

$$U_{Auto}(D, D') = GD(D, D') \tag{8}$$

where GD is the divergence between the original dataset distribution D and the anonymized dataset distribution D'. In our approach, we adopt the Earth Mover's distance (EMD, Wasserstein distance) as a divergence computation method because it defines the ground distance between any pair of values [15]. EMD is simply the minimum cost or work in (9) needed to match two distributions normalized by the weight of one of these distributions (D, D'), as shown in (10):

$$Work(D, D', F) = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij} \tag{9}$$

$$EMD(D, D') = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \tag{10}$$

where D is the first dataset distribution with m elements, D' is the second dataset distribution with n elements, $[d_{ij}]$ is the ground distance matrix where d_{ij} is the ground distance between clusters D_i and D'_j, and F represents a flow matrix $[f_{ij}]$ among the two dataset distributions D and D' with f_{ij} the flow between D_i and D'_j, that minimizes the overall cost to get the optimal flow F. We recall that in our scenario $m = n$ because the dataset D' has been produced by anonymizing each element of the dataset D applying

Instead, the *Differential Privacy* technique does not anonymize the original dataset before the training process, but it adds the noise to the gradients during gradients computation. Thus, the method we use to compute the *Utility Loss* compares the accuracy of the classification results of the original data analysis model against the accuracy of the classification results obtained from the differential private model and converts this percentage reference into a decimal number, following the approach proposed in [21]. The resulting *Utility Loss* measure is shown in (11) and it falls in the range $0 - 1$:

$$U_{DP}(D, \lambda, \lambda') = Acc(D, \lambda) - Acc(D, \lambda') \tag{11}$$

where D is the original dataset, λ is the analysis model, λ' is the differential private analysis model that adds noise to the gradients during training, and Acc is the accuracy of the results of the analysis model applied to the dataset.

C. Explainer and Explainability Gain

A relevant requirement of the reference scenario concerns the understanding of why the model has produced a given prediction, e.g., the classification of a face image as a "happy" facial expression. For instance, the prediction could have been made due to edges at certain positions of the image, brightness in some areas, or some other aspects. XAI provides the possibility to understand this by producing explanations of why a specific prediction has been made.

In our approach, we produce these explanations in the form of Saliency Maps, using the *SmoothGrad* explainability method described in Section II-B. The *SmoothGrad* method takes an input image, generates n sample images by adding Gaussian noise, computes a Saliency Map for each of these sample images, and computes a single Saliency Map as an average of the previous ones. Despite adding the same amount of Gaussian noise to the same image, the resulting n samples are different due to the random nature of the noise addition process. This process is controlled acting on the number of samples used during the averaging process and on the Gaussian noise level parameter used for map sharpening. In summary, the computation of a saliency map for each data instance of any image dataset used by a classification model λ is shown (12):

$$\hat{\alpha}_l(d) = \frac{1}{n} \sum_{1}^{n} \alpha_l(d + N(0, \sigma^2)) \tag{12}$$

where d is a data instance representing an input image, l is the class predicted for an input image among a set of classes L, n is the number of sample images used to produce the final saliency map, $\alpha(d + N(0, \sigma^2))$ is a Saliency Map for a sample image generated by adding Gaussian noise $(N(0, \sigma^2))$ to the original one. The saliency map of a sample image is computed by differentiating α_l concerning the input image d based on the analysis function λ, as shown in (13):

$$\alpha_l(d) = \frac{\partial S_l(d)}{\partial d} \tag{13}$$

where S_l is the class scoring function. *SmoothGrad* method uses (12) to enhance originally generated saliency maps by basing a visualization on a smoothing of gradient $\partial S_l(d)$ with a Gaussian kernel instead of basing it directly on $\partial S_l(d)$. Detailed equations explanation is presented in [28].

SmoothGrad mechanism improves the quality of the saliency maps generated by increasing the degree of Gaussian noise used to smooth the map compared to the original non-smoothed map as indicated in [28]. Thus, to compute the *Explainability Gain* $EG(d)$ for the smoothed saliency map, we use (14):

$$EG(d) = \frac{n}{\max(n)} \times N(0,\sigma^2)_d \times 100 \tag{14}$$

where d is a data instance (an input image), n is the number of sample images used for averaging saliency maps to produce one saliency map, $\max n$ is the maximum number of used samples, and $N(0,\sigma^2)$ is the Gaussian noise.

However, this equation needs to be modified in case a privacy mechanism has been used, since the privacy mechanisms applied in our methodology either modifies the analysis model in the way it computes gradients when using *Differential Privacy* or alters the input dataset in case of *Autoencoders*. Therefore, to quantify *Explainability Gain* EG_{DP} when applying *Differential Privacy* mechanism, we need to take into consideration the model performance change represented by model accuracy, because explainability mechanisms do explain model decisions. Thus, when the model's accuracy decreases due to the use of a differential private model, the *Explainability Gain* needs to be bounded to this change represented by U_{DP} in (11). *Explainability Gain* when using *Differential Privacy* mechanism is reported in (15), where d is a data instance (an input image), D is the Dataset, λ is the analysis function, and λ' is the differential private analysis function obtained from λ.

$$EG_{DP}(d,\gamma) = \left(\left(\frac{n \times N(0,\sigma^2)_d}{\max(n)}\right) - U_{DP}(\gamma)\left(\frac{n \times N(0,\sigma^2)_d}{\max(n)}\right)\right) \tag{15}$$

For *Autoencoders* privacy-preserving mechanism, (16), where $\theta = \{D, D'\}$ is used to compute the *Explainability Gain*, where $\gamma = \{D, \lambda, \lambda'\}$, by taking into consideration the change percentage between images of the original dataset and anonymized dataset, since saliency maps are generated on these input image. Any change to the input image affects the output saliency map and the $EG(d)$. This change percentage is measured by the geometric divergence using the EMD method.

$$EG_A(d,\theta) = \left(\left(\frac{n \times N(0,\sigma^2)_d}{\max(n)}\right) - GD(\theta)\left(\frac{n \times N(0,\sigma^2)_d}{\max(n)}\right)\right) \tag{16}$$

Providing decision explanations for classifications made on image datasets means highlighting the pixels that have the most influence on the model prediction in the generated saliency maps. For instance, in the facial expression recognition case, the pixels representing eyes and mouth are highlighted, as shown in the maps of Fig. 12 for the FER dataset.

As a matter of fact, controlling the noise parameter affects saliency map explanations. The higher the value we define for the Gaussian noise, the more explainable the analysis result is. Therefore, the *Explainability Gain* is quantified using the noise value.

V. Proposed Methodology

This section describes the data format and the methodology we defined to properly set up the data analysis process, i.e., to find the values for the *Privacy Degree* and for the *Explainability Degree* parameters (see Section IV-A), to balance preserved data privacy, explainability of the model and classification analysis accuracy. As discussed in Section III, we are considering a scenario where several stakeholders produce image datasets consisting of faces taken from surveillance cameras and they want to exploit them to perform *collaborative analysis*, by training a common analysis model which would give more accurate results than a model built from the dataset of a single user only. In particular, each stakeholder defines their own privacy and explainability requirements in terms of constraints on the values of *Privacy Degree* and *Explainability Degree* parameters, which must be respected for the stakeholder to consent to the use of their dataset. In this scenario, the proposed methodology is aimed at finding a proper configuration of the data analysis process which allows for obtaining an analysis model that respects the stakeholders' requirements while balancing preserved data privacy, explainability level, and classification accuracy.

A. Data Format

In this paper, we are focusing on *image* datasets, i.e., datasets where each element is a picture. Typical features of the elements of such datasets are thus pixel-related, such as: *color*, *brightness*, *edges*. A further feature is the image *resolution*, n, which is equal to $x \times y$ (where x and y are the dimensions of the image) in the case of gray-scale images and to $x \times y \times 3$ in the case of colored images. In particular, we are considering two labeled gray-scale image datasets, FER and MNIST, in addition to a labeled colored image dataset, CIFAR-10. The FER dataset is relevant for the reference scenario because the images in the dataset are people's faces, and the labels paired with the images represent the facial expression category and are classified into seven emotions (classes). In the MNIST dataset, instead, the images are handwritten digits and the label paired with each image corresponds to the digit represented by the image. Hence MNIST images are classified into ten classes, with labels ranging from 0 to 9. This dataset has been chosen because it is a well-known dataset typically used as a benchmark to validate proposed approaches. In the CIFAR-10 dataset, instead, the images are of various objects and the label paired with each image corresponds to the object represented by the image. Hence, the images are classified into ten classes. This dataset has been chosen because it is a more complex well-known dataset and is used usually as a benchmark for validation. All datasets are divided into a *training set* and a *testing set*. Each dataset is thus processed by the multi-class classifier λ, one image at a time to be divided into pixels based on the image resolution as an input layer in the neural network.

B. Optimal Trade-Off Computation Through Compatibility Matrix

As discussed, two possible techniques can be used for performing the analysis in a privacy-preserving way, namely *Autoencoders* and *Differential Privacy*. When the *Autoencoders* technique is used, our methodology determines the optimal values of the *Privacy Degree* and *Explainability Degree* parameters as follows. First of all, the stakeholders apply the autoencoder technique to their datasets. In particular, each stakeholder computes a set of anonymized datasets varying the value of the *Privacy Degree* parameter according to the requirements they defined, and computes locally both the *Privacy Gain* and the *Utility Loss* for each of these datasets. Then, each stakeholder shares requirements, the set of anonymized datasets, and the related *Privacy Gains* and *Utility Losses* with the server. The server executes the data analysis process on each of the received datasets, applies the *SmoothGrad* technique for a set of values of the *Explainability Degree* and for each of them computes the related *Explainability Gain*. Finally, using the values of the *Privacy Gain*, *Utility Loss*, and *Explainability Gain* computed varying the values of the *Privacy Degree* and *Explainability Degree* parameters, the server computes the *trade-off* scores to fill the compatibility matrix.

If the *Differential Privacy* technique is selected, then the analysis process is performed on the stakeholder side and only analysis results will be shared with the server. The values of *Privacy Gain*, *Explainability Gain*, *Utility Loss*, and *trade-off* score are computed on the stakeholder's side for each configuration of *Privacy Degree* and *Explainability Degree* satisfying the stakeholder's requirements. Afterward, all these values, along with the stakeholder's requirements, are shared with the server to find the optimal *trade-off* score.

Hence, as a first step of our methodology, we define the formula to calculate the *trade-off* among the aforementioned three measures, i.e., $T_{DP}(D, \lambda, \lambda')$ (shown in (17)) when the *Differential Privacy* technique is applied or $T_{Auto}(D, D', \lambda)$ (shown in (18)) when the *Autoencoders* technique is applied. The *trade-off* formulas enable us to balance the conflicting objectives of *privacy*, *explainability*, and *utility loss* by aggregating them into a single score. The numerator of the formula captures the desirable goals of *privacy* and *explainability*, while the denominator represents the undesirable objective of *utility loss*. Dividing the desirable objectives by the undesirable objective yields a *trade-off* score that reflects the optimal balance between these objectives.

$$T_{DP}(D, \lambda, \lambda') = \frac{PG_{DP}(D, \lambda) + EG_{DP}(d, D, \lambda, \lambda')}{2 + U_{DP}(D, \lambda, \lambda')} \quad (17)$$

where D is the original dataset, λ is the analysis function, λ' is the differential private analysis function, $PG_{DP}(D, \lambda)$ is the *Privacy Gain* defined by (7a), $EG_{DP}(d, D, \lambda, \lambda')$ is the *Explainability Gain* defined by (15), and $U_{DP}(D, \lambda, \lambda')$ is the *Utility Loss* defined by (11);

$$T_{Auto}(D, D', \lambda) = \frac{PG_{Auto}(C, D) + EG_{Auto}(d, D, D')}{2 + U_{Auto}(D, D')} \quad (18)$$

where D is the original dataset, D' is the anonymized dataset, λ is the analysis function, C is the Code size used for the *Autoencoders* technique, $PG_{Auto}(C, D)$ is the *Privacy Gain* defined by (7b), $EG_{Auto}(d, D, D')$ is the *Explainability Gain* defined by (16), and $U_{Auto}(D, D')$ is the *Utility Loss* defined by (8).

By first performing discretization on the intervals defined for *Privacy Degree* and *Explainability Degree*, we transform these continuous intervals into two sets of discrete values across the original interval range. Thus, for any specific dataset, privacy parameter value, and explainability value in the discretized ranges, the related *trade-off* is computed. Considering the available datasets, the computed *trade-off* values can be represented schematically by means of a tri-dimensional *compatibility matrix*.

C. Compatibility Matrix

The compatibility matrix has been defined in [4] as a mechanism to compute the best *trade-off* between data utility and privacy in a multi-stakeholder collaborative analysis problem. Each element of the compatibility matrix represents, in fact, the *trade-off* value computed on a specific dataset (row) for a given privacy configuration (column). In this work, we are considering a further dimension: explainability. Hence, we define a tri-dimensional compatibility matrix which on the x dimension reports the values obtained from the discretization of the *Privacy Degree* (ϕ_k) interval, on the y dimension it reports the available datasets (D_i, generally one per stakeholder), and on the z dimension the values obtained from the discretization of the *Explainability Degree* (ω_j) interval. Each element $CM_{i,k,j}$ represents the *trade-off* value computed with the specified combination of the three parameters, i.e., the *trade-off* computed on dataset D_i, with *Privacy Degree* ϕ_k according to the selected privacy-preserving mechanism, and with the *Explainability Degree* ω_j.

The proposed methodology is always able to find an optimal *trade-off* point for a given setup, given that the domain we define is finite in two dimensions (privacy and explainability) for each dataset/stakeholder and discretized into points. In particular, the compatibility matrix, ensures a finite number of configurations couples (privacy degree, explainability degree) for each dataset and hence a finite number of *trade-off* values to be computed by the algorithm. Consequently, the algorithm consistently identifies the optimal solution choosing the maximum *trade-off* score among the computed ones. For the sake of performance a trade off is also to be found in the discretization interval, in order to minimize the discretization error while looking for the maximum, still being able to compute it in a feasible time.

In the image analysis problem that we are considering, we are using two different privacy mechanisms, whose *trade-off* values are not comparable, since they are computed in different ways. To this end, two compatibility matrices will be defined respectively for the *Differential Privacy*-based analysis, and for the *Autoencoders*-based one. Thus, the compatibility matrix representing the *Differential Privacy* mechanism uses (17) to compute the *trade-off*, while the compatibility matrix representing the *Autoencoders* mechanism uses (18). Fig. 9 reports the structure of a compatibility matrix.

If the values of either ϕ_k or ω_j do not respect specific requirements expressed on the dataset D_i by the owning stakeholder, the corresponding element $CM_{i,k,j}$ is set to 0, representing thus

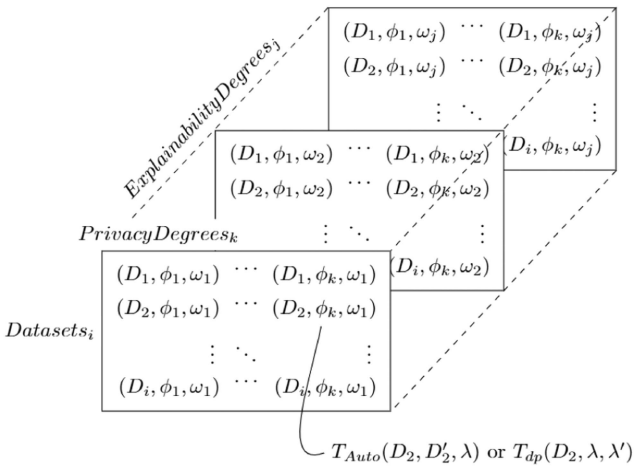


Fig. 9. Tri-dimensional compatibility matrix.

the impossibility of computing a *trade-off* for that specific configuration. Once the compatibility matrix has been computed, it is reduced to 2 dimensions by aggregating the dataset dimension. In particular, the *trade-off*s obtained for distinct datasets with the same *Privacy Degree* and *Explainability Degree* are averaged to compute the optimal *trade-off*, according to (19) for *Differential Privacy* mechanism and (20) for *Autoencoders* mechanism:

$$\bar{T}_{DP}(\bar{D}, \lambda, \lambda') = \sum_{i=1}^{m} w_i T(D_i, \lambda, \lambda') \tag{19}$$

$$\bar{T}_{Auto}(\bar{D}, \bar{D}', \lambda) = \sum_{i=1}^{m} w_i T(D_i, D'_i, \lambda) \tag{20}$$

where w_i is the inverse of the number of datasets for which the *trade-off* value is not 0.

Finally, after obtaining the best *trade-off* score using the concept of *linear objective optimization*, the corresponding privacy and *Explainability Degrees* are used.

D. Applicative Example

Let us consider a case where three stakeholders, S_1, S_2, and S_3 are willing to perform collaborative analysis. The stakeholders request an analysis function on their datasets, D_1, D_2, and D_3, respectively, from a central server using the architecture defined in Section III, where $D_1 \cup D_2 \cup D_3$ represent the whole dataset D.

Each stakeholder shares its privacy and explainability requirements R_1, R_2, and R_3, where each requirement defines the privacy mechanism, and constraints on the *Privacy Degree* and on the *Explainability Degree* to be used. The *Privacy Degree* and *Explainability Degree* are both defined by each stakeholder to be higher than a specific threshold or to obtain the maximum achievable degree. Hence, the requirement items provided by stakeholder S_i are as follows:

- Pm_i*:* Privacy mechanism selected, which is either *Autoencoders* or *Differential Privacy* in this work.
- ϕ_i*:* Minimum *Privacy Degree*, which is the code size for *Autoencoders* mechanism and *Privacy Degree* parameter for *Differential Privacy*.
- ω_i*:* Minimum *Explainability Degree*.

TABLE I
STAKEHOLDER'S REQUIREMENTS

Stakeholders	ϕ	ω
Stakeholder1	> 1.7	> 0.6
Stakeholder2	> 1.5	> 0.6
Stakeholder3	Empty	Empty

To find the best *trade-off* score, all returned configurations *trade-off* scores are investigated for the datasets using (19) or (20), which refers, respectively to the *Differential Privacy* and the *Autoencoders* privacy-preserving techniques. Then the *trade-off* score that has the maximum value representing the optimal solution is returned.

As an example, taking into account the reference scenario described in Section III, in the following we consider 3 stakeholders having an image dataset of 20,000 images each. These stakeholders want to analyze these datasets with the help of a central server. To do so, each stakeholder selects the privacy mechanism of interest and defines his own privacy and explainability requirements with specific degrees for each of them. For instance, we assume that all stakeholders have chosen the same privacy technique, e.g., *Differential Privacy*, and that they defined their privacy and explainability requirements as shown in Table I. For instance, $Stakeholder1$ wants that the *Privacy Degree*, ϕ, is greater than 1.7, and that the *Explainability Degree*, ω, is greater than 0.6. Instead, $Stakeholder3$ does not specify any requirement, meaning that he is interested in getting the optimal value among all.

From the above-mentioned details, the compatibility matrix shown in Fig. 10 is created. The x dimension of the matrix consists of 21 values for *Privacy Degree* in the interval $[0.0 - 2.0]$ with step 0.1, the y dimension of the matrix consists of 3 rows representing the 3 stakeholders' datasets, while the z dimension consists of 11 possible degrees of explainability ω of the *SmoothGrad* mechanism in the interval $[0.0 - 1.0]$ with step 0.1. Hence, the size of the compatibility matrix is $21 \times 3 \times 11$ and is computed following the below steps:

Step 1: For $Stakeholder1$, *trade-off* score is calculated for all possible combinations of *Privacy Degrees* $1.8 \leq \phi \leq 2.0$ and *Explainability Degrees* $0.7 \leq \omega \leq 1.0$ and reported in the related cells of the compatibility matrix;

Step 2: For $Stakeholder2$, *trade-off* score is calculated for all possible combinations of *Privacy Degrees* $1.6 \leq \phi \leq 2.0$ and *Explainability Degrees* $0.7 \leq \omega \leq 1.0$ and reported in the related cells of the compatibility matrix;

Step 3: For $Stakeholder3$, *trade-off* score is calculated for all possible combinations of *Privacy Degrees* ϕ and *Explainability Degrees* ω and reported in the related cells of the compatibility matrix;

Step 4: The remaining empty cells of the compatibility matrix are filled with 0, because the corresponding combination of *Privacy Degrees* and *Explainability Degrees* do not satisfy the requirements defined by stakeholders. The resulted tri-dimensional matrix is represented in Fig. 10;

Step 5: The averaged *trade-off* score is calculated on the y dimension by aggregating the *trade-off* values obtained for the 3 datasets using (19) with their respective weights. For instance, the *trade-off* score cells for $\phi = 1.0$ and $\omega = 0.5$ would

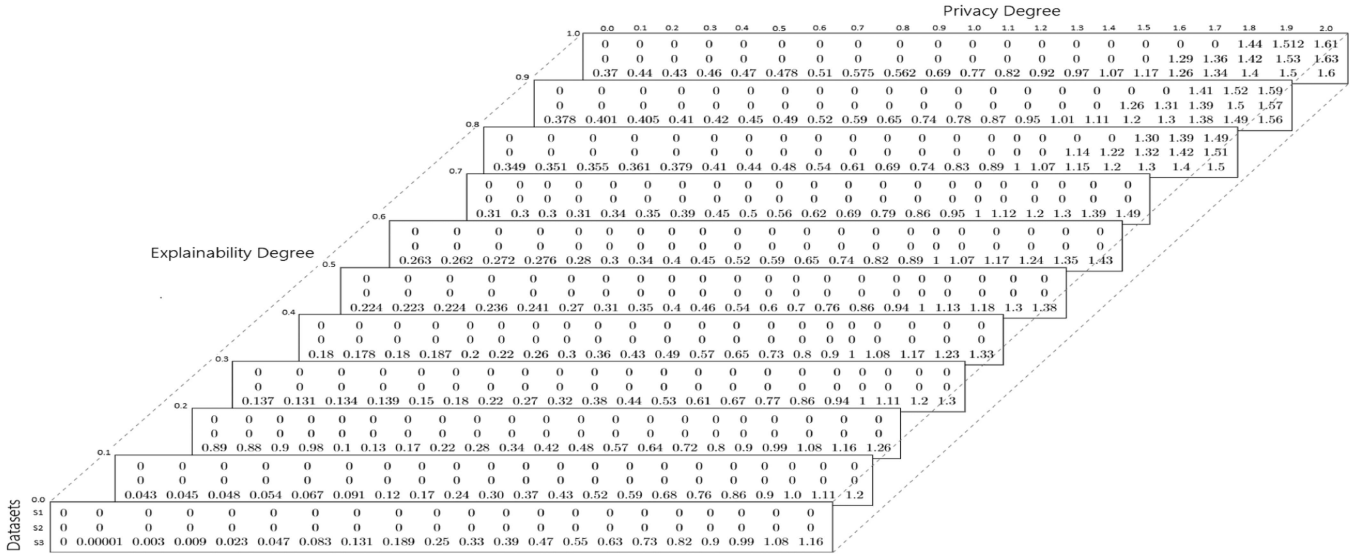


Fig. 10. Tri-dimensional compatibility matrix for three stakeholders using DP mechanism example.

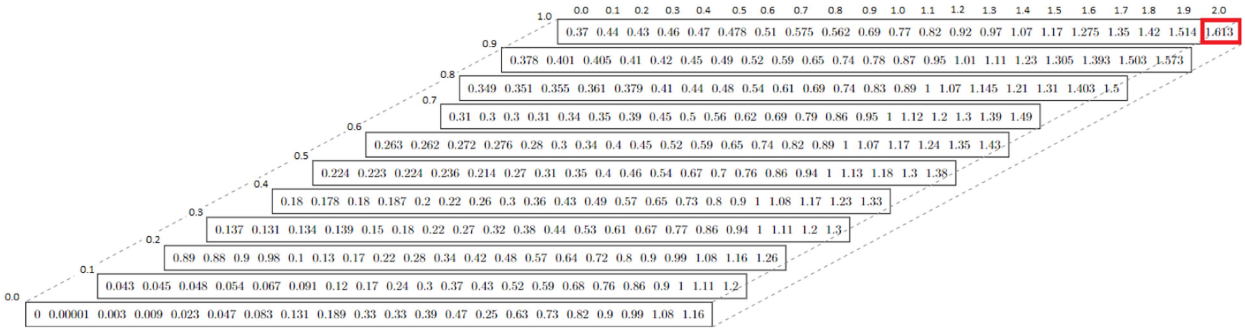


Fig. 11. Averaged tri-dimensional compatibility matrix for three stakeholders using DP mechanism example.

have the value 0 for $Stakeholder1$ and $Stakeholder2$, while it would be greater than 0 for $Stakeholder3$. Therefore the *trade-offs* aggregation equation will have the following weights: $w_1 = 0, w_2 = 0, w_3 = 1$. On the other hand, the *trade-off score* cells for $\phi = 1.8$ and $\omega = 0.7$ would have values greater than 0 for $Stakeholder1$, $Stakeholder2$, and $Stakeholder3$ since they have all defined requirements having these values for privacy and *Explainability Degrees*. Therefore the *trade-off* aggregation equation will have the following weights: the $w_1 = 1/3, w_2 = 1/3, w_3 = 1/3$. The averaged compatibility matrix is represented in Fig. 11;

Step 6: Finally, the combination of privacy and *Explainability Degrees* yielding the best-averaged *trade-off* score is considered the optimal solution, which is at $\phi = 2.0$ and $\omega = 1.0$ in our example as it returns the maximum *trade-off* score highlighted in Fig. 11.

VI. Use Cases and Experiments

This section reports the classification experiments performed on the three image datasets described in Section V-A, i.e., MNIST, FER, and CIFAR-10, using the two privacy-preserving techniques shown in Section II-A, i.e., *Autoencoders*, and *Differential Privacy*, and the explainability technique described in Section II-B, i.e., *SmoothGrad*. In the experiments, we use the metrics described in Section IV to measure the *Privacy Gain*, the *Explainability Gain*, and the *Utility Loss*, and then we choose the solution providing the best *trade-off* according to (19) or (20). We are considering a use case in which the stakeholders do not specify any privacy and explainability requirements.

The MNIST dataset is a well-known dataset of gray-scale handwritten digit images of size 28×28 pixels, classified into 10 classes, i.e., the digits from 0 to 9. The dataset consists of 70,000 images divided into a labeled training dataset of 60,000 images and a labeled test dataset of 10,000 images [38]. The FER dataset consists of gray-scale images of size 48×48 pixels, representing faces. This dataset is meant for facial expression recognition and the images are classified into 7 classes: angry, disgusted, fear, happy, neutral, sad, and surprised. The dataset consists of 32,298 images divided into a labeled training dataset of 28,709 images and a labeled test dataset of 3,589 images [39]. The CIFAR-10 dataset consists of colored images of size 32×32 and 3 channels, representing objects. This dataset is meant for object recognition and the images are classified into 10 classes: Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship, and Truck. The dataset consists of 60,000 images, with 6000 images per class. It is divided into a labeled training dataset of 50,000 images and a labeled test dataset of 10,000 images. For each of the three datasets, the experiments have been conducted by computing the compatibility matrix both for the *Differential*

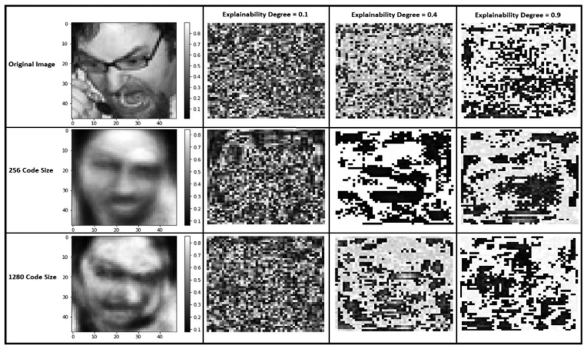


Fig. 12. *Autoencoders* with different *privacy degrees* and code (latent space representation) sizes on FER dataset.

Privacy and the *Autoencoders* techniques. For the former, the value of the noise parameter we used ranges from 0.0 to 2.0.

For the latter, since the dimensions and number of color channels of the images of the datasets are different, three different ranges for the code size have been chosen: for the MNIST dataset, the code size ranges between 1.02% and 78.57% of the original image size. For the FER dataset, the code size ranges between 0.35% and 89.58%. The CIFAR-10 dataset code size ranges between 8.33% and 83.33%. The Gaussian noise in *SmoothGrad* is set to be in [0,1] for all experiments. Fig. 12 presents a sample image of the FER dataset with the use of different privacy degrees of the Autoencoder method and several explainability degrees and the effect they have on the anonymized image and the resulting saliency maps.

A. Experimental Model Architecture

We point the reader to Appendix A for the description of the classifiers used to analyze the 3 datasets.

For conducting the experiments using the *Autoencoders* technique, we implemented our own prototype in Python. The original image is transformed by this prototype into the latent representation, whose dimension is used to control the *Privacy Degree*. The output is a reconstructed image perturbated based on the latent space size.

For *Differential Privacy*, the original image is processed using Tensorflow privacy framework.[6] It uses data patches and clipping for gradients computation and it averages multiple mini-gradient updates from the data mini-patches. Each mini-gradient value is clipped to restrict its individual impact, then these mini-gradients are averaged together, so no individual example will be memorized by the model. In the experiments, the sensitivity parameter default value is set to $s = 1.5$ like applied in Tensorflow-Privacy, and it represents the influence each training point in the mini-batch has on the gradient computation and its effect on the model parameters update. The clipping value is set to 1 and we define multiple noise values to control the added noise. The model returns the ϵ value as the privacy budget. Moreover, mini-batches are used to select a subset of the predefined size of the dataset during model training to compute gradients, error, and model parameters instead of using patches

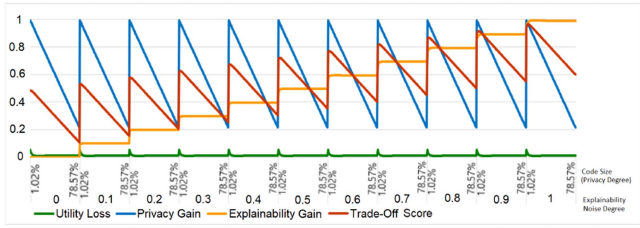


Fig. 13. Results computed on the MNIST dataset using autoencoders and SmoothGrad techniques.

that pass the entire dataset to the model to enhance performance and randomness.

Additionally, tf-explain framework[7] is used for enhancing the explainability using the *SmoothGrad* mechanism by defining multiple values of the noise in the interval [0,1]. For each classifier, this approach results in two matrices for each of the privacy mechanisms, each of them has a privacy and explainability parameter defined, *Privacy Gain*, *Explainability Gain*, and *Utility Loss* for the defined requirements. Afterward, a *trade-off* score is calculated for each row in these matrices based on the *trade-off* equation defined previously. Then the maximum *trade-off score* for each privacy-preserving mechanism is selected and the scores are compared between the two mechanisms with their parameters.

B. Experiments Adopting the Autoencoders Technique for Preserving Data Privacy

Fig. 13 present the results in a 2D representation of the experiment performed on the handwritten digit classification problem of the MNIST dataset, using the autoencoder technique for data privacy and the *SmoothGrad* mechanism for explainability. On the X-axis of Fig. 13 we report the two input parameters: the explainability noise (below) and the code size (above). For each value of the explainability noise, ranging from 0 to 1 with step 0.1, we computed the *trade-off* varying the value of the code size in the range from 1.02% to 78.57% of the original image size. In Section B of the Appendix, a 3D representation of Fig. 13 is provided.

The results show that there is a positive relationship between *Explainability Gain* (orange line) and the Explainability Noise Degree parameter, and a negative relation between the code size and the *Privacy Gain* (blue line). As a matter of fact, the *Explainability Gain* increases when the noise introduced by the *SmoothGrad* technique increases. The *Privacy Gain* (blue line), instead, is high when the code size is low, and it decreases with the code size. Moreover, we can see that the value of the explainability noise does not affect the *Privacy Gain*. The *Privacy Gain* plot is represented by a constant function, since the explainability mechanism is applied to investigate the models predictions after the analysis finishes. The observed trends of the *Explainability Gain* and of the *Privacy Gain* is observed in all the performed experiments, not reported here for the sake of brevity.

The *Utility Loss* (green line) is calculated using the EMD divergence between the histograms of the original dataset and

[6]https://github.com/tensorflow/privacy

[7]https://tf-explain.readthedocs.io/en/latest/

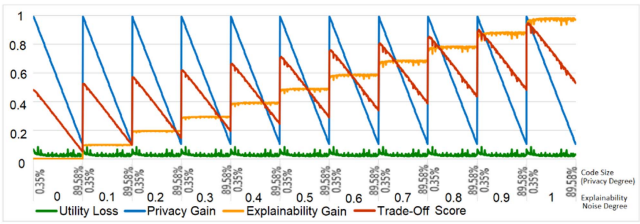


Fig. 14. Results computed on the FER dataset using autoencoders and SmoothGrad techniques.

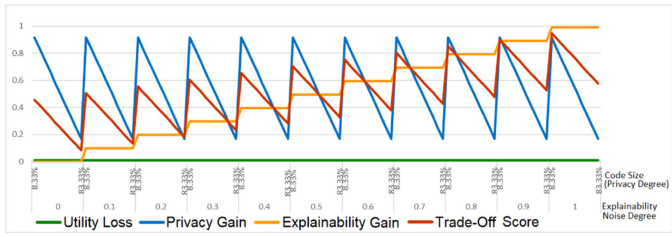


Fig. 15. Results computed on the CIFAR-10 dataset using autoencoders and SmoothGrad techniques.

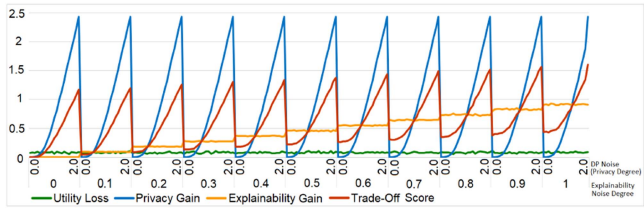


Fig. 16. DP with SmoothGrad on MNIST dataset results.

the anonymized dataset using (10). Its value is almost close to 0 for all the values of explainability noise and of code size, thus indicating that the *Utility Loss* is almost negligible and it is not significantly affected by such parameter values when the *Autoencoders* technique is applied.

Finally, the trend of the *trade-off* scores (red line) shows that such score is negatively affected by the code size and positively affected by the explainability noise parameter. Hence, the lower the code size and the higher the noise, the higher the *trade-off*, which means that the greater compression rate and privacy implemented on the data with the greater explainability parameter applied, return the better *trade-off* score.

Fig. 14 presents the results of the experiment conducted on the FER dataset for facial expression recognition in a 2D representation, where the autoencoder technique was used for data privacy preservation, and the *SmoothGrad* mechanism was used for explainability. For each value of the explainability noise, ranging from 0 to 1 with step 0.1, we computed the value of the *trade-off* varying the code size, ranging from 0.35% to 89.58% of the original image size. The trends of the *Explainability Gain* and of the *Privacy Gain* are as expected. The *Utility Loss* (green line) is calculated using the EMD divergence between the histograms of the original dataset and the anonymized dataset. In this experiment too, its value is close to 0 for all the values of explainability noise and code size, thus indicating that the *Utility Loss* is almost negligible and it is not significantly affected by such parameters.

The *trade-off* score (red line) is negatively affected by the code size and positively affected by the explainability noise parameter, as also observed in the previous experiment. Hence, in this case, too, the lower the code size and the higher the explainability noise are, the higher the *trade-off* is, meaning that the greater compression rate and privacy implemented with the greater explainability parameter applied, returns a better *trade-off* score. In our experiment, the best *trade-off* score is obtained with the explainability noise degree equal to 1 and the code size equal to 0.35% of the original image size.

Fig. 15 present the results of the experiment conducted on the CIFAR-10 dataset for object recognition in a 2D representation, where the autoencoder technique was used for data privacy preservation, and the *SmoothGrad* mechanism was used for explainability. For each value of the explainability noise, ranging from 0 to 1 with step 0.1, we computed the value of the *trade-off* varying the code size, ranging from 8.33% to 83.33% of the original image size.

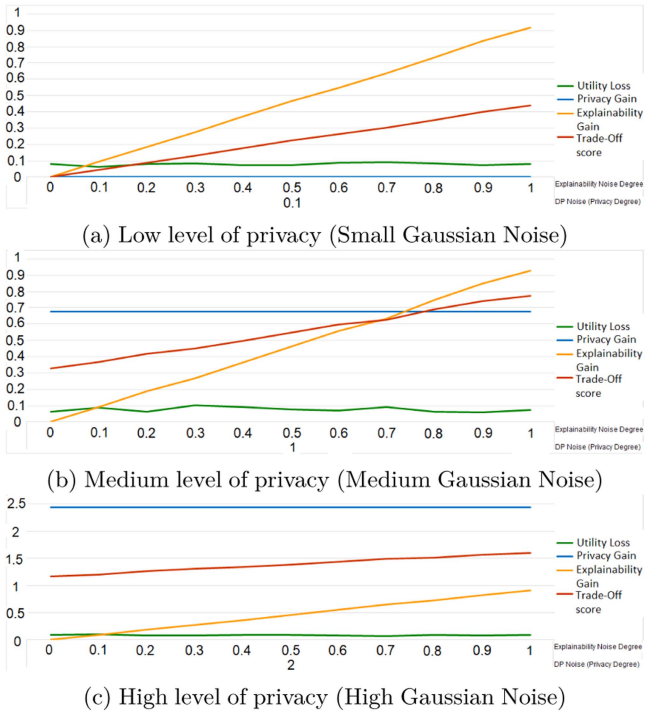
The *Explainability Gain* and of the *Privacy Gain* trends are as expected. The *Utility Loss* (green line) is calculated using the EMD divergence between the histograms of the original dataset and the anonymized dataset. Its value is close to 0 for all the values of explainability noise and code size and is not significantly affected by such parameters.

The *trade-off* score (red line) is negatively affected by the code size and positively affected by the explainability noise parameter, as observed in the previous experiments.

C. Experiments Adopting the Differential Privacy Technique for Preserving Data Privacy

Fig. 16 reports, in a 2D representation, the results of an experiment that has been conducted on the same testbed of the first experiment of the previous section (handwritten digits classification problem, the MNIST dataset, and the *SmoothGrad* technique for increasing explainability), with the only difference being that we used the *Differential Privacy* technique to protect data privacy. On the X-axis we report the two input parameters: the explainability noise (below) and the *Differential Privacy* noise multiplier (above). For each value of the explainability noise, ranging from 0 to 1 with step 0.1, we computed the *trade-off* value varying the *Differential Privacy* noise multiplier from 0 to 2 with step 0.1. The DP noise multiplier value of 0 means no privacy restriction is applied, resulting in a returned ϵ value of infinity according to (3). Hence, the higher the noise value, the lower the ϵ value, and the higher the *Privacy Gain*. Therefore, we set the maximum value of the DP noise multiplier to 2, as it results in ϵ value close to 0 and a high *Privacy Gain* accordingly. In Section B of the Appendix, a 3D representation of Fig. 16 is provided. Fig. 17(a)–(c) report the results of the same experiment, each for a single value of the *Differential Privacy* noise multiplier (i.e., 0.1 in Fig. 17(a), 1.0 in Fig. 17(b), and 2.0 in Fig. 17(c)) and varying the value of the explainability noise in the previously mentioned range.

In this experiment, since we used the *Differential Privacy* technique, the *Utility Loss* depends on the accuracy loss, and the *Privacy Gain* depends on the value of ϵ, as shown in (7a). Hence, Fig. 18(a) shows the values of ϵ returned applying the *Differential Privacy* technique varying the value of the *Differential*



(a) Low level of privacy (Small Gaussian Noise)

(b) Medium level of privacy (Medium Gaussian Noise)

(c) High level of privacy (High Gaussian Noise)

Fig. 17. DP and *SmoothGrad* results on MNIST dataset with varying privacy levels.

Privacy noise multiplier in the same range as Fig. 16. When the *Differential Privacy* noise multiplier is set to 0, which means no privacy restriction is applied, the returned epsilon value is infinity according to (3). The values of the *Privacy Gain* shown in Fig. 16 are then obtained by applying (7a) to the values of ϵ in Fig. 18(a). Hence, the higher the value of the noise, the lower the value of epsilon, and the higher is the *Privacy Gain*, as shown by the blue line in the graph of Fig. 16. Differently from Fig. 14, in Fig. 16 the privacy degree has a different semantic, in fact, a higher code size (Fig. 14) implies a lower privacy level, whilst a higher DP-Noise (Fig. 16) implies a higher privacy level. This has been discussed in detail in Section IV-A. The *Utility Loss* (green line in Fig. 16) does not increase significantly when increasing the explainability and privacy parameters. The results comparison on the classification accuracy on both the original dataset and the dataset anonymized through *Differential Privacy* and the *SmoothGrad* techniques is reported in Fig. 18(b) with *Utility Loss* represented as a percentage.

For what concerns the *trade-off* score, from the graph of Fig. 16 (red line) we observe that it increases with the *Differential Privacy* noise multiplier parameter and the explainability noise parameter as from Fig. 17(a)–(c). In our experiment, the best *trade-off* score occurs when the explainability noise parameter is set to 1 and the *Differential Privacy* noise multiplier is set to 2. This means that the higher *Privacy Degree* and the higher explainability parameter values return a better *trade-off* within the value range defined.

The next experiment was conducted on the FER Dataset for facial expression recognition problem using the *Differential Privacy* technique to preserve data privacy and the *SmoothGrad* mechanisms for increasing explainability (Fig. 19). As in the previous experiment, for each value of the explainability noise, ranging from 0 to 1 with step 0.1, we computed the *trade-off* value varying the value of the *Differential Privacy* noise multiplier within the interval [0,2] with step 0.1.

The *Privacy Gain* depends on the values of ϵ computed by the *Differential Privacy* technique, which are reported in Fig. 20(a). When the *Differential Privacy* noise multiplier is set to 0, which means no privacy restrictions, the *Differential Privacy* technique returns an epsilon value of infinity. The higher value of the noise is defined, the lower value of epsilon is returned, as shown in Fig. 20(a), which means better *Privacy Gain* (Fig. 19).

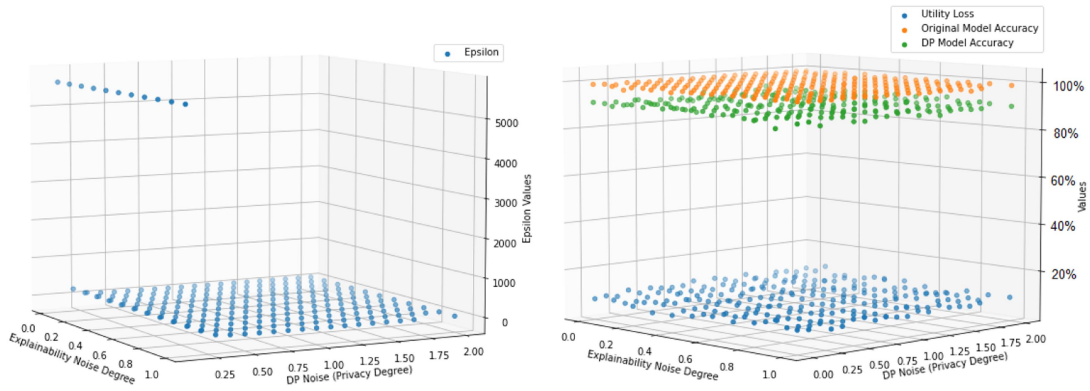
The *Utility Loss* (green line in Fig. 19) is negatively affected when the value of the noise multiplier is greater than zero. However, it does not increase significantly when increasing the noise multiplier as well as the explainability parameter. The accuracy of the results applying the classification function to the original dataset and to the anonymized dataset by applying the *Differential Privacy* technique and *SmoothGrad* is reported in Fig. 20(b) with *Utility Loss* represented as a percentage. We observe that there is a decrease in the classification accuracy after applying *Differential Privacy* and *SmoothGrad*, and we noticed that the loss of utility in percentage is lower than 48% for all degrees of privacy and explainability. Finally, the *trade-off* score (red plot in the graph of Fig. 19) increases with the *Differential Privacy* noise multiplier parameter and the explainability noise parameter. The best *trade-off* score occurs when the explainability noise parameter is set to 1 and the *Differential Privacy* noise multiplier is set to 2. This means that the higher privacy restriction and the higher explainability parameter values return a better *trade-off* within the value range defined.

The last experiment was conducted on the object recognition problem applied to the CIFAR-10 dataset using the *Differential Privacy* technique to preserve data privacy and the *SmoothGrad* mechanisms for increasing explainability. The results are reported in Fig. 21 in a 2D representation. For each value of the explainability noise, ranging from 0 to 1 with step 0.1, we computed the *trade-off* value varying the value of the *Differential Privacy* noise multiplier within the interval [0,2] with step 0.1. Section B of the Appendix contains a 3D visualization of Fig. 21.

The *Privacy Gain* depends on the values of ϵ computed by the *Differential Privacy* technique, which are reported in Fig. 22(a). In this case too, when the *Differential Privacy* noise multiplier is set to 0 the *Differential Privacy* technique returns an epsilon value of infinity. The higher value of the noise is defined, the lower value of epsilon is returned, as shown in Fig. 22(a), which means better *Privacy Gain* (Fig. 21).

The *Utility Loss* (green line in Fig. 21) is negatively affected when the noise multiplier is greater than zero, but it does not increase significantly when increasing the noise multiplier as well as the explainability parameter. The accuracy of the results applying the classification function to the original dataset and to the anonymized dataset by applying the *Differential Privacy* technique and *SmoothGrad* is reported in Fig. 22(b) with *Utility Loss* represented as a percentage. We observe that there is no major decrease in the classification accuracy after applying *Differential Privacy* and *SmoothGrad*.

Finally, the *trade-off* score (red plot in the graph of Fig. 21) increases with the *Differential Privacy* noise multiplier parameter



(a) Epsilon Results on MNIST Dataset

(b) Utility Loss on MNIST Dataset

Fig. 18. Utility loss and epsilon results on MNIST dataset.

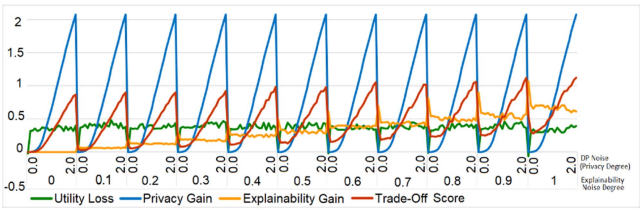


Fig. 19. DP with SmoothGrad on FER dataset results.

and the explainability noise parameter. The best *trade-off* score occurs when the explainability noise parameter is set to 1 and the *Differential Privacy* noise multiplier is set to 2. This means that the higher privacy restriction and the higher explainability parameter values return a better *trade-off* within the value range defined.

D. Results Discussion

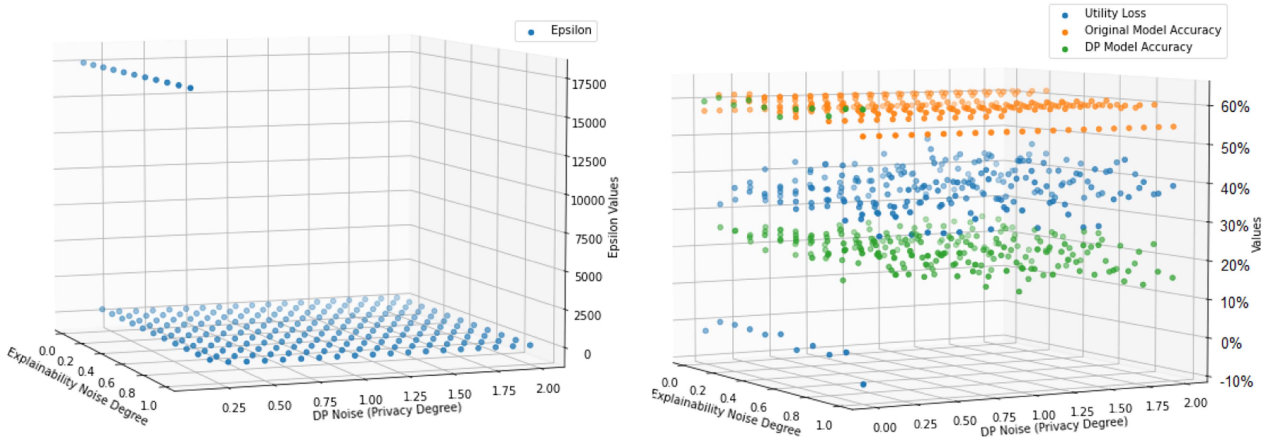
The main objective of this study is to propose a comprehensive model for image data analysis that places equal importance on data *privacy*, *explainability*, and *data utility*. In our experiments, we investigated the effect of varying *Privacy* and *Explainability Degrees* on the *trade-off score*, which is defined as a function of data *Utility Loss*, *Privacy Gain*, and *Explainability Gain*. In the three classification use cases investigated in our experiments and for both *Autoencoders* and *Differential Privacy* techniques, we found out that using a privacy mechanism resulted in a *Utility Loss*, but the loss does not increase significantly when the value of the *Privacy Degree* parameter is increased. Interestingly, the impact on the *Utility Loss* of the adoption of the *Differential Privacy* technique was more significant on the facial expression recognition model than on the digit classification model and the object recognition model. The reason behind this is that, on the one hand, the facial expression recognition model uses face attributes, such as the shape of the eyes, that in many cases might be identifying or sensitive for different individuals. On the other hand, the *Differential Privacy* technique preserves data privacy by minimizing model memorization of such identifying face attributes, thus affecting the accuracy of the analysis function, as explained in Section IV-A.

The *trade-off* between *privacy*, *explainability*, and *data utility* is a crucial consideration when configuring our approach for specific applications. Our experimental results indicate a relevant positive relationship between the *trade-off score* and the increase in both the *privacy degree* and *explainability degree*. This relationship is primarily attributed to the improved *privacy gain* and *explainability gain* achieved as the values of these parameters are incremented. However, this increase of the *trade-off* is restricted by the decrease in data utility, that generally starts with a relatively high loss with the utilization of a privacy mechanism with a low privacy degree, then it remains somehow steady with the increase of the privacy degree.

Another aspect to be analysed of the *trade-off* model we defined is its sensitivity to the variations of input parameters values. In particular, from the experimental results we can observe that, as we tune the *privacy degree* parameter to increase the privacy level, a relevant increase in the *trade-off* value occurs. For instance, in Fig. 16, choosing 0.5 as value for the *explainability noise degree*, we can see that increasing the *DP noise* value from 0.0 to 2.0 we have an increase in *trade-off score* from 0.2 to 1.3. Similarly, when we increase the *explainability degree* parameter a meaningful increase of the *trade-off* score can be observed. For instance, in Figs. 16 and 17(b) we can see that, choosing 1.0 as value for the *DP noise* degree (Medium Gaussian Noise), increasing the *Explainability noise* degree value from 0.0 to 1.0 results in an increase in *trade-off score* from 0.32 to 0.78.

In summary, our experiments reveal that the privacy and explainability degree parameters play a pivotal role in shaping the outcomes of our methodology. Achieving a balance between privacy, explainability, and data utility is a non-trivial task and should be tailored to the specific requirements of each use case and per-user requirements. These *trade-offs* necessitate careful consideration, and the choice of parameter settings should align with the specific objectives of the application. Furthermore, practical implementation often involves domain-specific requirements and constraints, making parameter tuning a nuanced process. Thus, it is recommended to experiment with different parameter configurations and assess their impact on privacy, explainability, and data utility to find the optimal balance for different use cases.

Our method formalizes the common elements of trustworthy AI, *data privacy*, *data utility*, and *model explainability*, and can be applied to various data analysis types without losing generality. While we cannot elaborate on the applicability of



(a) Epsilon Results on FER Dataset

(b) Utility Loss on FER Dataset

Fig. 20. Utility loss and epsilon results on FER dataset.

TABLE II
COMPARISON WITH THE EXISTING METHODOLOGIES ADDRESSING COMBINED ASPECTS OF PRIVACY, UTILITY, AND EXPLAINABILITY

Reference	Privacy	Explainability	Utility	*trade-off*	Collaborative Learning	Dataset	Experiment
[49]	Yes	Yes	No	No	No	Image	Yes
[50]	Yes	Yes (explain AI system's privacy guarantees only)	Yes	No	No	General	No
[51]	Yes	Yes	Yes	No	No	Image	No
[12]	Yes	Yes	Yes	Yes (Sample-based on pairs of trust aspects only)	No	Image	Yes
[32]	Yes	Yes	Yes (Accuracy)	No	No	Image	Yes
[52]	Yes (for explanations sharing only)	Yes	Yes (Accuracy & F1 Score)	No	No	Image	Yes
[53]	Yes	Yes	No	No	No	Tabular	Yes
[14]	Yes	Yes	Yes	Yes	Yes	Tabular	Yes
Our approach	Yes	Yes	Yes	Yes	Yes	Image	Yes

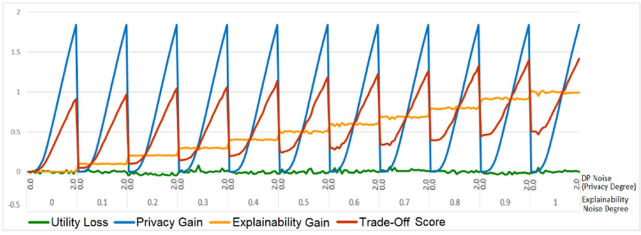


Fig. 21. DP with SmoothGrad on CIFAR-10 dataset results.

our method to all image analysis models, we offer the necessary tools and methodologies, as demonstrated by our experiments on three distinct problems. Our method can serve as a guideline for designing customized image analytics systems that align with user requirements and specifications for a safer, more explainable, and trustworthy analytics experience.

VII. Related Work

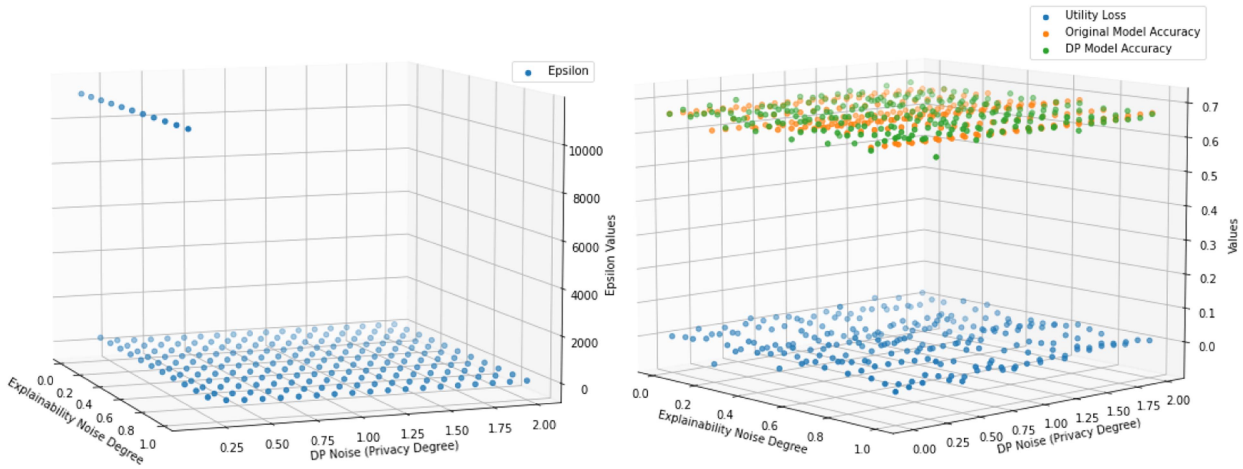
This section reviews related work in Privacy-Preserving Machine Learning, Explainable Artificial Intelligence, and Privacy-preserving AI systems with explainability and possible *trade-off* computation.

A. Privacy-Preserving Machine Learning

Sample anonymization methods for images include resolution quality reduction techniques [40], image obfuscation [41], image blurring [42], face swapping [43], and image altering with the K-Same algorithm [44]. Nevertheless, these methods can cause a major drop in the accuracy of the model results or do not provide strong privacy guarantees. Autoencoder Neural Networks are used as a compression mechanism and also for privacy protection [22]. Similarly, in [26], a method has been proposed to privacy-preserving surveillance video streams anomaly detection. It employs a privacy-preserving mechanism based on autoencoder neural networks implemented with three different types of differential private optimizers. Furthermore, the prevailing approach for assuring the privacy guarantees of a deep learning model predominantly relies on the application of optimization techniques that enforce *Differential Privacy*. *Differential Privacy* has proven to be a successful defence against several models' privacy attacks, but its drawback lies in the degradation of the models' performance. In [36], the authors assess the effectiveness of the DP-SGD algorithm in comparison to conventional optimization approaches coupled with regularization techniques. They analyze the resulting models' utility, training performance, and the effectiveness of membership inference and model inversion attacks against the learned models.

B. Explainable Artificial Intelligence (XAI)

XAI emerged to produce human-level explanations for complex AI models that are not transparent by design [45]. Examples of such complex models include Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) [46]. XAI methods can be applied



(a) Epsilon Results on CIFAR-10 Dataset

(b) Utility Loss on CIFAR-10 Dataset

Fig. 22. Utility loss and epsilon results on CIFAR-10 dataset.

pre-model to explore the dataset used, *in-model* as in interpretable by design models, or *post-hoc* for methods that are applied to complex models. From a model scope perspective, XAI methods are either *model specific* applicable to a single model or a type of models [47], or *model agnostic* being applicable to any model [48]. Furthermore, explanation methods can produce *local explanation* related to a certain prediction, or *global explanations* enforced in the model structure and are referred to as dataset-level explanations since the techniques applied to produce such explanations examine the model used in the entire dataset [6].

C. *Privacy-Preserving AI Systems With Explainability and Possible* Trade-Off *Computation*

In [49], a *Differential Privacy* mechanism combined with feature-based model explanation techniques and gradient descent algorithm was proposed as an adaptive approach for privacy protection in local model-agnostic explanations. In this approach, the analysis accuracy was not considered in the evaluation, and no optimization of the considered evaluation metrics was performed. In [52], the authors discuss that explainable privacy-preserving systems can be improved by defining the privacy guarantees that are being treated as constraints to be optimized, where the optimization function is the *Utility Loss*. This method has been proposed as a theory with no application, no *trade-off* was presented between the privacy and *Utility Loss*, and explainability was presented only to explain what privacy guarantees a system offers, not the prediction an AI system makes. Furthermore, in [51] a fuzzy logic system has been proposed for indoor gestures classification as a replacement to the traditional AI systems, with a close accuracy but completely interpretable compared to the traditional systems. This method can be applied to only one type of analysis method which is fuzzy logic, it does not consider controlling the privacy, utility, or explainability, and does not optimize them or do a *trade-off* on these parameters as a result. The authors in [12] propose a solution to investigate the *trade-off* between explainability, privacy, and accuracy. The solution deploys a family of inherently explainable models that use Locally Linear Maps (LLM) for mapping between input data and the class score function in the classification problem. DP was leveraged to ensure privacy at local and global levels of explanation. The proposed approach was adopted only on simple datasets with the *trade-off* and optimization criteria performed on Sample-based on pairs of trust aspects only. In [32], the authors introduced a methodology aimed at enhancing the trustworthiness of AI systems. Their approach incorporates principles of privacy, explainability, and fairness into the design, specifically applied in the context of face recognition. The authors employed Homomorphic Encryption for privacy protection, provided explainable predictions with local and global explanations, and ensured fairness by learning a fair representation from the data. Fairness constraints are set to measure the impact on result accuracy and corresponding explanations, employing Demographic Parity as the fairness metric. However, the study lacked explicit control over the levels of explainability and privacy, and did not introduce a *trade-off* criterion for balancing these crucial aspects.

According to [52], case-based explanations are frequently applied in domains featuring sensitive visual data, such as medical diagnosis. They involve retrieving instances of similar disease-matching images as the entered image for diagnoses, thus yielding additional insights to explain a diagnosis. However, retrieving an image containing sensitive identity information from a private dataset has the potential to compromise the privacy of the individual depicted. Therefore, an anonymization process is needed to ensure the removal of identifying attributes from the image before for sharing as an explanation. The authors formalize the generation of Privacy-Preserving Case-Based Explanations as a multi-objective problem for future work, aiming to minimize the inclusion of identity information while maintaining realism and preserving explanatory evidence within the images. In [53], the authors investigated the impact of privacy-preserving methods, specifically masking techniques, employed in machine learning algorithms on tabular datasets. Their focus was on explainability, gauged through Shapley values. The study involved two analyses: one examining differences in Shapley values and another assessing the rank correlation of these values.

The findings suggest that, under certain assumptions, explainability and privacy are not incompatible. Nevertheless, it is important to note that the paper did not measure privacy gain, utility loss, or explore the *trade-off* and optimization aspects in this context. The authors of [14] propose an approach leveraging machine learning techniques tailored for tabular datasets. The model establishes a comprehensive *trade-off* optimization criterion that considers data privacy, model explainability, and data utility. It involves regulating the privacy parameter within the privacy-preserving mechanism applied to analysis algorithms and explainability techniques. The method explores various configurations for the privacy parameter, identifying the optimal setup that maximizes privacy gain and explainability similarity while minimizing the impact on data utility. Table II provides a general comparison of our proposed approach with existing methodologies addressing the combined trust aspects of privacy, utility, and explainability.

VIII. Conclusion

In this paper, we have proposed a model which pushes toward the direction of Trustworthy AI, considering altogether a set of measures that, up to now, have been mainly considered as separate problems, representing privacy, data utility and model explainability. We proposed three use cases based on a novel deep learning-based approach for image datasets analysis, which preserves data privacy and achieves model explanation with an optimal compromise among data privacy, data utility, and model explainability. The experiments have demonstrated how it is possible to improve, by regulating noise levels, both the level of privacy and of explainability, with minimal impact on data utility. As a future extension, we plan to apply different perturbation mechanisms like the Generative Adversarial Network (GAN), and investigating trade off options for the utilization of cryptographic methods.

References

[1] D. W. Chadwick et al., "A cloud-edge based data security architecture for sharing and analysing cyber threat information," *Future Gener. Comput. Syst.*, vol. 102, pp. 710–722, 2020.

[2] Z. Xu, "An empirical study of patients' privacy concerns for health informatics as a service," *Technological Forecasting Social Change*, vol. 143, pp. 297–306, 2019.

[3] M. Strobel, "Aspects of transparency in machine learning," in *Proc. 18th Int. Conf. Auton. Agents MultiAgent Syst.*, 2019, pp. 2449–2451.

[4] M. Sheikhalishahi, A. Saracino, F. Martinelli, and A. L. Marra, "Privacy preserving data sharing and analysis for edge-based architectures," *Int. J. Inf. Secur.*, vol. 21, pp. 79–101, 2021.

[5] M. Al-Rubaie and J. M. Chang, "Privacy-preserving machine learning: Threats and solutions," *IEEE Secur. Privacy*, vol. 17, no. 2, pp. 49–58, Mar./Apr. 2019.

[6] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, 2021, Art. no. 18.

[7] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, vol. 11700. Berlin, Germany: Springer, 2019.

[8] L. Sweeney, "k-Anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 05, pp. 557–570, 2002.

[9] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, pp. 3–es, 2007.

[10] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, 2007, pp. 106–115.

[11] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Computation*, 2008, pp. 1–19.

[12] F. Harder, M. Bauer, and M. Park, "Interpretable and differentially private predictions," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4083–4090.

[13] R. Shokri, M. Strobel, and Y. Zick, "On the privacy risks of model explanations," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2021, pp. 231–241.

[14] W. Abbasi, P. Mori, and A. Saracino, "The explainability-privacy-utility trade-off for machine learning-based tabular data analysis," in *Proc. 20th Int. Conf. Secur. Cryptography*, 2023, pp. 511–519.

[15] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[16] W. H. L. Pinaya, S. Vieira, R. Garcia-Dias, and A. Mechelli, "Autoencoders," in *Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 193–208.

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

[18] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. Annu. Int. Conf. Theory Appl. Cryptographic Techn.*, 2006, pp. 486–503.

[19] S. Raskhodnikova, A. Smith, H. K. Lee, K. Nissim, and S. P. Kasiviswanathan, "What can we learn privately," in *Proc. 54th Annu. Symp. Foundations Comput. Sci.*, 2008, pp. 531–540.

[20] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2009, pp. 19–30.

[21] B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *Proc. 28th USENIX Secur. Symp.*, 2019, pp. 1895–1912.

[22] O. Hajihassani, O. Ardakanian, and H. Khazaei, "Latent representation learning and manipulation for privacy-preserving sensor data analytics," in *Proc. IEEE Second Workshop Mach. Learn. Edge Sensor Syst.*, 2020, pp. 7–12.

[23] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Mobile sensor data anonymization," in *Proc. Int. Conf. Internet Things Des. Implementation*, 2019, pp. 49–58.

[24] M. D'Souza et al., "Autoencoder as a new method for maintaining data privacy while analyzing videos of patients with motor dysfunction: Proof-of-concept study," *J. Med. Internet Res.*, vol. 22, 2020.

[25] J. Liu, J. Liu, P. Li, and Z. Kuang, "Embedded autoencoders: A novel framework for face de-identification," in *Proc. Int. Cogn. Cities Conf.*, 2019, pp. 154–163.

[26] G. Giorgi, W. Abbasi, and A. Saracino, "Privacy-preserving analysis for remote video anomaly detection in real life environments," *J. Wireless Mobile Netw. Ubiquitous Comput. Dependable Appl.*, vol. 13, no. 1, pp. 112–136, 2022.

[27] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[28] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[29] K.-Y. Lam, X. Lu, L. Zhang, X. Wang, H. Wang, and S. Q. Goh, "Efficient FHE-based privacy-enhanced neural network for trustworthy AI-as-a-service," *IEEE Trans. Dependable Secure Comput.*, early access, Jan. 12, 2024, doi: 10.1109/TDSC.2024.3353536.

[30] K. Sahinbas and F. O. Catak, "Secure multi-party computation-based privacy-preserving data analysis in healthcare iot systems," in *Interpretable Cognitive Internet of Things for Healthcare*. Berlin, Germany: Springer, 2023, pp. 57–72.

[31] D. Melanson, R. Maia, H.-S. Kim, A. Nascimento, and M. D. Cock, "Secure multi-party computation for personalized human activity recognition," *Neural Process. Lett.*, vol. 55, no. 3, pp. 2127–2153, 2023.

[32] D. Franco, L. Oneto, N. Navarin, and D. Anguita, "Toward learning trustworthily from data combining privacy, fairness, and explainability: An application to face recognition," *Entropy*, vol. 23, no. 8, 2021, Art. no. 1047.

[33] Q.-X. Huang, W. L. Yap, M.-Y. Chiu, and H.-M. Sun, "Privacy-preserving deep learning with learnable image encryption on medical images," *IEEE Access*, vol. 10, pp. 66345–66355, 2022.

[34] J. Chen, W. H. Wang, and X. Shi, "Differential privacy protection against membership inference attack on machine learning for genomic data," in *Proc. BIOCOMPUTING: Proc. Pacific Symp.*, 2021, pp. 26–37.

[35] P. Stock, I. Shilov, I. Mironov, and A. Sablayrolles, "Defending against reconstruction attacks with Rényi differential privacy," 2022, *arXiv:2202.07623*.
[36] E. Lomurno and M. Matteucci, "On the utility and protection of optimization with differential privacy and classic regularization techniques," in *Proc. Int. Conf. Mach. Learn., Optim., Data Sci.*, 2022, pp. 223–238.
[37] 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 *Proc. IEEE Symp. Secur. Privacy*, 2019, pp. 739–753.
[38] Y. LeCun, C. Cortes, and C. J. Burges, "Mnist handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist
[39] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.*, 2013, pp. 117–124.
[40] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, "Privacy-preserving human activity recognition from extreme low resolution," in *Proc. 31st AAAI Conf. Artif. Intell.*2017, pp. 4255–4262.
[41] J. Wang, B. Amos, A. Das, P. Pillai, N. Sadeh, and M. Satyanarayanan, "A scalable and privacy-aware IoT service for live video analytics," in *Proc. 8th ACM Multimedia Syst. Conf. Assoc. Comput. Machinery*, 2017, pp. 38–49.
[42] D. J. Butler, J. Huang, F. Roesner, and M. Cakmak, "The privacy-utility tradeoff for remotely teleoperated robots," in *Proc. ACM/IEEE 10th Annu. Int. Conf. Hum.-Robot Interaction*, 2015, pp. 27–34.
[43] Y. Zhong, R. Arandjelović, and A. Zisserman, "Faces in places: Compound query retrieval," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 56.1–56.12.
[44] E. M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 2, pp. 232–243, Feb. 2005.
[45] A. Rai, "Explainable AI: From black box to glass box," *J. Acad. Marketing Sci.*, vol. 48, no. 1, pp. 137–141, 2020.
[46] F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA, USA: Harvard Univ. Press, 2015.
[47] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS One*, vol. 10, no. 7, 2015, Art. no. e0130140.
[48] K. N. Ramamurthy, B. Vinzamuri, Y. Zhang, and A. Dhurandhar, "Model agnostic multilevel explanations," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 5968–5979.
[49] N. Patel, R. Shokri, and Y. Zick, "Model explanations with differential privacy," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, 2022, pp. 1895–1904.
[50] J. Ramon and M. Basu, "Interpretable privacy with optimizable utility," in *Proc. ECML/PKDD workshop eXplainable Knowl. Discov. Data Mining*, 2020, pp. 492–500.
[51] J. Rožman, H. Hagras, J. A. Perez, D. Clarke, B. Müller, and S. F. Data, "Privacy-preserving gesture recognition with explainable type-2 fuzzy logic based systems," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2020, pp. 1–8.
[52] H. Montenegro, W. Silva, A. Gaudio, M. Fredrikson, A. Smailagic, and J. S. Cardoso, "Privacy-preserving case-based explanations: Enabling visual interpretability by protecting privacy," *IEEE Access*, vol. 10, pp. 28333–28347, 2022.
[53] A. Bozorgpanah, V. Torra, and L. Aliahmadipour, "Privacy and explainability: The effects of data protection on Shapley values," *Technologies*, vol. 10, no. 6, 2022, Art. no. 125.

Wisam Abbasi is currently working toward the PhD degree with the Computer Science Program, Pisa University. She is also a research fellow with CNR-IIT, Pisa. Her research interests include design and development of trustworthy AI models that implement privacy-preserving data analysis mechanisms and interpretable (explainable) techniques.

Paolo Mori received the PhD degree. He is currently a senior researcher in the Trust, Security, and Privacy research unit with CNR-IIT. He has co-authored more than 150 papers on trust, security, and privacy in distributed systems, such as Cloud, IoT, and mobile devices. His research interests include Blockchain technology and its applications. Since 2016, he has been the program co-chair of the International Conference of Information System Security and Privacy. He is usually actively involved in EU-funded research projects on information and communication security.

Andrea Saracino received the PhD degree. He is associate professor of applied cybersecurity and AI with Scuola Superiore Universitaria Sant'Anna. He co-authored more than 80 papers on access control, ML applied to cybersecurity, mobile and distributed system security. He coordinates and participates to EU and national research project. He co-chairs the Cybersecurity and AI group in the Italian Association for Artificial Intelligence.