Received September 28, 2021, accepted October 24, 2021, date of publication November 2, 2021,
date of current version November 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3124844

Privacy-Preserving Generative Adversarial Network for Case-Based Explainability in Medical Image Analysis

HELENA MONTENEGRO, (Member, IEEE), WILSON SILVA, (Student Member, IEEE), AND JAIME S. CARDOSO, (Senior Member, IEEE)
Faculty of Engineering, University of Porto, 4200-465 Porto, Portugal
INESC TEC, 4200-465 Porto, Portugal

Corresponding author: Helena Montenegro (up201604184@edu.fe.up.pt)

This work was supported in part by the Project TAMI—Transparent Artificial Medical Intelligence by the ERDF—European Regional Fund through the North Portugal Regional Operational Program—NORTE 2020 under Grant NORTE-01-0247-FEDER-045905, in part by the Portuguese Foundation for Science and Technology—FCT through the Carnegie Mellon University (CMU)—Portugal International Partnership, and in part by the Portuguese Foundation for Science and Technology—FCT within Ph.D. degree under Grant SFRH/BD/139468/2018.

ABSTRACT Although Deep Learning models have achieved incredible results in medical image classification tasks, their lack of interpretability hinders their deployment in the clinical context. Case-based interpretability provides intuitive explanations, as it is a much more human-like approach than saliency-map-based interpretability. Nonetheless, since one is dealing with sensitive visual data, there is a high risk of exposing personal identity, threatening the individuals' privacy. In this work, we propose a privacy-preserving generative adversarial network for the privatization of case-based explanations. We address the weaknesses of current privacy-preserving methods for visual data from three perspectives: realism, privacy, and explanatory value. We also introduce a counterfactual module in our Generative Adversarial Network that provides counterfactual case-based explanations in addition to standard factual explanations. Experiments were performed in a biometric and medical dataset, demonstrating the network's potential to preserve the privacy of all subjects and keep its explanatory evidence while also maintaining a decent level of intelligibility.

INDEX TERMS Case-based interpretability, deep learning, generative adversarial networks, privacy-preserving machine learning, medical image analysis.

I. INTRODUCTION

Deep Learning has achieved outstanding results in most image classification tasks, including medical imaging [1], [2]. Nonetheless, most of these models are "black-boxes" whose predictions are difficult for humans to understand and, consequently, trust. Moreover, their outstanding performance sometimes relies on confounding factors rather than application-related features [3], [4]. Due to the lack of interpretability of Deep Learning algorithms, their application in real-world contexts, namely in clinics, is hindered. To overcome this problem, several interpretability methods have been proposed to enhance the transparency in the decision-making process and improve the trust in the results of the models [5].

Case-based interpretability techniques are very much in line with human reasoning, providing intuitive explanations through the presentation of representative examples [6]–[8]. These examples are selected using retrieval systems that can find the most semantically similar cases from a pool of well-curated candidates, explaining the observation under analysis [9]–[12]. However, these methods cannot be applied to contexts where the data exposes identity, such as in the medical scene, where privacy is a major concern. In [13], the authors show the weaknesses in the application of current privacy-preserving methods to medical data. Most of the current strategies fail to preserve relevant semantic features that serve as explanatory evidence in the context of case-based explanations. Furthermore, some privacy-preserving methods

The associate editor coordinating the review of this manuscript and approving it for publication was Diego Oliva.

also fail to ensure privacy for all the subjects in the training data. This fact inhibits the use of these methods in the privatization of medical case-based explanations and highlights the need for having new privacy-preserving approaches.

Montenegro *et al.* [13] also propose that privacy-preserving explanations must preserve: privacy, explanatory evidence, and intelligibility. Regarding privacy, the authors argue that explanations must protect the privacy of all data subjects through a privatization mechanism that is independent of the training data. When it comes to explanatory evidence, the privatized explanations must contain the explanatory features that allow humans to understand their similarity to the case being explained. Finally, the explanations must be intelligible, as they are intended for humans.

In this work, we develop a privacy-preserving generative model that privatizes case-based explanations taking into account the above requirements. We guarantee privacy for the entire training data using a multi-class identity recognition network to promote a uniform identity distribution in the privatized images. The model also ensures the preservation of explanatory evidence by reconstructing relevant explanatory features obtained using interpretability saliency maps. Since the main domain motivating our research is medicine, we also adapt this model to situations where the data lacks images per identity (something pervasive in the medical field). For such scenarios, we use a Siamese identity recognition network [14] to aid privatization. Finally, we extend our model by generating counterfactual explanations based on the privatized factual explanations. In order to have a robust evaluation, we validate our model using the dataset Warsaw-BioBase-Disease-Iris v2.1 [15], [16], which is both medical and biometric and thus has well-defined identities.

The main contributions of this work are:

- The proposal of a privacy-preserving generative model capable of privatizing case-based explanations in a clinical setting, enabling their use in real-world contexts.
- The generation of counterfactual explanations that increase the explanatory value of the deep learning system.

II. RELATED WORK

This section briefly introduces case-based interpretability, followed by a literature review of the current privacy-preserving methods for visual data. Since deep learning privacy-preserving methods use generative networks, we will also introduce some background on relevant deep generative models.

A. CASE-BASED INTERPRETABILITY

Case-based interpretability focuses on retrieving cases from the data, which may or may not be used to perform the predictive task, as explanations for a model's decision. There are various types of explanations that these methods can provide. Methods that establish a similarity metric to compare the data and retrieve the most similar cases produce factual explanations by similar examples [9]–[12]. The most well-known example of such a distance-based method is the traditional K-Nearest Neighbors algorithm. Prototype-based methods, which define prototypes representative of the data, produce explanations by typical examples [17]–[22]. Some methods produce counterfactual examples, whose purpose is to explain the alterations that should occur in the images to change their prediction [23], [24]. These methods usually generate the counterfactual explanations based on the original image rather than retrieving a case from the data. Finally, semi-factual examples have the same class as the original image but are closer to the decision boundary than the most similar case. Semi-factual explanations can be generated based on the original image [23] or retrieved from the data as a sample that is closer to a decision boundary than the case under analysis [25].

B. PRIVACY-PRESERVING METHODS FOR VISUAL DATA

Privacy-preserving methods have been applied in medical imaging with the purpose of increasing the availability of medical data to train artificial intelligence algorithms [26]. Anonymization and pseudonymization techniques remove or alter metadata associated with the medical images (e.g., the patients' names). However, the images themselves expose identity, which can be used to identify the patients through re-identification techniques [27]. Encryption [28] results in unintelligible images that cannot be shown to humans as case-based explanations. Other privacy-preserving techniques avoid disclosing sensitive information about the data during a model's training. For instance, Federated Learning [29] consists of training the models in the data owners' servers to avoid sharing the private medical data [30], [31]. Differential privacy [32] has also been applied to hide contributions of individual patients during a model's training [33]. Nevertheless, these techniques cannot be applied to privatize case-based explanations, which are meant to be exposed to humans, as they act on the model and not on the data itself.

No privacy-preserving method for medical imaging considers altering the image to remove a patient's identifiable features while preserving disease-related information and the image's intelligibility. In this section, we review state-of-the-art privacy-preserving methods capable of generating intelligible privatized images, that have been applied in domains other than the medical field. We discuss the methods in regards to their application to case-based explanations. Furthermore, we consider that identity-related features in the images may be entangled with explanatory features that must be preserved. We distinguish these methods in traditional and Deep Learning methods.

Traditional privacy-preserving methods are applied over the whole input, as they cannot identify sensitive image regions. These methods require an additional pre-processing step to locate the image regions that need to be privatized. The most well-known traditional method consists of applying filters such as blur to an image [34]. The most significant issue in this type of method is that relevant explanatory features are lost at the same rate as identity features. As such, privatized images with acceptable degrees of privacy do not preserve explanatory evidence [13]. Another famous class of privacy-preserving techniques is the K-Same-based family [35], [36], which were developed for face de-identification. In these methods, the privatized images are an average of various training images, guaranteeing K-Anonymity, where the highest probability of a person being recognized in the image is $\frac{1}{K}$. This technique imposes limitations in privacy, as the privatization process directly uses images from other subjects in the database, and in explanatory evidence preservation. An alternative to those methods is face-swapping [37], which consists of replacing the faces in an image with models from a public database. Although this method guarantees privacy, if identity-related features and explanatory features are entangled, the replacement of the image regions that contain identity-related features will result in the loss of the associated explanatory features.

In Deep Learning, privacy-preserving models usually comprise a generative network responsible for generating privatized images and an identity recognition network that guides the privatization process. Some models directly obtain identity vectors from the images by disentangling identity-related features from the remaining features, as is the case with the CLEANIR model [38] and the R^2VAE model [39]. These identity vectors can then be altered to hide the original identity of the images. Other privacy-preserving strategies focus on creating privatized images that do not share the same identity as the original images by using a Siamese identity recognition network [14] to guide the generation of privacy-preserving images [40], [41]. These networks ensure image utility by maximizing the structural similarity between the original and privatized images.

The biggest problem in the previous deep learning methods is that none guarantees the preservation of relevant semantic features needed for a particular classification task. Privacy-preserving methods that preserve task-related features use a task-related classifier to ensure the feature preservation process. PPRL-VGAN [42] was developed for privacy-preserving facial expression recognition. It privatizes images through identity replacement. Although this model successfully hides the identity from the original image, it exposes identities from other subjects in the data. As such, this model still violates privacy. Furthermore, this model only preserves the task-related class of the original image and not its explanatory features.

In general, none of the privacy-preserving models explores the explicit preservation of the original images' explanatory evidence. Furthermore, some of the models still possess privacy issues as they directly use training data in the privatization process.

C. GENERATIVE MODELS

Generative models model the probability distribution of the data and allow the generation of new data by sampling from the learned distribution. The most relevant generative models for this work are Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).

GANs [43] capture the data distribution implicitly, through a minimax game between a generator and a discriminator. The generator is responsible for generating new samples, while the discriminator is responsible for the binary classification task of distinguishing between real and fake images. The goal of the generator is to trick the discriminator into classifying the generated images as real.

VAEs [44] explicitly learn an approximation of the data distribution. These models maximize data likelihood using an encoder-decoder architecture. The encoder maps an image in the original data space into its representation in a latent space with a simpler distribution. The decoder learns to map points from the latent space into the original data space by reconstructing the original image from its representation. Then, to generate new images, we can sample points from the simple distribution used in the latent space and use the decoder to visualize them in the original data space. VAEs are trained using a regularization loss based on the Kullback-Leibler (KL) divergence, to approximate the distribution of the latent representation to the original one, and a reconstruction loss.

These generative models have been applied in a multitude of domains, including in the medical field, with applications in medical image synthesis [45]–[48], segmentation [49]–[51], and detection [52], [53]. GANs have also been used in medical image classification tasks [54], [55].

III. PROPOSED METHODOLOGY

Given an image that serves as a Deep Learning model's explanation, we aim to produce an intelligible image that does not expose the original image's identity and, at the same time, preserves the image's explanatory evidence.

We build our approach on top of one existing privacy-preserving model: the PPRL-VGAN [42]. This work is motivated by [13], where the authors analyze the PPRL-VGAN model and its applicability to the privatization of case-based explanations, highlighting its potential but also identifying its weaknesses. In this work, we introduce new modules and loss functions to improve the PPRL-VGAN in terms of privacy, intelligibility, and preservation of explanatory evidence, in the context of medical imaging.

A. PPRL-VGAN MODEL

The PPRL-VGAN model proposed by Chen *et al.* [42] comprises a GAN with a conditional VAE as the generator and a multi-task discriminator, as shown in Figure 1. The generator is conditioned into generating an image with the target identity replacement c. The multi-task discriminator contains a real/fake classifier to promote realistic synthetic images, a multi-class identity recognition network to guide the privatization, and a task-related classification network to ensure the preservation of the original image's class.

There are various weaknesses in this model that prevent its use for the privacy-preservation of case-based explanations.

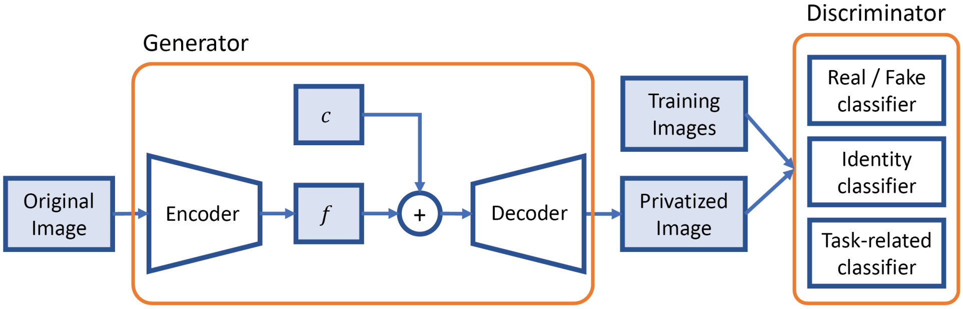


FIGURE 1. **Overview of the PPRL-VGAN model's architecture. *c* represents the replacement identity, and *f* is the representation of the original image in the latent space of the generator.**

The most critical ones are the privacy violation inherent to using identity replacement as the privatization mechanism and the non-preservation of explanatory evidence as it exists in the original image [13]. Regarding applying this model to medical data, the model also has difficulty in disentangling identity-related factors in cases where most subjects only have images from one disease-related class [13]. Moreover, a multi-class identity recognition network is challenging to train when the data only has a small number of images per identity, as frequently happens in the medical context.

B. PRIVACY-PRESERVING NETWORK WITH MULTI-CLASS IDENTITY RECOGNITION

Using the PPRL-VGAN model as a base, we defined a novel privacy-preserving network for the privatization of case-based explanations.

To ensure that privacy is preserved for every subject in the training data, we removed the replacement identity given to the decoder. Instead of creating an image that looks like the replacement identity, we try to keep the identity recognition close to random guessing (i.e., close to a uniform distribution).

By promoting a uniform distribution across identities, the generative task became more complex, leading to poor image quality and mode collapse problems. We pre-trained the identity recognition model and the task-related classifier on the dataset used to train the privacy-preserving model, to facilitate the generative task and improve image quality. In PPRL-VGAN, the mode intentionally collapsed to the identity given as replacement and to the task-related class from the original image. However, in our case, the mode collapse was unintentional and affected the explanatory value of the images, as they all looked identical. To fix this problem and improve image quality, we replaced the generative framework with a WGAN-GP network [56], using Wasserstein loss with gradient penalty to stabilize the discriminator.

We explicitly preserve explanatory evidence by using interpretability saliency maps to reconstruct relevant task-related features in the privatized images. In specific, we use Deep Taylor [57] to create masks containing the relevant image features. We input these masks into the generative network and concatenate them with the original images inside the VAE's encoder, after feature extraction and before calculating the parameters of a Gaussian distribution. In the loss function, we use the squared L2 loss to reconstruct relevant features. We also ensure that the privatized images are assigned the same classification score as the original images to aid the preservation of explanatory features.

We summarized the changes introduced to the PPRL-VGAN model in Figure 2. With these changes, we obtained a privacy-preserving model with three modules: a generative module, a privacy module, and an explanatory module.

1) GENERATIVE MODULE

The generative module is responsible for the generation of intelligible images, given an image I from the original data space's probability distribution p_d. It is composed of a GAN with a VAE as the generator G. The discriminator, D, is trained using Wasserstein loss and gradient penalty, as shown in Equation 1, where $\hat{x}$ corresponds to random samples and λ is the weight associated with the gradient penalty term. In the generator, there are two terms: a realness term to promote the generation of realistic images (Equation 2), and a regularization term in the VAE. The regularization term, shown in Equation 3, consists of approximating the prior distribution on the latent space $p(f(I))$, where $f(I)$ corresponds to the image I's latent representation, and the conditional distribution $q(f(I) \mid I)$ parameterized by the encoder.

$$\begin{aligned}\mathcal{L}_D &= E_{I\sim p_d(I)}[D(G(I))] - E_{I\sim p_d(I)}[D(I)] \\ &\quad + E_{\hat{x}\sim p_{\hat{x}}}[\lambda(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2] \qquad (1)\\ \mathcal{L}_{realness} &= E_{I\sim p_d(I)}[-D(G(I))] \qquad (2)\\ \mathcal{L}_{reg} &= E_{I\sim p_d(I)}[KL(q(f(I) \mid I)||p(f(I)))] \qquad (3)\end{aligned}$$

2) PRIVACY MODULE

The privacy module is responsible for anonymizing the images, guaranteeing privacy for the subjects in the image and in the database. Using a pre-trained multi-class identity recognition network D_{id}, we promote a uniform identity distribution in the privatized images. As such, the generator

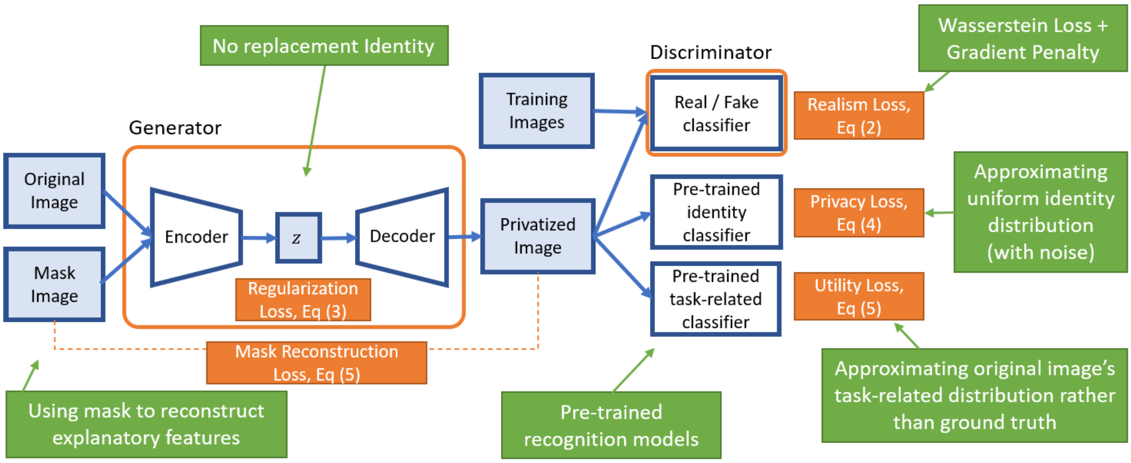

No replacement Identity
Generator
Original Image
Mask Image
Encoder
z
Decoder
Regularization Loss, Eq (3)
Mask Reconstruction Loss, Eq (5)
Using mask to reconstruct explanatory features
Training Images
Privatized Image
Discriminator
Real / Fake classifier
Pre-trained identity classifier
Pre-trained task-related classifier
Pre-trained recognition models
Realism Loss, Eq (2)
Privacy Loss, Eq (4)
Utility Loss, Eq (5)
Wasserstein Loss + Gradient Penalty
Approximating uniform identity distribution (with noise)
Approximating original image's task-related distribution rather than ground truth

FIGURE 2. **Overview of our privacy-preserving model's architecture. We included in green a summary of the changes that occurred to the original PPRL-VGAN model, to apply it to the domain of case-based interpretability.**

contains a privacy term in the loss function, represented in Equation 4. In this equation, U represents a uniform distribution with noise.

$$\mathcal{L}_{privacy} = E_{I\sim p_d(I)}[-D_{id}(G(I))log(U)] \quad (4)$$

3) EXPLANATORY MODULE

The explanatory module is responsible for guaranteeing the privatized images' explanatory value. We preserve the explanatory evidence through the reconstruction of explanatory features in the images, using Deep Taylor saliency maps, M, obtained by applying the task-related classifier D_{exp} on the original images. We also approximate the privatized images classification score to the one in the original images. The generator loss terms representative of this module are shown in Equation 5.

$$\mathcal{L}_{exp} = E_{(I,M)\sim p_d(I,M)}[\lambda_3 D_{exp}(I)log(D_{exp}(G(I))) + \lambda_4(I \times M - G(I) \times M)^2] \quad (5)$$

Finally, the entire generator's loss is depicted in Equation 6. λ_x are parameters to control the importance of each loss term x.

$$\mathcal{L}_G = \lambda_1\mathcal{L}_{realness} + \lambda_2\mathcal{L}_{privacy} + \mathcal{L}_{exp} + \lambda_5\mathcal{L}_{reg} \quad (6)$$

C. PRIVACY-PRESERVING NETWORK WITH SIAMESE IDENTITY RECOGNITION

As it stands, our privacy-preserving model cannot be used in domains where the number of images per subject is scarce, which is frequently the case in the medical context, since a multi-class identity recognition network is hard to train in these scenarios. To widen the range of application of our model, we replace the multi-class identity recognition network with a Siamese network [14], pre-trained on the dataset used to train the privacy-preserving model.

The Siamese identity recognition network compares the original image with its privatized version and computes their identity-related distance, which can be used to classify whether the images belong to the same identity or not. We trained this network using a contrastive loss [58], represented in Equation 7. In this equation, m represents a margin to limit the distance between images, Y represents the label assigned to an image pair (1 when the images belong to the same identity, and 0 otherwise), and ED represents the Euclidean Distance between the image pair embeddings.

$$ContrastiveLoss = \frac{1}{2} \times Y \times ED^2 + \frac{1}{2} \times (1 - Y) \times [\max(0, m - ED)]^2 \quad (7)$$

By using this network, we ensure that the privatized image is different from the original image in terms of identity. To guarantee that the generated images also do not look like the images of the other identities present in the dataset, we use the Siamese network to increase the identity-related distance between the privatized image and the images from each of the subjects present in the database. In practice, at each epoch during training, we randomly select one image from each of the identities, and promote that this image is far from the privatized image.

The privacy term of the generator loss function, when using the Siamese network, is represented in Equation 8, where N is the number of identities that exist in the dataset.

$$\mathcal{L}_{privacy} = E_{(I,N)\sim p_d(I,N)}[\lambda_2[\max(0, m - ED(I, G(I)))]^2 + \lambda_6 \sum_{i=0}^{N} \frac{[\max(0, m - ED(G(I), I_N))]^2]}{N}] \quad (8)$$

D. GENERATION OF COUNTERFACTUAL EXPLANATIONS

We also apply our model to the generation of counterfactual explanations. We add a counterfactual generation module to the previously defined privacy-preserving network in the form of a counterfactual decoder responsible for mapping an image's latent representation to its counterfactual.

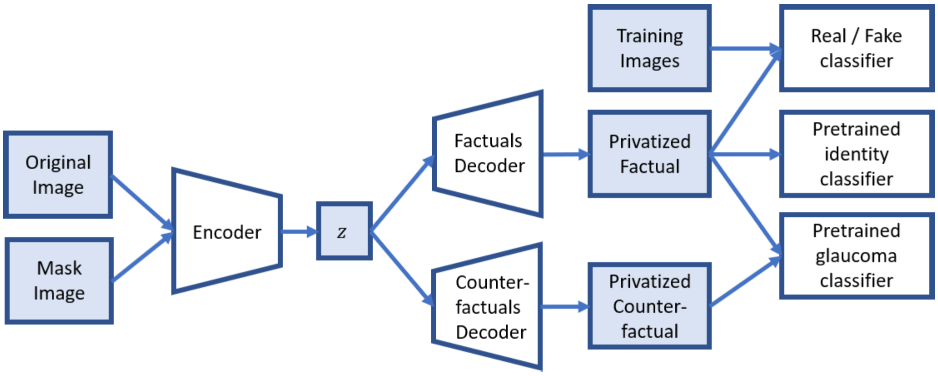


FIGURE 3. Architecture of the privacy-preserving model with generation of counterfactual explanations.

To generate counterfactual explanations, we aim to perform the smallest number of alterations to the privatized factual explanations to change their predicted class. As such, the counterfactuals' decoder is trained to minimize the pixel-wise distance between the factual and counterfactual explanations while changing the original image's task-related prediction. We use the saliency masks with the explanatory features to promote changes in the image regions relevant to the explanatory classification task while preserving the remaining image parts. This network's architecture is shown in Figure 3.

Regarding the training approach, we first train the factual decoder as in the previously presented networks, with the counterfactual decoder frozen. Then, we freeze the factual decoder and transfer its weights to the counterfactual decoder to train it. The generator's loss function used to train the counterfactual decoder is represented in Equation 9. In this equation, $F(I)$ and $C(I)$ denote the privatized factual and counterfactual explanations, respectively.

$$\mathcal{L}_C = E_{(I,M)\sim p_d(I,M)}[\lambda_7(F(I)\times(1-M)-C(I)\times(1-M))^2 + \lambda_8 D_{exp}(I) log(1-D_{exp}(C(I)))] \tag{9}$$

IV. EXPERIMENTS

For the experiments, we used the medical and biometric dataset Warsaw-BioBase-Disease-Iris v2.1 [15], [16], composed of 2,996 iris images with various eye pathologies acquired from 115 different patients. We only used the 1,795 images taken from the device IrisGuard AD100, and we focused on one of the pathologies, glaucoma. The images were labeled according to the presence or absence of glaucoma. In the pre-processing stage, we cropped the images to remove labels in their lower corners, horizontally flipped the patients' right eye images, and centered the iris of the eye in the middle of the image. The images' resolution was set to 64 × 64 and they were split into 65% for training, 15% for validation, and 20% for testing. To obtain masks with relevant glaucoma features located inside the iris, we generated iris segmentation masks and performed an AND operation between the Deep Taylor saliency maps and the iris segmentation masks.

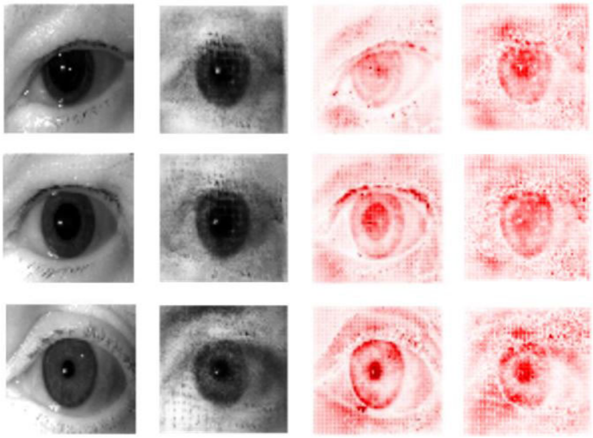


FIGURE 4. Results of privacy-preserving model with multi-class identity recognition. The first and second columns represent the original images and their privatized versions, respectively. The third and fourth columns contain Deep Taylor saliency maps obtained from the original and privatized images, respectively.

A. PRIVACY-PRESERVING MODEL WITH MULTI-CLASS IDENTITY RECOGNITION

In the privacy-preserving model with multi-class identity recognition, we used as parameters $\lambda_1 = 0.4$, $\lambda_2 = 1$, $\lambda_3 = 2$, $\lambda_4 = 0.001$ and $\lambda_5 = 0.002$. We used $\lambda = 10$ in the discriminator's loss, as suggested in the original WGAN-GP paper [56]. We used the Adam optimizer with a learning rate of $2e^{-5}$. The model was trained for 1,184 epochs. The results are presented in Figure 4. Although the images possess some visible noise, they can be considered intelligible. We notice that the network has some difficulty creating a realistic eye structure surrounding the iris. In the visual results, the privatized image's Deep Taylor saliency maps closely resemble the ones from the original images, evidencing the correct preservation of explanatory evidence.

We include in Table 1 the results achieved with this network. The identity recognition network's accuracy is evaluated at recognizing the subject from the original image. To evaluate privacy at the whole dataset's level, we analyze the maximum score that the identity recognition model assigns to an identity when making a prediction about a

TABLE 1. Results of privacy-preserving model with multi-class identity recognition. We expect low values in the privacy-related metrics and high values in glaucoma recognition accuracy. The best results for each metric are highlighted in bold.

Dataset	Identity Recognition Accuracy	Maximum Identity Score	Average KL Divergence	Glaucoma Recognition Accuracy
Original testing set (baseline)	89.71%	88.22%	4.24	100.00%
Privatized set with explanatory evidence	**0.88%**	**33.15%**	**2.53**	**91.47%**
Privatized set without explanatory evidence	**0.88%**	34.49%	2.60	89.41%

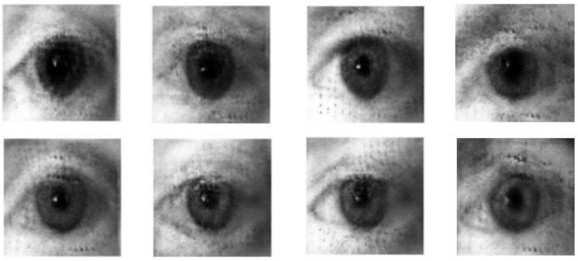


FIGURE 5. Comparison between results from the network when explanatory evidence is considered (first row) and not (second row).

privatized image. We also evaluate the divergence between the privatized images' identity distribution and the uniform distribution, using KL Divergence. Finally, we assess the Glaucoma Recognition network's accuracy at detecting the original images' glaucoma score in the privatized images.

The low accuracy in identity recognition suggests that the privacy-preserving model succeeds at privatizing the images. The values for the maximum identity score and KL divergence suggest that the network has difficulty recognizing any identity, as these values are significantly lower than the baseline. Furthermore, the high values in glaucoma recognition accuracy advocate for the network's high capacity of preserving explanatory evidence.

During the network's development, the most significant challenge we came across was to manage the trade-off between privacy, intelligibility and explanatory evidence. In most cases, improving one of these dimensions would result in worsening the remaining ones. In our model, the most sacrificed dimension was intelligibility, as the generated images have poorer quality than the original ones. When we try to remove one of the other dimensions, the image quality improves. For instance, removing explanatory evidence results in the higher-quality results shown in Figure 5.

B. PRIVACY-PRESERVING MODEL WITH SIAMESE IDENTITY RECOGNITION

Using the privacy-preserving model with Siamese identity recognition, with parameters $\lambda_1 = 0.4$, $\lambda_2 = 5$, $\lambda_3 = 2$, $\lambda_4 = 0.001$, $\lambda_5 = 0.002$ and $\lambda_6 = 10$, we obtained the results shown in Figure 6. The model was trained for 900 epochs.

This model provides higher-quality images than the previous multi-class identity recognition model. Nonetheless, the model also suffers from a trade-off between privacy, intelligibility, and explanatory evidence. For instance, when we remove the overall privacy term ($\lambda_6 = 0$), we obtain privatized explanatory features that resemble more closely the ones from the original images, as shown in Figure 7.

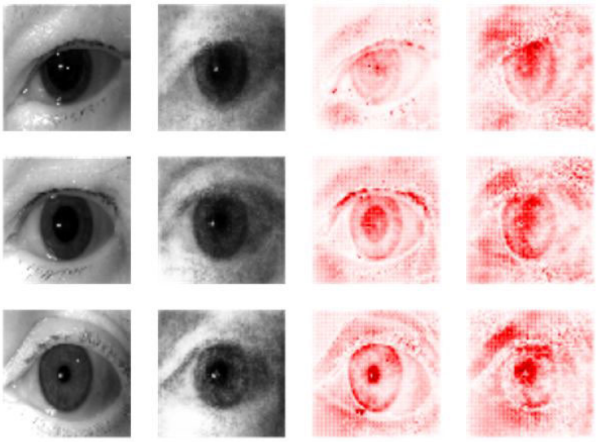


FIGURE 6. Results of privacy-preserving model with Siamese identity recognition. The first and second columns represent the original images and their privatized versions, respectively. The third and fourth columns contain Deep Taylor saliency maps obtained from the original and privatized images, respectively.

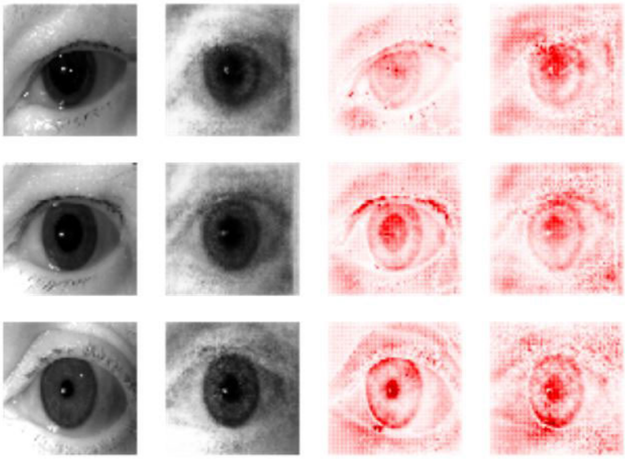


FIGURE 7. Results of privacy-preserving model with Siamese recognition, not considering overall privacy. The first and second columns represent the original images and their privatized versions, respectively. The third and fourth columns contain Deep Taylor saliency maps obtained from the original and privatized images, respectively.

Table 2 exposes the results obtained with this model. To evaluate privacy, we use the previously developed multi-class identity recognition model as an evaluation network. Then, we use the Siamese identity recognition model's accuracy at recognizing that the original and privatized images belong to different identities. To calculate this accuracy,

TABLE 2. Results of the privacy-preserving model with Siamese identity recognition. The best results for each metric are highlighted in bold.

Dataset	Multi-class Identity Recognition Accuracy	Siamese Identity Recognition Accuracy	Siamese Recognition Accuracy (Whole Dataset)	Average Number of Real Pairs	Glaucoma Recognition Accuracy
Original testing set (baseline)	89.71%	83.80%	-	-	100.00%
Privatized set with no overall privacy ($\lambda_6 = 0$)	**0.88%**	89.41%	78.91%	22.99	88.53%
Privatized set with overall privacy ($\lambda_6 = 10$)	1.76%	**92.65%**	**91.99%**	**8.74**	**91.47%**

we verify whether the distance between image pairs is higher than 0.777, corresponding to the average distance value obtained when using the Siamese network on image pairs from the original testing set. To evaluate the privacy in the whole dataset, we obtain the identity recognition accuracy when comparing the privatized images with an image from each identity available in the dataset. We also evaluate the number of pairs that are considered to be from the same identity (real pairs). In this table, we expect to achieve low values in multi-class identity recognition and average number of real pairs, and high values in the remaining metrics.

We obtained a higher privacy degree by considering overall privacy, as seen by the lower accuracy in multi-class recognition and higher accuracy in Siamese identity recognition. Furthermore, when we consider overall privacy, there are fewer images from the dataset's subjects that are considered to be from the same identity as the privatized images. The privatized set with overall privacy also achieved higher glaucoma recognition accuracy.

C. COUNTERFACTUAL GENERATION

By adding a counterfactual generation module to the privacy-preserving model with multi-class identity recognition, with parameters $\lambda_7 = 0.001$ and $\lambda_8 = 1$, we were capable of inverting the glaucoma classification of the original image with 90.29% accuracy. With the model that uses Siamese identity recognition, we achieved 90.88% accuracy in inverting the images' glaucoma classification. Furthermore, in both models, the differences between the factual and the counterfactual explanations are located mainly in the iris region. An example of the obtained results using the Siamese identity recognition model is shown in Figure 8.

In this experiment, we used the Deep Taylor glaucoma masks to promote changes located inside the iris and, thus, avoid alterations in zones that are irrelevant to glaucoma classification that may occur as an adversarial attack. However, even with these masks, the counterfactual decoder may be performing an adversarial attack on the glaucoma classification network, tricking it into misclassifying the samples and generating adversarial samples instead of counterfactual explanations.

D. ABLATION STUDY

To verify how the generator used in the privacy-preserving models fares in comparison with other state-of-the-art architectures, we replaced it with a ResNet VAE, which contains ResNet [59] as the encoder and decoder, and with a UNET architecture [60]. We performed this experiment with the multi-class identity recognition version of the privacy-preserving model. The results are shown in Table 3 and in Figure 9.

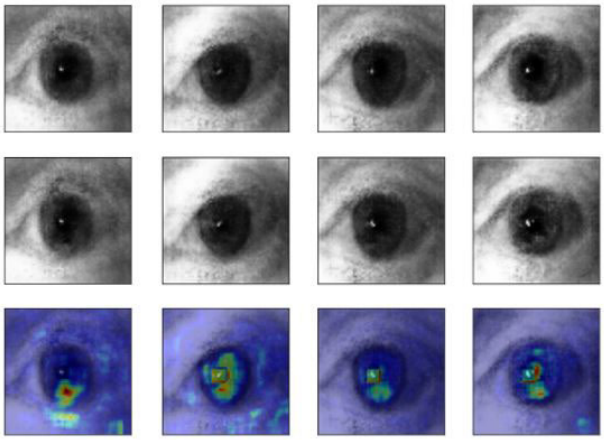


FIGURE 8. Results of counterfactual generation using privacy-preserving model with Siamese identity recognition. The first and second rows represent privatized factual and counterfactual explanations, respectively. The final row contains a map with the differences between the factual and counterfactual explanations.

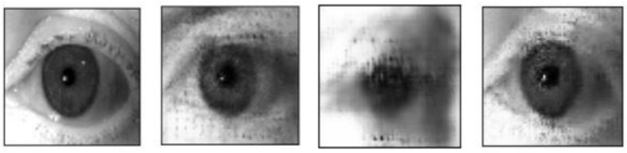


FIGURE 9. Results obtained by replacing the generator with other architectures. The first image is the original one, and the following images are privatized images using the original VAE, the ResNet VAE, and UNET, respectively.

Although the results obtained with the ResNet VAE present higher privacy, the images lack intelligibility and explanatory value, hindering their use as explanations. The UNET has a higher capacity to preserve features, as verified by the higher accuracy in identity recognition and glaucoma recognition. Furthermore, since the image generated by the UNET is extremely similar to the original one, this network might be performing an adversarial attack on the identity recognition network instead of an adequate anonymization.

Given these results, we can conclude that the original generator with a standard convolutional VAE is the one that provides better and more balanced results, guaranteeing both privacy and the explanatory value of the images.

TABLE 3. Results of replacing generator with ReNet VAE and UNET. For convenience, the first lines repeat the results shown in Table 1.

Dataset	Identity Recognition Accuracy	Maximum Identity Score	Average KL Divergence	Glaucoma Recognition Accuracy
Original testing set (baseline)	89.71%	88.22%	4.24	100.00%
Privatized set with original VAE	0.88%	33.15%	2.53	91.47%
Privatized set with ResNet VAE	**0.59%**	**9.0%**	**0.74**	84.41%
Privatized set with UNET	4.41%	26.86%	2.03	**91.76%**

TABLE 4. Comparison between the results of our privacy-preserving models and results of state-of-the-art models obtained from [13] in regards to privacy and preservation of explanatory evidence.

Method	Identity Recognition Accuracy	Glaucoma Recognition Accuracy
Blurring	19.41%	75.59%
K-Same-Select	**0.88%**	77.06%
PPRL-VGAN	1.76%	86.56%
Ours (Multi-class)	**0.88%**	**91.47%**
Ours (Siamese)	1.76%	**91.47%**

E. STATE-OF-THE-ART COMPARISON

In this section, we compare our privacy-preserving models with the state-of-the-art methods blurring, K-Same-Select [35] and PPRL-VGAN [42]. These methods had previously been applied to the Warsaw-BioBase-Disease-Iris v2.1 dataset in [13]. The results in terms of identity recognition and glaucoma recognition are summarized in Table 4.

Our privacy-preserving models have a higher capacity to preserve explanatory features than the methods from the literature while obtaining comparable results in identity recognition. Furthermore, our models promote privacy for every patient in the dataset, unlike K-Same-Select and PPRL-VGAN, which directly use identities from the dataset in the privatization process (through image averaging or identity replacement). As such, our privacy-preserving models are the most appropriate to be applied to the domain of medical case-based explanations.

V. DISCUSSION AND CONCLUSION

In this paper, we developed a privacy-preserving model to privatize case-based explanations. The model tackles the most significant weaknesses of current privacy-preserving models, guaranteeing privacy, intelligibility, and preservation of explanatory evidence. At first, we used a multi-class identity recognition model to guide image privatization. Then, we widened the range of application of our model by using a Siamese identity recognition network to guide the privatization, enabling the model to be used when medical data only has a small number of images per subject.

Our approach regarding the preservation of explanatory evidence consisted of using interpretability saliency maps to reconstruct relevant features. However, *post hoc* techniques are often criticized for not reflecting a model's real reasoning [22]. As such, using these methods to preserve explanatory features when privatizing explanations obtained through intrinsic interpretability methods clashes with the intrinsic methods' goal of providing accurate representations of a model's reasoning. In such cases, if the intrinsic interpretability method defines a similarity measure to semantically compare two images, it should be possible to use this measure to approximate the privatized image to the original image in regard to explanatory features.

We have also applied the model to generate counterfactual explanations based on the privatized factual explanations. The counterfactual explanations highlight the changes in an image that would lead to a reversal of the class prediction. We used interpretability saliency maps to promote changes in image regions related to the classification task. Nonetheless, the resulting explanations may be adversarial examples whose alterations are not related to the concepts associated with the classification task. Even though we only considered a binary classification task in our work, the approach is generalizable to the multi-class scenario. To apply the counterfactual generation model to multi-class classification problems, the counterfactual decoder could be trained to receive the latent representation of an image and the target class of the counterfactual, allowing to retrieve counterfactual explanations representative of each class.

Future work should consider integrating privacy in the image retrieval process to optimize the selection of explanatory cases and using causality to ensure that features preserved in the privacy-preserving explanations are causally related to the explanatory task.

In conclusion, this work contributes to enabling the use of case-based explanations in contexts where the data violates the privacy of individuals, like in medical imaging.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[2] S. M. McKinney, M. Sieniek, and V. Godbole, "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, Jan. 2020.

[3] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Mach. Intell.*, vol. 3, no. 7, pp. 610–619, May 2021.

[4] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek, "Analyzing classifiers: Fisher vectors and deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2912–2920.

[5] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, Jul. 2019.

[6] E. M. Kenny, C. Ford, M. Quinn, and M. T. Keane, "Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies," *Artif. Intell.*, vol. 294, May 2021, Art. no. 103459.

[7] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.

[8] D. B. Leake, "CBR in context: The present and future," in *Proc. Case-Based Reasoning, Exper., Lessons, Future Directions*, 1996, pp. 3–30.

[9] M. T. Keane and E. M. Kenny, "How case-based reasoning explains neural networks: A theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems," in *Case-Based Reasoning Research and Development*. Cham, Switzerland: Springer, 2019, pp. 155–171.

[10] E. M. Kenny and M. T. Keane, "Twin-systems to explain artificial neural networks using case-based reasoning: Comparative tests of feature-weighting methods in ANN-CBR twins for XAI," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 2708–2715.

[11] N. Papernot and P. D. McDaniel, "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning," *CoRR*, vol. abs/1803.04765, pp. 1–18, Mar. 2018.

[12] W. Silva, A. Poellinger, J. S. Cardoso, and M. Reyes, "Interpretability-guided content-based medical image retrieval," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Lima, Peru: Springer, 2020, pp. 305–314.

[13] H. Montenegro, W. Silva, and J. S. Cardoso, "Towards privacy-preserving explanations in medical image analysis," in *Proc. 1st Workshop Interpretable Mach. Learn. Healthcare (IMLH)*, 2021, pp. 1–7.

[14] J. Bromley, J. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Sackinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, p. 25, Aug. 1993.

[15] M. Trokielewicz, A. Czajka, and P. Maciejewicz, "Assessment of iris recognition reliability for eyes affected by ocular pathologies," in *Proc. IEEE 7th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2015, pp. 1–6.

[16] M. Trokielewicz, A. Czajka, and P. Maciejewicz, "Implications of ocular pathologies for iris recognition reliability," *Image Vis. Comput.*, vol. 58, pp. 158–167, Feb. 2017.

[17] P. Angelov and E. Soares, "Towards deep machine reasoning: A prototype-based deep neural network with decision tree inference," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 2092–2099.

[18] P. Angelov and E. Soares, "Towards explainable deep neural networks (xDNN)," *Neural Netw.*, vol. 130, pp. 185–194, Oct. 2020.

[19] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32. Red Hook, NJ, USA: Curran Associates, 2019, pp. 8930–8941.

[20] X. Gu and W. Ding, "A hierarchical prototype-based approach for classification," *Inf. Sci.*, vol. 505, pp. 325–351, Dec. 2019.

[21] B. Kim, C. Rudin, and J. Shah, "The Bayesian case model: A generative approach for case-based reasoning and prototype classification," in *Proc. NIPS*. Cambridge, MA, USA: MIT Press, 2014, pp. 1952–1960.

[22] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in *Proc. AAAI*, 2018, pp. 3530–3537.

[23] E. M. Kenny and M. T. Keane, "On generating plausible counterfactual and semi-factual explanations for deep learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 13, pp. 11575–11585.

[24] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, "Generative counterfactual introspection for explainable deep learning," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2019, pp. 1–5.

[25] C. Nugent, D. Doyle, and P. Cunningham, "Gaining insight through case-based explanation," *J. Intell. Inf. Syst.*, vol. 32, no. 3, pp. 267–295, Jun. 2009.

[26] G. A. Kaissis, M. R. Makowski, D. Ráckert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Mach. Intell.*, vol. 2, no. 6, pp. 305–311, Jun. 2020.

[27] C. G. Schwarz, W. K. Kremers, T. M. Therneau, R. R. Sharp, J. L. Gunter, P. Vemuri, A. Arani, A. J. Spychalla, K. Kantarci, D. S. Knopman, R. C. Petersen, and C. R. Jack, "Identification of anonymous MRI research participants with face-recognition software," *New England J. Med.*, vol. 381, no. 17, pp. 1684–1686, Oct. 2019.

[28] L.-B. Zhang, Z.-L. Zhu, B.-Q. Yang, W.-Y. Liu, H.-F. Zhu, and M.-Y. Zou, "Medical image encryption and compression scheme using compressive sensing and pixel swapping based permutation approach," *Math. Problems Eng.*, vol. 2015, Aug. 2015, Art. no. 940638.

[29] J. Konecny, B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," in *Proc. 8th NIPS Workshop Optim. Mach. Learn.*, 2015, pp. 1–5.

[30] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen, and S. Bakas, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, Art. no. 12598.

[31] H. R. Roth, K. Chang, P. Singh, N. Neumark, W. Li, V. Gupta, S. Gupta, L. Qu, A. Ihsani, B. C. Bizzo, and, "Federated learning for breast density classification: A real-world implementation," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning* (Lecture Notes in Computer Science). Lima, Peru: Springer, 2020, pp. 181–191.

[32] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.

[33] A. Ziller, D. Usynin, R. Braren, M. Makowski, D. Rueckert, and G. Kaissis, "Medical imaging deep learning with differential privacy," *Sci. Rep.*, vol. 11, no. 1, Dec. 2021, Art. no. 13524.

[34] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent, "Large-scale privacy protection in Google street view," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2373–2380.

[35] R. Gross, E. Airoldi, B. Malin, and L. Sweeney, "Integrating utility into face de-identification," in *Privacy Enhancing Technology*. Berlin, Germany: Springer, 2016, pp. 227–242.

[36] 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.

[37] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: Automatically replacing faces in photographs," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–8, 2008.

[38] D. Cho, J. H. Lee, and I. H. Suh, "CLEANIR: Controllable attribute-preserving natural identity remover," *Appl. Sci.*, vol. 10, no. 3, p. 1120, 2020.

[39] M. Gong, J. Liu, H. Li, Y. Xie, and Z. Tang, "Disentangled representation learning for multiple attributes preserving face deidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 19, 2020, doi: 10.1109/TNNLS.2020.3027617.

[40] W. Oleszkiewicz, T. Wlodarczyk, K. Piczak, T. Trzcinski, P. Kairouz, and R. Rajagopal, "Siamese generative adversarial privatizer for biometric data," in *Proc. ACCV*, Apr. 2018, pp. 482–497.

[41] Y. Wu, F. Yang, Y. Xu, and H. Ling, "Privacy-protective-GAN for privacy preserving face de-identification," *J. Comput. Sci. Technol.*, vol. 34, no. 1, pp. 47–60, 2019.

[42] J. Chen, J. Konrad, and P. Ishwar, "VGAN-based image representation learning for privacy-preserving facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1570–1579.

[43] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27. Red Hook, NJ, USA: Curran Associates, 2014, pp. 2672–2680.

[44] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, 2013, pp. 1–14.

[45] V. V. Laptev, O. M. Gerget, and N. A. Markova, *Generative Models Based on VAE and GAN for New Medical Data Synthesis*. Cham, Switzerland: Springer, 2021, pp. 217–226.

[46] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018.

[47] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, and H. Nakayama, "GAN-based synthetic brain MR image generation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 734–738.

[48] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, "Generalization of deep neural networks for chest pathology classification in X-rays using generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2018, pp. 990–994.

[49] M. Rezaei, H. Yang, and C. Meinel, "Whole heart and great vessel segmentation with context-aware of generative adversarial networks," in *Bildverarbeitung Medizin*, A. Maier, T. M. Deserno, H. Handels, K. H. Maier-Hein, C. Palm, T. Tolxdorff, Eds. Berlin, Germany: Springer, 2018, pp. 353–358.

[50] J. Son, S. J. Park, and K. Jung, "Retinal vessel segmentation in fundoscopic images with generative adversarial networks," 2017, *arXiv:1706.09318*.

[51] E. M. Yu, J. E. Iglesias, A. V. Dalca, and M. R. Sabuncu, "An auto-encoder strategy for adaptive image segmentation," in *Proc. 3rd Conf. Med. Imag. Deep Learn.*, 2020, pp. 1–11.

[52] V. Alex, M. Safwan, S. S. Chennamsetty, and G. Krishnamurthi, "Generative adversarial networks for brain lesion detection," *Proc. SPIE, Med. Imag., Image Process.*, vol. 10133, Feb. 2017, Art. no. 101330G.

[53] X. Chen and E. Konukoglu, "Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders," in *Proc. MIDL Conf. Book*, 2018, pp. 1–9.

[54] X. Yi, E. Walia, and P. Babyn, "Unsupervised and semi-supervised learning with categorical generative adversarial networks assisted by Wasserstein distance for dermoscopy image classification," 2018, *arXiv:1804.03700*.

[55] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1038–1042.

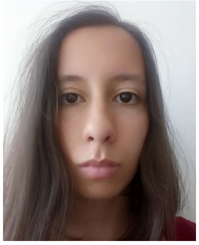
[56] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5769–5779.

[57] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017.

[58] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1735–1742.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[60] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.



HELENA MONTENEGRO (Member, IEEE) was born in Vila Nova de Famalicão, Portugal, in 1998. She received the M.Sc. degree in informatics and computing engineering from the Faculty of Engineering, University of Porto, in 2021. She is currently a Research Assistant with INESC TEC. Her main research interests include machine learning and computer vision, with a special focus on privacy-preserving methods for visual data and interpretability.

WILSON SILVA (Student Member, IEEE) received the integrated B.Sc. and M.Sc. degree in electrical and computer engineering from the Faculty of Engineering, University of Porto (FEUP), in 2016, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. He was a Research Assistant at INESC TEC, where he is associated with the Visual Computing and Machine Intelligence (VCMI) Group and the Breast Research Group. His main research interests include machine learning and computer vision, with a particular focus on explainable artificial intelligence and medical image analysis.

JAIME S. CARDOSO (Senior Member, IEEE) is currently a Full Professor with the Faculty of Engineering, University of Porto (FEUP), and a Coordinator of the Centre for Telecommunications and Multimedia, INESC TEC. From 2012 to 2015, he served as the President of the Portuguese Association for Pattern Recognition (APRP), affiliated in the IAPR. He has coauthored more than 300 articles, more than 90 of which in international journals, which attracted more than 5800 citations, according to Google Scholar. His research can be summed up in three major topics: computer vision, machine learning, and decision support systems. Image and video processing focuses on medicine and biometrics. The work on machine learning cares mostly with the adaptation of learning to the challenging conditions presented by visual data, with a focus on deep learning and explainable machine learning. The particular emphasis of the work in decision support systems goes to medical applications, always anchored on the automatic analysis of visual data.