FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

A privacy-preserving framework for case-based interpretability in machine learning

Maria Helena Sampaio de Mendonça Montenegro e Almeida

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Jaime S. Cardoso

Second Supervisor: Wilson Silva

July 24, 2021

A privacy-preserving framework for case-based interpretability in machine learning

Maria Helena Sampaio de Mendonça Montenegro e Almeida

Mestrado Integrado em Engenharia Informática e Computação

July 24, 2021

Abstract

In recent years, deep learning has become state-of-the-art for predictive tasks such as image classification, achieving excellent results, sometimes even surpassing human ability. Deep learning consists of the use of deep artificial neural networks, which were developed to mimic the human brain, containing units representative of neurons and connections between units that propagate signals.

Due to its complexity, the calculations made by a network to achieve a decision are not intuitive and are quite hard to understand or even compute from a human's point of view. As such, these models are considered "black-boxes", severely lacking in interpretability. The difficulty in understanding a deep learning algorithm's reasoning makes it unfeasible to use it in a real-life scenario where unacceptable results have significant consequences, such as in the medical field. As a result, interpretability has become a trending topic in research regarding deep learning, intending to improve the acceptance of deep learning models in society's various contexts.

One way to add interpretability to a model is through case-based explanations, where the goal is to find data samples or prototypes that are similar to the observation under analysis to use as explanations. In a medical context, this would mean showing images of patients similar to the image being analysed by the medical specialist. Even though this approach has great explanatory value, it raises concerns regarding the violation of patients' privacy.

This dissertation's primary goal is to solve the privacy issue by generating synthetic images based on the patients' images, which preserve features relevant to the medical diagnosis but do not disclose information about the patient's identity. This goal can be achieved using deep generative models, which are state-of-the-art in image generation, aiming to maximise the images' explanatory evidence while minimising the identity leak.

To conclude, the generation of case-based explanations with synthetic images that preserve privacy will help increase the acceptance of deep learning models in the medical field, allowing us to take full advantage of the scientific advances in deep learning to improve the quality of medical diagnosis.

Keywords: Deep Learning, Interpretability, Privacy-preserving Machine Learning, Generative Models

Resumo

Deep learning tornou-se, recentemente, no estado da arte para tarefas preditivas como classificação de imagens, tendo conseguido obter resultados excelentes, por vezes até capazes de ultrapassar a capacidade humana. Deep learning consiste no uso de redes neuronais artificiais profundas, que foram desenvolvidas para imitar o cérebro humano, contendo unidades, representativas de neurónios, e ligações entre as mesmas que propagam sinais.

Os cálculos feitos por uma rede no processo de tomada de decisão não são intuitivos e são difíceis de compreender, ou até computar, do ponto de vista de um humano. Como tal, estes modelos são considerados "caixas negras", caracterizados por uma severa falta de interpretabilidade. A dificuldade em compreender o raciocínio de um algoritmo de deep learning inviabiliza o seu uso num cenário real, especialmente quando resultados inaceitáveis possuem consequências significativas, como é o caso da área da medicina. Por estas razões, a interpretabilidade tornou-se num tópico de tendência na investigação sobre deep learning, com o objetivo de melhorar a aceitação destes modelos em vários contextos da sociedade.

Uma forma de adicionar interpretabilidade a um modelo é através de explicações baseadas em casos, cujo objetivo é encontrar exemplos de dados, ou protótipos, semelhantes à observação em análise para usar como explicações. Num contexto médico, isto significa mostrar imagens de pacientes que são semelhantes à imagem a ser analisada por um especialista médico. Apesar de esta abordagem ter grande valor explicativo, levanta preocupações no que toca à violação da privacidade dos pacientes cujas imagens são mostradas.

O objetivo principal desta dissertação é resolver o problema de privacidade ao gerar imagens sintéticas, baseadas nas imagens dos pacientes, que preservam características relevantes ao diagnóstico médico mas que não revelam informação sobre a identidade do paciente. Este objetivo pode ser alcançado com o uso de modelos generativos profundos, que são o estado da arte em geração de imagens, visando maximizar a capacidade explicativa das imagens e minimizar o reconhecimento do paciente.

Para concluir, a geração de explicações baseadas em casos através de imagens sintéticas que preservam privacidade ajudará a aumentar a aceitação de modelos deep learning na área da medicina, permitindo tirar vantagem dos avanços científicos em deep learning para melhorar a qualidade de diagnósticos médicos.

Keywords: Deep Learning, Interpretabilidade, Privacidade, Modelos Generativos

Acknowledgements

My first word of thanks goes to professor Jaime Cardoso and PhD student Wilson Silva, who accompanied the dissertation since the beginning as its supervisors. Professor Jaime and Wilson, thank you for all the ideas, insightful discussions, and all the support, which kept me motivated throughout the whole dissertation writing process.

I would like to extend my gratitude to the VCMI group at INESC TEC, with a special focus on the interpretability group, with whom I discussed the results of this dissertation and who were always available to offer insights and put things into perspective. In the VCMI group, I would also like to thank Diogo Pernes, Mohsen Saffari, and Wilson Silva, with whom I prepared a presentation on deep generative models, for helping me deepen my understanding of these models.

I am also thankful to the Transparent Artificial Medical Intelligence (TAMI) project at INESC TEC, under which I received a research grant. Being associated with this project was very motivational, and it allowed me to work in collaboration with one other institution associated with the project: Carnegie Mellon University. I would like to thank professors Asim Smailagic and Matt Fredrikson, and PhD student Alex Gaudio, from Carnegie Mellon University, for agreeing to collaborate with us on developing a survey regarding the topics of case-based interpretability and privacy, directly related to the contents of this dissertation.

Finally, I would like to thank my family and friends for all the support they have given me throughout my integrated master's degree.

Helena Montenegro

"Privacy is not something that I'm merely entitled to, it's an absolute prerequisite."

Marlon Brando

Contents

List of Figures

List of Tables

Abbreviations

AE	Autoencoder
cGAN	Conditional Generative Adversarial Network
CNN	Convolutional Neural Network
COLE	Contributions Oriented Local Explanations
DCGAN	Deep Convolutional Generative Adversarial Network
DkNN	Deep K-Nearest Neighbours
EOR	Explanation Oriented Retrieval
GAN	Generative Adversarial Network
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HP	Hierarchical Prototype
IAF	Inverse Autoregressive Flow
IG-CBIR	Interpretability-guided Content-Based Image Retrieval
IoU	Intersection over Union
JS	Jensen-Shannon
KL	Kullback-Leibler
KNN	K-Nearest Neighbours
LSTM	Long Short-Term Memory
MAF	Masked Autoregressive Flow
MRF	Markov Random Field
PGM	Probabilistic Graphical Model
PIECE	PlausIble Exceptionality-based Contrastive Explanations
PP-GAN	Privacy-Protective Generative Adversarial Network
PPRL-VGAN	Privacy-Preserving Representation Learning Variational Generative Adversarial Network
ProtoPNet	Prototypical Part Network
R2VAE	Replacing and Restoring Variational Autoencoders
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SGAP	Siamese Generative Adversarial Privatiser
SN-GAN	Spectral Normalisation Generative Adversarial Network
SSIM	Structural Similarity Index Measure
Tanh	Hyperbolic Tangent
VAE	Variational Autoencoder
VQ-VAE	Vector Quantized Variational Autoencoder
WGAN	Wassertein Generative Adversarial Network
xDNN	Explainable Deep Neural Network

Chapter 1

Introduction

1.1 Context

In the medical field, image processing and analysis is often used to diagnose a patient's condition. For instance, breast cancer, which is the leading cause of death by cancer in women and the second leading cause among both genders [13], can be detected through the analysis of mammographies [67]. Another example is the detection of glaucoma, an eye disease caused by intraocular pressure, responsible for the deterioration of optic nerves leading to vision loss and even blindness [61], which can be done through the analysis of digital fundus images [21]. On top of the concerning consequences seen in these diseases, their early diagnosis, which is critical for a successful treatment, is complicated due to lack of symptoms [61]. Since conditions such as these have severe consequences, the respective diagnosis must be timely and accurate, which prompts the need for having tools to quickly and easily identify these pathologies.

When it comes to diagnosis through image analysis, deep learning has recently achieved state-of-the-art results, sometimes even surpassing human capacity [43, 68]. Recently, McKinney *et al.* [68] have developed a deep learning model for breast cancer screening whose predictions exceed those of human experts. As such, deep learning has shown the potential to provide accurate predictions that can aid medical experts in the decision-making process, especially when dealing with ambiguous diagnostic cases.

The problem that blocks deep learning from being used in real-life scenarios is that many of these models are "black-boxes" whose decisions are hard to understand. Deep learning uses deep neural networks, which contain various units representative of neurons from the human brain. The units are organised in many layers, with connections between them that propagate signals, mimicking the behaviour of synapses from the human brain. Each unit performs an operation over its inputs and propagates the results to the succeeding units. As such, the whole model performs a massive amount of calculations, making it difficult for a human to replicate or even understand the computations done. The lack of interpretability is a barrier that stops these models from being

applied to real scenarios, especially where unacceptable results have significant consequences [28]. Such is the medical field's case, where a wrong diagnosis may lead to a serious disease not being treated or to healthy patients receiving dangerous treatments for a disease they do not possess.

To be able to use these models to improve medical diagnosis, helping doctors achieve a decision even in ambiguous cases that are difficult to diagnose, we need to provide them not only the network's predictions but also explanations that support the predictions. When it comes to image classification, one method to provide intuitive explanations is presenting cases similar to the image being analysed [102, 101]. There are various case-based interpretability methods that allow retrieving images from databases as explanations. However, there are privacy concerns when it comes to using images from patients as explanations.

Currently, the image retrieval process to obtain explanations by examples is characterised by an explanation consumer who submits a medical image to the image retrieval system and receives a similar case acting as an explanation (Figure 1.1). The problem in the current system is that explanation consumers without authorised access to the system's data cannot take advantage of the retrieval system as it would violate the privacy of the patients present in the database. Consumers without authorised access to the data may include patients, medical interns, and even medical specialists who are not accompanying the patient in the retrieved images. To use explanations by example in a medical scenario to support medical experts' decisions and ensure transparency in the decision-making process, there is a need to privatise the retrieved case-based explanations before providing them to the explanation consumers.

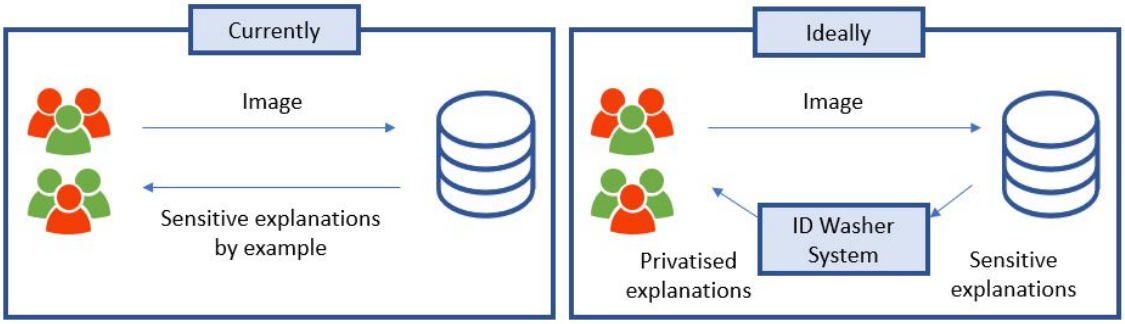


Figure 1.1: Diagram that represents how case-based retrieval systems currently work. Red users represent explanation consumers that do not possess authorised access to the retrieval system's database, such as patients or medical interns. Green users represent explanation consumers that have authorised access to the system's database, such as doctors.

1.2 Motivation

The use of deep learning algorithms, which have the potential to surpass human experts in medical diagnosis [68], can facilitate doctors' decision-making process, helping to achieve decisions even in complicated and ambiguous situations. Although deep learning models make decisions, only

doctors can introspect these decisions and put them into practice. Providing intuitive explanations for the predictions of deep learning models ensures that doctors can trust the models' predictions and use them to obtain additional insights about diagnostic cases.

To take advantage of deep learning's potential to improve medical diagnosis, a model must provide explanations, which can take the form of examples of similar cases. To provide these types of explanations, we need to attend to the privacy concerns associated with showing patients' images. By providing privacy-preserving case-based explanations to justify deep learning models' decisions, we can increase these models' acceptance in real-world contexts.

Furthermore, the privatisation of case-based explanations would allow showing these explanations to a broader audience. Doctors could even show the privatised explanations to the patients to explain the reasons behind their diagnosis. Therefore, privacy-preserving case-based explanations can help to increase the transparency of medical diagnosis.

Even though interpretability and privacy-preserving image generation are trending topics in the research community, their fusion into privacy-preserving explanations has not yet been addressed in the literature. Thus, this dissertation's development can strongly contribute to the scientific community through research on the integration of privacy into explainable artificial intelligence.

1.3 Objectives

This dissertation aims to enable case-based interpretability in the medical scene through the privatisation of visual explanations. For this purpose, we aim to investigate and discuss the current literature in interpretability and privacy to outline the requirements that privacy-preserving case-based explanations must fulfil, considering the characteristics of medical data. After outlining these requirements, we aim to develop a deep generative model capable of generating synthetic images that preserve a patient's privacy while also preserving the explanatory evidence used to understand the reasons behind a deep learning model's decisions. Furthermore, we aim to use this model to improve a classification model's interpretability by generating counterfactual explanations.

Finally, we aim to call the scientific community's attention to the need to privatise case-based explanations and promote discussion on the aforementioned topics. This dissertation will serve as a first step towards the integration of case-based interpretability and privacy.

1.4 Main Contributions

This dissertation's main contributions are:

- We explore the application of current privacy-preserving models to the domain of case-based interpretability for medical image analysis, reflecting on the weaknesses of privacy-preserving approaches. We elaborated a research paper [76] regarding this topic which was published at a workshop on Interpretable Machine Learning in Healthcare, held as part of the ICML conference (ICML 2021 IMLH).

- We propose a novel privacy-preserving model to privatise case-based explanations in the medical scene, which addresses some of the weaknesses of the existing privacy-preserving methods. This novel model was built on top of one of the most promising privacy-preserving architectures available in the literature, improving it from three perspectives: privacy, intelligibility and explanatory value. Furthermore, we also propose an approach to use this model to generate counterfactual explanations. We submitted a research paper [75] with this work to the Winter Conference on Applications of Computer Vision (WACV 2022).

- We survey state-of-the-art case-based interpretability methodologies and privacy-preserving methods, reflecting on the integration of these two research topics. We propose guidelines to guide future work on the novel research topic of Privacy-preserving Case-based Interpretability. We submitted a white paper with this survey's proposal to the IEEE SPM Special Issue on Explainability in Data Science: Interpretability, Reproducibility, and Replicability, which was accepted. Currently, we are preparing the submission of the full paper [77].

1.5 Document Structure

This document contains various chapters that fall under two categories: literature review and experimental work. The literature review chapters present all the relevant information regarding the topics of this dissertation. Chapter 2 introduces background concepts about Deep Learning, giving particular attention to Convolutional Neural Networks, which are widely used in computer vision tasks. Furthermore, this chapter presents a literature review on Deep Generative Models, which will be the focus of this dissertation, as we will develop a generative model to generate privatised images. Chapter 3 reviews the literature on interpretability in machine learning, with a special focus on case-based interpretability. Then, Chapter 4 reviews the literature about current privacy-preserving methods for visual data. Regarding chapters that focus on the experimental work, we start by formalising this dissertation's problem and introducing our approach to fix this problem in Chapter 5. This chapter includes an introduction to the datasets used and some preliminary experiments that precede our privacy-preserving models' development. Chapters 6 and 7 present the experiments that led to the development of two privacy-preserving models applied to the privatisation of case-based explanations. Chapter 8 focus on using the privacy-preserving models to generate counterfactual explanations. Finally, Chapter 9 concludes this document with some final remarks and future work proposals. In Appendix A, we included some visual results to complement the ones exposed in the experimental work.

Chapter 2

Background: Deep Learning

The exponential growth of the information available and accessible at a worldwide level following the appearance of the web has led to the existence of massive amounts of information that can aid decision-making in various real-life scenarios. However, exploring such a significant amount of data requires great amounts of computational resources, which we lack. Facing such limitations, the field of Data Mining has evolved to provide processes and methodologies to extract knowledge from data and apply them to real-world contexts. One of the most significant Data Mining areas in current research is Machine Learning, composed of a family of algorithms that have the ability to learn to uncover patterns in data without being explicitly programmed. Machine learning algorithms usually have two phases: a training phase and a testing phase. During training, the models learn patterns in data, which allow them to perform specific tasks. For instance, in a classification task, where the goal is to assign a label to an object, models are given a set of inputs and the classes they belong to in order to learn how to map the classes to the inputs. The testing phase is used to evaluate the performance of the models. During testing, the models are provided with only inputs to which they guess the respective labels. By comparing them to the real labels, it is possible to evaluate their accuracy and other metrics.

Deep learning is a subset of machine learning algorithms that can learn to extract features according to a given task, unlike traditional machine learning methods where feature extraction must be done separately. The automatic feature extraction process has facilitated tasks such as image classification, where manual feature extraction is hard. As such, deep learning has recently become state-of-the-art in various tasks, including image classification, having achieved excellent results, sometimes even surpassing the human ability [43, 68]. To understand what deep learning is, we first need to understand the concept of artificial neural networks.

An artificial neural network is a model of machine learning developed to mimic the human brain. It is composed of units representing neurons and connections between units, which, similarly to synapses, propagate signals. The most basic version of a neural network is the Perceptron [97], composed of a single unit, which is represented in Figure 2.1. As internal parameters,

a unit contains a set of weights, applied to its inputs, and a bias b, which is a constant value. The perceptron linearly combines its weighted inputs, adds the bias, and applies an activation function h over the result. The resulting output can be seen in Equation 2.1.

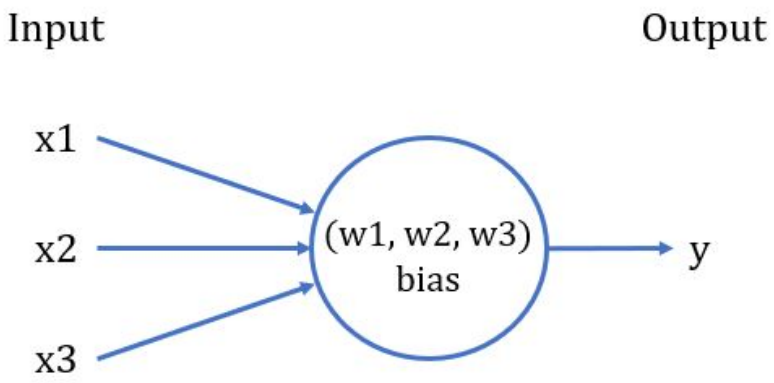


Figure 2.1: Representation of perceptron with 3 input values.

$$y = h(\sum_{i=0}^{n} x_i \times w_i + b) \tag{2.1}$$

The activation function defines the output of the unit. Some examples of commonly used activation functions are [47]:

- **Sigmoid**: this function outputs a number between 0 and 1, which can be thought of as a probability. The derivative of this function is always a value inferior to 1. This function often leads to a problem during the network's training, called the vanishing gradients, where the gradients become so small that they are practically zero, which prevents the unit from optimising the values of its weights during training.

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.2}$$

- **Hyperbolic Tangent (Tanh)**: this function is very similar to the sigmoid function. However, it outputs a value between -1 and 1, and it is centred around zero, which provides stronger gradients that vary in the interval [0, 1]. Similar to the sigmoid function, the Tahn function also suffers from the vanishing gradient problem.

$$f(x) = \frac{2}{1+e^{-2x}} - 1 \tag{2.3}$$

- **Rectified Linear Unit (ReLU)**: this function is very simple to compute. Its gradient is either 0, for negative input values, or 1 if the input is positive. This function leads to a problem called the dead ReLU, characterised by the non-activation of units containing negative inputs since the function outputs 0 in these cases.

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases} \tag{2.4}$$

- **Leaky Rectified Linear Unit (Leaky ReLU)**: this function was defined to fix the dead ReLU problem on the ReLU function. It uses a linear function with a small slope for negative inputs, which results in small gradients capable of adjusting the unit's weights even when the input is negative.

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha x, & \text{if } x \leq 0 \end{cases} \tag{2.5}$$

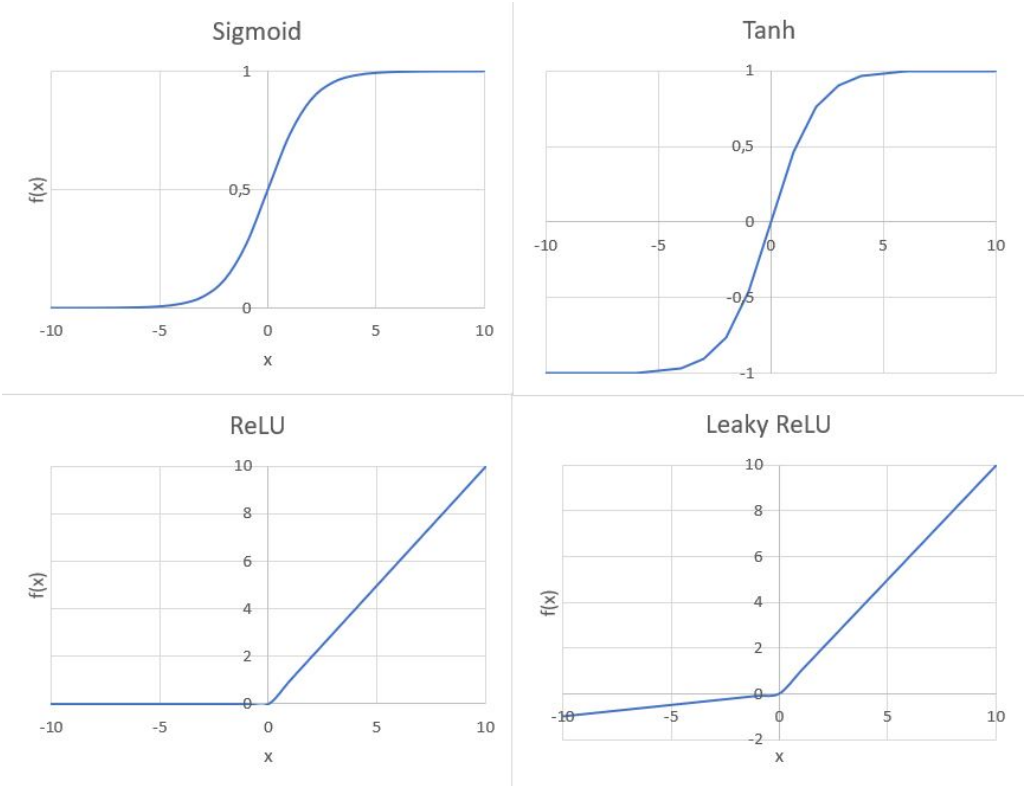


Figure 2.2: Activation function graphs.

A single perceptron can only be applied to a binary classification problem, where there are only two classes to assign to the data instances. We can organise multiple perceptrons in one single layer where all the units are connected to all the network's inputs to solve linearly separable problems, such as multiclass classification, where each unit is a binary classifier capable of identifying one class. Once we add different layers of units with connections between each other, we obtain a multi-layer network capable of solving more difficult, non-linearly separable problems. The number of hidden layers of a multi-layer network defines its depth. Deep learning refers to the use of deep artificial neural networks, characterised by a high number of hidden layers.

There are two types of artificial neural networks:

- **Feed-forward neural networks**: the information, in the form of signals, flows in one direction only, from the input to the output layer, with each layer propagating signals to the layer that follows.
- **Recurrent neural networks (RNNs)**: units can be connected to other units of previous layers. These networks are usually applied to tasks with sequential inputs, such as speech or text [60].

A feed-forward neural network has the capacity to improve itself by tampering with its internal parameters. One method that can be used to optimise a network's weights during training is backpropagation. The backpropagation algorithm is composed of two phases: the forward phase where the network propagates signals from the input layer until the output layer, in a forward manner, and the backward phase, which propagates gradients used to update the weights of the units in a backward manner, as can be seen in Figure 2.3. During the training process, the network goes over the training set multiple times, each corresponding to an epoch, and calculates a loss function for each training sample. The loss function measures how the network is performing regarding the predictive task that it aims to achieve. The loss is minimised through a gradient descent algorithm and propagated to the previous units in the backward phase to compute the units' gradient, which is used to update the respective weights.

One challenge in defining a neural network model is the definition of an appropriate loss function since the model's performance directly depends on this function's quality. The loss function is a manifestation of what we want the model to do, and it is a value that we want to minimise during the training process. For example, if we want a model to classify an input, the simplest loss function we can think is the number of misclassified samples, which is called the zero-one loss function. However, this function's optimisation is computationally complex [80]. A well-known loss function used in classification is the cross-entropy, which is differentiable and decreases as the training samples' probability of being classified as the correct class increases. However, even this function, which is so widely used and has achieved outstanding results, may be lacking in some contexts, such as when there is label noise in the training data [29]. Therefore, it is continuously being improved in the scientific community.

One other challenge in deep learning networks is overfitting, where the network becomes incapable of generalising to new data, achieving low error rates on the training data but high error

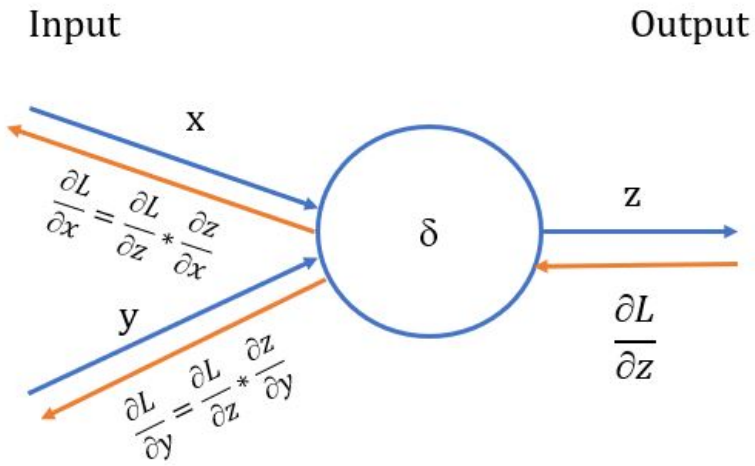


Figure 2.3: Backpropagation on perceptron.

rates on validation data, which is the part of the dataset used to assess the model's capacity to generalise. This issue happens because the number of parameters of the network, given by the number of units, is often larger than the training data. Various regularisation techniques have been introduced in the literature to solve this problem:

- **Early stopping**: consists of stopping the training process before the performance of the network drops for the validation set.
- **Dropout** [44]: consists of randomly dropping units of the network during training, causing perturbations in the network.
- **Data augmentation**: consists of creating new data samples for training using translations and other types of transformations on the training data, increasing the size of the dataset.
- **L1 and L2 Regularisation**: consists of adding a regularisation term to a model's loss function that penalises the models' weights.

2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are state-of-the-art for tasks regarding multidimensional data such as images, thanks to their capacity to extract features through convolution and pooling operations. Image classification, which is a predictive task where the goal is to assign a class to an image, is one example of the tasks that use CNNs.

Various CNNs have been developed and are widely used in the scientific community for image classification, such as VGG [105], Inception [109] and ResNet [42]. Additionally, these models have been trained on generic datasets, such as ImageNet [25]. As such, their weights can be

reused in various domains through transfer learning, improving the network's training and its performance.

CNNs are composed of convolutional and pooling layers used alternately to detect and extract features in multidimensional data like images. In each layer, they can extract different features starting with simple ones such as lines in the first layer and evolving to more complex features that combine the ones defined in the previous layer. Like other neural networks, convolutional neural networks can also use backpropagation for training. The next sections explain the types of layers used in these networks.

2.1.1 Convolutional Layer

A convolutional layer is composed of filters, or kernels, used on an input image through a convolution operation to generate a feature map, also called an activation map. The filter is typically smaller than the input image and acts like a sliding window that slides over the image to build the feature map, as shown in Figure 2.4. Each element of the feature map is calculated by the scalar product between the filter and the part of the image overlapped by the filter. Figure 2.4 illustrates the first iterations of the process of calculating the values of the feature map by applying a filter with dimensions 3x3, which slides over an input image with dimensions 6x6.

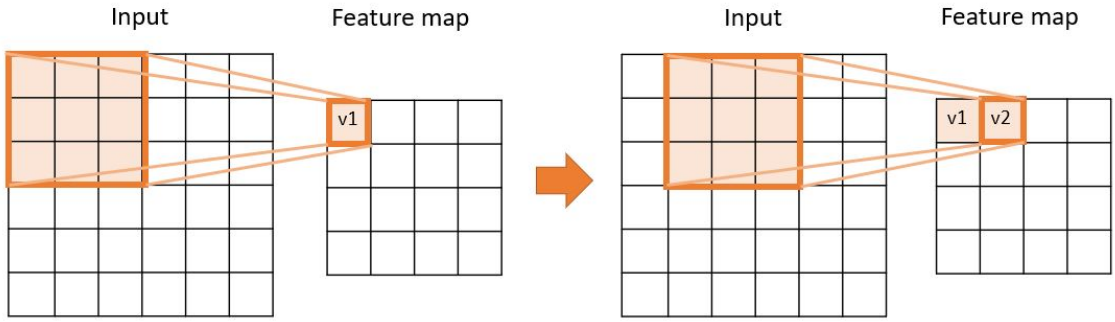


Figure 2.4: Convolution operation done in convolutional layer.

The feature map is smaller than the original input image since the filter is not applied to the image's borders. On a convolutional layer, we can use multiple filters, each one representing one feature. For each filter we use, we obtain a feature map representing the existence of the feature represented in the respective filter in the input image.

Convolution operations can also be applied with the purpose of up-sampling the data through Transposed Convolutional Layers. In this case, we add columns and rows full of zeros in the input's borders to make it bigger, and we apply the convolution operation, obtaining an output bigger than the input.

2.1.2 Pooling Layer

The pooling layer applies a commutative operation, such as the maximum operation or an average, over an input, using a sliding window that slides over the whole input. Usually, the window's stride, which is the number of pixels that are skipped ahead when the window moves over the image, is the same size as the window. Thus, each pixel of the input image only contributes to the value of one activation map element. This layer allows obtaining smaller and more compact representations of an input by down-sampling the input image. Figure 2.5 shows an example of what happens in a Max Pooling layer, where the maximum operation is applied over the input of size 4x4, in windows of size 2x2 with a stride of 2. In each element of the feature map, the corresponding value is being calculated considering only the input pixels of the same colour.

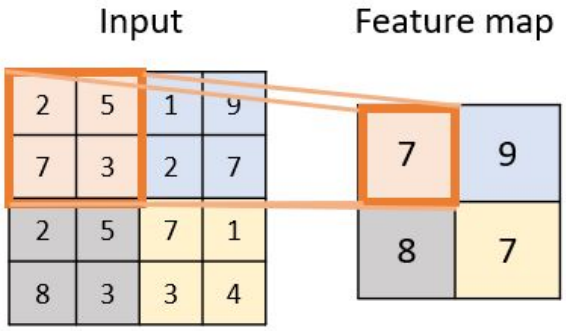


Figure 2.5: Example of max pooling operation.

2.1.3 Fully Connected Layer

A fully connected layer contains its units connected to all the units of the previous layer. Usually, this layer is used after the convolutional and pooling layers, taking the features extracted in these layers as inputs to make decisions in tasks such as classification.

2.2 Generative Models

Generative models learn the data probability distribution, which can be used to generate new data samples that look like they belong to the training data used to train the model. These models learn patterns in the data that allow them to generate new data based on these patterns.

Before introducing the deep generative models, it is important to introduce the following concepts:

- **Kullback-Leibler (KL) Divergence** [59]: asymmetric function that measures how a probability distribution p diverges from a second distribution q. D_{KL} reaches zero when the probability distributions are equal, and it tends to infinity when the probabilities are disjoint [117].

$$D_{KL}(p||q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx \tag{2.6}$$

- **Jensen-Shannon (JS) Divergence** [64]: symmetric function to measure the similarity between two probability distributions. When both probabilities overlap, this metric is not differentiable [117].

$$D_{JS}(p||q) = \frac{1}{2} D_{KL}(p||\frac{p+q}{2}) + \frac{1}{2} D_{KL}(q||\frac{p+q}{2}) \tag{2.7}$$

- **Wasserstein Distance**: also called Earth Mover's distance since it indicates how much mass must be moved to transform the probability distribution p in q [9]. Compared to the previous metrics, the Wasserstein distance is differentiable for all points and contains a value representative of the distance between the distributions, even when these are disjoint [117]. In Equation 2.8, $\prod(p,q)$ refers to the set of all the possible joint probability distributions between p and q.

$$D_W(p||q) = inf_{\gamma \in \prod(p,q)} \mathbb{E}_{(x,y)\sim\gamma}[||x-y||] \tag{2.8}$$

- **Likelihood**: probability that the model assigns to the training data assuming independence of the samples. It is the product of the probabilities of the m training samples, as can be seen in Equation 2.9.

$$Likelihood = \prod_{i=1}^{m} p_{model}(x_i) \tag{2.9}$$

- **Maximum Likelihood Estimation**: consists of choosing the parameters of a model that maximise the likelihood of the training data. Maximising the likelihood is equivalent to minimising the KL Divergence between the data distribution and the generative model's distribution.

Generative models differ in the way they learn the data distribution. In Figure 2.6 there is a taxonomy proposed by Goodfellow [35]. Models that can draw samples from a probability distribution without directly defining it are considered implicit density models, as is the case of Generative Adversarial Networks. On the other hand, models that explicitly define the data distribution, and that can therefore maximise the likelihood directly, are considered explicit density models and can be further divided into two groups: models where the probability distribution is

intractable, which need to make approximations to maximise the likelihood, like Variational Autoencoders, and models that explicitly define the distribution to be computationally tractable, like autoregressive models and flow-based models. The main problem in developing explicit models is that it is difficult to obtain a model that successfully captures the data distribution's complexity while preserving computational efficiency [35].

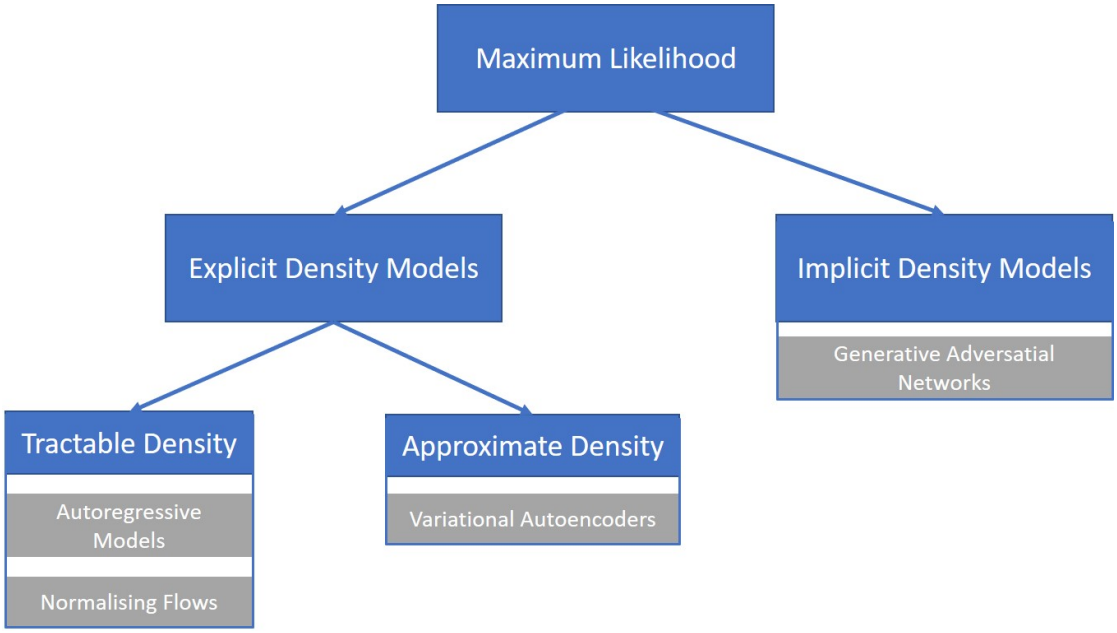


Figure 2.6: Taxonomy for generative models introduced by Goodfellow.

The following subsections start by introducing traditional generative models, which do not use deep learning models and are mostly applied to tasks other than data generation, followed by the introduction of the four most popular deep generative models: Generative Adversarial Networks, Variational Autoencoders, Autoregressive Models and Normalising Flows. Table 2.1 presents an overview of how each of the introduced models captures the data's probability distribution.

Table 2.1: Overview of Deep Generative Models

Model	**Taxonomy**	**How it captures data probability distribution**
GAN	Implicit	Minimax game between generator and discriminator makes the probability distribution in the generated data converge to the real data distribution.
VAE	Explicit	Maximises data likelihood using encoder-decoder architecture: encoder approximates the true posterior $p(z \mid x)$ and decoder models the likelihood $p(x \mid z)$.
Autoregressive Model	Explicit	Computes the joint distribution of the data by calculating the product of the conditional distributions of each data dimension.
Normalising Flow	Explicit	Transforms a simple data distribution into a more complex distribution through a series of invertible transformations.

2.2.1 Traditional Generative Models

Traditional Generative Models focus on estimating the probability density through two techniques: parametric, which estimates the parameters of a known density that better approximate the training data, and non-parametric, which does not assume the form of the data density.

The models introduced in this section are generative models that were not specifically designed for data generation. They can be applied to the tasks of classification, clustering, anomaly detection, among others.

2.2.1.1 Parametric Techniques

The simplest example of a generative model is the Naive Bayes model, used in classification, which learns the joint probability $p(x,y)$ between an input x and a class y. To calculate the likelihood of a class given an input, the algorithm calculates the posterior $p(y \mid x)$ through the Bayes theorem, expressed in Equation 2.10, assuming independence between attributes of an instance. The label with the highest likelihood is then assigned to the input.

$$p(y \mid x) = \frac{p(x \mid y) \times p(y)}{p(x)} \tag{2.10}$$

The Naive Bayes model belongs to a family of models called the Probabilistic Graphical Models (PGMs). These models are represented by graphs composed of nodes that represent variables and edges that express dependencies between variables [108]. There are two main types of PGMs: Bayesian networks, characterised by directed graphs, and Markov networks, characterised by undirected graphs.

Bayesian classifiers, such as the Naive Bayes classifier, apply the Bayes theorem to estimate each class's probability. In these models' graphs, an edge's target node's probability is conditioned by the edge's source node. The graph of a Naive Bayes classifier has nodes representative of classes with edges pointing towards attributes, as can be seen in Figure 2.7, where C represents a class, and A_i represents data attributes.

Various models are based on Markov chains, which are stochastic models composed of states and transitions between states. Each transition between state s_A and s_B is characterised by the state's probability of changing from s_A to s_B, which is the conditional probability $p(s_B \mid s_A)$. These models satisfy the Markov property since the probability of the state's change depends only on the current state, not on the states that preceded it. Markov chains repeatedly update the states and respective probability distributions and generate samples by retrieving x given an input y from the conditional probability $p(x \mid y)$. The convergence towards the real data distribution is slow, making these models inappropriate for high-dimensional data spaces [35].

Markov chains are the basis of the Hidden Markov Model (HMM), which is a particular case of a Bayesian network [108]. HMM contains observable states, represented by data that we want to classify, and hidden states, which are not observable and that commonly refer to labels when

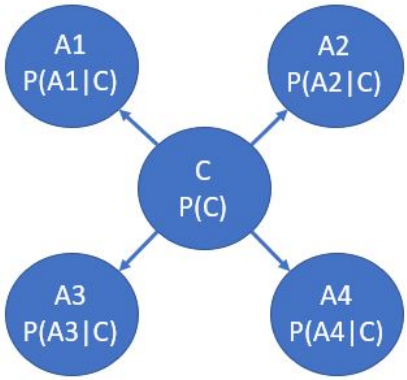


Figure 2.7: Example of Naive Bayes graph.

it comes to classification problems. The probability of an observable state, or observation, is conditioned by the probability of the underlying hidden state. During training, the model learns the probability of a data sample x conditioned by a state y: $p(x \mid y)$, called the emission probability, and the probability of a state y_1 being followed by the state y_2 in the sequence: $p(y_2 \mid y_1)$, called the transition probability. HMMs have found many applications, including speech recognition to understand what word was uttered in a speech signal, where each label represents a phoneme [91].

Markov networks refer to Markov Random Fields (MRFs), where the states that make up the graph can be organised in a chain or a grid. Each state's probability is conditioned solely by the respective neighbours, meaning that a state is conditionally independent of all others given its neighbours [108].

Another type of parametric models is the Gaussian Mixture Models (GMMs) [69, 70]. GMMs are a type of generative model which linearly combines different Gaussian densities, originating complex data distributions [10]. The density estimation for one observation is calculated through Equation 2.11, where K is the number of components of the mixture model, equivalent to the number of clusters when considering a clustering task, $\mathcal{N}(x \mid \mu_k, \Sigma_k)$ is a Gaussian distribution with mean μ and covariance Σ, and π_k is the weight assigned to the respective distribution. GMMs are trained through maximum likelihood estimation, using optimisation techniques such as Expectation-Maximisation [10].

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) \tag{2.11}$$

2.2.1.2 Non-parametric Techniques

There are three non-parametric techniques that focus on calculating the data density through Equation 2.12, where K is the number of observations inside a region R, N is the total number of observations and V is the volume of the region R.

$$\hat{p}(x) = \frac{K}{NV} \tag{2.12}$$

The three methods differ in the definition of the region R and subsequent calculation of the variables K and V:

- **Histogram**: This method divides each attribute into M bins, where each bin quantifies the occurrence of samples. As such, in the calculation of $\hat{p}(x)$, K refers to the number of samples inside the bin that contains the sample x and V refers to the respective bin's volume.

- **Parzen Windows**: This method defines R as a region centred on the sample x with a set volume V. It uses a kernel function to calculate the number of samples inside the region K. If we interpret this region as a hypercube, a possibility of a kernel function is expressed in Equation 2.13, which checks whether its input u is contained inside a unitary hypercube centred on the origin. This kernel function results in the density estimation characterised by Equation 2.14, where h represents the size of the hypercube's edge and D represents its dimensions [10].

$$k(u) = \begin{cases} 1, & \text{if } |u_i| \leq \frac{1}{2}, \quad i = \{1,...,D\} \\ 0, & \text{otherwise} \end{cases} \tag{2.13}$$

$$\hat{p}(x) = \frac{1}{Nh^D} \times \sum_{n=1}^{N} k(\frac{x - x_n}{h}) \tag{2.14}$$

- **K-Nearest Neighbours**: Similarly to Parzen Windows, this method defines a region R around x, but instead of having a fixed volume and calculating K, it does the opposite, calculating the volume V of the region around x that contains its K nearest neighbours [10].

Compared with parametric techniques, non-parametric ones have the advantage that they do not assume the form of the data distribution, enabling them to model complex target densities. However, these algorithms are sensitive to some parameters, such as the number of used bins or the regions' size, and the data dimensions.

2.2.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [34] became the most well-known method in image generation, thanks to their high capacity of generating realistic images. GANs are composed of two adversarial networks, which compete with each other: the generator and the discriminator. The generator generates realistic synthetic images starting with an input, often defined as random noise, while the discriminator discriminates between real and synthetic images. The goal of the generator

is to trick the discriminator into thinking that the generated image is real. The discriminator is a binary classification network that classifies the images as either real or fake.

The loss function for a GAN is defined in Equation 2.15, where x represents the real samples and z represents the random noise. $G(z)$ represents the synthetic image originated by the generator. The generator and the discriminator play a two-player minimax game where the generator wants to minimise the loss function, minimising the number of correct classifications in the discriminator, while the discriminator intends to maximise it.

$$V(D,G) = E_x[\log D(x)] + E_z[\log(1 - D(G(z)))] \tag{2.15}$$

The training process of a GAN involves training the generator and the discriminator simultaneously, using backpropagation to minimise the loss, as can be seen in Figure 2.8. The loss function approximates the JS Divergence [35].

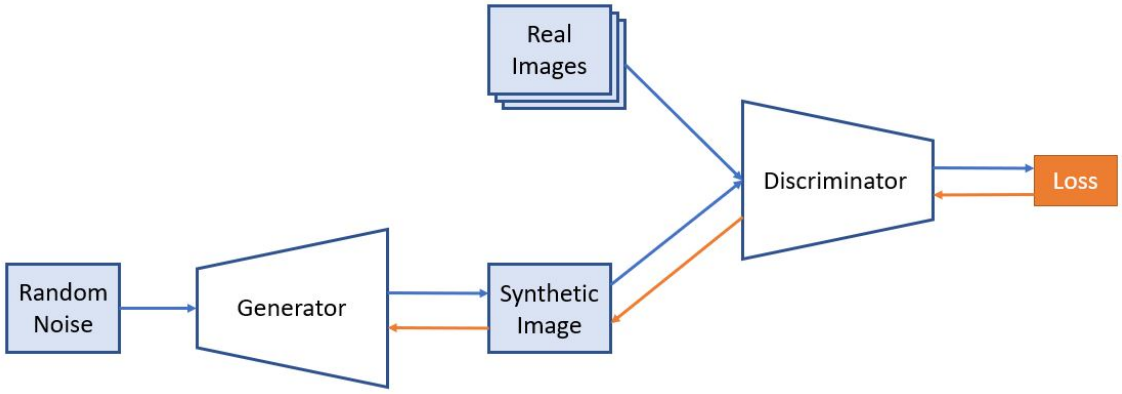


Figure 2.8: Training process of a GAN.

Regarding the probability distribution captured by this generative model, the generator implicitly defines a data distribution p_g as the distribution of the samples it generates. As proven by Goodfellow *et al.* [34], the minimax game between the discriminator and the generator possesses a global optimum for $p_g = p_{data}$, which leads the generator's distribution to converge to the real data distribution p_{data} if the generator and discriminator have enough capacity.

GANs possess various problems:

- **Vanishing Gradient**: occurs when the network's gradients are very close to zero, which prevents it from optimising its weights. In the context of GANs, this problem usually happens when the discriminator is too good and successfully classifies the samples that come from the generator as fake since the gradient of the cross-entropy loss approaches zero [35].

- **Mode Collapse**: happens when the GAN learns to always output the same data point for different inputs [35]. In practice, it is more common for partial mode collapse to occur,

where instead of the same output point, the GAN generates points that always contain the same set of features.

- **Difficulty in achieving Nash Equilibrium**: training a GAN requires finding a Nash Equilibrium where both the discriminator and the generator do not change their strategies regardless of what the other network will do. This means that the opponent player's action in the minimax game does not change the game's outcome. Once this state is achieved, the probability distribution of the model has converged to the real data's probability distribution. The problem is that this state is difficult to achieve [35].

Despite the disadvantages, GANs are known to generate high-quality synthetic images. Different research lines in GANs focus on tackling the problems frequently seen in GANs and improving image quality. One research line investigates loss functions that can be applied to GANs and improve their performance. For instance, Arjovsky *et al.* [9] proposed the Wasserstein GAN (WGAN), which uses the Wasserstein distance in the loss function to fix mode collapse and stabilise the network. More recently, Xiangli *et al.* [120] proposed the RealnessGAN, which produced very realistic results, as can be seen in Figure 2.9. This network represents realness as a distribution rather than a scalar. In practical terms, this means that the discriminator outputs a distribution instead of a scalar, using a softmax activation in the decision layer. In the loss function, this network uses the Kullback-Leibler Divergence between gaussian distributions for real and fake images and the network's output. Additionally, RealnessGAN takes advantage of spectral normalisation, proposed in Spectral Normalisation GAN (SN-GAN) [73], which is applied to the discriminator's layers to stabilise its training.

Figure 2.9: Example of images obtained with RealnessGAN. Source: [120]

Other lines of research investigate GANs' architecture. For instance, Deep Convolutional Generative Adversarial Networks (DCGAN) [92] use a deconvolutional neural network, composed of transposed convolutional layers, in the generator. Other architectures like SAGAN [115] and

BigGAN [14] use self-attention mechanisms in the generator and discriminator's architectures, capable of capturing global dependencies in the data. Additionally, many variations of GANs have been developed in the past few years to adapt them to different tasks. One variation of interest for the context of privacy-preserving image generation is the conditional generative adversarial network (cGAN) [72], which applies restrictions to GANs, such as the initial image used by the generation process instead of random noise. The Pix2Pix GAN [46] is an example of a cGAN that performs image-to-image translation. In addition to the discriminator's loss, this network uses L2 Normalisation to approximate the generated image to a given ground truth.

2.2.3 Variational Autoencoders

Autoencoders (AEs) are networks composed of two neural networks: an encoder and a decoder, as can be seen in Figure 2.10. The encoder performs an operation of dimensionality reduction by creating a representation of the input image in a low dimensional latent space. The decoder, given the data sample in the encoder's latent space, tries to reconstruct the original input image. The reconstruction error given by the distance between the original input image and the synthetic image outputted by the AE is used as a loss function, updating the weights of the network through the backpropagation process. As such, it forces the encoder to learn to codify as much information as possible in the low dimensional space so that the decoder has enough information to rebuild the original image. Autoencoders are generally used for data dimensionality reduction or data denoising. To use AEs with images as input, we can define the encoder and decoder's architectures as convolutional neural networks.

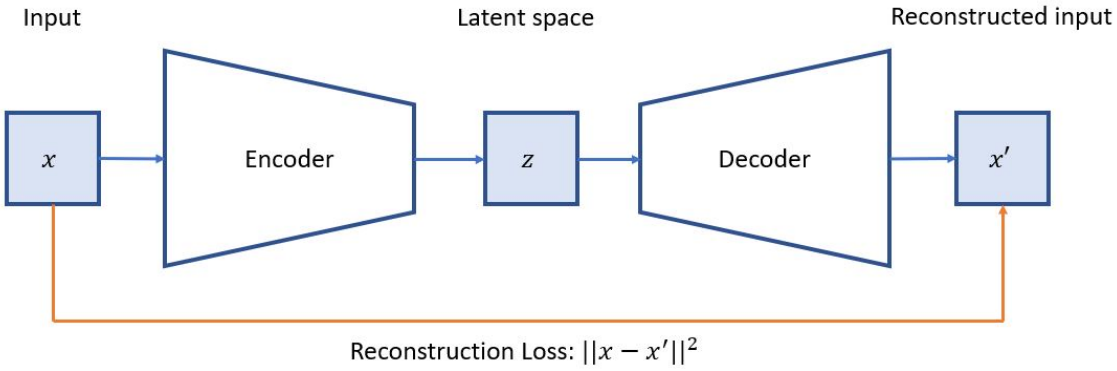


Figure 2.10: Architecture of autoencoder.

Variational Autoencoders (VAEs) [56] are a variant of AEs that can be applied to new data generation. Instead of mapping an input image to single points in the latent space, this network maps the input into a simple distribution over the latent space, such as a normal distribution characterised by a mean μ and a standard deviation σ. The encoder maps the input x into a latent representation z, yielding an approximate distribution $q(z \mid x)$ in the latent space. The decoder reconstructs the original input data based on the latent representation z, resulting in a likelihood

distribution $p(x \mid z)$. Figure 2.11 displays an overview of this model's architecture. New images are obtained from decoding an observation sampled from the low dimensional space.

The loss function in VAEs, represented in Equation 2.16, contains two parts: a reconstruction loss and a regularisation loss. The reconstruction loss is the negative likelihood, used to ensure that the encoder codifies as much useful information as possible so that the decoder can use that information to reconstruct the original input and to encourage the decoder to obtain a synthetic image as similar to the original one as possible. The regularisation loss uses the KL divergence to measure the distance between the encoder's distribution $q_\theta(z \mid x)$ and the original distribution $p(z \mid x)$ [83], quantifying the information that is lost in the latent representation of the original data. As such, by minimising the loss, the VAE estimates the parameters of the latent space's distribution that better approximates the real data distribution through the process of maximum likelihood estimation performed by the minimisation of the KL divergence.

$$L = -E_{z\sim q_\theta(z|x)}[\log p(x \mid z)] + D_{KL}(q_\theta(z \mid x)||p(z \mid x)) \tag{2.16}$$

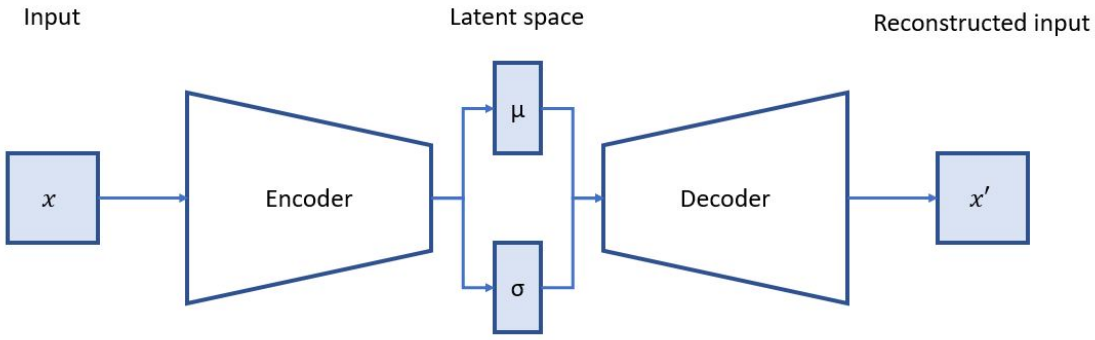


Figure 2.11: Architecture of variational autoencoder with normal distribution.

VAEs have a great capacity to learn representations of data. Depending on the loss function used, these models can preserve only certain features that are important for a particular task. For example, suppose we connect a classification model to the VAE's output and define the network's loss as the cross-entropy loss instead of the distance between the original image and the reconstructed one. In this case, the VAE can learn to represent only the semantic features needed for the classification task, discarding others.

Since VAEs are probabilistic models, they often result in blurry images when applied to image generation. However, recently there have been some developments towards producing high-quality images such as the ones seen in Figure 2.12. These images were obtained with the model VQ-VAE-2 [93], which combines a Vector Quantized Variational Autoencoder (VQ-VAE), originally proposed by Oord *et al.* [113], to produce a discrete latent space using vector quantisation techniques, with an autoregressive model, to learn a prior from which we can sample variables from the latent space to generate new images.

Figure 2.12: Example of images obtained with VQ-VAE-2. Source: [93]

2.2.4 Autoregressive Models

Autoregressive models are generative models where the generation of a new sample of data depends on the previously generated samples [35]. These models are instrumental in the generation of sequential data, where an element in the sequence depends on the elements that precede it, as can be seen in Equation 2.17.

$$p_{model}(x) = \prod_{i=1}^{n} p_{model}(x_i \mid x_1, ..., x_{i-1}) \tag{2.17}$$

Images can be interpreted as a pixel sequence, where each pixel is conditioned by the pixels previously generated. This is the basis of the PixelRNN model [112] developed for image generation. PixelRNN is a model based on recurrent neural networks, specifically Long Short-Term Memory (LSTM) networks, which are known to excel at sequence problems, applied to model natural images. It generates one pixel at a time, predicting its conditional distribution over the different possibilities of values conditioned by the context already generated. This model applies convolutions to compute all the image pixels along one spatial dimension of the data, such as a row or a diagonal. The model possesses the following architectures:

- **Row LSTM**: this architecture uses one-dimensional convolutions to capture features in a whole row, which are used to define the context used to compute the probability distribution of a pixel. The context captured by this layer is a triangular set of pixels that are positioned above the pixel under analysis.

- **Diagonal BiLSTM**: this architecture, unlike the one before, can capture the entire context of the image by diagonally scanning the image, from the top left corner until the bottom

right corner of the image.

- **PixelCNN**: this architecture, instead of an RNN, uses a CNN with multiple convolutional layers which compute features for all pixel positions at once, using masks to avoid seeing the future context. This CNN does not possess pooling layers in order to preserve the spatial resolution of the data.

The main problem in these networks is that the generation of samples cannot be parallelised since the elements in a sequence depend on previously generated elements, leading to a slow and inefficient generation process.

2.2.5 Normalising Flows

Normalising flows [94] are generative models that transform a simple distribution of probabilities into a more complex one through a sequence of invertible and differentiable transformations. Starting with a simple distribution, like a Gaussian distribution, normalising flows apply a series of invertible transformations to approximate the real data distribution as illustrated in Figure 2.13.

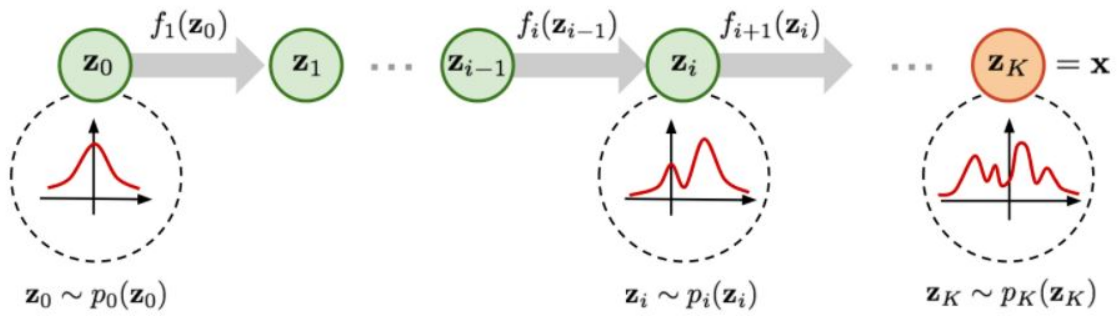


Figure 2.13: Transformation of a simple distribution into a more complex one. Source: [118]

More formally, given a variable z, whose probability distribution function p_z is known, and an invertible function g, whose inverse function is f, a new variable in a more complex data space can be obtained using Equation 2.18. The probability distribution of the new variable is obtained using Equation 2.19, where $Df(x)$ refers to the Jacobian of the inverse function f [57].

$$x = g(z) \tag{2.18}$$

$$p_x(x) = p_z(f(x)) \times |detDf(x)| \tag{2.19}$$

The invertible function g is called the generator since it generates a data sample from a complex distribution based on a random sample from a simple data distribution. The inverse function f can

be applied to normalise the data distribution, turning a complex distribution into a simpler one [57].

An overview of the architecture of these models can be seen in Figure 2.14. In the normalising direction, the normalising flow composed of multiple transformations is applied to the real data to transform it into a latent space with a simpler probability distribution. Through the generative direction, we sample a point from the latent space and apply the inverse flow to obtain the synthetic image in the original data space.

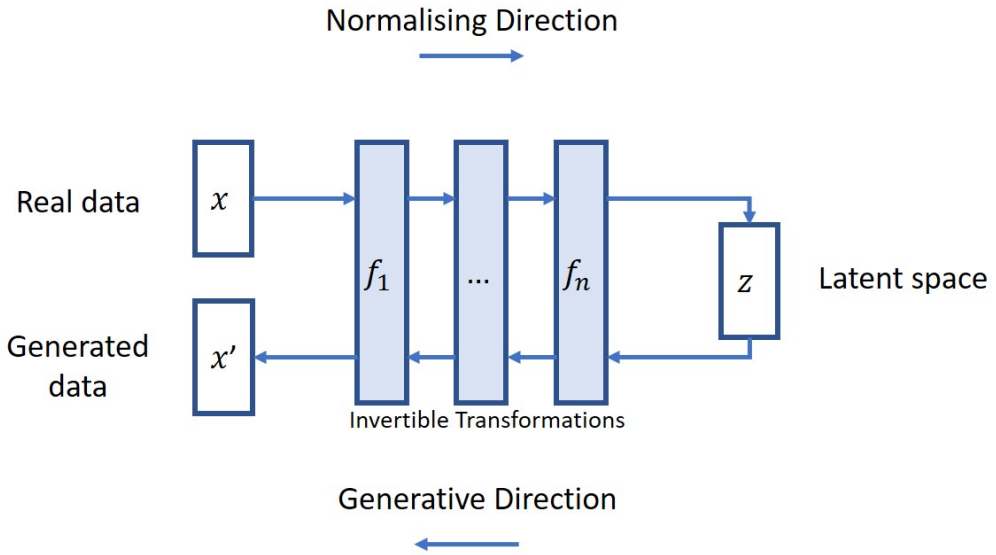


Figure 2.14: Overview of Normalising Flows Architecture.

During training, a normalising flow minimises the negative log-likelihood over the training dataset D, which is used as the loss function, represented in Equation 2.20.

$$L(D) = -\frac{1}{|D|}\sum_{x}\log p(x) \tag{2.20}$$

A normalising flow must be [57]:

- **Invertible**: since for sampling, we need the flow function, and to calculate the probability distribution of the target complex distribution, we need its inverse.
- **Expressive**: to be able to represent the distribution of interest, despite its complexity.
- **Computationally Efficient**: when it comes to the computation of both the flow and its inverse, as well as the respective Jacobians.

These networks possess various advantages, as they achieve the difficult task of explicitly learning the data distribution. Unlike GANs and VAEs, these models can evaluate the probability distribution of new points by reverting them into the simpler distribution [57]. They also have high

expressive power since they can model any distribution even if the base distribution is very simple [86]. However, these networks possess one drawback: they impose restrictions on the choice of function g, which must be invertible, restricting the dimension of the latent variables that must have the same dimensionality as the original input [35].

The different types of flows that can be used to model a normalising flow are [57]:

- **Elementwise flows**: apply non-linear transformations to each variable in the flow, assuming independence between variables. The problem with these flows is that they do not express possible correlations between variables in the data.
- **Linear flows**: apply linear transformations between variables in the flow. These flows lack the expressiveness needed to model complex distributions.
- **Planar and radial flows**: apply non-linear transformations with the disadvantage that the respective inverse functions are hard to compute. Planar flows contract and expand the distribution along a plane, while radial flows distort the distribution around a point in the data space. Both of these were proposed by Rezende and Mohamed [94].
- **Coupling and Autoregressive Flows**: use coupling functions to build invertible non-linear transformation with high expressive power.
- **Residual flows**: use invertible residual networks as flexible transformations.
- **Infinitesimal flows**: are extensions of residual flows that apply to continuously dynamic systems.

In the following subsections, we introduce a more detailed description of the most common flow architectures: coupling flows and autoregressive flows.

2.2.5.1 Coupling Flows

Coupling Flows were introduced by Dinh *et al.* [26]. Given a bijection $h(\cdot,\theta)$ and a disjoint partition of the data x in two subsets (x^A, x^B), a coupling flow g is defined by Equation 2.21, where $\Theta(x^B)$ is called the conditioner and is an arbitrary function applied to x^B [57]. The bijection is often called *affine coupling layer* in the literature.

$$\begin{aligned} y^A &= h(x^A;\Theta(x^B)) \\ y^B &= x^B \end{aligned} \tag{2.21}$$

The coupling flow g is invertible if h is invertible, and its inverse is obtained by Equation 2.22.

$$\begin{aligned} x^A &= h^{-1}(y^A;\Theta(y^B)) \\ x^B &= y^B \end{aligned} \tag{2.22}$$

Both the coupling flow's Jacobian and its determinant are easy to compute, as the Jacobian Dg is a triangular matrix where the diagonal contains the Jacobian of h, Dh, and the identity matrix, and the determinant of Dg is the determinant of Dh [57]. The advantage of using coupling flows is that $\Theta(x^B)$ can be arbitrarily complex since we don't need to compute its inverse to obtain the inverse of the coupling flow.

One model that uses coupling flows is the RealNVP model proposed by Dinh *et al.* [27], which possesses a series of affine coupling layers where the conditioner function Θ consists in a scale and shift transformation, as can be seen in the coupling flow's equation in Equation 2.23, where $s()$ represents a scale operation and $t()$ represents a translation. In this equation, we can see that the first d elements remain unchanged. In order to avoid this and make sure that all input values have a chance of being altered, the order of the inputs is inverted on each layer on the network through an inverse permutation.

$$\begin{aligned} x_{1:d} &= y_{1:d} \\ x_{d+1:D} &= y_{d+1:D} \odot \exp(s(x_{1:d})) + t(x_{1:d}) \end{aligned} \tag{2.23}$$

Kingma *et al.* [54] proposed the Glow model to simplify the architecture of the RealNVP by replacing the reverse permutation done on every coupling layer by an invertible convolution 1x1. This model generates realistic images, as can be seen in Figure 2.15, which presents an example of face images generated with it.

Figure 2.15: Example of images generated with Glow Architecture. Source: [54]

2.2.5.2 Autoregressive Flows

Autoregressive flows are applied to sequential data, imposing the restriction that an output value depends only on the values observed in the past. These flows use autoregressive models, such as

the ones seen in Section 2.2.4, as invertible functions. Images can be interpreted as sequences of pixels, where the probability distribution of a pixel being drawn depends on the pixels that have already been drawn.

Putting autoregressive flows into more formal terms, given a bijection $h(\cdot,\theta)$, an autoregressive model is a function g whose output is conditioned by the previous inputs, defined by the Equation 2.24, where the function Θ_t is called the conditioner [57].

$$y_t = h(x_t;\Theta_t(x_{1:t-1})) \tag{2.24}$$

The inverse of the function g can be calculated recursively through Equation 2.25. The most significant disadvantage in autoregressive flows is that the inverse function is a sequential operation that cannot be parallelised, resulting in a slow and computationally expensive generative process.

$$\begin{cases} x_1 = h^{-1}(y_1;\theta_1) \\ x_t = h^{-1}(y_t;\theta_t(x_{1:t-1}), \quad t = 2,\ldots,D \end{cases} \tag{2.25}$$

The use of autoregressive models in a network to create an autoregressive flow was first introduced in the Masked Autoregressive Flow (MAF) by Papamakarios *et al.* [87].

In order to make the generative process faster, Kingma *et al.* [55] introduced the Inverse Autoregressive Flow (IAF), where instead of conditioning an input x_t on the previous inputs $x_{1:t-1}$, it conditions x_t based on the outputs of the previous entries $y_{1:t-1}$, as can be seen in Equation 2.26. This function is equivalent to the inverse function of a normal autoregressive flow, hence the name "Inverse Autoregressive Flow", making the computation of the flow sequential and, therefore, slow. However, since the inverse function of the IAF is the direct autoregressive flow, its generative process can be computed efficiently.

$$y_t = h(x_t;\Theta_t(y_{1:t-1})) \tag{2.26}$$

The choice between using autoregressive flows or inverse autoregressive flows needs to consider whether to prioritise the efficiency of the generative process, encouraged in IAF, or of the computation of the probability density estimation, encouraged in direct autoregressive flows.

2.3 Summary

Deep Learning has achieved state-of-the-art results for Computer Vision tasks. Deep Learning models are based on artificial neural networks, which imitate the human brain through a structure composed of units that propagate numerical signals. Computer vision tasks work with visual data, where feature extraction is difficult to optimise. Convolutional neural networks were developed to

facilitate feature extraction in multi-dimensional data through the alternation of convolution and pooling operations.

One field of research in Deep Learning with particular interest for privacy-preserving case-based interpretability concerns Deep Generative Models, capable of fulfilling the unsupervised task of data generation. Generative Models learn the distribution of the data, which is used to generate new data that follows the same distribution. These models are often used in deep learning privacy-preserving approaches to generate privacy-preserving images. Generative Models differ in how they model the data distribution. In traditional Machine Learning, generative models can be parametric, approximating the data density by estimating the parameters of a known probability distribution, or non-parametric, which can model more complex data distributions since they do not assume the form of the data density.

Regarding Deep Generative Models, GANs model the data density implicitly through a minimax game between two networks: a generator responsible for generating data and a discriminator which distinguishes between real and fake data samples. This adversarial training promotes the generation of realistic images in the generator, tricking the discriminator into thinking the generated samples are real. VAEs model the data density explicitly by learning an approximation of the real data distribution using a simple data distribution that enables sampling. These models possess an encoder that learns to map samples in the original data space to images in a latent space that follows the simple data distribution. Then, there is a decoder that learns to map observations in the latent space to the original data space, allowing to visualise instances obtained through sampling from the latent space. Autoregressive Models and Normalising Flows define a tractable data density. Autoregressive Models interpret images as pixel sequences and compute the joint distribution of the data by calculating the product of the conditional distributions at the pixel level. Normalising flows transform a simple data distribution into a more complex one through a sequence of invertible and differentiable transformations. The Normalising Flows' architecture possesses a normalising flow that maps data in the original space to the latent space through the transformations and a generator which maps data in the latent space to the original data space through the inverse operations of the normalising flow. Like in VAEs, we can sample from the simple data distribution and visualise it through the generator.

The development in the area of Deep Generative Models has enabled the generation of realistic images. Nevertheless, the achievement of this challenging unsupervised learning task requires a significant amount of computational resources.

Chapter 3

Literature Review: Interpretability in Machine Learning

One of the biggest problems in developing machine learning algorithms that can learn without being explicitly programmed is that many are not interpretable. For instance, deep learning algorithms are often described as "black-boxes" since their large number of units makes it impossible for a human to replicate or even interpret the network's computations to achieve a decision. This lack of interpretability is a barrier that does not allow these algorithms to be used in real-life scenarios, where unacceptable results have significant consequences [28], due to a lack of trust in the algorithms and the respective results. This lack of acceptance led to a rise in research about interpretability in machine learning in recent years, which will be reviewed in this section, focusing on algorithms used for classification tasks that generate explanations by similar examples.

First, there is a need to define interpretability in the context of machine learning. Interpretability has many definitions. Doshi-Velez and Kim define it as "the ability to explain or present in understandable terms to a human" [28], while Miller defines interpretability as the "degree to which an observer can understand the cause of a decision" [71]. All definitions circle around the same question: why does a model make a specific decision? In this context, we can distinguish two terms, both referring to the concept of interpretability: interpretation and explanation. We may define interpretation as the capacity to understand a model's behaviour, which aligns with Kim's statement that "a method is interpretable if a user can correctly and efficiently predict the method's results" [52]. An explanation is the means used to interpret a model or its decisions, which will be provided to the users to make them understand the model or its decisions.

There are many taxonomies used to group interpretability strategies. The one used in this work considers intrinsic methods, where we build inherently interpretable models, and *post hoc* methods, where we generate explanations after the model has been built [18].

Intrinsic interpretability can be achieved through two means: the design of entirely interpretable models or the addition of constraints or simplifications to the models that restrict their

behaviour, making them more interpretable. An example of a family of models that are interpretable by design is the rule-based models, which learn a set of rules that are used to make predictions, like the famous decision trees. A decision tree model takes the form of a tree where each node represents a specific feature in the data, and the respective edges are rules that apply to the features. It is intrinsically interpretable since a user can replicate the model's behaviour when making a decision simply by following its rules. Regarding the addition of constraints to a model to make it more interpretable, one possible approach often applied to neural networks is monotonicity, which consists of adding constraints to features known to have monotonic relationships [40]. Models can also be simplified through regularisation techniques like L1 Regularisation [102, 101]. Regularisation and monotonicity do not generate tangible explanations. These techniques are considered intrinsic interpretability methods since the restrictions they impose on models make their behaviour simpler and, therefore, more interpretable.

Post hoc strategies require the development of an explanatory model separate from the classification model used to classify the samples. For this reason, it is argued that these strategies do not reflect the real reasoning behind the models' decisions [63]. These methods can be applied to three different contexts: model explanation, to explain the behaviour of the model, outcome explanation, to explain the reasoning behind one decision, and model inspection, to inspect the model and understand some of its properties [38].

A model explanation is achieved by developing a simpler interpretable model based on the original one, aiming to mimic its behaviour [38]. The main problem in this approach is that it might be difficult, or even impossible, to define a simpler model that behaves like the original one. It is also hard to verify if the simpler model is a good representation of the original [101].

Outcome explanation intends to explain the reasoning behind a single decision, without the need to understand the model's general behaviour [38]. One possible approach is through gradient-based methods [104, 106, 99, 122, 107], which identify the data features that contribute the most to the final decision [102]. An example of such method is the Grad-CAM algorithm [99], which assigns different degrees of importance to different units in the network based on the respective gradients and feature maps, or saliency maps, which highlight the features that contributed the most to the decision made by the model [38]. Another approach introduced in the literature is deconvolution [121], which consists of developing a deconvolution network composed by deconvolution and unpooling operations, representing the inverse of the original convolutional neural network used for classification. By applying the deconvolution to the features obtained in the original network, we can reconstruct and visualise the features that the network has learnt.

Model inspection intends to inspect the model to understand one of its properties or decisions through a visual or textual representation [38]. This can be achieved through sensitivity analysis, by perturbing a model's input and analysing how the output changes [102, 101]. One example of what could be done to perturb an input based on images could be the respective partial occlusion [30, 121, 89]. The Grad-CAM algorithm mentioned above can also be considered a model inspection algorithm since it presents a visual representation of the features that contribute the most to a decision.

In this dissertation, we will explore case-based interpretabilty, which produces visual explanations by example, that can be compared with the observation under analysis. The following sections detail case-based interpretability methods.

3.1 Case-based Interpretability

Case-based interpretability models enable the retrieval of cases from the training data as explanations for models decisions. These explanations are very intuitive and easy to understand, as they approximate human reasoning by analogy. The types of explanations that case-based interpretability methods can retrieve are:

- **Similar / Factual examples**: examples from the training data that are the most similar to the current case and that share the same class as the current case.
- **Typical examples**: prototypes from the training data that best represent the case under analysis. These examples are usually obtained in models that perform prototypical learning, which define clusters represented by prototypes. The advantage of typical examples in comparison with similar examples is that typical examples for different predictions are typically more distinct [48].
- **Counterfactual examples**: examples that are the most similar to the case under analysis but that belong to a different class. Counterfactual explanations aim to explain what changes should be made to a sample so that the machine learning algorithm outputs a different prediction. These examples do not necessarily have to be a sample from the training data, as they can be generated based on the sample under analysis. Together with factual explanations, counterfactual explanations help to understand the boundaries between two different classes in a classification task.
- **Semi-factual examples**: examples that belong to the same class as the original sample but that are closer to the decision boundary. The goal of these explanations is to indicate changes that, even if they were made to the original sample, they still wouldn't change its prediction. The generation of semi-factual explanations involves making the biggest possible alteration to a sample without changing its prediction [51].

In the context of case-based interpretability, intrinsic approaches involve the design of inherently interpretable models with case-based or prototypical reasoning. *Post hoc* methods use the original decision model as a similarity metric to compare the new observation with the training data and retrieve the training sample with the most explanatory value [17]. Counterfactual and semi-factual explanations can also be generated from an example retrieved from the case-based interpretability method, independently from the training data. We consider the generation of these types of explanations to be a *post hoc* interpretability method, as this process uses the original decision model after it has been built to guide the generation process.

The most significant problem in case-based explanations in the medical scene is that the retrieved examples expose the patients' identity, limiting their usability in a real scenario. For instance, these explanations could not be shown to patients or other people who do not have authorised access to the training data.

The following sections introduce state-of-the-art case-based interpretability methods in traditional machine learning and in deep learning.

3.1.1 Case-based Interpretability in Traditional Machine Learning

In traditional machine learning methods for case-based interpretability, the feature extraction process is separated from the decision process. As such, these methods are difficult to use with images as the extraction of features from multi-dimensional data is not trivial. Nevertheless, we can apply these methods on features that were previously extracted using deep learning methods.

3.1.1.1 Intrinsic Interpretability Methods

In traditional machine learning, the most well-known model that allows using similar examples as explanations for decisions is the K-Nearest Neighbours (KNN) model [31]. This classification model classifies an observation according to the K training samples that are the closest to it. During training, the model memorises the position of all the training samples in space. To classify a new observation, the model calculates the distance from the observation to all training samples to obtain the closest neighbours. Once these are identified, the observation is classified with the majority of the neighbours' labels. To provide explanations by similar examples using this model, we can fetch the nearest neighbours used to classify an observation. In this case, neighbours of the same class as the decision can be retrieved as similar examples, while neighbours from a different class can be retrieved as counterexamples. Figure 3.1 exemplifies this retrieval process, where the red sample in the centre is the observation that must be classified as either "x" or "o" according to its nearest neighbours.

One other method is the Bayesian case model [53], an intrinsically interpretable model that organises the data in clusters. The clusters are represented by a prototype, the training data sample that is the best representation of the cluster's data, and a subspace of the features needed to characterise each cluster. For classification, a new observation is mapped into a cluster and classified according to its prototype, which is then used to explain the model's decision. The explanations provided by this method are typical examples.

3.1.1.2 *Post hoc* Interpretability Methods

In terms of *post hoc* techniques to generate explanations by similar examples in traditional machine learning, these can be applied to models such as decision trees, which are used as distance metrics for the retrieval of training samples that are similar to an observation, as suggested by Caruana *et al.* [17]. To use models like decision trees for case-based explanations, the models need to memorise the training samples instead of discarding them once the model has been built so that

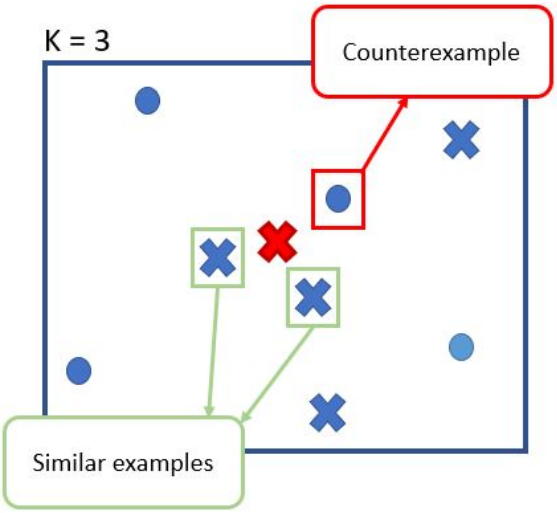


Figure 3.1: Example of the process of extracting explanations from KNN with K=3.

the explanatory model can retrieve them. In a decision tree, similar data samples are samples that conform to the same rules in the tree, which end up gathering in the same leaves. In this case, as can be seen in Figure 3.2, we could retrieve two training samples in the same leaf of the new observation in the tree to use as explanations. A sample with the same class as the new observation can be provided as a similar example, and one with a different class can be provided as a counterexample.

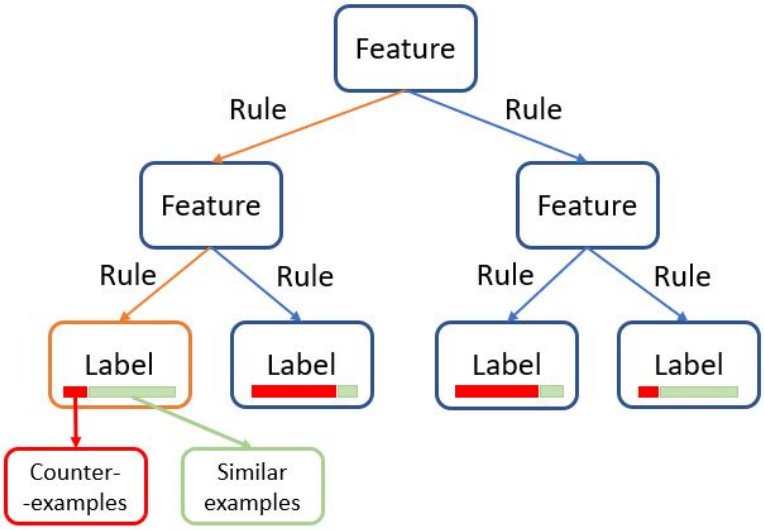


Figure 3.2: Example of the process of extracting explanations from decision tree model.

One framework that was proposed to obtain *post hoc* case-based explanations is Explanation Oriented Retrieval (EOR) [82]. The authors, Nugent *et al.*, argue that explanatory cases should

be retrieved based on their utility as explanations. The EOR method tries to obtain convincing explanations rather than the simple similar example. The approach starts by obtaining the nearest neighbours of a sample, and performing classification according to the KNN algorithm. Then, it selects an explanation utility measure based on the classification task, and uses it to reorder samples of the same class. The explanations obtained with this method are semi-factuals, since they are cases that are closer to the decision boundary and thus are more convincing than the most similar case.

3.1.2 Case-based Interpretability in Deep Learning

Deep learning methods present the advantage of an automatic feature extraction process optimised according to the target classification task. Furthermore, deep learning opens new possibilities in the retrieval of case-based explanations with the generation of explanations using generative models. As such, in this section, we will not only expose intrinsic and *post hoc* methods, but also methods that can be used to generate counterfactual and semi-factual explanations.

3.1.2.1 Intrinsic Interpretability Methods

The Explainable Deep Neural Network (xDNN) network [8] is a feedforward neural network that defines prototypes as local peaks in the data density and classifies the observation according to the most similar prototype. During training, its network is organised in 5 layers, which accomplish the following tasks: the feature layer is responsible for the extraction of features, which are then represented in a latent space that is fed to the density layer, used to calculate the proximity between images, and to the "*typicality*" layer which calculates a probability distribution of the data in the latent space. Then, the network uses a prototype layer to select the prototypes representative of each class, forming data clouds, which define each prototype's area of influence. Finally, the "*MegaClouds*" layer joins prototypes that are neighbours and belong to the same class. To classify the new observation, the network calculates the similarity between the observation and each prototype and calculates the most similar prototypes of each class, assigning the most similar prototype's class to the observation. This last prototype, used to classify the new observation, is given as a typical example to explain the prediction.

More recently, the authors of the xDNN network introduced the network Deep Machine Reasoning (DMR) [7], which improves the xDNN for more complex multi-class problems where the data is not balanced. This network synthesises data by creating linear interpolations between perturbed data samples around prototypes, in order to balance classes. This process of data synthesis appears in the augmented prototypes layers, which succeeds the prototype layer in the original xDNN network. Additionally, the network uses a decision tree which performs pairwise comparison between the top two classes to determine the class label that is assigned to the sample under analysis. During inference, the prototypes are compared in regards to minimum error in training. Similarly to the previous network, the prototype of the winning pair whose class was chosen in

the inference process can be used as a typical example, while the prototype whose class was not chosen can be used as a counterexample.

The intrinsically interpretable method developed by Li *et al.* [63] generates prototypes that best represent the training data. As can be seen in Figure 3.3, the model is composed by an autoencoder followed by the prototype classifier. The encoder extracts features into a latent space, which can be used to calculate the distance between instances. The decoder learns to map latent vectors into images in the original data space, by being trained to reconstruct the training data instances in the latent space. The prototype classifier has an initial layer which learns the prototypes, represented as the Prototype Layer in Figure 3.3. There may be a higher number of prototypes than the number of classes in the classification task. The model is trained to approximate the prototypes to training data instances. Finally, the prototype classifier performs the classification task. To obtain explanations by typical examples, we can retrieve the prototypes that are the closest to the instance under analysis in the model's latent space. Since the prototypes are generated instances learnt in the latent space, they can be visualised as images in the original data space through the decoder.

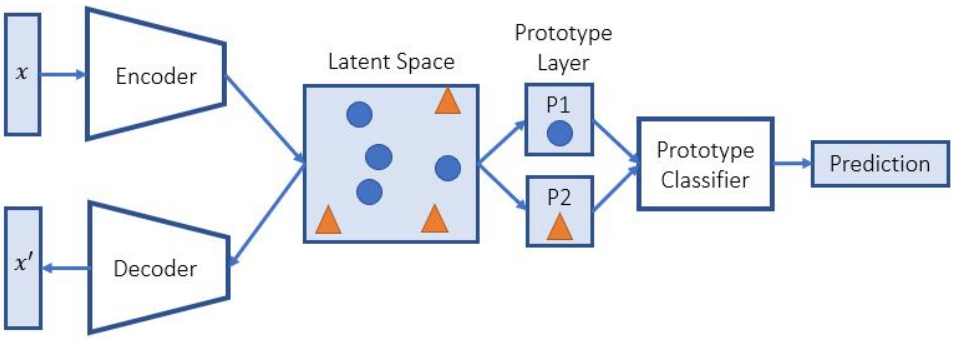


Figure 3.3: Overview of Li *et al.* [63] model.

The Deep k-Nearest Neighbours network (DkNN) [88] is an intrinsically interpretable classification model that aims to achieve high confidence in its predictions, interpretability, and robustness against adversarial attacks. In each layer, this network calculates the new observation's representation, searches for the k nearest neighbours in the training dataset whose output in the layer is the most similar to the observation's output, and collects the respective labels. In the end, all the labels collected from the nearest neighbours throughout the layers of the network are used to compute the final label assigned to the new observation, ensuring that the prediction conforms with the intermediate computations performed in the hidden layers. Interpretability is achieved using the training samples that support the prediction as explanations by similar examples for the network's decision.

The Prototypical Part Network (ProtoPNet) [19] is also a prototype-based network where prototypes are parts of the training data images. When making a prediction for a new observation, the network finds prototypes similar to parts of the new image, computing the respective similarity

scores and combining them to make the decision. The network contains a CNN for feature extraction, connected to a prototype layer, to compare the new observation to each class's prototypes, and finally, a fully connected layer to classify it. In this work, multiple types of explanations are provided, such as activation maps, which show the parts of the image that were used to achieve the prediction, and similar examples, which correspond to images from the training data that contain prototypes that contributed to the classification of the new observation.

The Hierarchical Prototype (HP) classifier [37] is a prototype-based classifier which defines prototypes in a hierarchy where higher levels contain more abstract prototypes, and lower levels contain detailed prototypes that are more similar to the training data samples. The prototypes represent local peaks in multimodal distributions derived from the data. In this approach, a new observation is classified by choosing a layer and calculating the most similar prototype to the observation, assigning its class to it. This prototype, which is used to classify the observation, can be used as an explanation of the decision process.

3.1.2.2 *Post hoc* Interpretability Methods

In terms of *post hoc* techniques, non-case-based learning models are used as distance metrics to verify similarity between data samples and choose the samples that are the most similar to the new observation, to provide as explanations, as mentioned by Caruana *et al.* [17].

Silva *et al.* [102, 101] created a classification model that applies the restriction of monotonicity as an intrinsic method of interpretability and generates explanations through *post hoc* methods. Regarding monotonicity, the model first applies a monotonic network to features that are known to be monotonic and a non-monotonic network to the remaining features. Both these networks are then concatenated into a monotonic network which creates a latent space where all the features are monotonic. The latent space is then used to calculate the distance between a new observation and the training data points and retrieve the closest data samples. The closest training sample of the same class as the one predicted for the new observation is used as a similar case, while the closest sample of the opposite class is used as a counterexample. In this approach, the prediction is not made using the closest samples of the new observation, but these are retrieved to explain the model's decisions.

One method that has been recently introduced by Chen *et al.* [22] for intrinsic interpretability is Concept Whitening. This method adds a concept whitening module to a classification network to organise its latent space according to high-level concepts. These concepts can be the original labels of the classification problem or any other concept introduced in an auxiliary dataset. For instance, in medical data, these concepts can be symptoms that identify the potential existence of a disease or even patients' characteristics. Using this method, we can obtain case-based explanations through a *post hoc* method where we use the latent space organised by the concept whitening module to measure the distance between data points. To obtain similar examples, we can identify the samples that are the closest to the observation under analysis. The proximity between samples can be analysed regarding all the latent space's dimensions or regarding only a specific set of concepts usually correlated with the observation's class.

One *post hoc* interpretability approach that was originally developed for medical image retrieval is Interpretability-guided Content-Based Image Retrieval (IG-CBIR) [103]. The method consists of using interpretability saliency maps to obtain the image regions that are the most relevant for a classification task. These regions are used to measure the semantic similarity between images, allowing to retrieve semantically similar examples as explanatory cases. This work concluded that the use of saliency maps obtained with methods such as Deep Taylor [74] produces improved similarity measures, increasing the quality of medical image retrieval.

Finally, there are also *post hoc* methods that develop interpretable surrogate models based on the original "black-box" classification model. Liu and Arik [65] proposed the approach illustrated in Figure 3.4, based on unsupervised clustering. The method uses layer activations from the base classification model, which are encoded into a latent representation using an encoder, and clustered based on euclidean distance. Each cluster is represented by a centroid. During inference, the model calculates the probability of the sample belonging to each cluster for each layer, and assigns to the sample the results of the weighted average of the predictions made at each layer. The explanations provided are similar training samples that are the most similar to the test sample in the model's latent space.

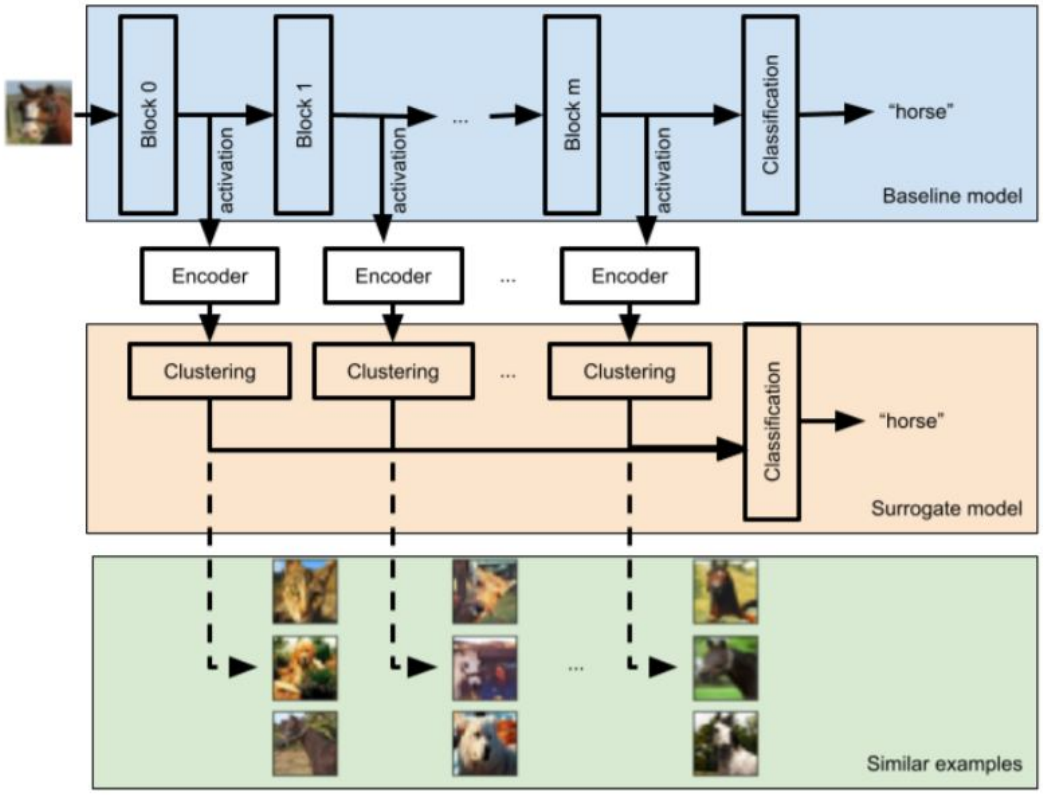


Figure 3.4: Illustration of the *post hoc* interpretability method based on unsupervised clustering proposed by Liu and Arik. Source: [65].

The Twin Systems framework introduced by Kenny and Keane [50, 49], also uses a surrogate interpretable model to explain the original one, but instead of unsupervised clustering, it uses KNN. The method extracts features from a base "black-box" model and applies the interpretable model KNN over these features to obtain explanations for the base model's decisions. In their work, the authors listed various feature extraction methods, such as sensitivity and perturbation-based methods, and introduced a new method: Contributions Oriented Local Explanations (COLE). COLE uses methods that generate saliency maps, such as DeepLift [100], to extract feature contributions to the decision-making process. This method obtains explanations by similar examples, as the retrieved samples are the nearest neighbours according to the chosen feature extraction approach.

3.1.2.3 Generation of Counterfactuals and Semi-factuals

PlausIble Exceptionality-based Contrastive Explanations (PIECE) [51] is a method for generating counterfactual and semi-factual explanations for a model's decisions. The method focus on generating plausible explanations, where the generated images are sufficiently different from the original ones so that the explanation consumer can understand the explanation. In counterfactual explanations, the method starts by identifying the target class (*i.e.* the class to which the counterfactual will belong to). When the test case's prediction is wrong, the target class is the test case's actual label. Otherwise, the target class is found by maximising the Equation 3.1, where $S(G(z))$ is the prediction of the classification network in regards to classifying the counterfactual and Y_c is the prediction of the test case. In the generative process, the method identifies features in the original image whose probability of occurring in the target class is low and uses a GAN to modify them into the expected values for the target class. For semi-factuals, the process is the same but the feature modifications are stopped before the model's prediction changes from the original class to the counterfactual class.

$$||S(G(z)) - Y_c||_2^2 \tag{3.1}$$

3.2 How to choose a model for explanations by example?

To develop a classification model capable of generating explanations by example, a developer must consider the type of model that should be implemented to achieve the data mining goals set for the respective project. The best model depends on the type of data that is being used and its domain, as well as the type of explanations to be generated. The domain on which the model will be applied should be carefully studied to define priorities regarding the model's confidence, interpretability, robustness, and other characteristics. For instance, in a medical scenario, the model's confidence is critical since a wrong prediction could have serious consequences, such as a patient who does not need treatment getting a treatment that could be detrimental to their health, or a patient who needs treatment not getting it. Interpretability in a medical scenario is also very

important since it increases the trust in the machine learning models used, which allows applying them in real situations. Considering the characteristics of the problem that we want to solve, we can choose between traditional machine learning models and deep learning models and between intrinsic methods and *post hoc* methods for interpretability.

Traditional machine learning models have the disadvantage that feature extraction must be done separately from the target task, making it hard to optimise the feature extraction process to fit the task, especially for image data. On the other hand, deep learning performs feature extraction automatically, optimising this process by considering the prediction task at hand. For image analysis, deep learning is the obvious choice.

An advantage of intrinsic methods for achieving interpretability is that the explanations generated by these are a direct consequence of the process used to make the decision. In contrast, in *post hoc* methods, the explanation generated may not reflect the real reasons behind models' decisions [63]. One reason why *post hoc* might be preferable is that, in many cases, "black-box" models achieve better performance (in terms of accuracy and other metrics of interest). When choosing between intrinsic and *post hoc* methods, the developer must consider whether to prioritise interpretability or the model's performance. The ideal scenario would be to develop an intrinsically interpretable model with high confidence in its predictions.

For medical data, the model's performance is critical since its decisions have significant consequences. Therefore, the development of a high confidence model with the posterior addition of interpretability through *post hoc* methods is a valid approach. However, interpretability is also critical to increase the medical community's trust in deep learning models and the respective application in real life. Since one of the major problems stopping deep learning models from being used is the lack of trust derived from lack of interpretability, the explanations generated must be trustworthy and a reflection of the model's reasoning, leading us to intrinsic methods of interpretability.

Chapter 4

Literature Review: Visual Privacy

Although providing explanations by showing examples of similar cases is a reasonable method to achieve interpretability in machine learning models, it raises several privacy concerns when the data is sensitive, such as biometric or medical data. Images that showcase sensitive data such as the patients' identity need to lose identity features to be used as examples to explain a model's decisions. This chapter will focus on visual privacy, which applies the concept of privacy to images and videos.

Given an image containing information about a person's identity, such as an eye image, our goal is to prevent this person from being recognised by the human eye or by identity recognition algorithms while preserving utility features in the image. For example, in images of the eye iris for glaucoma detection, a doctor should be able to identify the eyes' characteristics that discern glaucoma without identifying the patient to whom the image belongs. In this section, we discriminate between two groups of methods to preserve privacy: traditional and deep learning methods. In this work, we refer to traditional methods as methods that are applied over the whole input, requiring a preprocessing step to identify the image parts that must be privatised. In contrast, deep learning methods can learn to identify sensitive regions and privatise them in the same network.

4.1 Traditional privacy-preserving methods

Traditional methods contain two phases: identifying sensitive regions in images and modifying these regions. The modification of image regions can be grouped in three groups: image filtering, which applies filters to the sensitive parts in order to hide them; image de-identification, which alters properties in the images to conceal a person's identity while keeping the images' intelligibility; object removal or replacement, that removes objects from the image, filling the gap with the background of the image or with predefined visual models [85]. One other type of methods often used in visual privacy to protect biometric data templates is cryptography, which consists of

encrypting the sensitive parts of images with a key so that only entities that possess the key can decrypt and obtain the original data. However, these methods are not appropriate for the preservation of semantic features as the images become unintelligible.

The most common image filtering methods seen in the literature are blurring and pixelating. Blurring consists of applying a mask, often with a gaussian distribution, over the sensitive parts of an image, which alters each pixel according to the neighbouring pixels. Pixelating consists of dividing an image into a grid, associating the image's pixels to the grid's cells, and computing for each cell the average of the pixels it contains. Frome *et al.* [32] apply image filters to protect sensitive information such as faces in Google Street View images. In specific, they detect faces in images using a sliding window approach and apply gaussian blur to the detected faces to achieve anonymisation. The biggest problem with these methods is that, while they may preserve privacy with high degrees of blurring or pixelating, the image's utility is also lost. This issue was demonstrated by Neustaedter *et al.* [78] who studied the privacy-utility trade-off in blurring approaches by blurring images at different degrees and studying whether a group of people could identify the people in the images and the actions they were doing.

For image de-identification, one well-known method is the K-Same algorithm [79], originally applied to face images. This algorithm is applied over a set of normalised pictures where the different parts of the face, such as eyes, nose, and mouth, are roughly in the same locations. After the preprocessing stage of recognising, cropping, and normalising the faces in the images, the dataset is organised into k-sized clusters, where the respective samples are replaced by the image resulting from the average of all the samples in the cluster. This algorithm guarantees k-anonymity, as the probability of a person being recognised in the image is $\frac{1}{k}$. The K-Same method forms clusters by calculating the proximity between data samples using a distance measure such as the Euclidean distance between pixels, disregarding whether images belong to the same class. As such, this method does not guarantee the resulting images' utility. To fix this problem, Gross *et al.* [36] proposed the K-Same-Select algorithm. This algorithm divides the dataset into different subsets using a utility function, creating subsets that are similar regarding the images' utility. Then, it applies the K-Same algorithm to each subset. One example of a face privatised using this algorithm can be seen in Figure 4.1. This figure clearly shows a privacy-intelligibility trade-off, with higher values of k resulting in blurred images. One problem with both these methods is that when the dataset possesses more than one image of the same person, the respective images might end up in the same cluster. As a result, this person contributes more to the cluster's averaged image than other people in the cluster, increasing the probability of this person being recognised.

Other types of face de-identification methods include face swapping, as proposed by Bitouk *et al.* [11], that estimates the pose of the face in the image and replaces it with a face image from a public library. This method could be extended to work with images of other body parts, as long as a model of the body part is available, to estimate the respective pose. However, this method does not preserve the semantic features needed to fulfil an artificial intelligence task, as the relevant features may be located in close proximity or even entangled with the identity-related features which are replaced by the model.

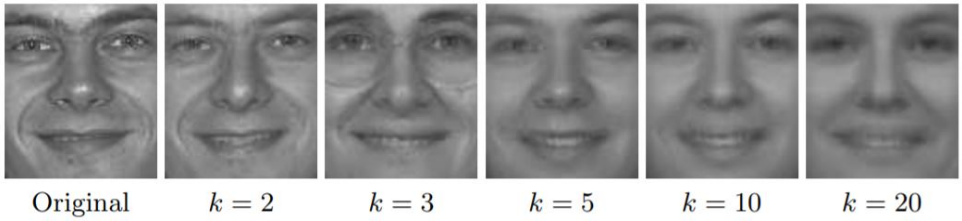


Figure 4.1: Examples obtained from K-Same-Select algorithm. Source: [36]

Finally, object removal/replacement methods remove an object from an image, filling the gap left with either the background of the image, through an inpainting strategy that reconstructs the missing parts based on information around the missing area, or with a model contained in a database. Like in face-swapping methods, object removal/replacement discards relevant task-related features.

4.2 Deep learning privacy-preserving methods

In Deep Learning, there are two general approaches that use identity recognition networks to guide the image privatisation process. The first approach consists of performing Disentangled Representation Learning to explicitly disentangle identity features from features that are independent of identity, obtaining an identity feature vector that can be altered to privatise the image [23, 33]. The second approach consists of using the identity recognition network to train a generator model to generate images that do not possess the same identity as the original one [20, 84, 119].

Most works in the current literature about privacy preservation are applied to face images since the face is the most recognisable part of the human body.

In regards to the preservation of task-related features in the privatised images, we can divide existing deep learning privacy-preserving methods into the following groups:

- **Task-independent methods**: focus on removing identity features while preserving every other feature, independently of a classification task.
- **Task-dependent methods**: remove identity features while explicitly preserving the attributes needed to fulfil a task, using the model developed to achieve the task to guide the semantic feature preservation process.

The following sections introduce the existing deep learning privacy-preserving methods, reflecting on their strengths and weaknesses when considering their application to case-based explanations.

4.2.1 Task-independent Methods

Task-independent methods remove identity features independently from a data mining task while preserving identity-independent features that might be useful for any task. Since these methods

do not consider a particular task in the preservation of an image's utility, the resulting privatised image may lose relevant task-related features that may be related to identity-related features.

CLEANIR [23], proposed by Cho *et al.*, is a network that removes identity features independently of any other feature existing in the image. Its architecture is based on a VAE, where the encoder is trained to disentangle identity features from a face image. During training, to enable the extraction of identity features, the network uses a pre-trained face embeddings extractor, based on the FaceNet network [98], that extracts the embeddings in the face images. This network is used to calculate an embedding loss, characterised by the distance between the identity vector obtained by the encoder and the identity vector obtained using FaceNet. The embedding loss is backpropagated to the encoder, allowing it to learn to represent identity vectors in its latent space. During the testing phase, the identity features vector z^i, extracted by the encoder, is converted into a modified identity vector z^m, whose identity is no longer recognisable. The Equation 4.1 is used to calculate z^m, where z^{90} corresponds to the 90-degree rotation of z^i, obtained through the Gram Schmidt process, and m corresponds to the degrees of identity modification. As such, the distortion of identity features is both controllable and independent of other training samples existing in the latent space. The architecture for the training and testing processes can be seen in Figure 4.2.

$$z^m = z^i cos(m) + z^{90} sin(m) \tag{4.1}$$

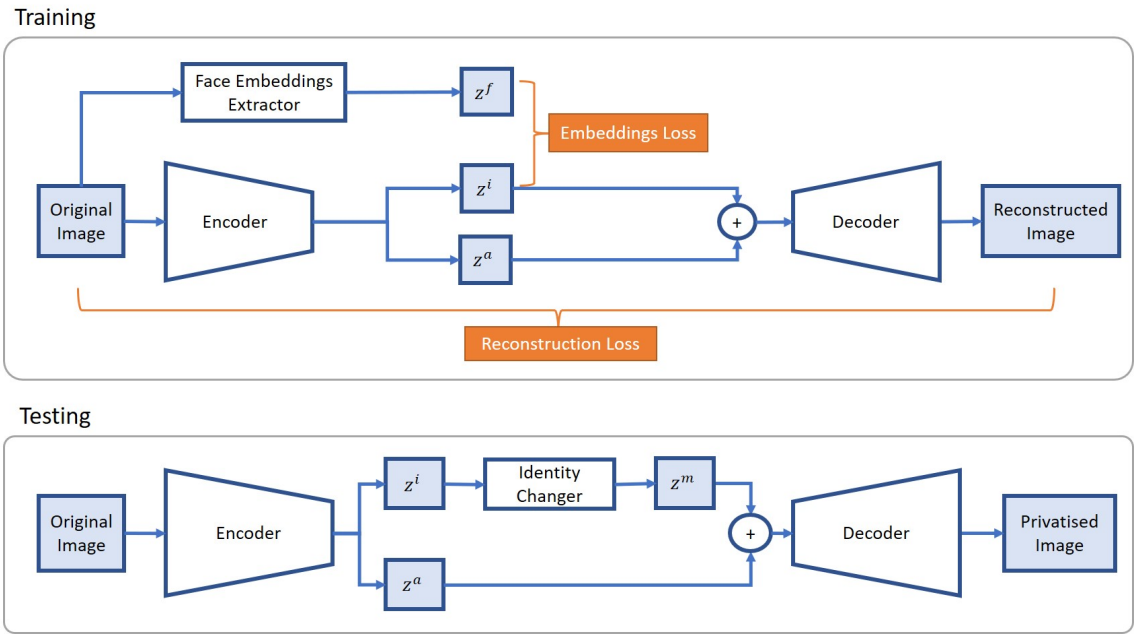


Figure 4.2: Architecture of CLEANIR network during training and testing.

In this model, the use of a variational autoencoder leads to blurry output images, as can be seen in Figure 4.3. Nonetheless, the images are intelligible. One drawback of the CLEANIR network is

that it does not guarantee that the vector z^a does not contain identity features. For instance, if the facial embeddings capture most or all the information needed to reconstruct the original image, then z^a could be correlated or even equal to the identity-related latent vector, leading to a potential identity leak in the privatised image. The main advantage is that identity removal in an image is independent of any other images from the training data. To apply this method in a medical context, we need to develop a model to extract embeddings from the respective medical images, as FaceNet does with face images.

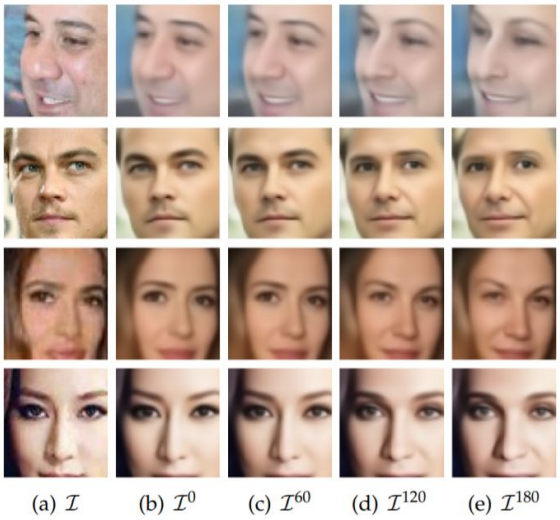


Figure 4.3: Example of privatised images obtained with CLEANIR. (a) corresponds to the original image and (b-e) are the privatised images obtained from applying transformations to identity features with 0, 60, 120 and 180 degrees, respectively. Source: [23]

One other task-independent method for generating privacy-preserving images through disentangled representation learning is called Replacing and Restoring Variational Autoencoders (R^2VAE) [33]. It is composed of three steps:

1. **Disentangled Representation Learning**: obtains the identity-related and identity-independent features. The model contains a VAE with two encoders, where Encoder 1 extracts identity-related features and Encoder 2 extracts the remaining image features, and a decoder that maps the latent representations of the encoders into one image in the original data space. During training, this generative model is organised in a chained architecture, R^2VAE, where the VAE is used twice in a row. First, the VAE produces an image P with identity features from one input image I and utility features from a second input image U. Then, it reconstructs the original image U using the utility features of the previously obtained image P with the identity-related features from U, to promote the preservation of relevant utility features. In the training process, there is also a discriminator to promote realism in the generated image P and an identity recognition network to ensure that P shares the same identity as image I.

2. **Identity Obfuscation**: consists of obtaining a vector of identity features that will be used to replace the identity features of the original image. The authors tested two methodologies: k-Random-Based Obfuscation and Identity-Prototype-Clustering-Based Obfuscation. K-Random-Based Obfuscation averages identity-related features obtained from applying Encoder 1 to k random samples from the training data, resulting in privatised images such as the ones in Figure 4.4. Identity-Prototype-Clustering-Based Obfuscation consists of creating clusters of the training samples organised by identity categories, obtained by passing the sample's identity features to a classifier. The identity features used to obfuscate the image's identity are then obtained by averaging the identity feature vectors of the samples in the cluster of the same identity category as the original image. The resulting vector of identity-related features is passed to a decoder, together with the identity-independent features extracted from the original image using Encoder 2, in order to obtain the de-identified image. As both methods depend on the training data, if a person is more represented in the database than others, this person may be more represented in the resulting images, which might threaten the person's privacy.

3. **Image Inpainting**: consists of filling the sensitive region of the image with the de-identified image obtained in the previous steps. The goal is to improve the quality of the resulting image while preserving the identity-independent features, alleviating pixel blur or colour discrepancy that might result from replacing the original sensitive part of the image with the de-identified one.

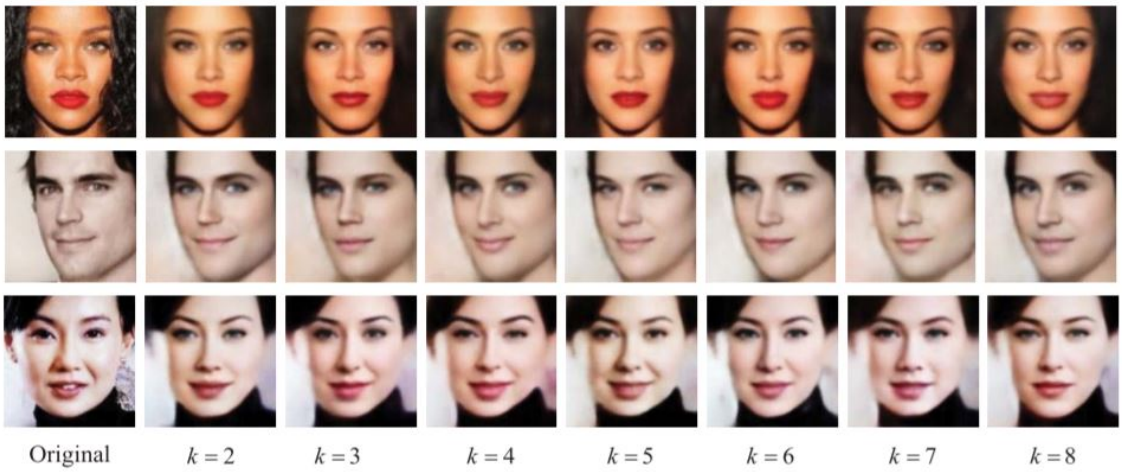


Figure 4.4: Example of privatised images obtained with the R^2VAE network. Source: [33]

One drawback in this approach is that the part of the image that contains identity features needs to be extracted before being fed to the described network, needing an additional preprocessing step focused on image recognition. Also, the identity obfuscation algorithm uses other samples from the training data to produce the synthetic image, restricting the privacy capabilities of the network.

Privacy-Protective GAN (PP-GAN) [119] is a privatisation method also applied to face images, that preserves general features while removing identity. It is composed of a GAN whose generator assumes a UNET architecture [96], following the idea of GANs used for image-to-image translation like the Pix2Pix GAN [46]. UNET possesses an architecture similar to an autoencoder, with an encoder and a decoder, but with skip connections between the encoder and the decoder. The discriminator is a Patch GAN that discriminates whether a patch of the image is real. Since there is no ground truth to train the images, the L2 Normalisation loss, usually used in Pix2Pix GAN, was replaced by two losses: the structural similarity index measure (SSIM) [116] and a contrastive loss. The SSIM loss guarantees that the privatised image is similar in terms of structure as the original image, to preserve the image's utility. The contrastive loss is used to train a siamese identity recognition network, by increasing the euclidean distance between the latent representations of two images with different identities, and decreasing this distance between images with the same identity. The contrastive loss guarantees the privacy of the subject in the original image. Some results from this work are illustrated in Figure 4.5.

Figure 4.5: Example of results obtained with the PP-GAN network. The top row represents the original images while the bottom row illustrates the respective privatised images. Source: [119]

Unlike the previous works, which focus on de-identification methods applied to face images, Siamese Generative Adversarial Privatiser (SGAP) [84] was applied to preserving privacy in biometric data. This task-independent model uses adversarial samples inspired by GANs to anonymise biometric data while maintaining the information's utility. The network is composed by a generator, which contains an autoencoder that tries to hide a person's identity, and a discriminator with a Siamese architecture that predicts the person's identity, as can be seen in Figure 4.6. The SGAP model identifies the parts of an image that hold higher discriminative power and perturbs them to privatise them by adding noise to the latent space created by the encoder in the generator. It also uses a distortion constraint, incorporated in its loss function, that ensures that the distance between the privatised image and the original image is lower than a constant to ensure that the synthesised image does not differ much from the original image. As the privatisation is done through the addition of noise to the image, it is independent of other samples in the training data, which is advantageous compared with algorithms that use the training data in the identity

obfuscation process. This model has been tested in images of fingerprints.

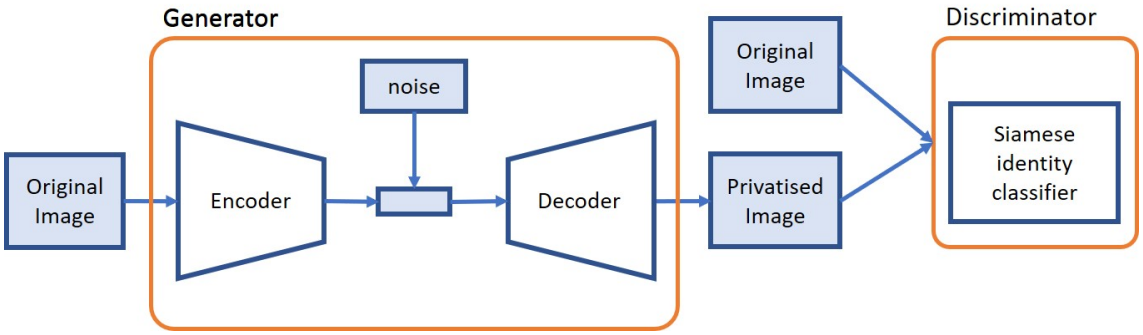


Figure 4.6: Architecture of SGAP network.

Most datasets used by the networks in this section are publicly available datasets of face images from various identities and that vary in characteristics like pose, illumination, age, ethnicity, among others. Some examples of datasets used were VGGFace2 [16], CelebA [66], and MORPH [95]. Each exposed method used a different dataset.

The limitation of identity remover algorithms independent of a data mining task is that, while these try to preserve the data utility, they might discard semantic features specific to the task. These methods assume that a task's semantic features and identity features are disjoint, which may not always be the case. As such, they always prioritise privacy in the privacy-utility trade-off, reducing the data utility.

4.2.2 Task-dependent Methods

Task-dependent methods tackle the de-identification problem directed towards specific classification tasks, using a classification network to ensure that semantic features are preserved.

Chen *et al.* [20] developed a model for privacy protection applied to the task of facial expression recognition: Privacy-Preserving Representation Learning Variational Generative Adversarial Network (PPRL-VGAN). The model replaces the identity in the original image with another identity from the training data. Its architecture, represented in Figure 4.7, consists of a GAN that contains a VAE as the generator, capable of learning identity-invariant representations. The generator competes with a multi-task discriminator composed of three different classifiers: a real/fake classifier to distinguish between real and synthetic images, an identity classifier to identify the person in the image, and an expression classifier, the task-dependent model used to identify the facial expression. The target identity c, which is used to replace the original identity, is given to the decoder, allowing it to learn to build an image with the given identity. The problem with identity replacement for privacy-preserving purposes is that it protects the person's identity in the original image by exposing the identity of the person who is used as a replacement. This method could only be genuinely a privacy-preserving method if it used predefined synthetic models in the database instead of images from real people to replace a person's identity.

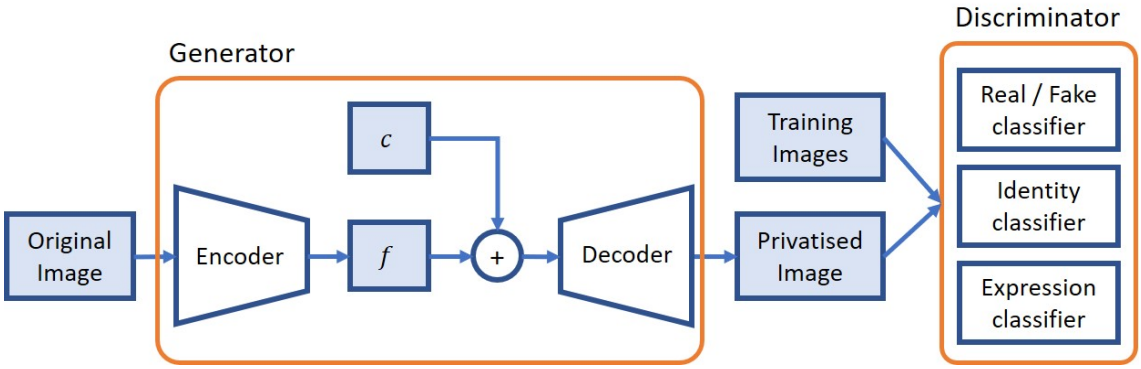


Figure 4.7: Architecture of PPRL-VGAN.

The PPRL-VGAN was trained on datasets for facial expression recognition, containing various different identities and different facial expressions. The datasets used were the FERG dataset [6], composed of animated characters, and the MUG dataset [3], representing real people.

The DeepObfuscator [62], proposed by Li *et al.*, is an adversarial training framework that obfuscates the identity in an image while preserving the semantic features needed by a classification task. This network, as can be seen in Figure 4.8, is composed of the following modules:

- **Obfuscator**: is an encoder composed of a CNN which is trained to hide identity-related features while preserving useful information for the intended classification task.
- **Classifier**: is the network that performs the task for which we want to preserve semantic features. By being trained jointly with the obfuscator, this network ensures that semantic features are retained in the privatised image.
- **Adversary Reconstructor**: is trained to reconstruct the original image based on the privatised image outputted by the obfuscator.
- **Adversary Classifier**: is trained to predict the private attributes based on the features outputted by the obfuscator to ensure that sensitive attributes are not leaked in the privatised image.

The adversarial training intends to maximise the reconstruction error in the adversary reconstructor and the classification error in the adversary classifier while minimising the classifier's classification error. Although this network preserves task-related features, it produces unrealistic images which may be difficult for humans to understand. Nonetheless, we expose this method as its architecture may be interesting for the generation of privacy-preserving explanations. For instance, it may be possible to add a discriminator to promote the generation of realistic privatised images in this model.

The biggest problem with task-dependent methods is that they solve problems for particular tasks, having a low generalisation capacity since they cannot be directly applied to other contexts

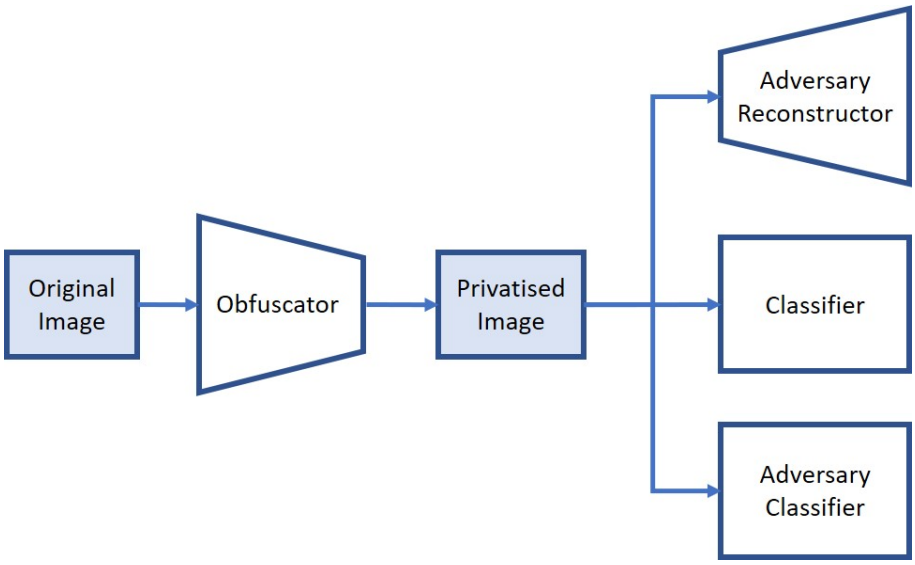


Figure 4.8: Architecture of DeepObfuscator.

or domains. Using these methods for tasks other than those these models were intended to implies replacing the task-specific model used in the network with another one and redo the whole training process.

4.3 How to ensure that visual privacy is protected?

There are two means to assess privacy in images:

- **Subjective means**: through inquiries or experiments where we have people looking at the privatised images and stating their opinions regarding if they can recognise the people contained in the images or not.

- **Objective means**: define quantifiable and objective measures to ensure privacy.

The limitation of subjective means to analyse visual privacy is that the assessment depends on the group of people used to analyse the synthetic images, which might not represent the group for which the images are intended. Generally, it is hard to gather a group of random people that can represent the population to which the privatised images are intended, and that can lead to a proper evaluation of privacy. As such, in this section, we will focus on objective means to ensure privacy.

To ensure that visual privacy is protected, we need a classification network capable of identifying the person to whom the image belongs. Two main types of networks can be used for this purpose: multiclass classification networks, where the labels refer to the identities of the people that can be detected in the image, or Siamese classification networks, which, given two images, identify whether these belong to the same person.

The works in Section 4.2 use an identity recognition network to evaluate privacy preservation and ensure privacy by backpropagating its loss. When using networks to measure the preservation of privacy, we need to consider that this measure is limited by the performance of the models used.

4.3.1 Multiclass classification networks for identity recognition

A multiclass classification network can be used to recognise the identity in an image by setting the labels as the person to whom the image belongs. Some works mentioned in Section 4.2 that use this type of classification networks are the PPRL-VGAN [20], which contains a person identifier in its discriminator, and the R^2VAE [33], which contains an identity classifier in its disentangled representation learning phase.

In this network, there can be two different approaches to check whether or not privacy is protected, which can be used simultaneously:

- **Wrong identity classification**: By checking if the identity outputted by the classification network differs from the true identity of the patient whose privatised image is being analysed.
- **Threshold**: By setting a threshold as the maximum value that the confidence of the privatised image being classified as its true identity can achieve. To set this threshold, we could average the prediction scores obtained for the not yet privatised image for all the identities except the real identity.

One problem with multiclass classification models is that they need to be trained with many samples belonging to each class, which means that each person needs to have multiple images in the training dataset. For medical and biometric data in a real context, there is a limit of images that can be obtained for each person, especially if these are obtained using x-rays, like in mammographies. Additionally, every time a new patient is introduced in the database, the network needs to be trained in order to be able to identify the new patient.

4.3.2 Siamese classification network for identity recognition

Siamese Networks were first introduced by Bromley *et al.* [15], for signature verification. A siamese classification network can be interpreted as a binary classification network that verifies whether two images belong to the same identity. As input, instead of receiving a sole image, it receives the pair of images that will be compared.

The architecture of a siamese network comprises two identical networks with shared weights, each responsible for extracting features from the respective input. By ensuring that both networks' weights are the same, similar images result in similar feature vectors [58]. Both networks' outputs are used to compute a score of similarity, that can be obtained by calculating the distance between the two feature vectors. For image recognition, the sub-networks can be built as convolutional neural networks [58]. The loss function for siamese networks must ensure that images of the same class have a high similarity score, while images from different classes have a low similarity score.

One function usually used in this context is the contrastive loss, initially proposed by Hadsel *et al.* [41]. The contrastive loss function minimises the euclidean distance between feature vectors from images of the same class while maximising this distance for images from different classes, as can be seen in Equation 4.2. In this equation, D represents the distance between the two feature vectors, Y is 0 if the images belong to the same class or 1 if the images belong to different classes, and $m > 0$ is a margin.

$$L = \frac{1}{2} \times (1-Y) \times D^2 + \frac{1}{2} \times Y \times \{max(0, m-D)\}^2 \tag{4.2}$$

Using this siamese network, we can input the original image and the privatised image and verify that these do not belong to the same identity to ensure that privacy is protected. This network's architecture is depicted in Figure 4.9. Two works from Section 4.2 that use a siamese classification network to guide the training process to produce privatised images are the SGAP [84] and the PP-GAN [119].

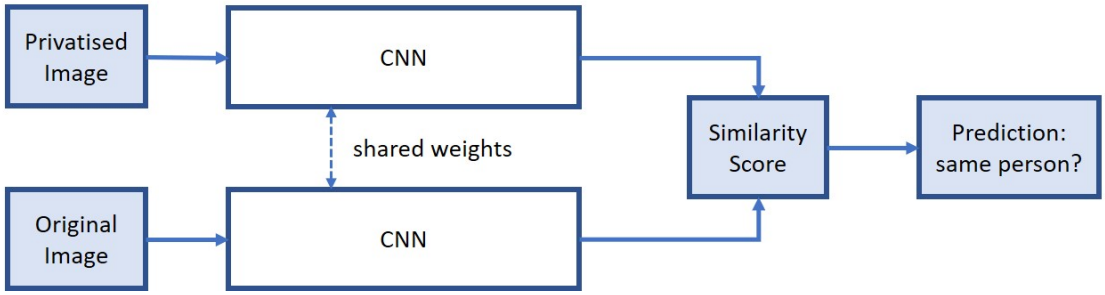


Figure 4.9: Architecture of Siamese Network for identity recognition.

One advantage of siamese networks in comparison with the multiclass classification networks seen in Section 4.3.1 is that, since they learn to compute the similarity between two images to verify if these belong to the same person, they do not need to be trained every time a new class, in this case, a new patient, is added to the database. Furthermore, this network can be used in data possessing a limited amount of images per subject, like medical data. On top of this, in large databases with thousands of patients, the output of a siamese network would be a binary output, unlike in multiclass classification networks where the output would be a vector with the number of predictions equivalent to the number of patients existing in the database [84].

4.4 How to select a method to preserve visual privacy in images?

In the previous sections, we have seen traditional methods, where the process to identify sensitive regions in images is distinct from the method to preserve privacy, and deep learning methods,

where the identification of the identity features that need to be privatised is made in the same network as the privatisation. Note that traditional methods can also be integrated with deep learning techniques. For instance, after finding the parts in an image that must be altered, we can apply traditional methods such as blurring. In this section, we will discuss how to choose a method for preserving visual privacy based on the developer's goals.

The first step towards defining a privacy-preserving method is defining the characteristics of the domain in which the final images will be used. For instance, in this dissertation, we want to use these images as explanations to justify the decisions made by a classification network in a medical context, where the subjects that will access the images are specialists such as doctors. As such, the privatised images must preserve the semantic features that the network uses to achieve the decision. Since the images will be shown to humans, they must be intelligible to humans, with some sense of realism. In medical images, the sensitive regions are often entangled with the semantic features that allow the patient's identification, making the process of preserving the utility of the images while discarding identity difficult.

When developing a method to preserve visual privacy in images, we need to make decisions regarding:

- Type of method to find sensitive regions in images.
- Type of method to make sensitive regions private.
- Method to assess privacy protection in the privatised images.

Regarding the method to find sensitive regions in images, deep learning privacy-preserving methods can identify identity features through the use of an identity recognition network to guide the feature disentanglement process. These methods are the most appropriate when semantic features and identity features are entangled or represented roughly in the same regions of the images. Generally speaking, deep learning methods achieve better results in privatising while preserving semantic features. The alternative methods to find sensitive regions in images find entire objects through object detection networks, which are more appropriate when the parts of the images we want to preserve are not related to sensitive regions. One example of such methods is the work done by Frome *et al.* [32], which hides faces from Google Street View images through blurring.

Regarding the method applied to privatise the image, some topics that guide the decision process are:

- **The expected quality of the privatised image**: if we want to have an intelligible privatised image, we can discard methods such as blurring or removal of the sensitive parts of the image. In general, deep learning methods, especially the ones that use GANs, have shown better results in the generation of high-quality images.
- **The use of samples from the training data in the method for privacy preservation**: using samples of the training data in the process of privatisation may put at risk the privacy

of the people to whom the training samples belong. We have seen methods such as K-Same-based methods, which average K different training samples to obtain K-anonymity, and methods that use identity replacement to protect one person's identity, sacrificing the other person whose image was used as a replacement. In comparison, methods that do not use training samples, which achieve privacy by applying transformations to the sensitive features or adding noise, are safer, guaranteeing full anonymity rather than K-anonymity.

- **The use of the final image on a specific task**: while most methods try to preserve attributes independent of the samples' identity, some of the methods explicitly use the model of the target task in the learning process to preserve semantic features. In contrast, others aim to preserve attributes independently of a specific task. The disadvantage of the task-independent methods is that they do not guarantee that the specific task's semantic features are preserved. The advantage is that these algorithms are more general and can be applied to various tasks. When choosing between these methods, we need to consider how important it is to ensure that semantic features are preserved.

In general, deep learning methods use an identity classification network to guide the process of privatisation. The use of these networks ensures privacy as they can also be used to evaluate the privacy-preserving capacities of the methods used. Identity classification networks can be of two types: multiclass classification models, which, given an input, predict the person's identity, and siamese classification networks, which predict whether two images given as input belong to the same person. Siamese networks possess the advantage of not having to be trained every time a new identity is added to the system and of being applicable to data with a low number of images per subject.

The preservation of visual privacy in medical images will allow us to use them as explanations to "black-box" models, increasing the trust in these algorithms and their acceptance in the medical community to improve medical diagnosis quality.

To conclude, we are currently preparing a paper [77] to submit to IEEE SPM Special Issue on Explainability in Data Science: Interpretability, Reproducibility, and Replicability. Upon submission of a white paper with this survey's proposal, we were invited to prepare the full paper for potential publication in this issue. This paper comprises a survey on case-based interpretability methods and privacy-preserving methods. We intend to reflect on the application of the privacy-preserving methods to privatise case-based explanations, exposing the main conclusions drawn from this chapter.

Chapter 5

Preliminary Experimental Work

As mentioned in Chapter 1, this dissertation aims to generate privacy-preserving images that keep their explanatory evidence, enabling their use as explanations of a deep learning classification algorithm. Given an input image in a medical setting, we want to obtain a synthetic version of it where the person in the image cannot be recognised. However, the explanatory evidence that can explain the model's results to medical specialists or other entities should be preserved. As such, to maximise the usability of the synthetic image and to make sure that it achieves its purposes, there are some prerequisites that it must fulfil:

- **Realism**: The image must be as realistic as possible to maximise its intelligibility and consequent understanding by the people who will be using the deep learning system. Furthermore, it should not be disturbing to the point of upsetting the consumers of the explanation.
- **Anonymity**: The anonymisation of the image should be independent of other patients' images to guarantee the privacy of all the patients available on the dataset used by the network. As we are dealing with very sensitive information, the privacy limits imposed by K-Anonymity may not be enough.
- **Explanatory Evidence**: As different patients with the same pathology might express different symptoms, the network must ensure the synthetic image's semantic features that serve as explanatory evidence are explicitly preserved. As such, a simple average of the most similar images of the same class for the sake of anonymisation is not enough. Besides, the explanation provided by the synthetic image must be a good representation of the reasons that led to the classification model's decisions.

Taking into consideration these requirements, the methodology that we adopted in the experimental work consists of performing various alterations to one of the visual privacy methods seen in Chapter 4, until we obtain a privacy-preserving model capable of satisfying these requirements. We decided to use PPRL-VGAN [20] as the base model since its implementation is publicly available [1] and it is the most promising privacy-preserving method. Unlike many other

privacy-preserving methods, PPRL-VGAN considers the preservation of task-related features. In the first phase, we tested this model with its original dataset and a medical dataset to empirically find its shortcomings and define the improvements that must be made to apply this model to the generation of privacy-preserving explanations.

In the following sections, we introduce the datasets used for the experiments and the respective preparation processes. We will also present a detailed overview of the PPRL-VGAN model's architecture and some preliminary results obtained from applying PPRL-VGAN to the datasets. Finally, we will gather the main conclusions from the preliminary experiments in a list of limitations of the PPRL-VGAN model that we intend to address in this dissertation.

5.1 Dataset Preparation

5.1.1 Original dataset: FERG-DB

The original dataset used by the authors of the PPRL-VGAN model is called FERG-DB [6]. As the original goal of PPRL-VGAN was to privatise images preserving facial expressions, this dataset is composed of 2D images belonging to 6 different characters and annotated for 7 different facial expressions. There are 55,767 images in the dataset, with roughly 9,000 images per character. The characteristics of this dataset significantly differ from the characteristics usually seen in medical data, as it contains many images per identity, obtained through video or fabricated with software to create 2D images, and a limited number of identities to guarantee computational efficiency. In medical data, we expect to see a higher number of identities corresponding to various patients. Furthermore, it is difficult to obtain multiple images in most medical cases, especially when images are obtained through x-rays, like mammographies. As such, medical datasets usually contain a minimal number of images per patient, unlike this dataset.

Regarding data pre-processing, we resized the images to $64 \times 64 \times 3$ and the data was split into 85% for training and 15% for testing, as was done in the original paper.

5.1.2 Medical dataset: Warsaw-BioBase-Disease-Iris v2.1

The medical dataset chosen for these experiments is called Warsaw-BioBase-Disease-Iris v2.1 [110, 111]. This dataset is composed of 2,996 eye iris images from 115 different patients, containing more than 20 different eye conditions. The Warsaw dataset is the ideal starting point of the experimental work in this dissertation, as its biometric nature facilitates the identity recognition process, and its medical nature allows to validate the performance of the developed privacy-preserving model in a medical context.

Since the images were taken from three different devices, we chose to work with images from only one device: IrisGuard AD100, constituting 1,795 images. Furthermore, the most predominant pathologies available in the dataset are cataract and glaucoma. For simplicity purposes, we focused only on glaucoma, labelling the images according to the presence and absence of glaucoma. The data is unbalanced since only 425 out of the 1,795 images contain glaucoma.

To prepare the data, we cropped the images to remove text in their lower corners since this text could be detected by identity or glaucoma recognition networks and be wrongly used to achieve the respective task. In the data normalisation process, we horizontally flipped right eye images so that every image looks like it is a left eye, and we centred the iris in the middle of the image. To centre the iris, we first had to detect the iris' central point in order to apply a translation that would move this central point to the middle of the image. As there were no iris segmentation ground truths available with the dataset, we chose to apply a traditional computer vision method for circle detection in images: Hough Circle Transform [45]. We used the implementation provided by OpenCV [12]. Before applying Hough Circle Transform, we pre-processed the images with steps based on thresholding and blur to improve the contrast between the iris and the remaining eyeball, as illustrated in Figure 5.1. Our approach to normalisation took the following steps:

1. **Thresholding**: we chose a threshold α for the image intensity over which the image becomes white and under which the image remains the same. We used $\alpha = 130$.
2. **Gaussian Blur**: we used blur to reduce noise in the image and facilitate edge detection in the Hough Transform algorithm (Figure 5.1(b)).
3. **Intensity Normalisation**: we normalised the intensities that were in the range [0, α] to be in the range [0, 255], improving contrast (Figure 5.1(c)).
4. **Thresholding**: we chose a new threshold, under which the pixels become black and over which the pixels becomes white. For this threshold, we used the same value as in the first thresholding step: $\alpha = 130$.
5. **Gaussian Blur**: blur was used once again to reduce noise (Figure 5.1(d)).
6. **Hough Circle Transform**: we applied the Hough Circle Transform algorithm to detect a circle in the image and obtain its centre coordinates and radius (Figure 5.1(e-f)).
7. **Translation**: we applied a translation to move the detected iris' centre to the middle of the image (Figure 5.1(g-h)).

We set a limit of 100 pixels on the displacement translation to deal with cases where the iris was not correctly detected and where this normalisation process would significantly change the position of the eye. Thus, any image movement that surpassed 100 pixels horizontally or vertically was ignored, and the image in question remained in its original position.

Finally, we set the images' resolution to 64×64 and split the data into 65% for training, 15% for validation, and 20% for testing.

5.2 PPRL-VGAN model

As mentioned in Section 4.2.2, PPRL-VGAN [20] was developed for privacy preservation in facial expression recognition through identity replacement. Before being inputted into the model, the data's values, which are in the interval [0, 255], are normalised into the interval [-1, 1].

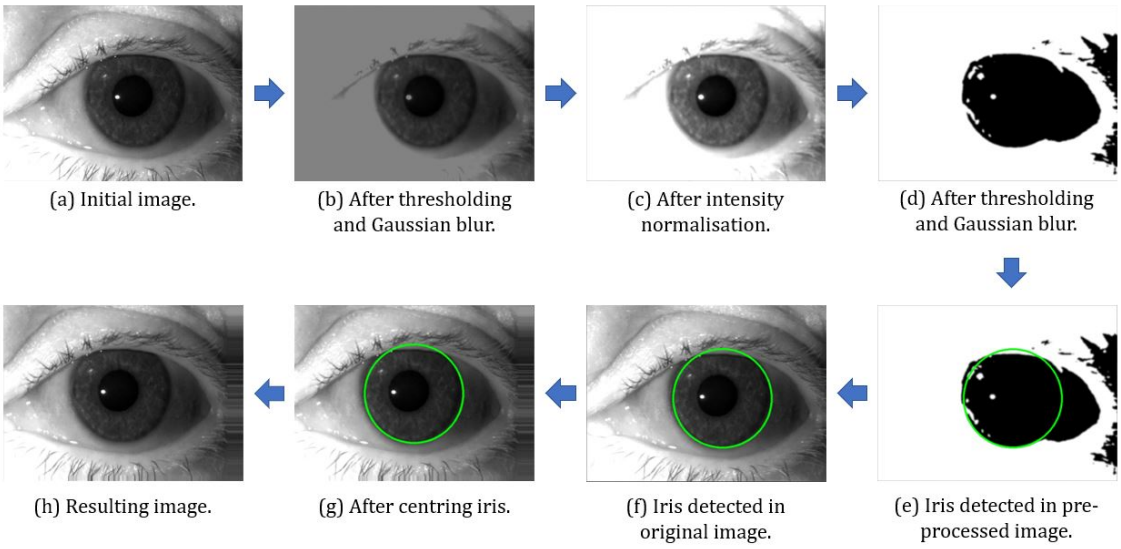


Figure 5.1: Normalisation process for Warsaw dataset.

The model's architecture contains a GAN with a conditional VAE as a generator and a multi-task discriminator. The conditional variational autoencoder is conditioned into outputting an image with the identity c, given to the decoder. As illustrated in Figure 5.2, the encoder of the VAE is composed of four blocks of strided convolution layers with LeakyReLU activation, Batch Normalisation and Dropout. Then, the model uses fully connected layers to calculate the mean μ and the deviation σ of a Gaussian distribution. The model proceeds to sample a latent representation z from the distribution, which is given to the decoder together with the identity c. The decoder starts with a fully connected layer to alter the size of the latent representation so that it can be reshaped into 3D data. Then, there are three blocks of transposed convolution layers with ReLU activation and Batch Normalisation, followed by a last transposed convolution layer to obtain the privatised image. Since the image pixels should contain values between -1 and 1, the last layer possesses a Tanh activation.

Since the VAE cannot be trained to reconstruct the input, as the final image is expected to be different from the original one, the model uses the discriminator to enable the generation of realistic images. Like most classification networks, the discriminator is composed of a CNN that extracts features, followed by a classifier responsible for performing the expected task. As shown in Figure 5.3, the CNN is composed of 4 blocks of strided convolution layers with LeakyReLU activation. The multi-task classifier is composed of one fully connected layer with LeakyReLU activation and Dropout, and a final decision layer for each task. The final decision layer is a fully-connected layer with sigmoid or softmax activation, depending on the number of classes in each task. When there are only two classes, like in the real/fake classifier, the decision layer uses sigmoid activation. When there are more than two classes, like in identity recognition, the network uses softmax activation.

The loss in the discriminator has three terms relative to each task. The discriminator is trained

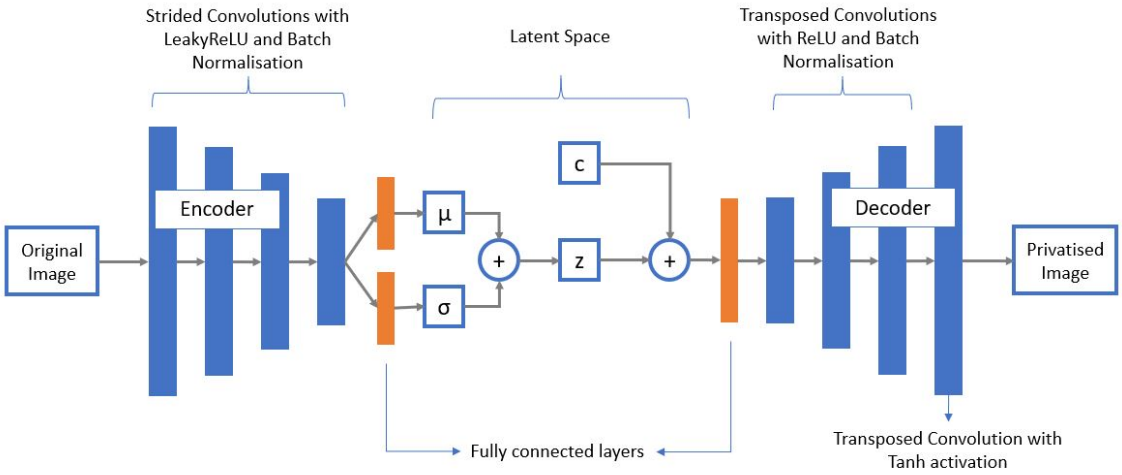


Figure 5.2: Architecture of the generator in the PPRL-VGAN model.

to distinguish between real and fake images with the real/fake discriminator D^1, to recognise identity with the identity recognition network D^2 and to perform one more task, like facial expression recognition, with the respective classification network D^3. As such, the discriminator intends to maximise the loss function defined in Equation 5.1. In this equation, λ_x^D are parameters to control the importance given to the task x, regarding real/fake classification, identity recognition and semantic classification. The variables y^{id} and y^e correspond to the target labels for identity recognition and for the semantic task, respectively, and I is an image from the original data space with probability distribution p_d.

$$L_D = \lambda_1^D \{E_{I\sim p_d(I)}[\log D^1(I)] + E_{I\sim p_d(I),c\sim p(c)}[\log(1-D^1(G(I,c)))]\} + E_{(I,y)\sim p_d(I,y)}[\lambda_2^D \log D^2_{y^{id}}(I) + \lambda_3^D \log D^3_{y^e}(I)] \tag{5.1}$$

In the original experiments, the discriminator was trained separately for real and fake images. When training with the fake images, the weights assigned to the identity and glaucoma loss terms were $\lambda_2^D = 0$ and $\lambda_3^D = 0$, respectively. These weights ensured that the identity and glaucoma recognition modules were trained using only the real images.

The loss in the generator contains four terms: a regularisation term to regularise the VAE's latent space and a term for each task performed by the discriminator. The generator's goal is to trick the discriminator into believing that the generated image is real, represents the replacement identity c and belongs to the same class as the original image regarding the semantic classification task. As such, the generator intends to minimise the loss function defined in Equation 5.2. In this equation, λ_x^G are parameters to control the importance of each term x in the loss function. In the regularization term, $p(f(I))$ corresponds to the prior distribution on the latent space, where $f(I)$ corresponds to the image I's latent representation, and $q(f(I) \mid I)$ is the conditional distribution parameterised by the encoder.

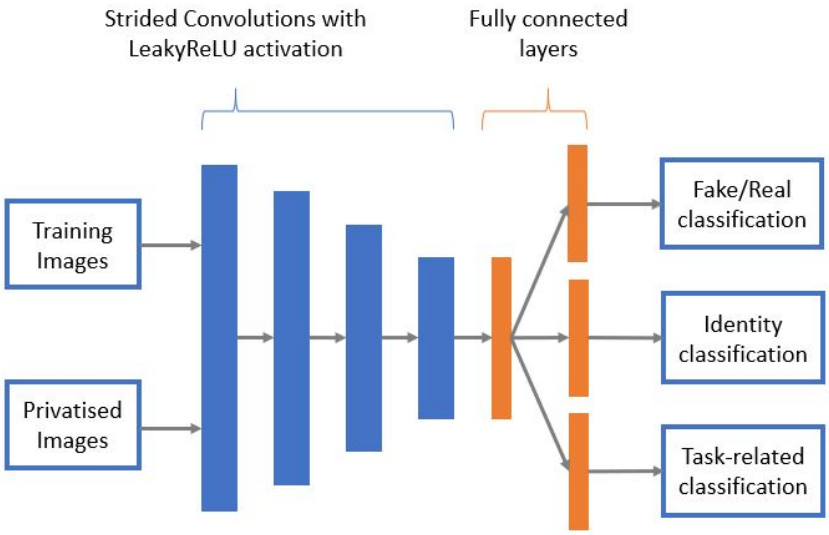


Figure 5.3: Architecture of the discriminator in the PPRL-VGAN model.

$$L_G = E_{(I,y)\sim p_d(I,y),c\sim p(c)}[\lambda_1^G \log(1 - D^1(G(I,c))) + \lambda_2^G \log(1 - D^2_{y'(c)}(G(I,c))) + \lambda_3^G \log(1 - D^3_{y^e}(G(I,c)))] + \lambda_4^G KL(q(f(I) \mid I)||p(f(I))) \quad (5.2)$$

To evaluate this network's capacity regarding privacy and preservation of semantic features, the authors included a classification network, which can be trained as an identity recognition model or as the task-related classifier used in the generative model. The architecture of this model is similar to the one used in the discriminator, with four blocks of convolutional layers followed by two fully-connected layers, as can be seen in Figure 5.4. The loss function used is the cross-entropy loss. Note that the privacy guarantees in this work are restricted by the quality of the identity recognition network used to evaluate the privacy-preserving models.

5.3 Preliminary Experiments

The purpose of the preliminary experiments is to analyse the PPRL-VGAN model's shortcomings that hinder its use as a privacy-preserving tool for case-based explanations. We will start by analysing the model's capacity in the setting it was originally developed for: facial expression recognition, using the FERG database. Then, we will apply the model to the medical Warsaw database to find its shortcomings and advantages in a medical scenario. Finally, we will compare the results of this model in a medical setting with results obtained from traditional privacy-preserving methods.

We compiled the experimental work in this section in a research paper [76] published at

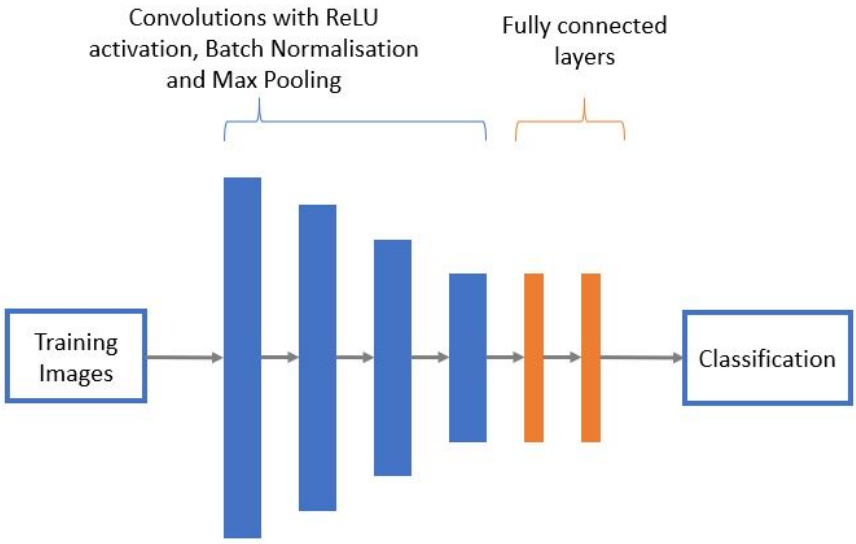


Figure 5.4: Architecture of the CNN provided along with the PPRL-VGAN model for evaluation purposes.

ICML's workshop on Interpretable Machine Learning in Healthcare (ICML 2021 IMLH). In specific, the published paper exposes the weaknesses of current privacy-preserving methods when applied to case-based interpretability in the medical scene. In this paper, we compare traditional and deep learning privacy-preserving methods, providing the empirical results obtained in Chapters 5.3.2 and 5.3.3 to support our claims.

5.3.1 Experiment with FERG Database

The experiment with the FERG database allows us to identify this network's capabilities of privacy-preserving image generation through an intuitive dataset that does not need a medical specialist to analyse its qualitative results.

We trained the network for 200 epochs with the RMSprop optimiser and a learning rate of $3e^{-4}$. We used the parameters that the authors described in the original paper: $\lambda_G^1 = 0.108$, $\lambda_G^2 = 0.6$, $\lambda_G^3 = 0.29$ and $\lambda_G^4 = 0.002$. As can be seen in the results expressed in Figure 5.5, there is no trace of the original character in the synthetic images, but the facial expression, representative of anger, is preserved. Although the first character's identity is not exposed, we can clearly see that the remaining characters' identity is exposed in the synthetic images, which does not solve the privacy issue in the images.

We noticed that when we try to replace the original identity with itself, while the feeling associated with the facial expression is preserved, its exact features are not, as can be seen in Figure 5.6, where the synthetic image contains a slightly opened mouth, with teeth showing, and with eyes more closed, unlike the original image where the mouth is closed. This is a limitation of preserving the facial expression that results from using the direct loss of the classification network

Figure 5.5: Results of training PPRL-VGAN network with FERG database for 200 epochs. (a) represents the original image and (b-f) represent the privatised images by replacing the original character with the other characters in the database.

for facial expression recognition to train the model. This loss does not guarantee that the original image's exact features are preserved. It only guarantees that the class to which the facial expression belongs is preserved. This mechanism for the preservation of task-related semantic features is not enough to ensure semantic similarity between the privatised image and the original one, which is a requirement for the privatisation of visual explanations. Furthermore, in medical data, different patients may have different symptoms of the same disease, which must be preserved as closely as possible.

Figure 5.6: Results of replacing character's identity with itself using PPRL-VGAN network. (a) represents the original image and (b) represents the synthetic image where the original identity was replaced with itself.

To assess this network objectively, we evaluated its capacity to preserve privacy and utility. To do so, we used the CNN classification network provided by the authors, which was trained first to recognise the images' identity and then to recognise facial expressions. We trained both the identity classification and the facial expression recognition networks with the original training set. Then, we used the encoder and decoder obtained in the PPRL-VGAN network to create a dataset of privatised images based on the test samples from the original dataset. For each test image, we privatised it by replacing the respective identity with all other available identities, generating 5 privatised images per sample. As such, the resulting generated dataset contains 42,260 images. With this generated dataset, we tested the identity recognition network's capacity to recognise the original identity and the original facial expression in the privatised images. Table 5.1 presents the results in terms of accuracy obtained with both networks applied to the testing set, and the

generated privatised set. In the identity recognition process with the privatised set, we assessed the privacy preservation of the original subject's privacy and the subjects used as a replacement, using the respective identity as a label.

Table 5.1: Results of experiment with FERG database on PPRL-VGAN network.

Dataset	**Identity Recognition**	**Expression Recognition**
Original testing set	100%	100%
Privatised set with original identity labels	0.57%	97.99%
Privatised set with replacement identity labels	97.35%	97.99%

From these results, we can conclude that facial expressions are preserved in the privatised dataset, as can be seen by the respective high accuracy. The privacy of the original character in the privatised image is preserved, making it very difficult for the identity recognition network to identify the respective identity, as can be seen by the significantly low accuracy. However, the identity of the character who was used as a replacement is sacrificed since the network can recognise it with very high accuracy. These results support this network's previously mentioned limitation regarding the lack of privacy for the characters used as a replacement.

In an attempt to use this model and still guarantee the privacy of all the characters in the dataset, we tried to average the synthetic images obtained by replacing the original identity with each character. Since all the characters present the same expression, the averaged image should also preserve this facial expression. This experiment's results can be seen in Figure 5.7 (a), where we can see that the resulting image is unintelligible. This happens because the eyes, nose, and mouth of the characters are located in different places of each synthetic image, pointing to the need to normalise the images by placing the different parts of the face roughly on the same spots. To see if the images become more understandable using fewer identities, we also tried averaging the faces of only three characters whose face parts are roughly on the same spots. In Figure 5.7 (b), we can see that this averaged image is slightly more intelligible and with a clearer facial expression. However, the facial expression is not completely clear. Additionally, since such a low number of identities was used, we can identify which characters were used to build this image if we know the dataset's characters. Therefore, this solution, as it is, is not appropriate to preserve privacy in this network. We also cannot average the latent representations of the generated images to produce a realistic privatised image mapped in the original data space through the decoder since the decoder was trained to receive both a latent representation and the class of the identity that should be used to replace the original identity. The only way we could use this network and still preserve the subjects present in the training data is if we had synthetic models that could be used to replace the original identity.

With this first experiment, we can conclude that this network possesses limitations in its privacy-preserving capabilities, as it does not protect the privacy of the subjects used as a replacement. Furthermore, the model is also lacking in regards to the preservation of explanatory

(a) (b)

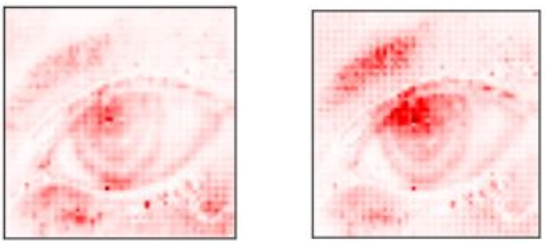
Figure 5.7: Results of averaging privatised images with 6 different identities (a) and 3 different identities (b) to preserve privacy in PPRL-VGAN network.

evidence, as the network does not preserve the exact semantic features from the original image, it only ensures that the semantic features discriminate the same class as in the original image.

5.3.2 Experiment with Warsaw Database

The experiment with the Warsaw database allows us to identify the PPRL-VGAN network's shortcomings when applied to medical data. Before applying the model to the data, we first made a quick analysis of how entangled the glaucoma-related features and the identity features are in the images. To do so, we used the CNN network that was introduced as an evaluation model in Section 5.2. The network was trained on the training set for identity recognition and for glaucoma recognition. Then, we applied a visual interpretability technique to visualise the parts of the images that are the most relevant to each of the recognition tasks, to see if they overlap. We used an implementation of Deep Taylor Decomposition [74], provided by iNNvestigate [4]. One example of results is available in Figure 5.8. In this figure, we can see that the identity features and glaucoma features are entangled as there are highlighted zones that are common in both images. To further evaluate how entangled the features are at the whole dataset's level, we investigated the intersection over union (IoU) score between the masks obtained from applying this interpretability technique with both networks. Ignoring values under a threshold of 0.1 in the images, we arrived at a IoU score of 37.43%. Additionally, the average percentage of pixels that appear in the intersection of identity and glaucoma-related features in the images is 12.37%. These results clearly show that there are features that are needed for both the glaucoma and identity recognition processes. However, as these values are not very high, there is also a component of the features that is exclusively relevant to either glaucoma recognition or identity recognition. The most significant challenge in the privatisation process with this data will be to manage how the entangled features are altered so that we keep enough disease-related information while discarding identity.

Regarding the application of the PPRL-VGAN network to the medical data, we trained the model for 730 epochs, where it obtained the best results regarding loss in identity recognition and glaucoma recognition. Like in the previous experiment, we used the RMSprop optimiser with a learning rate of $3e^{-4}$. Regarding the parameters used in the generator's loss, we used: $\lambda_G^1 = 0.5$,



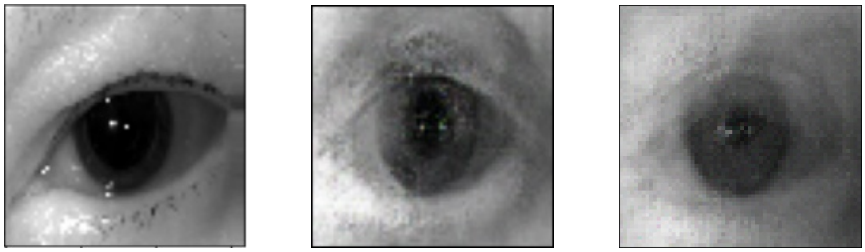
(a) (b)

Figure 5.8: Results of applying Deep Taylor Decomposition on the glaucoma recognition network (a) and on the identity recognition network (b).

$\lambda_G^2 = 0.5$, $\lambda_G^3 = 0.5$ and $\lambda_G^4 = 0.002$. For these experiments, we created several privatised sets using PPRL-VGAN on the testing set, to fully evaluate the model's performance:

- Privatised set with original identity: the images were privatised using the original identity from the input image as the replacement identity. This set evaluates the network's capacity to reconstruct the original image.
- Privatised set with random identities: the images were privatised using randomly selected replacement identities.
- Privatised set with identities sharing the same pathology as the original image: the images were privatised with replacement identities randomly selected from the pool of images with the pathology observed in the original image. An example of an image taken from this set is represented in Figure 5.9 (b).
- Privatised set with identities that do not share the pathology from the original image: the images were privatised with replacement identities taken from the pool of images that do not possess the pathology observed in the original image.
- Averaged privatised set: the privatised images result from averaging 6 privatised images that share the same pathology as the original one, including one image obtained from using the original identity as a replacement in the PPRL-VGAN model. An example of an image taken from this set is shown in Figure 5.9 (c).

To evaluate the network privatisation capabilities, we investigated the accuracy of an identity recognition network trained on the Warsaw dataset and evaluated on the testing set. Using the original testing set as the baseline, we evaluated the accuracy of the identity recognition network on each privatised set, expecting to get low accuracy results. Furthermore, we tested the accuracy of this network when applied to recognising the identity of the patients that were used in the privatisation process as a replacement. Regarding the preservation of explanatory evidence, we used a glaucoma recognition network trained on the Warsaw dataset and evaluated on the testing set, expecting to obtain high accuracy on the privatised sets. Since the glaucoma recognition



(a) (b) (c)

Figure 5.9: Example of results of the PPRL-VGAN method. (a) corresponds to the original image. (b) is a privatised image obtained with an identity that shares the same pathology as the original image. (c) corresponds to an averaged image obtained with 6 different privatised images.

network is a binary classifier, which classifies the images according to the presence or absence of glaucoma, we also used F-score to evaluate the preservation of glaucoma-related features. The results of the experiments are available on Table 5.2, with the two best results obtained for each metric highlighted in bold.

Table 5.2: Results of experiment with Warsaw database on PPRL-VGAN network.

Dataset	**Original Identity Recognition**	**Replacement Identity Recognition**	**Glaucoma Recognition**	**Glaucoma F-Score**
Original testing set	90.00%	-	93.24%	87.83%
Privatised set with original identity	81.76%	-	**89.12%**	**81.41%**
Privatised set with random identities	**0.50%**	74.68%	79.18%	62.22%
Privatised set with identities with the same pathologies	1.76%	78.35%	**86.56%**	76.01%
Privatised set with identities with different pathologies	**0.71%**	60.26%	65.06%	48.30%
Averaged privatised set	2.56%	**14.35%**	86.24%	**78.80%**

From the experiment with the set of images privatised with the original identity, we can conclude that there is a slight loss in the images' reconstruction, since the accuracy in identity recognition is lower than the baseline. This loss is not as accentuated in the glaucoma recognition results which were very high, proving the network's capacity to preserve glaucoma-related features when reconstructing an image. As such, the network may learn to prioritise preserving semantic features rather than anonymising the image through identity replacement.

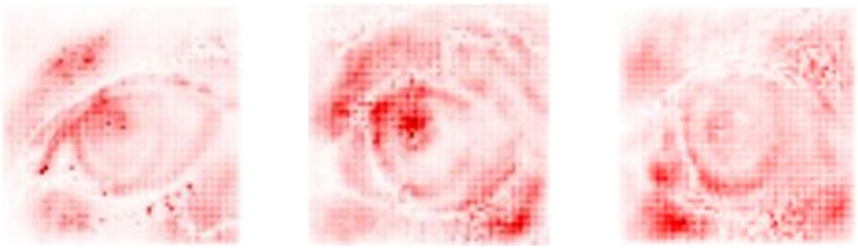
From the experiment with the set privatised with random identities, we can see that the network can successfully hide the identity from the patient in the original image, due to the significantly low identity recognition accuracy. However, the identity recognition network can identify the patient used as a replacement to privatise the image with very high accuracy, revealing the threat that PPRL-VGAN poses to this patient's privacy. Regarding the preservation of disease-related

semantic features, the low accuracy and F-score in glaucoma recognition points to a significant loss in explanatory evidence, unlike in the previous experiment with the FERG dataset. In the Warsaw dataset most patients are only associated to one pathology (either presence or absence of glaucoma), unlike the FERG dataset where every identity possessed images for all the facial expressions. As such, we inferred that the PPRL-VGAN model might have difficulties trying to recreate a pathology in a patient that originally does not possess this pathology. To confirm this theory, we used a set privatised with identities that share the same pathologies as the original image and a set privatised with identities that do not share the same pathology as the original image. From the experiments, we verified that the glaucoma recognition network achieves high accuracy and F-score when the images are privatised with identities with the same pathologies. On the other hand, the network achieves low accuracy and F-score when the identities used as a replacement do not possess the same pathologies as the original image. These results confirm the privatisation network's difficulty in recreating a pathology in patients that originally do not possess it. Furthermore, since the accuracy in the replacement identity recognition is lower for the privatised set with identities with different pathologies, we can infer that the existence or absence of glaucoma may be contributing to the identity recognition process.

Since the greatest weakness of the PPRL-VGAN model is the privacy violation of patients available in the dataset, due to its nature as an identity replacement network, we used the averaged privatised set to overcome this limitation. With the averaged privatised set, we obtained lower accuracy in the replacement identity recognition. The results show that this method ensures K-Anonymity, since the probability of identifying someone in the averaged image is slightly lower than $\frac{1}{K}$ ($\approx 16.7\%$). With this set, we achieved high accuracy and F-score in the glaucoma recognition network, proving this method's high capacity of preserving semantic features.

Furthermore, to evaluate the preservation of glaucoma-related semantic features, we used a visual interpretability method to check whether the image regions that contribute the most to the diagnostic decision are similar in the privatised and in the original images. We used an implementation of Deep Taylor Decomposition [74], provided by iNNvestigate [4], on the glaucoma recognition network. With this method, we obtained visual explanations where the highlighted regions of the images are relevant to the identification of glaucoma. The results of this experiment are shown in Figure 5.10. Similarly to the original image (a), the privatised image which was classified as glaucoma with high confidence (b) contains the upper region of the iris highlighted. Furthermore, this region is not highlighted in the privatised image which was classified as not having glaucoma (c). As such, we can conclude that glaucoma-related semantic features are preserved in the privatisation process.

With this second experiment, we confirmed that glaucoma-related features and identity features are entangled in the medical images. We also confirmed that most of the limitations identified in the previous experiment, such as the patient privacy violation, also apply to medical data. Furthermore, we identified one relevant drawback of this network when applied to medical data, which was not visible in the experiment with the FERG database: the network has difficulty in privatising an image using an identity that does not possess the pathology from the original image.



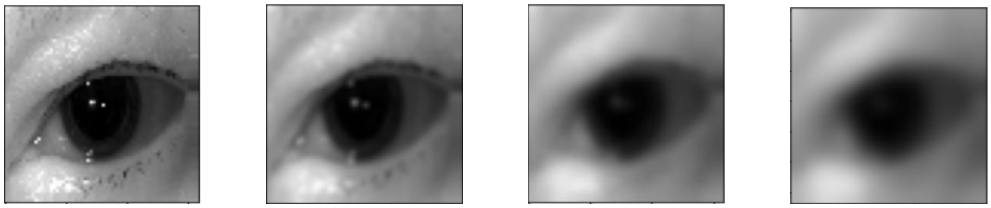
(a) (b) (c)

Figure 5.10: Results of applying Deep Taylor Decomposition in glaucoma recognition network. (a) represents the original image with glaucoma, (b) represents a privatised image with glaucoma, and (c) represents a privatised image without glaucoma.

5.3.3 Experiments with Traditional Privacy-preserving Methods

To further evaluate PPRL-VGAN as a deep learning approach for privacy, we compare it to two traditional privacy-preserving methods: blur and K-Same-Select [36]. The three methods are compared in terms of image intellegibility, privacy preservation and explanatory evidence preservation. For this comparison, we used the Warsaw database, to analyse each method in a medical context.

In the experiment with blur, we applied Gaussian kernels with different dimensions to the original test images. Bigger dimensions in the Gaussian kernels represent higher blurring degrees. The results of this method are represented in Figure 5.11. The highest level of privacy achieved in this method, with the highest blurring degree, translates into a significant loss in intelligibility, as can be seen in Figure 5.11 (d).

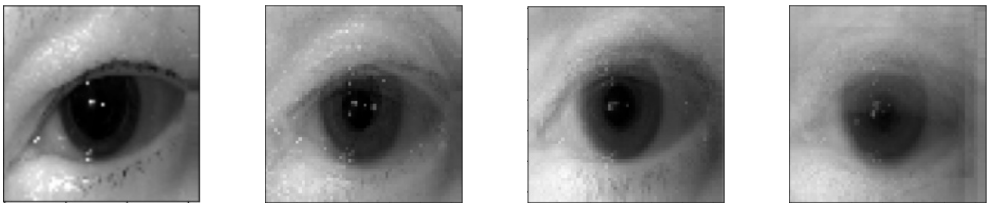


(a) (b) (c) (d)

Figure 5.11: Results of applying Gaussian blur on the images. (a) represents the original image and (b-d) represent images privatised with Gaussian kernels of 3, 9 and 15 dimensions, respectively.

In the experiment with K-Same-Select, we privatised the original testing set using different numbers of identities in the averaged images. We ensured that the images used in each privatised image belong to different patients, guaranteeing K-Anonymity. The results of this approach are represented in Figure 5.12. Since the images were normalised in the dataset's preparation, the images that result from this method have high intelligibility, especially when a small number of identities is used in each privatised image.

We have compiled the results from these experiments in Table 5.3, with the best results obtained in each metric highlighted in bold. From the deep learning method, we only included the privatised sets that achieved the best results in the previous experiment.



(a) (b) (c) (d)

Figure 5.12: Results of applying K-Same-Select on the images. (a) represents the original image and (b-d) represent images privatised with 3, 6 and 9 identities, respectively.

In the experiment with blurring, we observed that, as the blurring degree increases, it becomes more difficult for the identity recognition network to identify the patient. However, the highest level of privacy achieved is not enough to guarantee the patients' privacy. Overall, blurring achieved the worst results in identity recognition in comparison with the other methods. Furthermore, as the blurring degree increases, the glaucoma recognition accuracy decreases significantly. For the highest degree of privacy achieved, with a gaussian kernel of size 15, there is a significant loss in explanatory evidence and intelligibility which hinders this method's use as a privatisation tool for case-based explanations. The only advantage of this method is that an image's privatisation is independent of all other images in the dataset.

Table 5.3: Results from the experiments with K-Same-Select and Blur.

Experiment	**Dataset**	**Original Identity Recognition**	**Replacement Identity Recognition**	**Glaucoma Recognition**
Baseline	Original test set	90.00%	-	93.24%
PPRL-VGAN	Privatised set w/ identities w/ the same pathologies	**1.76%**	78.35%	86.56%
	Averaged privatised set	2.56%	**14.35%**	86.24%
Blurring	Privatised set w/ kernel size 3	69.41%	-	**93.24%**
	Privatised set w/ kernel size 9	31.76%	-	**88.82%**
	Privatised set w/ kernel size 15	23.24%	-	81.47%
K-Same-Select	Privatised set w/ 3 identities	7.06%	22.94%	82.35%
	Privatised set w/ 6 identities	2.94%	**14.41%**	81.76%
	Privatised set w/ 9 identities	**1.47%**	**14.41%**	78.53%

In the K-Same-Select method, the results from the column about replacement identity recognition correspond to the accuracy in recognising any identity used in the privatised images. As the number of identities used in the privatised images increases, the hardest it is for the identity recognition network to identify either the original patient or any of the patient's used to obtain the privatised image. With a significantly high number of identities, such as 9 identities, K-Same-Select achieves results that are comparable with the ones from the deep learning method in terms of privacy. However, regarding explanatory evidence, this method achieved poor results in comparison with the deep learning method.

Overall, the averaged privatised set using PPRL-VGAN achieved the most balanced results, with higher privacy and explanatory evidence preservation. However, this set suffers from the limits of K-Anonymity and the loss of the exact semantic features of one image that results from averaging different privatised images.

From these results, we can conclude that the deep learning method, PPRL-VGAN, has the highest capacity to preserve explanatory evidence from all the methods analysed. This method is the most suited to use as a starting point for the development of a more robust privacy-preserving approach that can be applied to medical data. The most significant weakness in this method in comparison to the traditional methods is the violation of privacy of subjects from the database, as concluded in the previous experiments.

5.4 Limitations of the PPRL-VGAN model

The development of the privacy-preserving framework for case-based interpretability in this dissertation will use the Deep Learning model PPRL-VGAN [20] as a starting point. In this chapter, we performed experiments to evaluate the work that must be done to adapt this model to obtain visual explanations using medical data. Using the FERG dataset, we identified drawbacks in the model when applied to fulfill its original goal of privatisation for facial expression recognition. Using the Warsaw dataset, we identified the model's limitations when applied to privatise medical data, where there are several subjects that do not possess all the pathologies identified in the diagnostic network, and a limited amount of images per subject.

The limitations that we concluded from these experiments were:

- **Anonymity**: The PPRL-VGAN network achieves the goal of preserving a subject's privacy through identity replacement. However, identity replacement does not guarantee privacy preservation for the subjects used as a replacement. Averaging several privatised images obtained with PPRL-VGAN to hide the identity of both the original subject and the subjects used as replacement negatively impacts the utility of the resulting image. Furthermore, this average imposes the limits of K-Anonymity on the privatised images. The privatisation process should be independent of the subjects present in the dataset to guarantee privacy preservation for all subjects. Additionally, the identity recognition network used in this model is a multi-class classification network which is difficult to train for medical data, as there is typically a limited number of images per patient.

- **Explanatory Evidence**: The use of a standard classification network for the semantic task only guarantees that the original image's class is preserved. It does not guarantee that the exact semantic features used by the classification network to classify the image are preserved. Although the averaged privatisation set preserves general semantic features, it loses the exact semantic features of the original image by mixing them with features from other images.

- **Realism**: The images possess a bit of noise in both experiments, however, the resulting images are intelligible. There is no current limitation regarding realism in the privatisation network.

Considering these limitations, the next chapters will focus on improving the PPRL-VGAN model to enable its use for the generation of privacy-preserving case-based explanations applied to medical data. The greatest challenge we expect to overcome in the experimental work that follows is to manage the trade-off between privacy, explanatory evidence and intelligibility. Since disease-related features are usually entangled with identity features, it might be difficult to find an equilibrium where we keep enough semantic features to preserve the images' explanatory value while also removing identity.

Chapter 6

Privacy-Preserving Model with Multi-class Identity Recognition

In this chapter, we propose privacy-preserving models for case-based explanations that use a multi-class identity recognition network to guide the privatisation process. We guide the development process through three main steps: improving privacy, improving realism and improving explanatory evidence in the privatised images, each represented in a section. The final section summarises the main observations taken from the experiments and compares our newly developed privacy-preserving model with the original PPRL-VGAN model [20].

The experiments in this chapter use the Warsaw dataset introduced in Chapter 5.1.2. We performed the experiments in Keras [24], with Tensorflow backend [2]. In this chapter, we expose a limited amount of visual results per experiment. More examples of visual results can be visualised in Appendix A.1, for a better assessment of image quality.

In the experiments, we use the identity and glaucoma recognition networks introduced in Chapter 5.2 as evaluation networks to measure privacy and preservation of explanatory evidence. At first, we only evaluate these networks' accuracy in the privatised images. As the network develops and its loss function changes, other metrics are introduced in each section to evaluate the privacy-preserving model's performance using these evaluation networks. In each experiment, we train the privacy-preserving network according to the experiment's description and save the models that obtain the best results in terms of glaucoma and identity recognition on the validation set. We always expose in the visual and tabular results the models saved at the epochs that provided the best results regarding glaucoma and identity recognition.

In all experiments, the data is organized into 65% for training, 15% for validation, and 20% for testing. During training, we mostly use Adam optimizer with learning rate of $2e^{-5}$, except in the first section, regarding the improvement of privacy, where we use RMSprop optimiser with a learning rate of $3e^{-4}$.

6.1 Improving privacy in the privacy-preserving model

The main drawback that hinders using the PPRL-VGAN model to anonymise medical data is the privacy violation inherent to using identity replacement as the privatisation process. To fix this issue, we remove the identity, c, which was previously given to the model's decoder. This first section aims to obtain images that preserve the privacy of all the patients in the training set.

6.1.1 Removing identity replacement from the PPRL-VGAN model

In order to privatise images independently from the training data, we turned the conditional VAE from the original network into a normal VAE, by removing the identity that was previously given to the decoder. The resulting network's architecture is shown in Figure 6.1.

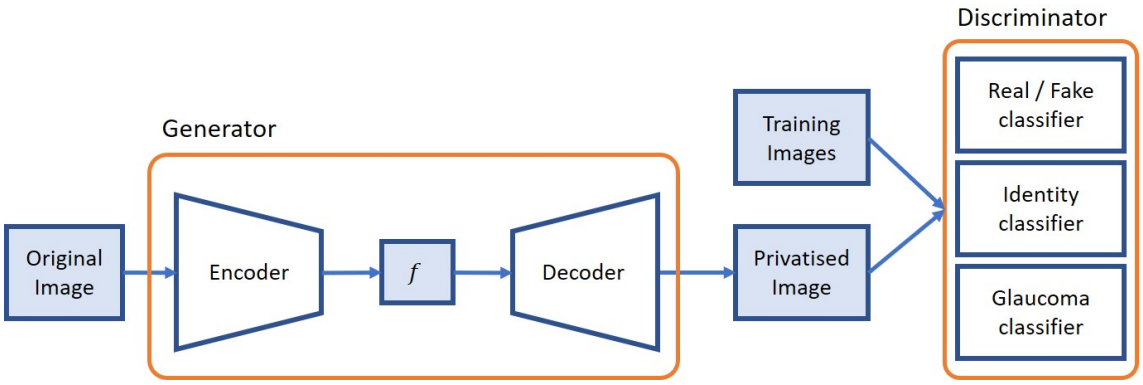


Figure 6.1: Architecture of the model based on PPRL-VGAN without identity replacement.

Since we no longer have a replacement identity, we can no longer approximate the privatised image's identity to the replacement identity in the loss function. As such, we encourage the generative model to make the identity recognition network fail to recognise the original identity by maximising the cross entropy between the original identity distribution and the privatised image's identity distribution. The discriminator is trained to maximise the loss function in Equation 6.1, while the generator is trained to minimise the loss function in Equation 6.2.

$$\begin{aligned} L_D = \lambda_1^D \{ E_{I\sim p_d(I)}[\log D^1(I)] + E_{I\sim p_d(I)}[\log(1 - D^1(G(I)))]\} + \\ E_{(I,y)\sim p_d(I,y)}[\lambda_2^D \log D^2_{y^{id}}(I) + \lambda_3^D \log D^3_{y^e}(I)] \end{aligned} \tag{6.1}$$

$$\begin{aligned} L_G = E_{(I,y)\sim p_d(I,y)}[\lambda_1^G log(1 - D^1(G(I))) + \lambda_2^G \log(D^2_{y^{id}}(G(I))) + \\ \lambda_3^G \log(1 - D^3_{y^e}(G(I)))] + \lambda_4^G KL(q(f(I) \mid I) || p(f(I))) \end{aligned} \tag{6.2}$$

The network was trained for 450 epochs, using a RMSprop optimiser with a learning rate of $3e^{-4}$. The results from this method can be found in Table 6.1, and some visual examples are shown

in Figure 6.2. The network obtained very high values in glaucoma recognition and low values in identity recognition, as expected of a privatisation method. However, we can see in Figure 6.2 that the privatised image (b) resembles an image (c) from the training data, even though the identity from this image was not explicitly given to the network. We can conclude from these results that the network still exposes someone's identity and that the metrics used in the results' table are not enough to guarantee privacy.

Table 6.1: Results of experiment of removing identity replacement from PPRL-VGAN network.

Dataset	**Identity Recognition**	**Glaucoma Recognition**
Original testing set	89.71%	92.94%
Privatised set	0.29%	89.12%

After further evaluating the results, we realised that 80.21% of the privatised images that contained glaucoma were classified with the identity in Figure 6.2 (c) by the identity recognition network. This suggests that this network is a victim to a well-known problem that often happens when training GANs: mode collapse.

(a) (b) (c)

Figure 6.2: Example of results of the privatisation method without identity replacement. (a) correspond to the original image. (b) is the privatised version of (a), and (c) is an image from the identity recognised by the identity recognition network.

In this experiment, we conclude that we need to explicitly train the network to generate images that do not resemble any of the training samples to preserve privacy in the whole dataset.

6.1.2 Approximating an uniform identity distribution in the privatised data

Since the previous network still exposed identities from subjects in the training dataset, we changed the loss function so that the generative network generates images that the identity classifier cannot recognise. As such, we approximate the identity distribution in the resulting images to a uniform distribution. The discriminator tries to maximise Equation 6.3, and the generator tries to minimise Equation 6.4, where U represents a uniform distribution. We trained the network with this loss function for 843 epochs, where we obtained the best results in identity and glaucoma recognition in the validation set.

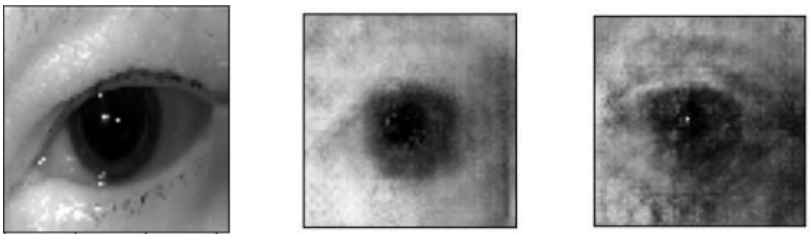
$$L_D = \lambda_1^D \{E_{I\sim p_d(I)}[logD^1(I)] + E_{I\sim p_d(I)}[log(1-D^1(G(I)))]\} + E_{(I,y)\sim p_d(I,y)}[\lambda_2^D logD^2_{y^{id}}(I) + \lambda_3^D logD^3_{y^e}(I)] \tag{6.3}$$

$$L_G = E_{(I,y)\sim p_d(I,y)}[\lambda_1^G log(1-D^1(G(I))) - \lambda_2^G D^2(G(I))log(U) + \lambda_3^G log(1-D^3_{y^e}(G(I)))] + \lambda_4^G KL(q(f(I)\mid I)||p(f(I))) \tag{6.4}$$

We also tried to approximate the images' resulting identity distribution to a uniform distribution using KL Divergence, as shown in Equation 6.5, representative of the generator's loss function. We trained the network with this loss function for 920 epochs.

$$L_G = E_{(I,y)\sim p_d(I,y)}[\lambda_1^G log(1-D^1(G(I))) + \lambda_2^G KL(U||D^2(G(I))) + \lambda_3^G log(1-D^3_{y^e}(G(I)))] + \lambda_4^G KL(q(f(I)\mid I)||p(f(I))) \tag{6.5}$$

The results from training the network with these loss functions are illustrated in Figure 6.3. The privatised images that resulted from this method severely lack in intelligibility.



(a) (b) (c)

Figure 6.3: Example of results of the privatisation method that approximates a uniform identity distribution. (a) is the original image. (b) and (c) are its privatised versions trained with entropy loss and KL divergence loss, respectively.

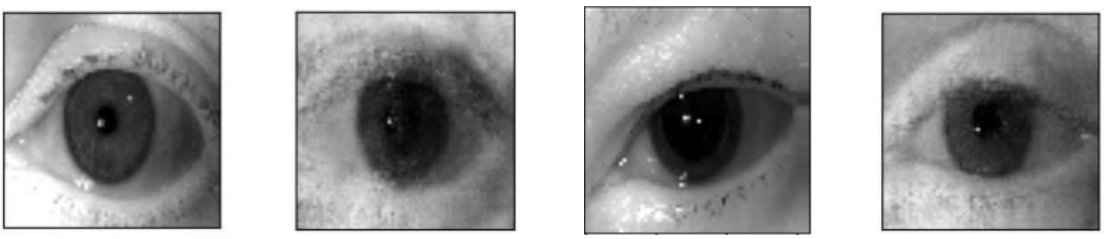
Since the previous experiment has an identity leak that was not recognisable with the evaluation metrics previously used, we added two new metrics to evaluate privacy at the whole dataset's level. First, we use the average maximum identity score, which corresponds to the maximum score that the identity recognition network assigns to an identity when making a prediction. Then, we use KL divergence to evaluate the distance between the predicted identity distribution and a uniform distribution. Note that images with identities that are not available in the dataset could still get high values in terms of KL divergence and maximum identity score in a multi-class recognition network. Nonetheless, these metrics serve to evaluate how good the generative network is at generating images that are difficult for the identity recognition model to recognise with high confidence, which is the goal of approximating the identity distribution to a uniform distribution.

The results are expressed in Table 6.2. In order to evaluate this method's results in terms of identity removal, we used as an additional baseline for KL divergence and maximum identity score the results of the model trained without the privacy term in the loss function (with $\lambda_2^G = 0$). We chose as the baseline the model that obtained the lowest maximum identity score and KL divergence. Our goal is to obtain values for these two metrics that are below this baseline. Figure 6.4 shows examples of results from this set that does not consider privacy. When we do not consider privacy, the synthetic images become intelligible, showing the privacy-intelligibility trade-off.

Table 6.2: Results of experiment that approximates uniform identity distribution for privacy.

Dataset	**Identity Recognition**	**Max Identity Score**	**Average KL Divergence**	**Glaucoma Recognition**
Original testing set (baseline)	89.71%	88.22%	4.24	92.94%
Generated set without considering identity (baseline)	2.35%	52.41%	3.26	92.65%
Privatised set w/ entropy loss	1.76%	**45.91%**	**2.93**	90.59%
Privatised set w/ KL loss	**1.18%**	57.31%	3.48	**91.18%**

The network that uses entropy loss to approximate a uniform identity distribution provided better results in terms of privacy than the one that used KL divergence. However, neither of the networks obtained satisfying results, as there is a high average maximum identity score and a high KL divergence in both the privatised sets. These results show that it is difficult for the network to generate realistic images that hide privacy. Regarding explanatory evidence preservation, both networks were capable of preserving glaucoma-related features in the images, as can be seen by the high accuracy obtained by the glaucoma recognition network on the privatised sets.



(a) (b) (c) (d)

Figure 6.4: Example of results of the PPRL-VGAN method without considering identity in the loss function. (a) and (c) are the original images. (b) and (d) are the respective synthetic versions.

With this experiment, the generative network was still not capable of generating privacy-preserving realistic images.

6.1.3 Using pre-trained identity and glaucoma recognition networks

The training of the previous network involves the complex task of optimising four components at the same time: a VAE generator, a real/fake discriminator, an identity classifier and a glaucoma

classifier. In order to simplify the training process, we pre-trained the models for identity and glaucoma recognition. As such, the network only has to optimise the components responsible for the generation of realistic images: the generator and the real/fake discriminator. The pre-trained models possess the architectures of the networks introduced in Chapter 5.2, which were used for evaluation purposes. The resulting network's architecture is illustrated in Figure 6.5. This network uses the entropy loss to approximate a uniform identity distribution, since, in the previous experiments, we achieved better results in terms of privacy using this loss term.

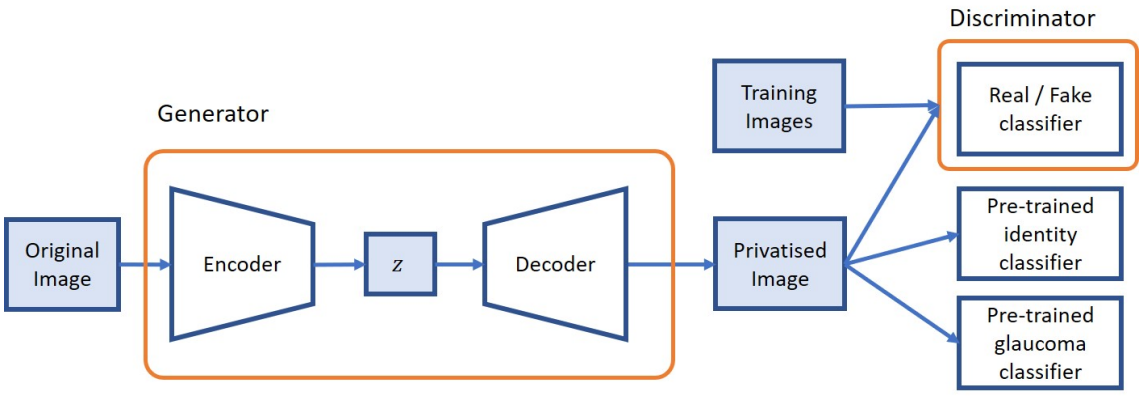
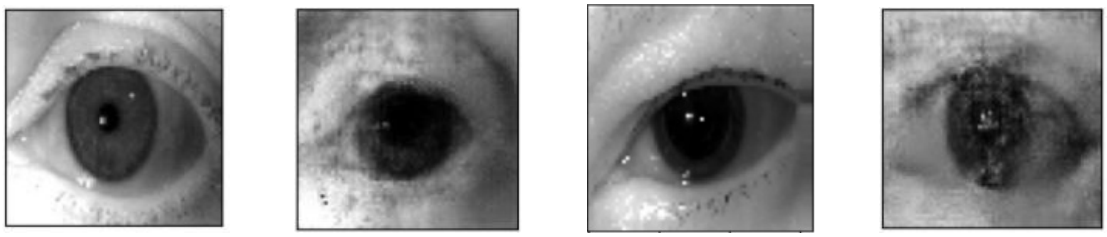


Figure 6.5: Architecture of PPRL-VGAN without identity replacement and with pre-trained identity and glaucoma recognition models.

The network was trained for 974 epochs, at which it provided the best results in terms of loss in identity and glaucoma recognition. Some examples of results are shown in Figure 6.6. From these results, we noticed various problems regarding realism. The generated images do not look real and may even upset doctors or patients that may have to look at them. Furthermore, we noticed that the two presented privatised images often appear in the privatised set, suggesting a mode collapse problem like in the previous experiment. The images that do not possess glaucoma generally collapse to the image represented by Figure 6.6 (b) while images with glaucoma collapse to Figure 6.6 (d).

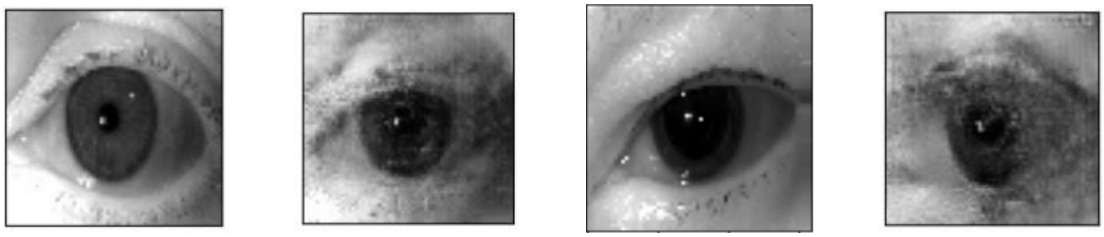


(a) (b) (c) (d)

Figure 6.6: Results of the privatisation method without identity replacement and with pre-trained models. (a) and (c) are the original images. (b) and (d) are the respective privatised versions.

We also collected results for the network trained after 971 epochs, where the network achieved

the best loss in terms of glaucoma recognition in the validation set. In this set, the images have higher quality than in the set collected at 974 epochs, as can be seen in Figure 6.7. Nonetheless, the intelligibility still neets to be improved, especially in the images that portray glaucoma, where the right side of the iris possesses a lot of noise.



(a) (b) (c) (d)

Figure 6.7: Example of results of the privatisation method without identity replacement and with pre-trained models, with higher image quality and preservation of glaucoma. (a) and (c) correspond to the original images. (b) and (d) are privatised versions of (a) and (c), respectively.

Table 6.3 expresses the results obtained with the privatised sets from the models trained at 971 and at 974 epochs. These results clearly show the trade-off between privacy and intelligibility. The network obtained at 971 epochs possesses lower privacy capabilities when compared with the one obtained at 974 epochs, since the results for maximum identity score and KL divergence are significantly higher at 971 epochs. However, at the cost of privacy, this privatised set presents higher-quality images and higher explanatory evidence preservation than the set obtained at 974 epochs. When we compare the newly developed network with the one from the previous experiment, we can see the great improvement in privacy that the new network has achieved at both epochs. Since the set saved at 971 epochs has considerably lower values in the privacy-related metrics than the baselines, it is better than the set saved at 974 epochs, as it contains higher-quality images and higher glaucoma recognition score.

Table 6.3: Results of experiment of removing identity replacement from PPRL-VGAN network by approximating uniform distribution and using pre-trained identity and glaucoma classifiers.

Dataset	**Identity Recognition**	**Max Identity Score**	**Average KL Divergence**	**Glaucoma Recognition**
Original testing set (baseline)	89.71%	88.22%	4.24	92.94%
Generated set without considering identity (baseline)	2.35%	52.41%	3.26	92.65%
Privatised set from previous experiment	1.76%	45.91%	2.93	**90.59%**
Privatised set with identity (971 epochs)	**0.59%**	23.12%	2.20	**90.00%**
Privatised set with identity (974 epochs)	1.47%	**6.36%**	**0.55**	87.94%

In this experiment, we obtained a network with high privatisation power that can preserve

the disease-related class of the original image. Pre-training the networks responsible for privacy and explanatory evidence preservation improves the model's results regarding privacy and image quality. We clearly identified the trade-off between privacy and intelligibility, which hinders the network's capacity to produce useful privatised case-based explanations. The most significant issues that need to be solved in this network are the lack of realism, as the network resulted in potentially upsetting images, and the mode collapse.

Regarding the mode collapse issue, we would like to point out that the original architecture suffered from an intentional mode collapse, where the mode collapsed to the identity given to the decoder. In this network, the mode collapse is not intentional and affects the network's capacity to recreate the exact glaucoma-related features from the original image, since all images with the same pathology look the same.

6.2 Improving realism in privacy-preserving model

To solve the mode collapse problem and improve the realism in the images, we tried various models to replace the current generative model. There are two types of changes that can be done to the current network: changes to the generator or changes to the discriminator. In the generator, we can change its architecture to produce higher quality images, while in the discriminator, we can change its loss function, in order to stabilise its training and fix the mode collapse problem. In this section, we explore both types of changes. First, we change the discriminator to battle the mode collapse problem and then we attempt to alter the generator's architecture to improve image quality. The experiments in this section are a continuation of the ones in the previous section. As such, we use the last privacy-preserving model obtained in the previous experiments, with pre-trained identity recognition and glaucoma recognition networks.

6.2.1 Fixing mode collapse with WGAN-GP

One model that we have successfully implemented to fix mode collapse is the WGAN-GP [39] network. In practice, this model changes the loss function of the GAN to the Wasserstein loss and penalises the gradients in order to enforce a Lipschitz constraint on the discriminator, rather than clipping its weights as suggested in the original WGAN [9]. The model's architecture remains the same as in the previous experiment. The only change in the discriminator is that, instead of a sigmoid activation, the decision layer possesses a linear activation. The goal of the Wasserstein loss is to maximise the difference between the values outputted for real and fake images. As such, the discriminator and generator are trained to minimise the loss functions in Equation 6.6 and Equation 6.7, respectively. In the discriminator, $\hat{x}$ corresponds to random samples which, in practice, are obtained by a weighted average between real and generated images.

$$L_D = E_{I\sim p_d(I)}[D(G(I))] - E_{I\sim p_d(I)}[D(I)] + E_{\hat{x}\sim p_{\hat{x}}}[\lambda(||\nabla_{\hat{x}}D(\hat{x})||_2 - 1)^2] \tag{6.6}$$

$$L_G = E_{(I,y)\sim p_d(I,y)}[-\lambda_1 D^1(G(I)) - \lambda_2 D^2(G(I))\log(U) + \lambda_3 \log(1 - D^3_{y^e}(G(I)))] + \lambda_4 KL(q(f(I) \mid I)||p(f(I))) \tag{6.7}$$

For the gradient penalty weight λ in the discriminator, we used the value 10, as suggested in WGAN-GP's original paper [39]. In this experiment, we tried different approaches to improve image quality and preserve privacy. In all approaches, we used the Adam optimiser with a learning rate of $2e^{-5}$. We generated the following privatised datasets as a result from the experiments:

- Privatised set: We used the previous model with pre-trained networks but with WGAN-GP, assigning the same degree of importance to each term in the loss function ($\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 1$). This network was trained for 952 epochs, at which it achieved the best results on the validation set. An example of a result is shown in Figure 6.8 (b).

- Privatised set with noise: We added gaussian noise with the distribution $N(0, \sigma^2)$ to the latent representations to improve image quality and robustness to small changes in the input images. As parameters of the gaussian noise's distribution, we used $\sigma = 0.02$. This network was trained for 1,223 epochs. An example is shown in Figure 6.8 (c).

- Privatised set with higher privacy degree: We changed the parameters to assign more weight to the privacy term in the loss function ($\lambda_1 = 0.4, \lambda_2 = 1, \lambda_3 = 0.6$). This network was trained for 1,166 epochs. An example from this set is illustrated in Figure 6.8 (d).

- Privatised set with random uniform identity distribution: We altered the identity term in the loss function so that instead of approximating a uniform distribution, where every identity would have the same probability of being recognised, we approximate a distribution randomly sampled from a uniform distribution. To achieve this, we sampled N values from the uniform distribution represented in Equation 6.8, where N is the total number of identities. To ensure that the sum of the values of the distribution is 1, we adjusted the values by calculating the difference between the total sum of the distribution and 1, and distributing this value between each identity in the identity distribution. The final value for each individual in the distribution is represented in Equation 6.9, where P_n is the base value for the individual n sampled from the uniform distribution U. One weakness of this method is that it is possible to obtain negative values using this equation if the average value of the probability distribution is higher than $\frac{3}{2N}$. Nonetheless, considering that this average is likely to attain values around the mean of the uniform distribution $\frac{5}{4N}$, it is very unlikely to obtain negative values. One alternative to normalize the values of this distribution is Equation 6.10, which was not used in this work and should be considered in the future. This method of defining the target identity distribution is equivalent to adding noise to the uniform distribution previously used. The parameters used to trained the network were $\lambda_1 = 0.4, \lambda_2 = 1, \lambda_3 = 0.6$. In this setting, the network achieved the best results in terms of identity and glaucoma recognition at 1,731 epochs. Figure 6.8 (e) represents an image obtained with this network.

- Privatised set with pre-trained generative network: We pre-trained the generative network without taking privacy into consideration ($\lambda_1 = 0.4, \lambda_2 = 0, \lambda_3 = 0.6$) in order to facilitate the network's task and generate higher quality images. After 500 epochs, we started incrementing the weight assigned to identity term loss λ_2 with increments of $\frac{1}{50}$, until it reached value 1. In this experiment, we also approximated the identity term loss to a distribution sampled from a uniform distribution, like in the previous privatised set. The network was trained for 1,397 epochs and obtained the results in Figure 6.8 (f).

- Privatised set with noise in latent representations and in uniform identity distribution: This set mixes the changes that occurred in the "Privatised set with noise" and in the "Privatised set with random uniform identity distribution". Since both these sets resulted in higher quality images, we decided to use both the changes in a new network to see if the resulting images had even better quality. As parameters, we used $\lambda_1 = 0.4, \lambda_2 = 0, \lambda_3 = 0.6$. The network was trained for 1,530 epochs, and obtained the image in Figure 6.8 (g).

$$U = \frac{1}{\frac{2}{N} - \frac{1}{2N}} \tag{6.8}$$

$$P_n^f = P_n + \frac{1 - \sum_{K=1}^{N} P_k}{N} \tag{6.9}$$

$$P_n^f = \frac{P_n}{\sum_{K=1}^{N} P_k} \tag{6.10}$$

We include more visual results in the Appendix, in Figure A.2. From the visual results, we can conclude that the network has difficulty in reproducing the eye parts that surround the iris. We found that adding gaussian noise to the latent representations (c) improves image quality. However, increasing the weight assigned to the privacy term in the loss function, which increases the image privacy (d), leads to a loss in image quality, evidencing the privacy-intelligibility trade-off. When we switch the privacy loss term from approximating an uniform identity distribution to approximating a distribution sampled from an uniform distribution (e), we see a clear improvement in image quality, with a clearer iris and eye structure even with higher weight in the privacy loss term than in the remaining terms. Injecting noise in deep neural networks is a technique that has been used in the literature to increase robustness and regularize models' training [5, 81, 90, 123]. In this work, we also verify that adding noise to the network results in its improved performance, since the privatised sets that present higher-quality images are the privatised set with noise (c) and the privatised set with random uniform identity distribution (e). However, mixing together the methods used to obtain these images (g) results in lower-quality images, as the high level of noise hinders the network's training. Further results from this experiment are available in Table 6.4.

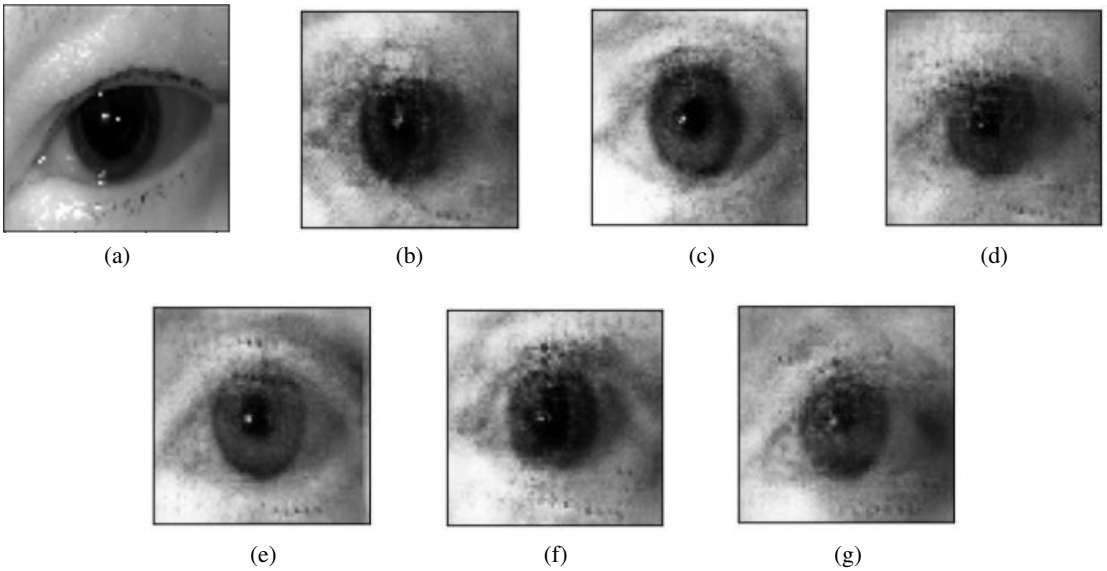


Figure 6.8: Example of results of the privatisation method using WGAN-GP model. The images include the original image (a) and an example of each of the privatised sets presented: privatised set (b), privatised set with noise (c), privatised set with higher privacy degree (d), privatised set with random uniform identity distribution (e), privatised set with pre-trained generative network (f), privatised set with noise in latent representations and in uniform identity distribution (g).

Table 6.4: Results of using WGAN-GP to solve mode collapse in privacy-preserving model.

Dataset	**Identity Recognition**	**Max Identity Score**	**Average KL Divergence**	**Glaucoma Recognition**
Original testing set (baseline)	89.71%	88.22%	4.24	92.94%
Generated set without considering identity (baseline)	2.35%	52.41%	3.26	92.65%
Privatised set from previous experiment	0.59%	23.12%	2.20	90.00%
Privatised set (b)	0.88%	29.08%	2.28	**91.18%**
Privatised set with noise (c)	**0.59%**	30.93%	2.24	89.18%
Privatised set with higher privacy degree (d)	1.18%	**26.90%**	**2.21**	89.41%
Privatised set with random uniform distribution (e)	0.88%	34.49%	2.60	89.41%
Privatised set with pre-trained generative network (f)	1.47%	29.03%	2.36	89.71%
Privatised set with noise (g)	1.18%	36.90%	2.68	88.82%

By increasing the privacy degree, through giving more weight to the identity loss term, the average maximum identity score and the KL divergence decrease. However, the image quality

decreases as well. In the privatised set with random uniform distribution, the maximum identity score is slightly higher than in the other sets, which may be justified by the fact that its identity loss term is not as strict as in the remaining sets. Overall, all the privatised sets produced good results regarding glaucoma recognition. Although none of the networks achieved as low KL divergence and average maximum identity score as achieved in the previous experiment, the values for these metrics achieved in this experiment are significantly lower than the presented baselines. Furthermore, it may be unrealistic to have an intelligible image where the identity recognition predictions follow a near uniform identity distribution. All the methods were capable of preserving the glaucoma class of the original images with high accuracy. Assigning a higher privacy degree in the parameters of the loss function results in higher privacy. In general, all the methods were capable of achieving acceptable privacy and explanatory evidence degrees. We chose to continue the experiments using the network (e), which approximates a noisy uniform identity distribution, as it clearly provided the best results in terms of image quality.

Using the privatised set with identity distribution sampled from a uniform distribution (e), we applied Deep Taylor Decomposition to further analyse the preservation of glaucoma-related features. An example of results is illustrated in Figure 6.9. In the visual results, we can observe that the privatised images' saliency maps are not very similar to the original images'.

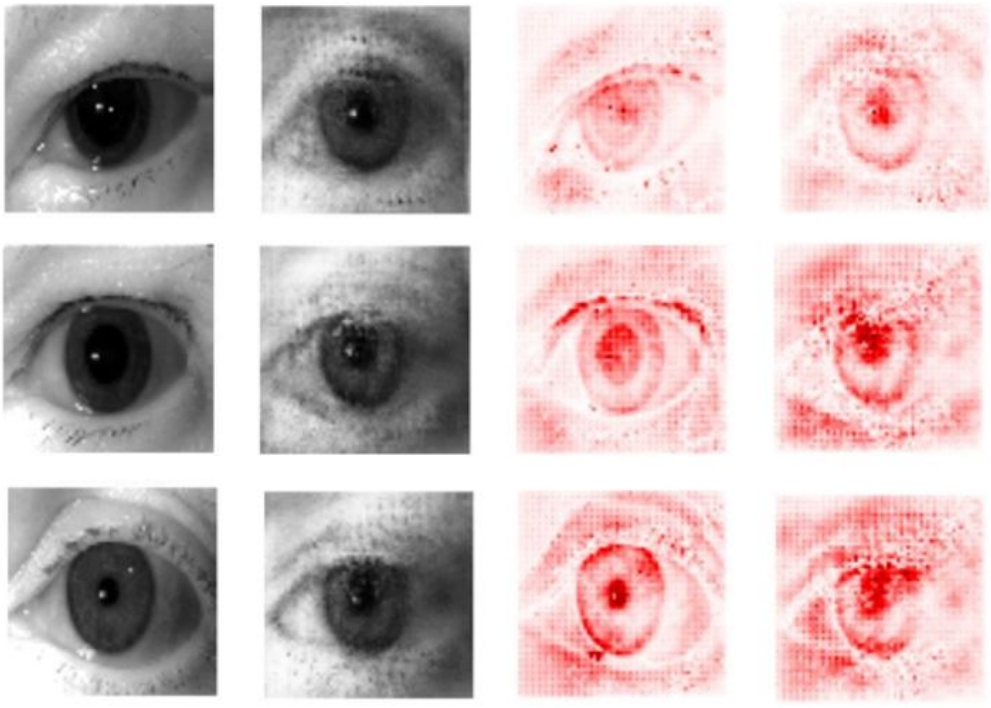


Figure 6.9: Results of applying Deep Taylor to images from WGAN-GP privacy-preserving model. The first two columns are the original images and their privatised versions. The last columns are the Deep Taylor saliency maps of the original and privatised images, respectively.

In this experiment, we achieved a privatisation network with intelligible images capable of

privatising the original images. The most significant issue in this network is that it fails to preserve the exact glaucoma-features of the original image, preserving only general features that allow to correctly classify the privatised images according to the presence or absence of glaucoma. We need to ensure that the original image's glaucoma-related features are explicitly preserved.

6.2.2 Improving image quality by changing the VAE architecture

We tested different architectures for the VAE in an attempt to further improve the quality of the images. In these experiments, we tried a ResNet VAE, whose convolutional layers have residual connections as proposed in the ResNet network, originally proposed for classification [42]. The architecture of the ResNet model is available in Figure 6.10. The VAE's encoder and decoder are composed of multiple convolutional blocks. There are three types of ResNet blocks that perform three different operations: downsampling, upsampling and an identity operation which preserves the data's dimensions. These blocks are represented in Figure 6.11. In the encoder, each strided convolutional block (Figure 6.11(b)) is followed by an identity convolutional block (Figure 6.11(a)). The same logic applies to the decoder where each transposed convolutional block (Figure 6.11(c)) is followed by an identity block (Figure 6.11(a)). The ResNet VAE is significantly deeper than the original convolutional VAE, containing the quadruple of the convolutional layers.

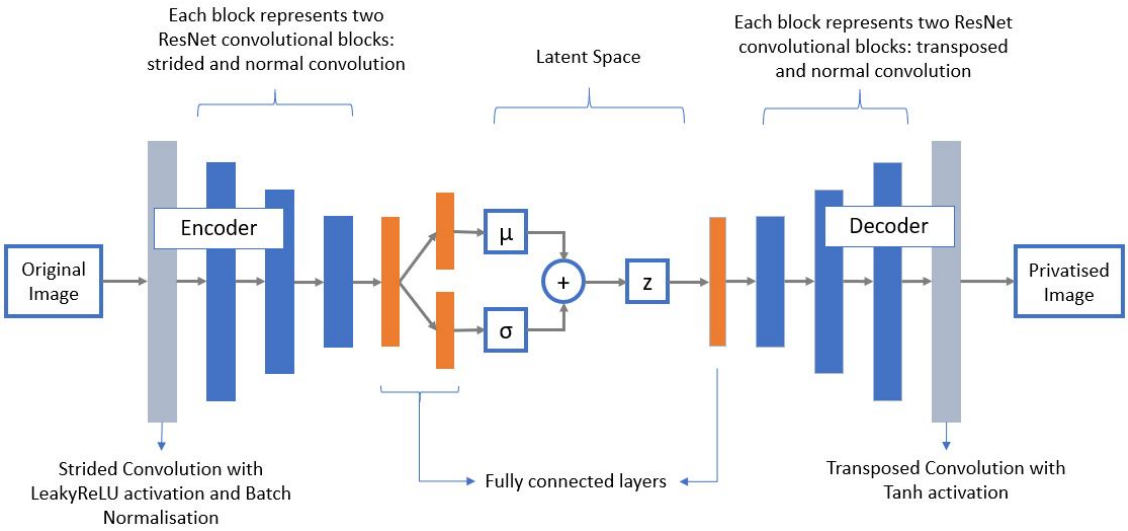
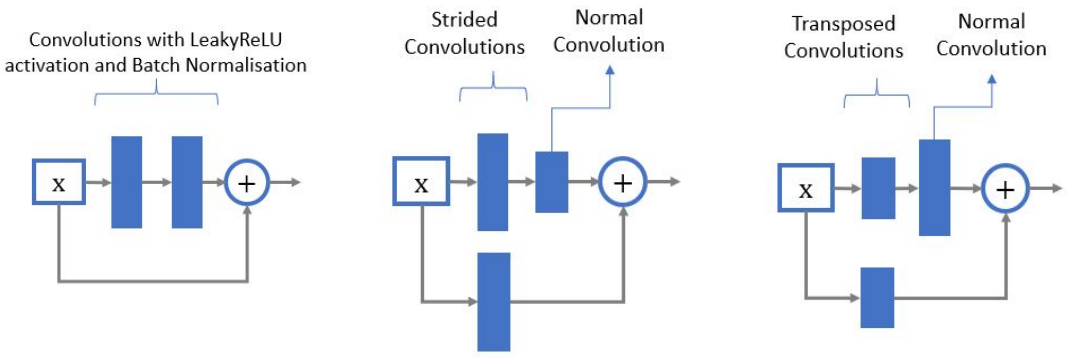


Figure 6.10: VAE with ResNet architecture as generator for the privatisation model.

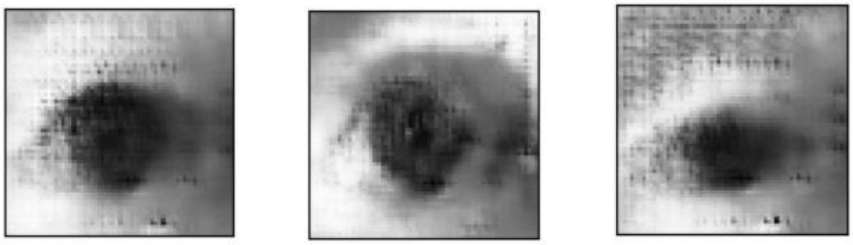
The network was trained for 1,801 epochs, where it provided the best results regarding glaucoma recognition and privacy. The network was trained using the Adam optimiser with a learning rate of $2e^{-5}$. As parameters, we used: $\lambda_1 = 0.4$, $\lambda_2 = 1$, $\lambda_3 = 0.6$, $\lambda_4 = 0.002$. We noticed that this network's training was unstable. Once the network starts obtaining somewhat intelligible images, the results revert to unintelligible images after a few epochs. An illustration of results obtained with the privatised model using the ResNet VAE as a generator is available in Figure 6.12. We



(a) (b) (c)

Figure 6.11: ResNet convolutional blocks. The first convolutional layer can be a normal convolution (a), a strided convolution for downsampling (b) or a transposed convolution for upsampling (c). All convolution operations are accompanied by LeakyReLU activation and Batch Normalisation.

can see that these results lack in quality when compared with the model that uses the original convolutional VAE as generator.



(a) (b) (c)

Figure 6.12: Examples of privatised images generated with the privatisation model with ResNet VAE as the generator.

We also tried a UNET architecture [96] in the generator. The UNET architecture was originally proposed for segmentation tasks and is often used in the context of image generation for image-to-image translation tasks. The UNET model possesses residual connections between the encoder and the decoder, encouraging feature preservation. We add noise to the data through dropout layers to instigate the loss of identity-related features. This network's architecture can be seen in Figure 6.13. Unlike the previous network, we do not use KL divergence to approximate the latent representations obtained with the UNET to a Gaussian distribution.

We trained this network for 1,556 epochs, where it provided the best results regarding glaucoma recognition and privacy. The network was trained using the Adam optimiser with a learning rate of $2e^{-5}$. Like in the previous experiments, we used the parameters: $\lambda_1 = 0.4$, $\lambda_2 = 1$, $\lambda_3 = 0.6$, $\lambda_4 = 0.002$. This method's results are available in Figure 6.14. The results of this network lack privacy, as the resulting images are very similar to the original ones.

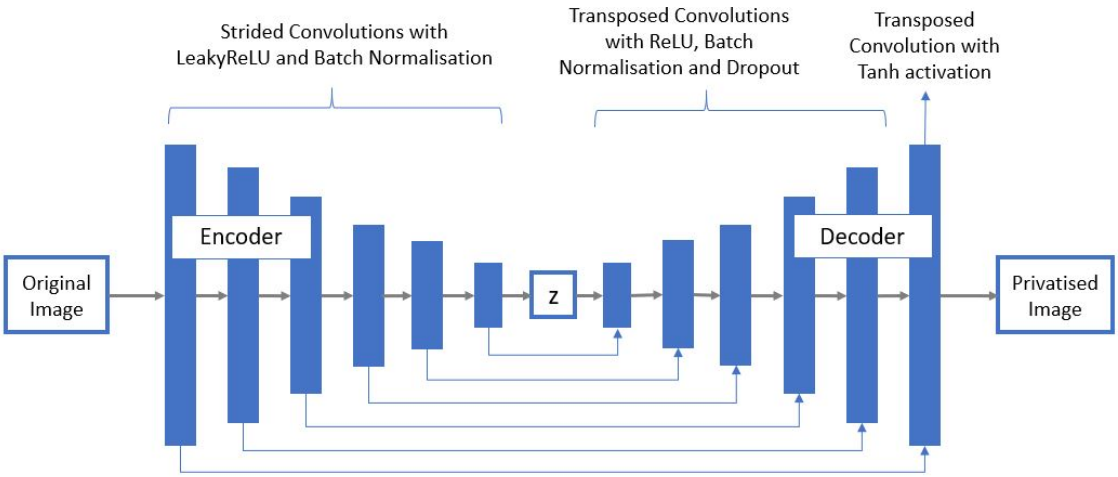
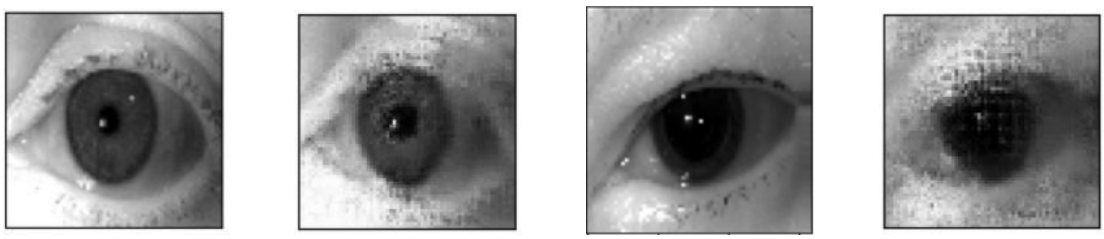


Figure 6.13: Architecture of UNET as generator for the privatisation model.



(a) (b) (c) (d)

Figure 6.14: Example of results of the privatisation method using UNET model as the generator. (a) and (c) correspond to the original images. (b) and (d) are privatised versions of (a) and (c), respectively.

We compiled the results from the privatised models with both architectures in Table 6.5. From these results we can once again see the trade-off between privacy, intelligibility and explanatory evidence. Using the ResNet architecture, the images attained very good results in terms of privacy, with a very low maximum identity score. However, this high degree of privacy comes at the cost of intelligibility, with low-quality images, and relatively low accuracy in glaucoma recognition.

In the experiment using the UNET architecture, the results in terms of identity recognition are higher than what we usually see with this privacy-preserving framework. However, these values are still low in comparison with the baseline corresponding to the original testing set. Through these results, we can infer that the UNET network is trying to perform an adversarial attack on the identity recognition network, where it tries to trick the network into wrongly classifying the image instead of removing identity features that would lead the network to have difficulty classifying it. The high data preservation capacity inherent to the UNET network hinders its privatisation capacity, requiring higher levels of noise in the model to aid the loss of identity features.

Neither of the networks used was capable of achieving higher quality than the original VAE while guaranteeing privacy. The ResNet VAE is very unstable to train, since it is significantly

Table 6.5: Results of using ResNet and UNET in privacy-preserving model.

Dataset	Identity Recognition	Max Identity Score	Average KL Divergence	Glaucoma Recognition
Original testing set (baseline)	89.71%	88.22%	4.24	92.94%
Generated set without considering identity (baseline)	2.35%	52.41%	3.26	92.65%
Privatised set using ResNet	0.59%	9.00%	0.74	84.41%
Privatised set using UNET	4.41%	26.86%	2.03	91.76%

deeper than the original VAE. The UNET generator was not capable of generating privacy-preserving images. As such, the experiments will procede using the network achieved in the previous section, with the original convolutional VAE.

6.3 Improving explanatory evidence preservation in privacy-preserving model

One aspect that is critical to the use of this privacy-preserving framework for case-based explanations is the preservation of disease-related features as they are in the original images. Only by preserving these features, we can guarantee the explanatory value of the privatised explanations. In the previous sections, we developed a privacy-preserving framework that preserves general disease-related features using a glaucoma recognition network to guide the explanatory evidence preservation process. However, this network only ensures the preservation of the class of the original image and not its exact features, diminishing its explanatory use. In this section, we aim to improve the quality of the privacy-preserving explanations produced by the privatisation model by improving the preservation of the original image's disease-related features. We evaluate the preservation of explanatory evidence through the comparison between Deep Taylor saliency maps obtained from the original images and their privatised versions.

6.3.1 Preserving explanatory evidence by explicitly preserving the iris of the eye

The first approach to the preservation of explanatory evidence is to explicitly preserve the parts of the images that contain the disease-related features. In the Warsaw dataset, we expect the glaucoma-related features to be in the iris of the eye. As such, we encourage the reconstruction of the iris of the eye in the privatised images. To do so, first, we obtained segmentation masks that identify where the iris is placed in the image.

To obtain iris segmentation masks, we used a process similar to the one used to normalise the dataset in Chapter 5.1.2, as can be seen in Figure 6.15. The only difference is that, instead of a translation to centre the iris, we create a mask where white represents the iris and black the absence of the iris. Furthermore, we also added a mechanism for when the iris is wrongly detected. We ensure that the mask is never too distant from the centre of the image, where the iris is usually

located. To do so, if the iris is recognised with its horizontal centre coordinate $x < 28$ or $x > 37$, we approximate the mask to the centre of the image horizontally, and we apply the same reasoning vertically. These values were obtained by manually analysing the iris detection method, where we verified that circles with $x < 28$ or $x > 37$ were too dislocated from the iris. We also limited the radius of the iris to $r < 17$, since upon analysis of the iris detection technique used, we verified that every image with $r >= 17$ defined a circle bigger than the iris.

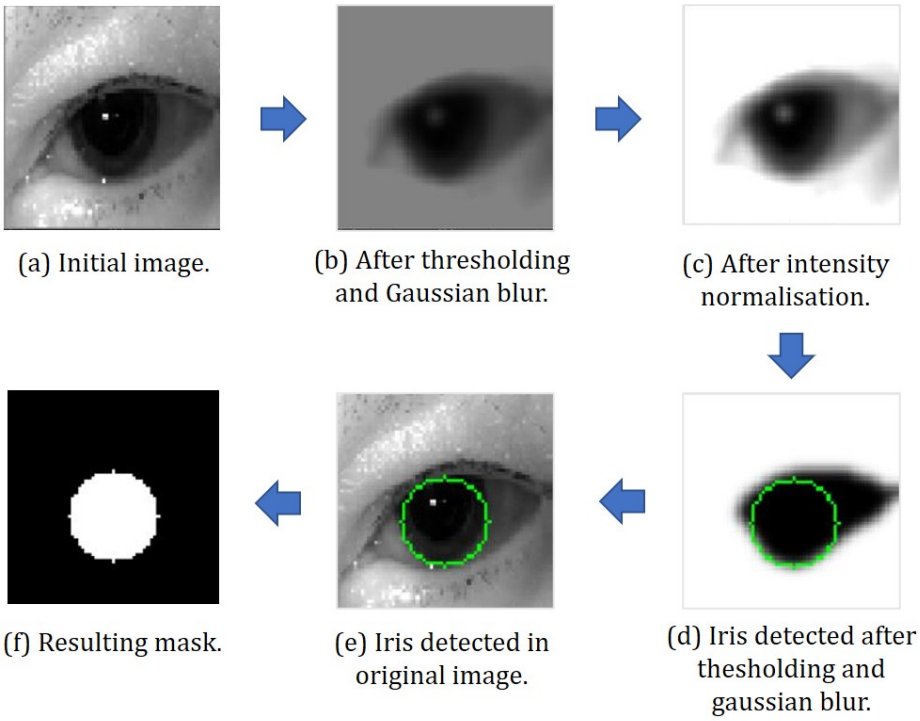


Figure 6.15: Process used to obtain iris segmentation masks.

The loss function in the discriminator remains the same as in the previous experiment. In the generator, we add a loss term for the iris reconstruction based on squared L2 normalisation loss, as can be seen in Equation 6.11. In this equation, U represents a uniform distribution with noise, which was obtained by sampling from a uniform distribution and adjusting the values so that the sum of all values in the distribution is 1, and M refers to the iris segmentation mask.

$$\begin{aligned} L_G = E_{(I,M,y)\sim p_d(I,M,y)}[-\lambda_1 D^1(G(I)) - \lambda_2 D^2(G(I))\log(U) + \\ \lambda_3 \log(1 - D^3_{y^e}(G(I))) + \lambda_4 KL(q(f(I) \mid I)||p(f(I))) + \\ \lambda_5 (I \times M - G(I) \times M)^2] \end{aligned} \tag{6.11}$$

The network was trained for 1,492 epochs, where we obtained the best results in glaucoma and identity recognition. As parameters, we used: $\lambda_1 = 0.4$, $\lambda_2 = 1$, $\lambda_3 = 0.6$, $\lambda_4 = 0.002$ and $\lambda_5 = 0.0002$. We achieved the results in Table 6.6. As baseline, we included the results from the

previous experiment (privatised set with random uniform distribution).

Table 6.6: Results of experiment using iris segmentation masks to preserve glaucoma.

Dataset	**Identity Recognition**	**Max Identity Score**	**Average KL Divergence**	**Glaucoma Recognition**
Original testing set (baseline)	89.71%	88.22%	4.24	92.94%
Generated set without considering identity (baseline)	2.35%	52.41%	3.26	92.65%
Privatised set from previous experiment	**0.88%**	34.49%	2.60	89.41%
Privatised set using iris masks	2.35%	**31.09%**	**2.38**	**90.59%**

From these results, we can see that the identity recognition network can recognise the original identity with slightly higher accuracy, but the average maximum identity score and the KL divergence score is lower than in the privatised set from the previous experiment. The set using iris segmentation masks also shows slight improvements in terms of glaucoma recognition accuracy, although the glaucoma accuracy is still not as good as the baseline sets which expose identity. In order to test whether the glaucoma-related features are preserved as they are, we use Deep Taylor Decomposition to compare the features relevant to the glaucoma recognition network in the images. The results are illustrated in Figure 6.16.

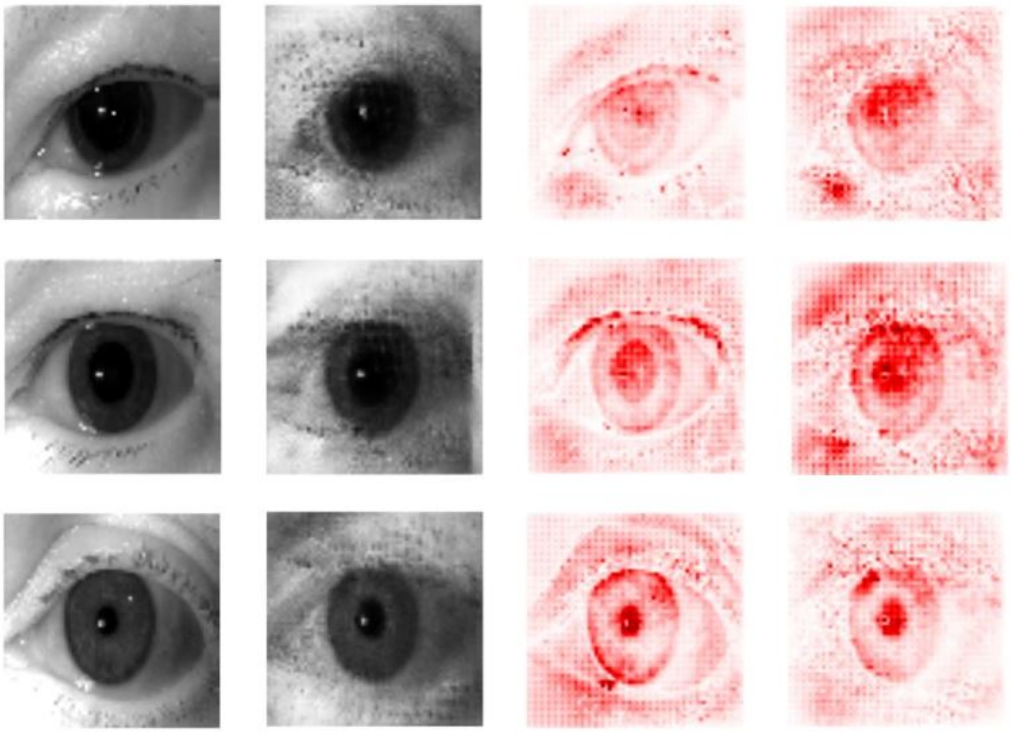


Figure 6.16: Results of using iris segmentation masks to preserve glaucoma features in the privatised images. The first two columns are the original images and their privatised versions. The last columns correspond to the Deep Taylor saliency maps of the original and privatised images.

In these results, we can see that this network has difficulty in generating an intelligible eye structure surrounding the iris. The iris is preserved as it is in the original image, with slight changes. Since we only use the masks in the loss function and we do not use them directly as input, the model learns to detect the eye iris in each image, in order to preserve it. Overall, the privatised images are more similar to the original ones than in the previous experiments. Regarding preservation of explanatory evidence, the privatised images' saliency maps are very similar to the original ones. As such, this network has a higher capacity to preserve the original images' disease-related features.

The problem with this approach is that, since our identity recognition network was trained with images of the whole eye, it uses not only the iris but also the structure of the eye in the recognition process, making it difficult to identify a patient using only the iris of the eye. As such, it is difficult to determine whether the privatised eye's iris still leaks identity using this network. It is well-known that the iris can also be used in identity recognition problems by itself, so preserving the whole iris may still leak identity. From the visual results, we can see the high similarity between the privatised images and the original ones, derived from the iris. Furthermore, although glaucoma-related features may be located in the iris, there may be segments in the iris that are not related to glaucoma and that are being preserved in this approach even though they may contribute to the identity leak.

6.3.2 Preserving explanatory evidence by explicitly preserving glaucoma-related features

Since the previous method unnecessarily exposes parts of the eye iris that may be irrelevant to the glaucoma recognition process, we develop in this experiment another approach to preserve the image parts that contain glaucoma-related features. We use saliency maps obtained with interpretability techniques to reconstruct the relevant explanatory features in the privatised images. In this experiment, we use Deep Taylor to obtain these masks.

In this experiment, we used two types of masks: one with the deep taylor features as they are, where the weight of each pixel's reconstruction is correlated with the respective relevance for the glaucoma recognition task, and a binary mask where all the glaucoma-related features are reconstructed with the same weight. Furthermore, we performed an AND operation between the generated masks and the iris segmentation masks, to ensure that only glaucoma-relevant features located in the iris are preserved, and to prevent the network from reconstructing the same eye structure as in the original image. Some examples of these masks are illustrated in Figure 6.17.

We trained the network using these masks with the Adam optimiser and a learning rate of $2e^{-5}$. As parameters, we used $\lambda_1 = 0.4$, $\lambda_2 = 1$, $\lambda_3 = 0.6$, $\lambda_4 = 0.002$ and $\lambda_5 = 0.001$. The network was trained for 1,500 epochs, where it provided the best results using both masks. From the results available on Table 6.7, we can conclude that the privatised sets using both types of masks have a similar privatisation capacity. The privatised set using binary masks obtained slightly higher accuracy in glaucoma recognition. Comparing the results of this experiment with the ones from

(a) (b) (c) (d)

Figure 6.17: Examples of masks obtained through an AND operation between Deep Taylor maps and iris segmentation masks. (a-b) are non-binary masks and (c-d) are binary masks.

the previous chapter, the identity recognition network has more difficulty in recognising identity in the experiments using the glaucoma masks, as can be seen by the lower values in maximum identity score and KL divergence. Furthermore, the glaucoma recognition accuracy has improved with this newly developed network. Some examples of visual results are available in Figure 6.18.

Table 6.7: Results of experiment to explicitly preserve glaucoma in privacy-preserving model using Deep Taylor masks.

Dataset	**Identity Recognition**	**Max Identity Score**	**Average KL Divergence**	**Glaucoma Recognition**
Original testing set (baseline)	89.71%	88.22%	4.24	92.94%
Generated set without considering identity (baseline)	2.35%	52.41%	3.26	92.65%
Privatised set from previous experiment (wgan-gp)	**0.88**%	34.49%	2.60	89.41%
Privatised set with non-binary masks	1.47%	**28.71**%	2.37	90.59%
Privatised set with binary masks	1.76%	29.62%	**2.31**	**92.35**%

From the visual results, we can see that the privatised set using the non-binary masks (second and fifth columns) has a higher difficulty in reconstructing the glaucoma-relevant features in the privatised images. Using the binary masks (third and sixth columns), the glaucoma-related features are being reconstructed closer to the original images, promoting a higher explanatory value in the privatised images. In practice, training the network with non-binary masks means that the network has to learn not only to identify the image parts that must be preserved but also to recognise their importance to the glaucoma recognition task, which is significantly harder than only locating and reconstructing relevant features. Regarding image quality, both methods resulted in images with lower quality than the results achieved in Chapter 6.2.1 with the privatised set using an identity distribution sampled from a uniform one. The third eye is the only one whose reconstructed versions have acceptable quality.

In this experiment, we can conclude that using binary Deep Taylor masks to reconstruct relevant image parts allows the network to preserve glaucoma-related features as they are in the

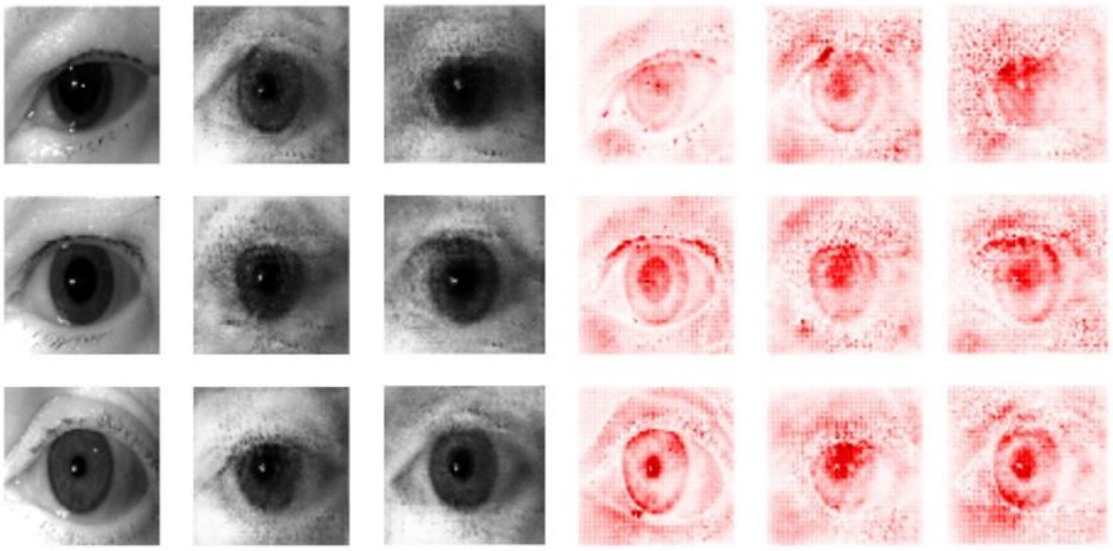


Figure 6.18: Results of using Deep Taylor masks to preserve glaucoma features in the privatised images. The first three columns correspond to the original images and their privatised versions using the non-binary masks and the binary masks, respectively. The three last columns are the saliency maps obtained with Deep Taylor of the original and privatised images, respectively.

original images. The non-binary masks did not produce good results in terms of preservation of explanatory evidence, as it becomes complicated for the network to learn to assign a degree of relevance to the relevant features. The images have lower quality at the exchange of higher explanatory evidence. The lack of quality may be aggravated by the growing difficulty of the generative task, since the generative network must learn to identify and reconstruct relevant glaucoma-related features in the images.

6.3.3 Using glaucoma masks directly in the generative model to preserve glaucoma-related features

In order to facilitate the generative model's task and improve image quality and glaucoma feature preservation, we performed experiments where we altered the network's architecture to receive not only the input image but also the Deep Taylor mask. We tried three different approaches to use the masks as input in the generative model. The approaches differ in where the concatenation between the input images and the masks happens in the generative network. This concatenation can happen before the images and masks are given to the encoder, inside the encoder, after feature extraction and before calculating the parameters of a Gaussian distribution, or in the latent space created by the encoder. These methods are illustrated in Figure 6.19.

Like in the previous experiments, we trained the three models using the Adam optimiser with a learning rate of $2e^{-5}$ and with the parameters: $\lambda_1 = 0.4$, $\lambda_2 = 1$, $\lambda_3 = 0.6$, $\lambda_4 = 0.002$ and $\lambda_5 = 0.001$. We perform this experiment using the binary masks, which achieved higher levels of explanatory value in the previous experiment. In order to visually compare the results from the

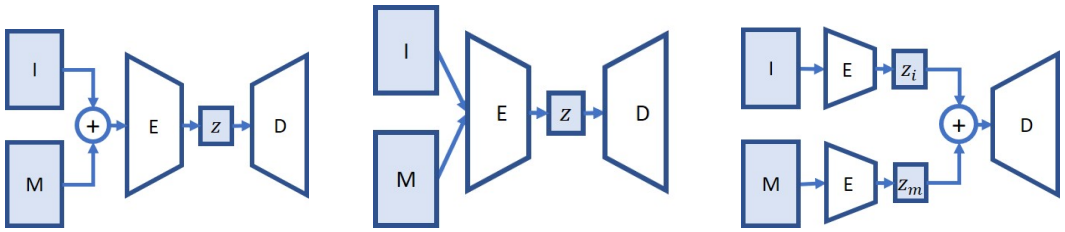


Figure 6.19: Architectures to input masks into the generator. (a) corresponds to the architecture where concatenation happens before the encoder. (b) corresponds to the architecture where concatenation happens inside the encoder. (c) is the architecture where the concatenation happens in the latent space created by the encoder.

three models, we included an example of a privatised image obtained with each of the mentioned architectures in Figure 6.20. From these images, we can clearly distinguish the first image for its lack of quality. As such, concatenating the input image with its Deep Taylor mask is not a good solution to preserve explanatory features along with image intelligibility. The other two methods resulted in similar-quality images.

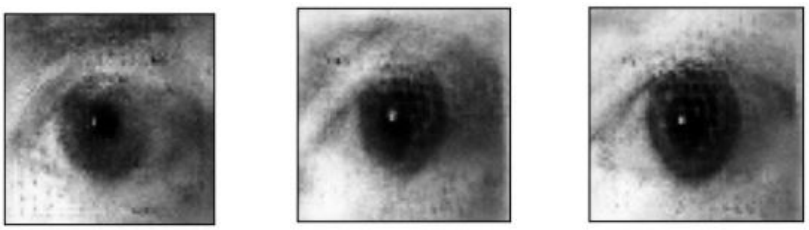


Figure 6.20: Examples of results obtained with each architecture that contains input masks in the generator. (a-c) are examples from the architectures displayed in Figure 6.19 (a-c), respectively.

After analysing the results obtained with each architecture, available and discussed below on Table 6.8, we arrived at the conclusions that the architecture (c) produces better results in terms of privatisation and that the architecture (b) possesses slighty better results in terms of explanatory evidence preservation. We chose to continue the experiments with the architecture (b), where the concatenation happens inside the encoder. For a more detailed overview of the generator in this network, we illustrate its detailed architecture in Figure 6.21.

The results of this experiment are available on Table 6.8. Through the analysis of the results, we concluded that inputting the masks resulted in lower accuracy in glaucoma recognition than in the previous experiments. By analysing the glaucoma-related results, we verified that the masks help preserve not the image's real glaucoma classification, but the glaucoma score that is assigned to the original images by the glaucoma recognition network, resulting in higher accuracy when this score is used as the ground truth. As such, in addition to the previously used metrics, we added one metric regarding glaucoma recognition using the scores assigned to the original images by the

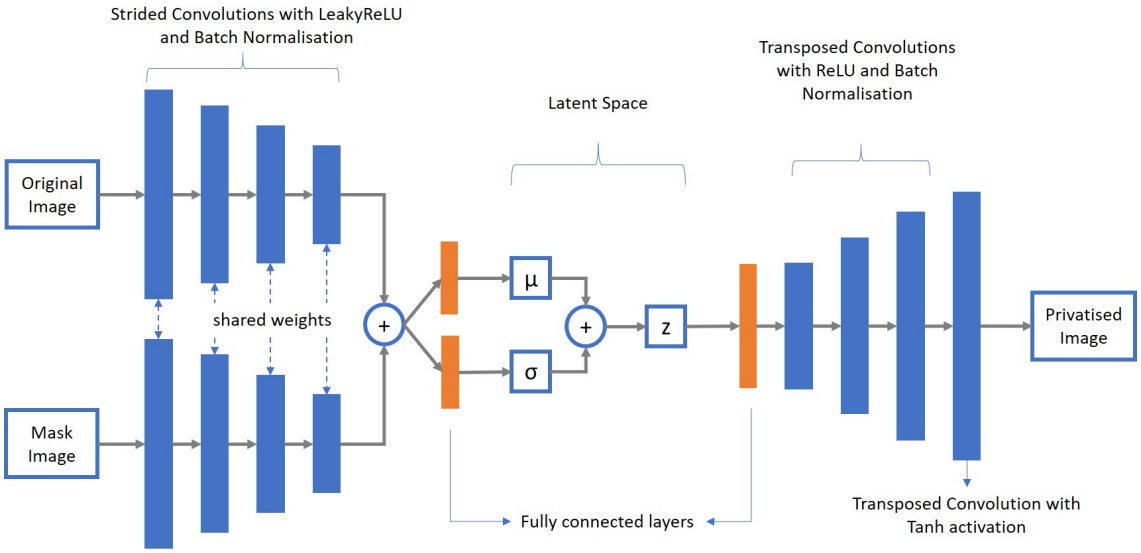


Figure 6.21: Architecture of the generator which receives a Deep Taylor mask as input and concatenates the mask with the original image inside the encoder.

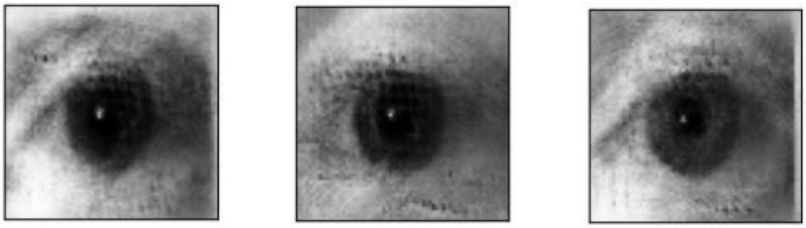
glaucoma recognition model as ground truth, in the column entitled "Assigned Glaucoma Recognition". Furthermore, as the results regarding explanatory evidence were slightly worse than in the previous experiment, we also tried assigning higher weight to the parameter λ_3, responsible for ensuring the preservation of general glaucoma features by minimising the crossentropy between the privatised image's glaucoma score and the real ground truth. Finally, we changed the parameter λ_5, responsible for the reconstruction of relevant glaucoma features in the privatised images, to see if we can improve the image quality and privacy while preserving explanatory evidence.

Table 6.8: Results of experiment inputting Deep Taylor masks into the generative model.

Dataset	**Identity Recognition**	**Max Identity Score**	**Avg KL Divergence**	**Real Glaucoma Recognition**	**Assigned Glaucoma Recognition**
Original testing set	89.71%	88.22%	4.24	92.94%	100.00%
Generated set without considering identity (baseline)	2.35%	52.41%	3.26	92.65%	-
Privatised set from previous experiment	1.76%	29.62%	2.31	92.35%	91.76%
Privatised set architecture (b)	2.35%	31.85%	2.49	86.47%	89.41%
Privatised set architecture (c)	2.06%	**29.90%**	**2.33**	83.24%	87.35%
Privatised set architecture (b) with $\lambda_3 = 2$	1.18%	32.63%	2.52	**88.82%**	**91.18%**
Privatised set architecture (b) with $\lambda_5 = 0.0005$	**0.88%**	30.94%	2.45	86.47%	88.82%

The privatised set obtained with architecture (c) resulted in slightly better results in terms of privacy when compared to the one obtained with architecture (b), with lower identity recognition

accuracy, maximum identity score and KL divergence. However, the results in terms of glaucoma recognition are low, in comparison with the privatised set obtained with architecture (b). The privatised set with $\lambda_3 = 2$ was capable of increasing the accuracy in the glaucoma recognition task using the real labels as ground truth and using the glaucoma score assigned by the glaucoma recognition network as ground truth. However, this method worsens the quality of the images, as can be seen in Figure 6.22 (b) where images from the sets obtained with architecture (c) are compared. The privatised set with $\lambda_5 = 0.0005$ improved the results in terms of privacy, with lower identity recognition accuracy. However, the results in terms of glaucoma recognition are slightly worse than in the other sets that use the same architecture. Regarding image quality, the results of this set are similar to the results with $\lambda_5 = 0.001$, as can be seen in Figure 6.22 (c).



(a) (b) (c)

Figure 6.22: Examples of results obtained with architecture (b) where input images and masks are concatenated inside the generator's encoder. (a) contains $\lambda_3 = 0.6$ and $\lambda_5 = 0.001$. (b) contains $\lambda_3 = 2$ and $\lambda_5 = 0.001$. (c) contains $\lambda_3 = 0.6$ and $\lambda_5 = 0.0005$.

To show this network's capacity to preserve explanatory evidence as is, we show some results with the respective Deep Taylor saliency maps in Figure 6.23. These results correspond to the privatised set with parameters $\lambda_3 = 0.6$ and $\lambda_5 = 0.001$, which had the highest-quality results in terms of image quality. In terms of image quality, this network produced better results than the previous network where the masks were only used in the loss function. Furthermore, this network is also capable of preserving relevant features for glaucoma recognition as in the original images.

With the experiments in this section, we conclude that we can preserve glaucoma-related features by using interpretability saliency maps to reconstruct relevant image regions in the privatised images. To achieve this purpose, we can use these masks only on the loss function, in a glaucoma reconstruction term, or by using the masks directly in the generative network. Using the masks solely on the loss function makes the network learn to recognise relevant features that must be preserved in the images. However, this process complicates the generative task, resulting in lower quality images. By inputting the masks into the generative network, we can generate higher-quality images that preserve explanatory evidence.

This method uses *post hoc* interpretability strategies for the preservation of relevant explanatory features. As such, it is highly appropriate to use when saliency maps are used for the retrieval of explanations, like in some of the interpretability methods mentioned in Chapter 3, such as IG-CBIR [103] and Twin Systems framework [50]. When applying this method to privatising explanations that result from intrinsic case-based interpretability methods, the use of *post hoc*

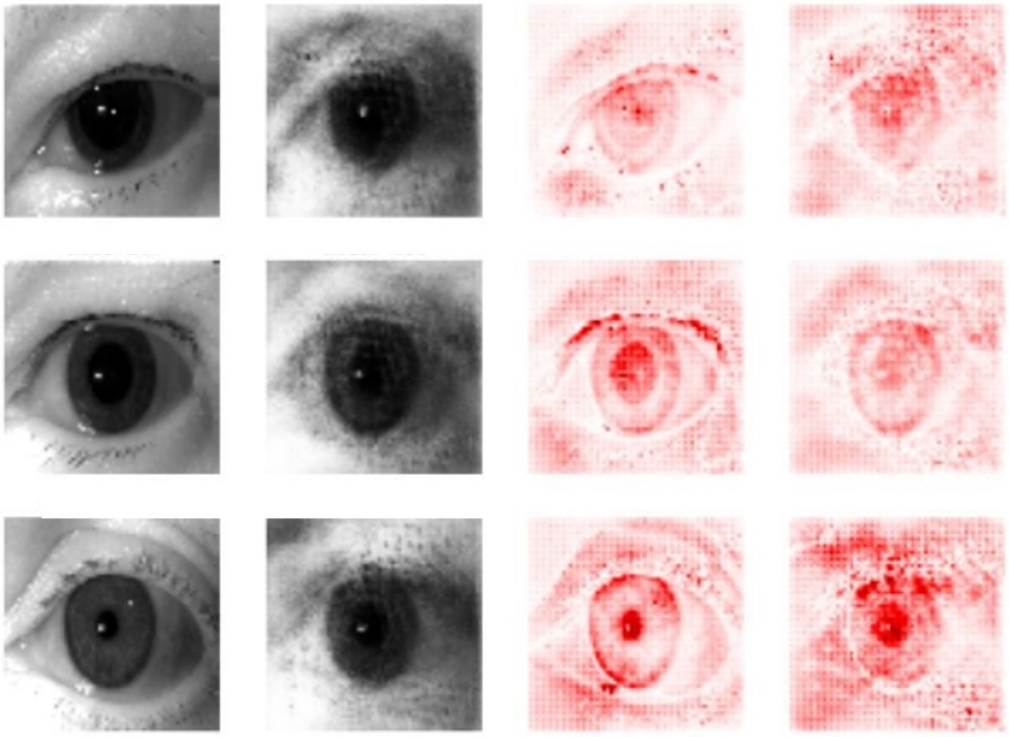


Figure 6.23: Results of using Deep Taylor masks as input in the generative network to preserve glaucoma features in the privatised images. The first and second columns correspond to the original images and their privatised versions using the binary masks, respectively. The two last columns are the saliency maps obtained with Deep Taylor of the original and privatised images, respectively.

strategies might clash with the intrinsic methods' goal of producing explanations that accurately explain the models' reasoning, since *post hoc* strategies are often criticised for not reflecting the models' real reasoning. As such, when intrinsic methods have a clearly defined similarity measure that can be used to semantically compare two images, it should be possible to use this measure to approximate features in the privatisation model, instead of using a method based on interpretability saliency maps.

6.3.4 Approximating glaucoma score in the original image instead of ground truth

In the previous experiments, we approximated the glaucoma score in the privatised images to the ground truth. However, in cases where a prediction is ambiguous, with a score close to 50%, using the glaucoma recognition network to approximate the ground truth might accentuate glaucoma-related features in the images, distorting them. Furthermore, in cases where the network makes a mistake in its prediction, using the ground truth might change the glaucoma-related features in the image to look more like what the network would classify as the opposite class. As such, instead of approximating the glaucoma score in the privatised images to the ground truth, we should approximate it to the glaucoma score obtained in the original image. Additionally, in the

previous experiment, we arrived at the conclusion that inputting masks with glaucoma-relevant features into the generative network made the privatised images have glaucoma scores that are closer to the ones assigned to the original images by the glaucoma recognition network, rather than the ground truth with the images' actual glaucoma score. As the explanations should explain a model's decisions, even if the model fails to recognise the condition in the image, the privatised model should be capable of preserving the glaucoma features that led to the score assigned by the recognition network, without accentuating them by approximating the image's real glaucoma score.

In this experiment, we change the generator's loss function, in order to approximate the glaucoma score assigned to the original images by the glaucoma recognition network. The generator's new loss function is represented in Equation 6.12.

$$L_G = E_{(I,M)\sim p_d(I,M)}[-\lambda_1 D^1(G(I)) - \lambda_2 D^2(G(I))\log(U) + \lambda_3 D^3(I)\log(D^3(G(I))) + \lambda_4 KL(q(f(I) \mid I)||p(f(I))) + \lambda_5 (I \times M - G(I) \times M)^2] \tag{6.12}$$

In this experiment, we trained the model twice, changing the parameter relative to the glaucoma score preservation, λ_3. The model was trained with the Adam optimiser with learning rate of $2e^{-5}$. The parameters used were $\lambda_1 = 0.4$, $\lambda_2 = 1$, $\lambda_4 = 0.002$ and $\lambda_5 = 0.001$. Table 6.9 contains the results. The glaucoma score in the results refers to the glaucoma recognition network's accuracy in recognising the glaucoma score assigned to the original image. Since approximating the image's glaucoma ground truth is no longer the network's goal, we no longer include the respective metric in the results.

Table 6.9: Results of experiment to approximate glaucoma score assigned by glaucoma recognition network to the original image.

Dataset	**Identity Recognition**	**Max Identity Score**	**Average KL Divergence**	**Glaucoma Recognition**
Original testing set (baseline)	89.71%	88.22%	4.24	100.00%
Generated set without considering identity (baseline)	2.35%	52.41%	3.26	92.65%
Privatised set from previous experiment	2.35%	31.85%	2.49	89.41%
Privatised set with $\lambda_3 = 0.6$	2.06%	**30.78%**	**2.39**	85.88%
Privatised set with $\lambda_3 = 2$	**0.88%**	33.15%	2.53	**91.47%**

From the results, we can see that the privatised set with $\lambda_3 = 2$ achieved better results in terms of privacy, with lower identity recognition accuracy, and in terms of preservation of explanatory value, with higher glaucoma recognition accuracy, than the privatised set with $\lambda_3 = 0.6$. Comparing these results with the previous experiment, we can see that the results are slightly better in terms of identity recognition accuracy and similar regarding the other privacy-related metrics. Regarding glaucoma recognition, the privatised set with $\lambda_3 = 2$ achieved better results than the

previous experiment. Figure 6.24 presents a visual representation of this experiment's results. We can see improvements in the glaucoma features preserved in these images, as there is a clearer boundary between the highlighted parts and the remaining parts of the images, in comparison to the previous experiments. In the privatised set where $\lambda_3 = 2$, the relevant parts are highlighted more strongly, with a stronger resemblance to the original image's features. In terms of image quality, the set with $\lambda_3 = 0.6$ seems to present slightly better results with more clearly defined eye structures, highlighting the trade-off between realism and explanatory evidence.

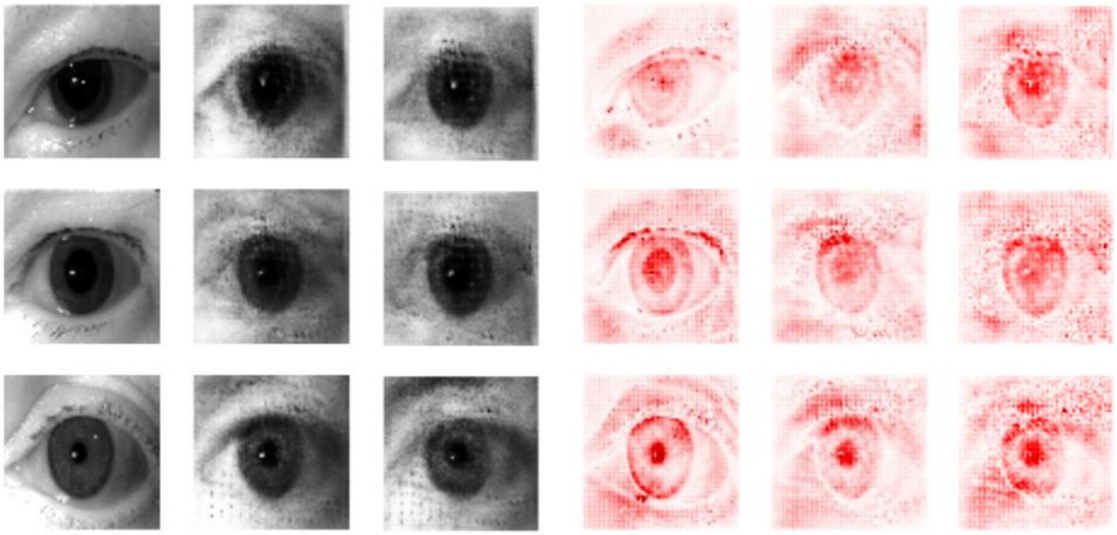


Figure 6.24: Results of approximating glaucoma score in the original image instead of ground truth. The first three columns correspond to the original images and their privatised versions with $\lambda_3 = 0.6$ and $\lambda_3 = 2$, respectively. The last columns are the saliency maps obtained with Deep Taylor of the original and privatised images, respectively.

With this experiment, we achieved a model capable of privatising eye images while preserving glaucoma-relevant features as they are in the original images. Using the glaucoma score identified by the glaucoma recognition network allows to preserve the explanatory features closer to what they are in the original images, since it does not accentuate features to force them into a class that may not coincide with the class recognised by the recognition network.

6.4 Main Conclusions

In this chapter, we defined a framework for the privatisation of case-based explanations. The framework is composed of three modules:

- **Generative Module**: The generative module is responsible for the generation of intelligible images. This module is composed of a deep generative model capable of transforming an image into its intelligible privatised version. In our privacy-preserving model, we used as the generative model a WGAN-GP network whose generator is a convolutional VAE.

- **Privacy Module**: The privacy module is responsible for removing the identity from an image. This module is secured by an identity recognition network which guides the privatisation process and ensures that no identity is leaked in the privatised image. In our model, we used a multi-class recognition network that preserves privacy at the level of the whole database by promoting an uniform identity distribution as the classification score assigned to the privatised images.

- **Explanatory Module**: The explanatory module is responsible for ensuring the privatised image's explanatory value, through the explicit preservation of relevant explanatory features. To identify relevant features, one can use *post hoc* interpretability methods capable of generating saliency maps to reconstruct the relevant features in the privatised image. In our model, we used Deep Taylor to obtain relevant features that should be preserved. In case the explanation retrieval model used to obtain the explanation before privatisation has a well-defined similarity measure to compare explanation candidates, it should be possible to directly use this measure to approximate features of interest.

To obtain a network that satisfies the requirements of privatised case-based explanations regarding realism, privacy and explanatory evidence, we performed several incremental changes to the PPRL-VGAN model [20]. Figure 6.25 introduces the architecture of the novel privacy-preserving model, highlighting the alterations that occurred to the PPRL-VGAN model in this process.

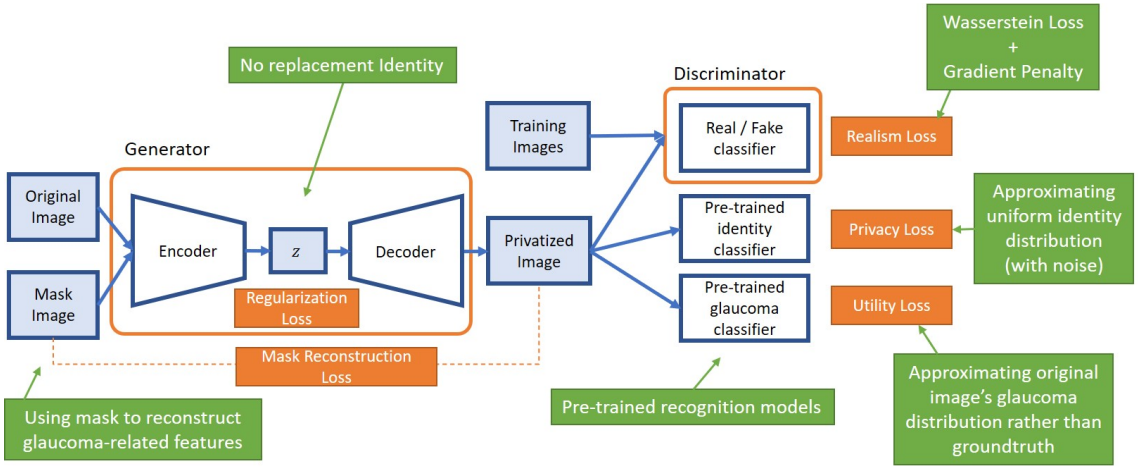


Figure 6.25: Illustration of the model's architecture, highlighting the differences between the initial PPRL-VGAN model and the accomplished privatisation model.

The loss functions used for the discriminator and generator are reflected in Equation 6.13 and Equation 6.14, respectively.

$$L_D = E_{I\sim p_d(I)}[D(G(I))] - E_{I\sim p_d(I)}[D(I)] + E_{\hat{x}\sim p_{\hat{x}}}[\lambda(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2] \tag{6.13}$$

$$L_G = E_{(I,M)\sim p_d(I,M)}[-\lambda_1 D^1(G(I)) - \lambda_2 D^2(G(I))\log(U) + \lambda_3 D^3(I)\log(D^3(G(I))) + \lambda_4 KL(q(f(I) \mid I)||p(f(I))) + \lambda_5 (I \times M - G(I) \times M)^2] \tag{6.14}$$

The main findings we came accross during the development of our privacy-preserving model are:

- We are able to privatise images by approximating a uniform identity distribution when we have a multi-class identity recognition network. Nonetheless, it is difficult to obtain a realistic image where the identity distribution assigned by an identity recognition network is very close to a uniform one. There is an evident privacy-intelligibility trade-off in this situation.

- Pre-training the recognition networks that guide the privatisation and the explanatory evidence preservation processes facilitates the generative network's training, improving the respective results.

- Using a GAN as the generative network for privatisation may lead to a mode collapse problem, where various privatised images generated for different identities look identical.

- WGAN-GP is an efficient generative model to solve mode collapse and improve image quality while also guaranteeing privacy.

- A ResNet VAE as the generator in the generative network is much harder to train than a simple convolutional VAE, as it is a significantly deeper network, leading to unstable training and lower-quality images. Using a ResNet VAE is not a good solution to improve image quality.

- A UNET architecture as the generator in the generative framework encourages the preservation of all features from the original image, leading the model to perform an adversarial attack on the identity recognition network which does not actually privatise the images.

- We can preserve explanatory evidence by reconstructing the general parts in the privatised images that contain a disease, such as the eye iris. However, this process unnecessarily preserves parts of the images that are unrelated to the disease recognition task and that needlessly contribute to an identity leak.

- We can preserve explanatory evidence by reconstructing disease-relevant features through saliency maps obtained with interpretability techniques. In this case, the masks should be binary, as these result in a higher preservation of explanatory features in the images.

- Providing glaucoma masks directly to the generative network promotes even further the preservation of explanatory features as they are in the original images and facilitates the generative network's task, resulting in higher-quality images.

- Using the original image's real disease label to preserve the privatised image's disease-related class may accentuate or even alter disease-related features in the privatised images. Instead, we should use as ground truth the disease scores assigned to the original images by a disease recognition network.

The trade-off between privacy, realism and explanatory evidence was present throughout the whole experimental process. We often came accross situations where to improve one of this modules, the others had to be sacrificed. In the final results, the dimension that was sacrificed the most seems to be image quality. When we remove one of the other dimensions, the image quality gets better. For instance, when we remove the privacy dimension, the images used can be the original images, which possess the highest quality. When we remove the explanatory evidence dimension, we have the results obtained in Chapter 6.2.1, which possess higher quality than the final results in Chapter 6.3.4, as can be seen in Figure 6.26.

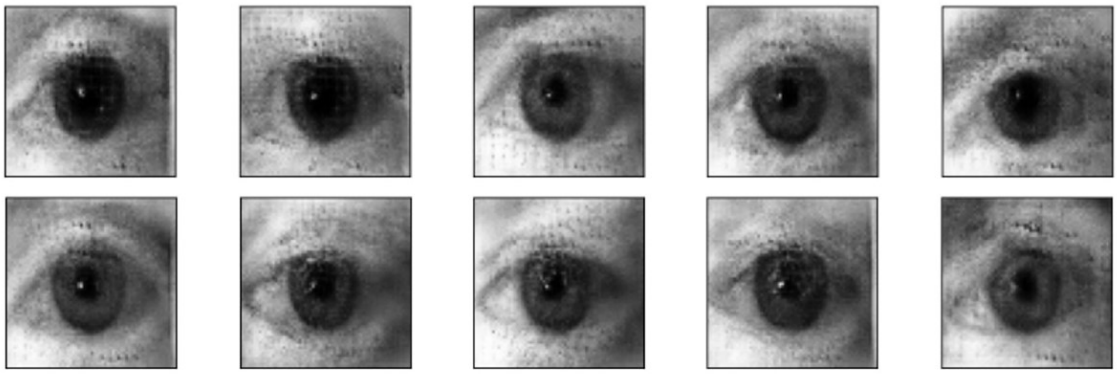


Figure 6.26: Comparison between images when we consider the preservation of explanatory evidence (first row) and when we do not (second row).

In comparison with the original PPRL-VGAN model, our model preserves the privacy of all subjects in the dataset, through the privacy loss term which approximates the privatised images' distribution to a noisy uniform distribution. Furthermore, our model has a higher capacity to preserve the original image's explanatory evidence through the reconstruction of relevant features, using interpretability saliency maps. Our model only lacks in terms of realism, as the PPRL-VGAN model generated higher-quality images.

By applying changes in an incremental manner, we obtained several networks at each point of the development process, where each one can be more appropriate than the others for certain privatisation problems, depending on the context. For example, if preservation of explanatory evidence as in the original image is not necessary, we could simply use the network obtained in Chapter 6.2.1, which has achieved the best results in terms of balancing image quality and privatisation.

Due to time constraints, hyperparameter selection was done manually, restricting the optimisation of hyperparameters and limiting the quality of the model and its results.

To conclude, we developed a privacy-preserving model capable of preserving privacy and explanatory evidence. This method can be used to privatise case-based explanations, enabling the use of these explanations in real-world contexts that deal with sensitive data, such as in the medical scene, given that the degree of realism in the images is acceptable to medical experts. This model is more adequate to use when the data contains many images per identity, as it uses a multi-class identity recognition network which would be difficult to train when there is a lack of images per subject. For cases when the data lacks images per patient, we should use a siamese identity recognition network instead of a multi-class one. The application of this privacy-preserving model to the use of a siamese identity recognition network is explored in the next chapter.

Chapter 7

Privacy-Preserving Model with Siamese Identity Recognition

In the previous chapter, we achieved a framework for the privatisation of case-based visual explanations that uses a multi-class identity recognition network. In a medical scenario, image acquisition might be a complicated process, especially if the images intend to show internal factors in the human body. As such, medical datasets often contain very few images per identity. In such cases, training a multi-class recognition network becomes unfeasible. To combat this issue, instead of a multi-class recognition network, we can use a siamese recognition network. This chapter describes the experimental work we developed to adapt the model achieved in the previous chapter to contexts with lack of data per identity, using a siamese identity recognition network. We start by introducing the developed siamese network and then we perform experiments by using it in the privacy-preserving model.

Similarly to the previous chapter, the experiments in this chapter used the Warsaw dataset and were performed in Keras [24], with Tensorflow backend [2]. As evaluation networks for the privacy-preserving model we use the previously introduced identity and glaucoma recognition networks, together with the siamese identity recognition model which will be introduced in the following section. To train the privacy-preserving model, we use the Adam optimiser with a learning rate of $2e^{-5}$.

7.1 Siamese Recognition Network

As explained in Chapter 4.3.2, a siamese network compares two images to check whether these have the same identity. This network computes a score that semantically compares two images, which can be used to reduce the similarity between two images in regards to identity. In our model, the comparison between images uses Euclidean distance. Our siamese model's architecture is shown in Figure 7.1. The network calculates the Euclidean distance between the embeddings of

the image pair, extracted using identical CNN networks. The Euclidean distance is then used in the loss function to approximate embeddings from the same identity and to increase the distance between embeddings from different identities.

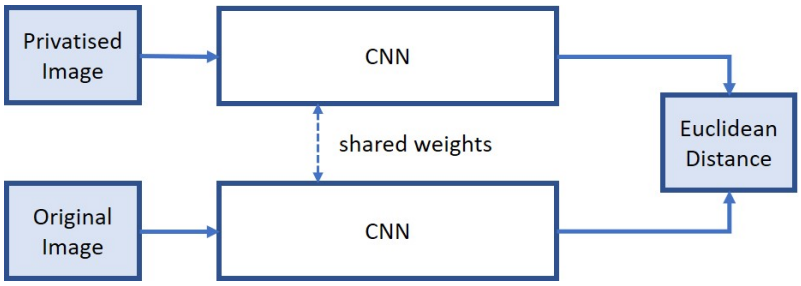


Figure 7.1: Architecture of the developed Siamese Network.

We used the CNN architecture shown in Figure 7.2. The network is composed of four blocks of strided convolutions which downsample the image, followed by global average pooling and a fully connected layer to obtain image embeddings.

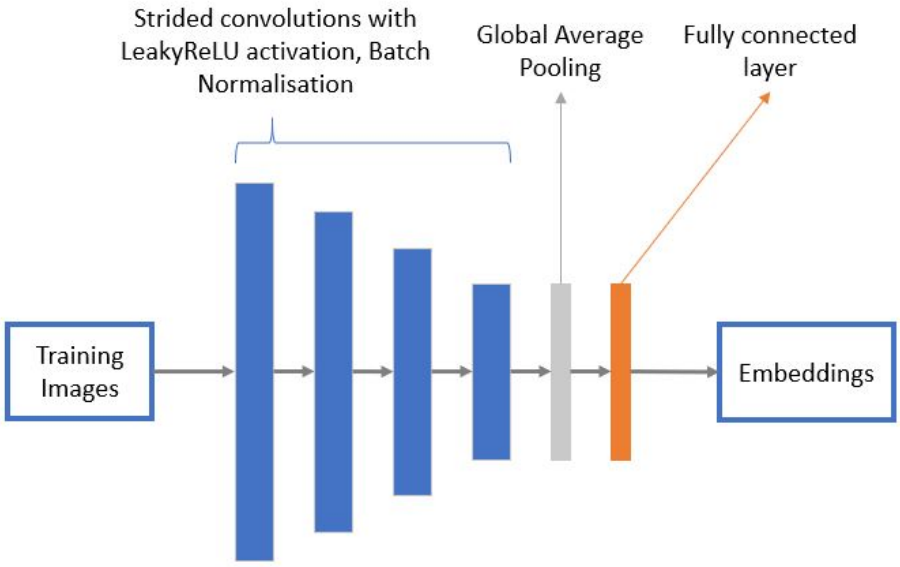


Figure 7.2: Architecture of the CNN model for the Siamese Network.

We trained the network with the contrastive loss, using a margin m of 1, as can be seen in Equation 7.1, where D represents the euclidean distance between the embeddings of two images. Y corresponds to the label assigned to the pairs, which is 1 when pairs share the same identity and 0 otherwise. This loss has two terms, one that is applied to image pairs from the same identity, where the distance between the respective images' embeddings is reduced. The second term only applies to image pairs from different identities and ensures that the distance between the images' embeddings increases. The margin m ensures that the distance between two images from different

identities does not get much bigger than the margin, since independently of how big this distance is, if it is larger than the margin then the loss is zero.

$$L = \frac{1}{2} \times Y \times D^2 + \frac{1}{2} \times (1 - Y) \times [\max(0, m - D)]^2 \tag{7.1}$$

As there is a much higher number of image pairs from different identities, when compared to images sharing an identity, we balanced the paired data so that there is the same number of real pairs (from the same identity) and fake pairs (from different identities). Since there is a low number of images for each patient, we paired each image with all other images sharing the same identity. In a second experiment, we added Gaussian noise to the images, to make the network more robust. One example of a noisy image, used to train the network, can be seen in Figure 7.3.

Figure 7.3: Image with noise to train robust siamese network.

In the experiments, we used the Adam optimiser with a learning rate of $1e^{-3}$, which was the optimiser with which we achieved the best results. The network was trained for 12 epochs with the data without noise, and for 22 epochs with the data with noise.

Our goal with this network is to ensure that the distance between embeddings of images from the same identity are close while embeddings of images from different identities are distant. This network will then be used to maximise the distance between the original image and the privatised image, to achieve privatisation. As such, to evaluate the results, we evaluate the average distance between images from the same identity and from different identities. Furthermore, we can consider the overall average distance between two images as the value used to classify two images as belonging to the same or to different identities. As such, distance values that are lower than the overall average distance can be considered real pairs (from the same identity), while distance values that are higher than the overall average can be considered fake pairs (from different identities). Using this concept, we can calculate the accuracy of the network. The results of this experiment are available on Table 7.1.

From these results, we arrived at the conclusion that the network trained with noisy images achieved better results, with a higher accuracy and distance between the average distances of real and fake pairs. Both networks seem to be worse at identifying pairs with images from different subjects than identifying pairs from the same identity.

Table 7.1: Results from experiment with siamese identity recognition network.

Dataset	Average Distance	Real Pairs		Fake Pairs		Distance between Averages	Accuracy
		Average Distance	% under Average	Average Distance	% over Average		
No noise	0.616	0.349	93.55%	0.884	81.16%	0.535	87.36%
With noise	**0.777**	0.389	93.97%	1.165	83.30%	**0.776**	**88.64%**

7.2 Privacy-preserving model with Siamese Recognition Network

In this section, we aim to develop a privacy-preserving model suitable for scenarios with a low number of images per subject. To achieve this, we use the siamese recognition network developed in the previous section to guide the privatisation process in the privacy-preserving model achieved in Chapter 6. We limited the amount of images we show per experiment. More visual results can be seen in Appendix A.2.

7.2.1 Replacing multi-class identity recognition model by siamese network

The architecture of the novel model which uses a siamese identity recognition network to privatise images is shown in Figure 7.4.

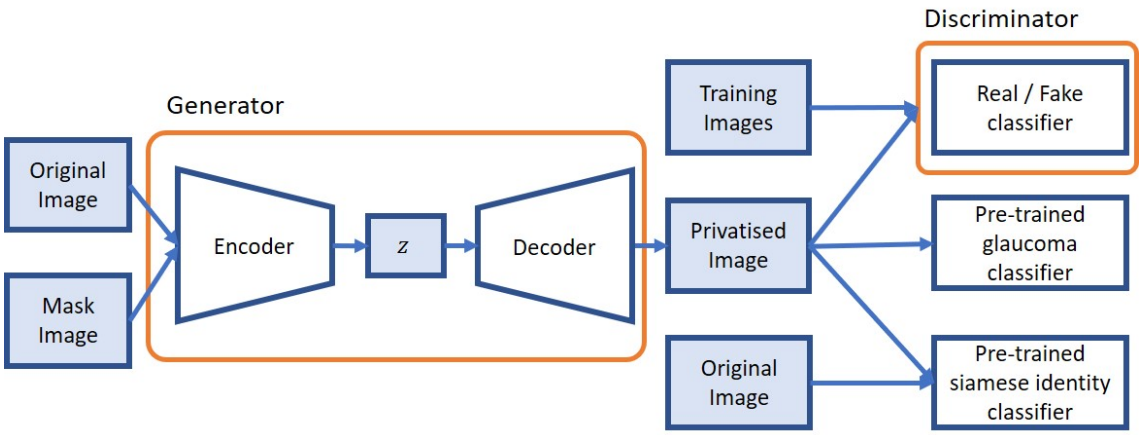


Figure 7.4: Architecture of the privacy-preserving model that uses a siamese identity recognition network.

The discriminator and generator's loss functions are represented in Equation 7.2 and Equation 7.3, respectively. In the generator's loss function, $ED(x,y)$ represents the Euclidean distance between x and y.

$$L_D = E_{I\sim p_d(I)}[D(G(I))] - E_{I\sim p_d(I)}[D(I)] + E_{\hat{x}\sim p_{\hat{x}}}[\lambda(||\nabla_{\hat{x}}D(\hat{x})||_2 - 1)^2] \tag{7.2}$$

$$L_G = E_{(I,M)\sim p_d(I,M)}[-\lambda_1 D^1(G(I)) + \lambda_2[\max(0, m - ED(I, G(I)))]^2 + \lambda_3 D^3(I)\log(D^3(G(I))) + \lambda_4 KL(q(f(I) \mid I)||p(f(I))) + \lambda_5(I \times M - G(I) \times M)^2] \tag{7.3}$$

In the first experiment, we trained the model assigning different weight values to the identity term in the generator's loss function (λ_2), in order to optimise this network's privacy degree, since the range of values obtained with the contrastive loss is different than the one obtained with the cross entropy loss used in the previous chapter. As the remaining parameters, we used: $\lambda_1 = 0.4$, $\lambda_3 = 2$, $\lambda_4 = 0.002$, $\lambda_5 = 0.001$.

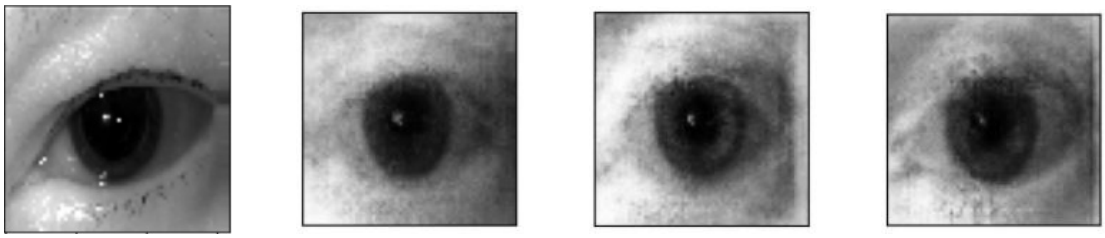
To evaluate this model's results, we kept the multi-class identity recognition as an evaluation network to test whether it can recognise the original identity in this model's privatised images. To further evaluate the generative network's privatisation capabilities, we include the average distance measured between the original images and their privatised versions, and the percentage of pairs whose distance surpasses the average value 0.777 obtained in Section 7.1. This percentage over average value can be interpreted as the siamese identity recognition network's accuracy in recognising that the images do not belong to the same subject. To evaluate the preservation of glaucoma-related features, we use the model's glaucoma recognition network, like in the previous chapter. Table 7.2 contains the results obtained with this model, with the best results highlighted in bold.

Table 7.2: Results of privacy-preserving model with siamese identity recognition.

Dataset	**Identity Recognition**	**Average Distance**	**% over Average**	**Glaucoma Recognition**
Original testing set (baseline)	89.71%	1.165	83.80%	100.00%
Privatised set: $\lambda_2 = 1$	4.12%	0.997	73.24%	87.94%
Privatised set: $\lambda_2 = 5$	**0.88%**	**1.165**	**89.41%**	88.53%
Privatised set: $\lambda_2 = 10$	1.18%	1.162	**89.41%**	**90.00%**

The set with the least privacy degree ($\lambda_2 = 1$) is the one that shows the worst results in terms of privacy, with relatively high identity recognition accuracy, and glaucoma recognition. The remaining sets contain similar values in the privacy-related metrics, with both surpassing the baseline in regards to the percentage of image pairs whose distance is larger than the average distance. The privatised set which was assigned the highest privacy degree ($\lambda_2 = 10$) provided the best results in terms of glaucoma recognition accuracy. The three privatised sets seem to be similar in regards to image quality and intelligibility, as can be seen in Figure 7.5, which shows an example of visual results taken from each privatised set.

To evaluate the results in regards to preservation of explanatory evidence, we illustrate in Figure 7.6 the result of applying Deep Taylor to the privatised images. We only included in these results the two privatised sets with higher privacy degree, since the privatised set with $\lambda_2 = 1$ does not fulfill the privacy requirements we seek in a privacy-preserving model. The saliency maps



(a) (b) (c) (d)

Figure 7.5: Examples of results obtained with generative model that contains a siamese recognition network. (a) is the original image and (b-d) are the respective privatised versions with the differing λ_2 values of 1, 5 and 10, respectively.

obtained with the privatised set with $\lambda_2 = 5$ seem to resemble the saliency maps of the original image more closely than the ones where $\lambda_2 = 10$.

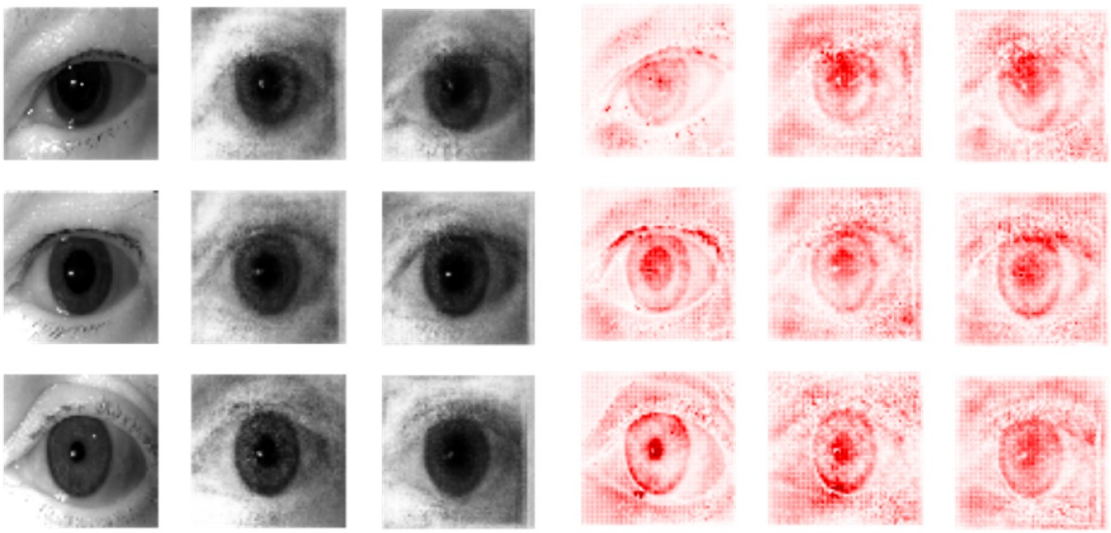


Figure 7.6: Results of using siamese identity recognition network in the privacy-preserving model. The first three columns correspond to the original images and their privatised versions with $\lambda_2 = 5$ and $\lambda_2 = 10$, respectively. The last columns are the saliency maps obtained with Deep Taylor of the original and privatised images, respectively.

The most significant problem in this approach is that we do not guarantee that the privatised images protect the privacy for all data subjects. In this model's loss function, we only guarantee that the privatised image looks sufficiently different from the original image so that they are not recognised as belonging to the same subject. With this loss function, the generative model could learn to generate images that are similar to other subjects in the training data. Figure 7.7 illustrates an example of the identity leak that occurs in this model, where a privatised image looks like an image from a subject from the dataset.

In this experiment, we obtained a privacy-preserving model that was capable of generating

(a) (b) (c)

Figure 7.7: Example obtained with siamese generative model that shows an identity leak. (a) is the original image, (b) is the privatised version of (a) and (c) is an image from the subject whose identity was recognised by the multi-class identity recognition network.

privatised images that hide the original subject's identity and that preserve explanatory evidence. However, this method still has an identity leak issue, where the model sometimes generates images that resemble a subject from the dataset.

7.2.2 Distancing privatised images from all subjects in the data

Since replacing the identity recognition model by a siamese network only guarantees the privacy of the original subject, we altered the model in order to ensure a higher degree of privacy at the whole dataset's level. In the new model, we distance the privatised image from every subject in the dataset through an additional term in the generator's loss function, represented in Equation 7.4. In this equation, $ED(x,y)$ represents the euclidean distance between x and y, m represents the margin used in the contrastive loss function, N is the number of subjects in the dataset and I_N is a training image from the subject N. The full generator's loss function is represented in Equation 7.5. In terms of the model's architecture, the only change is that training images are also fed to the siamese generative network, as illustrated in Figure 7.8. Since it would be impratical and time-consuming to compare every training image with the privatised image, we randomly choose one image from each subject at each epoch.

$$L_{privacy} = E_{(I)\sim p_d(I)}[\lambda_2[\max(0, m - ED(I, G(I)))]^2 + \lambda_6 \sum_{i=0}^{N} \frac{[\max(0, m - ED(G(I), I_N))]^2]}{N}] \tag{7.4}$$

$$L_G = E_{(I,M)\sim p_d(I,M)}[-\lambda_1 D^1(G(I)) + \lambda_2[\max(0, m - ED(I, G(I)))]^2 + \lambda_3 D^3(I)\log(D^3(G(I))) + \lambda_4 KL(q(f(I) \mid I)||p(f(I))) + \lambda_5(I \times M - G(I) \times M)^2 + \lambda_6 \sum_{i=0}^{N} \frac{[\max(0, m - ED(G(I), I_N))]^2]}{N}] \tag{7.5}$$

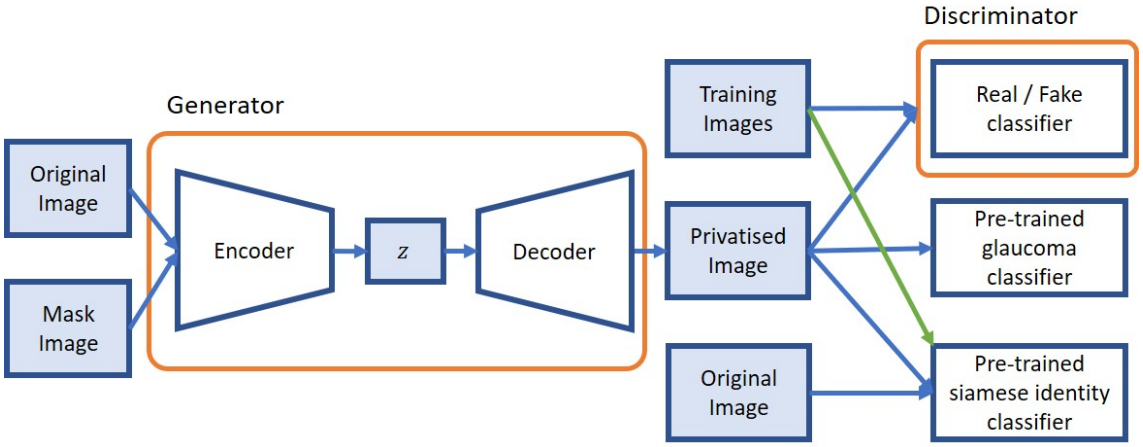


Figure 7.8: Architecture of the privacy-preserving model with a siamese identity recognition network that ensures privacy for all subjects.

We trained this model using the Adam optimiser with a learning rate of $2e^{-5}$. As parameters, we used: $\lambda_1 = 0.4$, $\lambda_2 = 0.001$, $\lambda_3 = 2$, $\lambda_4 = 0.002$ and $\lambda_5 = 0.001$. We tested the network at different levels of overall privacy, by training the network with different values in the parameter λ_6. This model was trained for 870 epochs when $\lambda_6 = 5$ and for 900 epochs when $\lambda_6 = 10$. The results are expressed on Table 7.3. In these results, although the accuracy in the multi-class identity recognition network has increased, the results using the siamese identity recognition network have improved, with significantly higher distance between original images and their privatised versions and higher percentage of pairs being recognised as fake (higher percentage over average). The results have also improved in terms of glaucoma recognition accuracy using the new model.

Table 7.3: Results of experiment using siamese identity recognition network to achieve privacy in the entire dataset.

Dataset	**Identity Recognition**	**Average Distance**	**% over Average**	**Glaucoma Recognition**
Original testing set (baseline)	89.71%	1.165	83.80%	100.00%
Privatised set from previous experiment	**0.88%**	1.165	89.41%	88.53%
Privatised set with overall privacy: $\lambda_6 = 5$	3.53%	1.255	90.59%	91.18%
Privatised set with overall privacy: $\lambda_6 = 10$	1.76%	**1.299**	**92.65%**	**91.47%**

The visual results obtained with this network are represented on Figure 7.9. In these results we can see that the glaucoma-relevant features are being preserved, with the regions with stronger highlights being around the same spots in the original and privatised images.

To evaluate privacy at the whole dataset's level, we used the siamese network to obtain the distance between embeddings from the privatised images and one image from each of the subjects

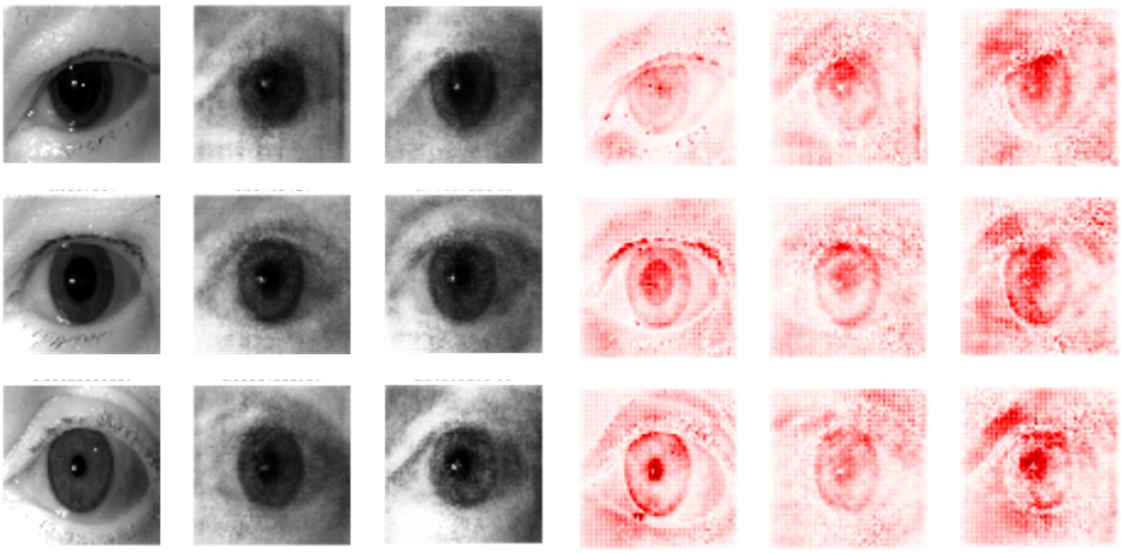


Figure 7.9: Results of using siamese identity recognition network in the privacy-preserving model to achieve overall privacy. The first three columns correspond to the original images and their privatised versions with $\lambda_6 = 5$ and $\lambda_6 = 10$, respectively. The last columns are the saliency maps obtained with Deep Taylor of the original and privatised images, respectively.

in the dataset. In the results, illustrated in Table 7.4, we included an additional metric other than the ones previously used to evaluate privacy using the siamese identity recognition model: average number of real pairs. This metric evaluates the average number of images from the dataset that the siamese recognition network recognised as belonging to the same identity as the privatised image.

Table 7.4: Results regarding privacy at the whole dataset's level in privacy-preserving model with siamese identity recognition.

Dataset	**Average Distance**	**% over Average**	**Average Number of Real Pairs**
Privatised set from previous experiment	1.113	78.91%	22.99
Privatised set with overall privacy: $\lambda_6 = 5$	1.320	89.81%	11.11
Privatised set with overall privacy: $\lambda_6 = 10$	**1.368**	**91.99%**	**8.74**

In the privatised set from the previous experiment, each privatised image was recognised with the same identity as 23 images. In this experiment, where we explicitly made the network generate images that are distant from images belonging to each identity in the training set, the value for the average number of real pairs significantly decreases. Furthermore, the values for average distance between privatised images and images from the dataset, and for the accuracy in recognising that the pairs did not belong to the same identity (% over Average) are significantly higher in this experiment. If we provide a bigger overall privacy degree, by increasing the parameter λ_6, the results in terms of privacy improve even further. As such, this experiment succeeded in improving

privacy at the whole dataset's level.

In this experiment, we conclude that we can preserve privacy for the whole dataset by distancing the privatised images to each subject in the dataset. We obtained a privacy-preserving model that uses a siamese recognition model to guide the privatisation process. This method, unlike the one achieved in Chapter 6 using a multi-class identity recognition network, can be used in contexts with low number of images per subject, which are very common in the medical scene.

7.3 Main Conclusions

In this chapter, we applied the framework defined in the previous chapter to the use of a siamese identity recognition network to guide the privatisation process, enabling the application of the developed privacy-preserving model to scenarios where there are few images per identity. This experimental process started with the definition of the siamese recognition network.

The siamese network semantically compares two images and calculates a distance between them in regards to identity. The bigger the distance, the smaller the likelihood of the images sharing the same identity. During the development of the siamese network, we concluded that augmenting the dataset by adding gaussian noise to the images improves the network's results.

Regarding the privacy-preserving framework from the previous chapter, we made alterations solely to the privacy module, responsible for the privatisation process. The explanatory module and the generative module were kept as they were in the previous chapter. Previously, the privacy module was composed of a multi-class identity recognition and a privacy loss term which approximated the identity distribution in the privatised images to a uniform one, ensuring privacy at the whole dataset's level. In this chapter, we replaced the multi-class identity recognition network by a siamese network. To ensure the privacy of the subject in the original image, we ensure in the privacy term of the loss function that the privatised image must be distant from the original image when it comes to privacy. We observed that solely comparing the privatised image with the original one was not enough to guarantee privacy for all the subjects in the dataset. To ensure privacy at the whole dataset's level, we added an additional term to the loss function where we compare the privatised image against images from all subjects in the training dataset, ensuring that these are distant in regards to identity features.

In comparison to the privacy-preserving model with multi-class identity recognition, this new model achieved higher-quality images, as can be seen in Figure 7.10, which shows examples of images generated with each model. The model that uses a siamese identity recognition network is capable of generating a more realistic eye structure surrounding the iris. When it comes to privacy, it is difficult to compare both models, as they have been evaluated using different metrics that are associated with the respective identity recognition models. Regarding preservation of explanatory evidence, both methods succeeded in preserving glaucoma-relevant features. The advantage of the model that uses a siamese recognition network is its applicability to a wider range of problems, as it can be applied in scenarios where the number of images per subject is low.

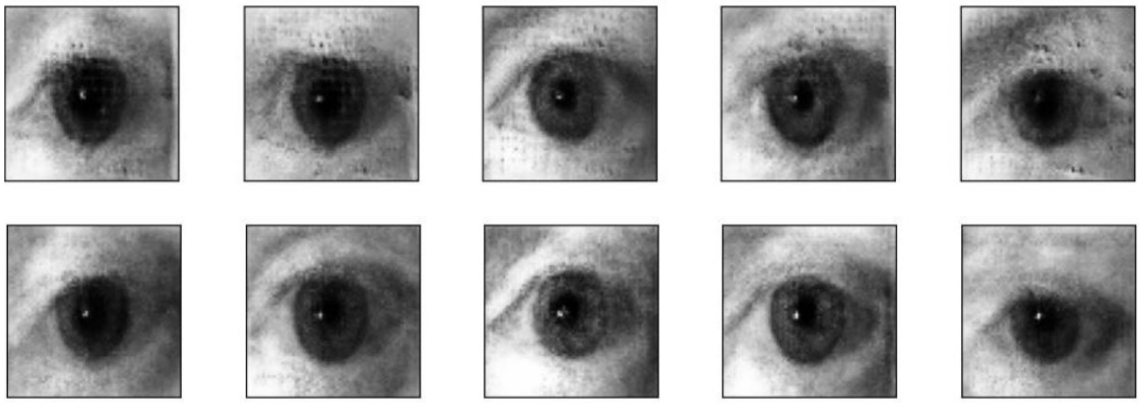


Figure 7.10: Example of images generated by the privacy-preserving models. The first row contains images generated by the model that uses multi-class identity recognition. The second row contains images generated by the model that uses a siamese identity recognition network.

To conclude, in this chapter we successfully developed a privacy-preserving model capable of preserving privacy, realism and explanatory evidence. This model can be applied to the privatisation of case-based explanations in the medical scene, where there is usually a small number of images per patient.

Chapter 8

Counterfactual Generation

The difficulty in classifying an ambiguous diagnostic case derives from the difficulty in identifying the decision boundary between the presence and absence of a disease. In such cases, factual explanations help build trust in a decision as their comparison with the original case allows a medical expert to understand the disease-related features that lead to the Deep Learning model's prediction. By introspecting these features, a medical expert can gain confidence and additional insights regarding a decision. However, the factual explanation does not make it clear where the decision boundary between the presence and absence of a disease is, as it only shows features that are relevant for one particular class. To make this boundary more evident, it is relevant to provide the changes to the medical image's features that would lead to a different prediction. This goal is easily achievable with counterfactual explanations, which highlight the alterations that an image has to suffer to change the prediction made by a Deep Learning model.

In the previous chapters, we focused our work on the privatisation of visual explanations for case-based interpretability. In this chapter, we will focus on applying the developed models to the generation of counterfactual explanations. Our goal is to obtain a network capable of privatising factual explanations and generating counterfactuals based on the privatised factuals. To do so, we add a counterfactual generation module to the privacy-preserving models developed in the previous chapters. The following sections present the experiments done to generate counterfactuals with each of the privacy-preserving models previously developed.

Like in the previous chapters, the experiments in this chapter used the Warsaw dataset and were performed in Keras [24], with Tensorflow backend [2].

8.1 Counterfactual Generation in Privacy-Preserving Model with Multi-class Identity Recognition

We developed a network capable of simultaneously generating privatised factual and counterfactual explanations, given an image that serves as a factual example. We used the network developed

in Chapter 6, capable of privatising factual explanations, as a base and added a component for the generation of counterfactuals. As can be seen in Figure 8.1, we added a decoder to the base model's generator, whose main purpose is to generate the counterfactuals. In this network, the factuals decoder is trained as in Chapter 6.

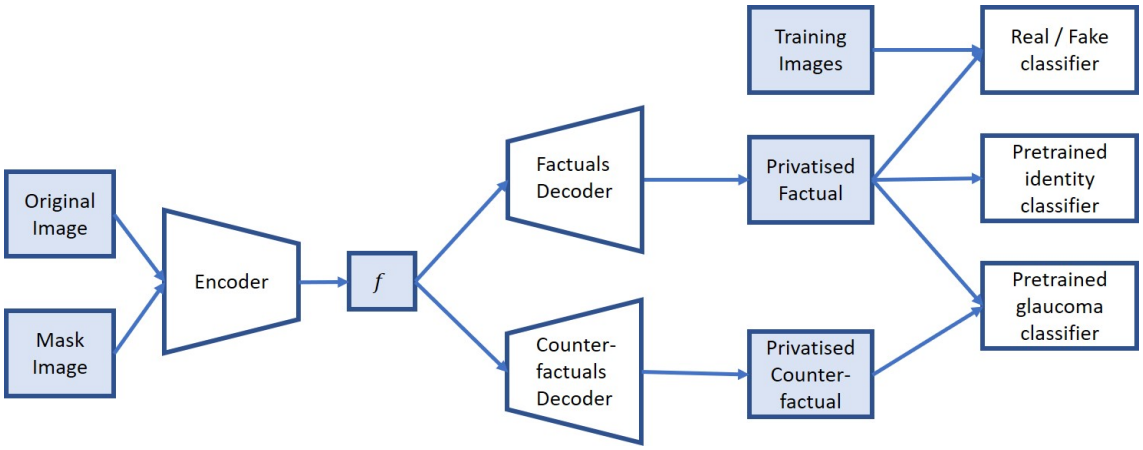


Figure 8.1: Architecture of the model to generate privatised factual and counterfactual explanations.

Our approach to the generation of counterfactuals consists of making the smallest number of alterations to the factuals to change their predicted class. As such, the counterfactuals decoder is trained to minimise the pixelwise distance between the factual and counterfactual explanations while changing the original image's glaucoma-related prediction. The pixelwise distance minimisation is achieved using the squared L2 Normalisation loss between the factual and the counterfactual images. The loss function used in the generator to update the counterfactuals decoder, C, is represented in Equation 8.1. In this equation, F represents the factuals decoder, and D^3 represents the glaucoma recognition network.

$$L_C = E_{(I)\sim p_d(I)}[\lambda_1 (F(I) - C(I))^2 + \lambda_2 D^3(I) \log(1 - D^3(C(I)))] \tag{8.1}$$

Regarding the training approach, we first train the factuals decoder with the counterfactuals decoder freezed, so that the squared L2 Normalisation used to approximate the counterfactual explanations to the factuals does not impact the performance of the factuals decoder. After training the factuals decoder up to a point that it does not improve after a predefined set of epochs, we freeze both the encoder and the factuals decoder, and train the counterfactuals decoder. In this process, we do transfer learning, where we reuse the weights of the factuals decoder on the counterfactuals decoder in order to facilitate its generative task.

We trained the counterfactuals decoder for 333 epochs, where it achieved the best results. The network was capable of inverting the glaucoma score assigned to the original images with 90.59%

accuracy. The visual results are represented in Figure 8.2. Since the counterfactuals and the factuals are identical, we expose the differences between these explanations in a saliency map. To calculate the differences between the two images, we used an implementation of Structural Similarity Index Measure (SSIM) [116] provided by scikit-image [114]. SSIM is a metric that evaluates the similarity between two images, taking into consideration structure, luminance and contrast. It is represented in Equation 8.2. In this Equation, μ represents an image's mean intensity, σ is the standard deviation used to estimate contrast, and C_1 and C_2 are constants to avoid instability.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{8.2}$$

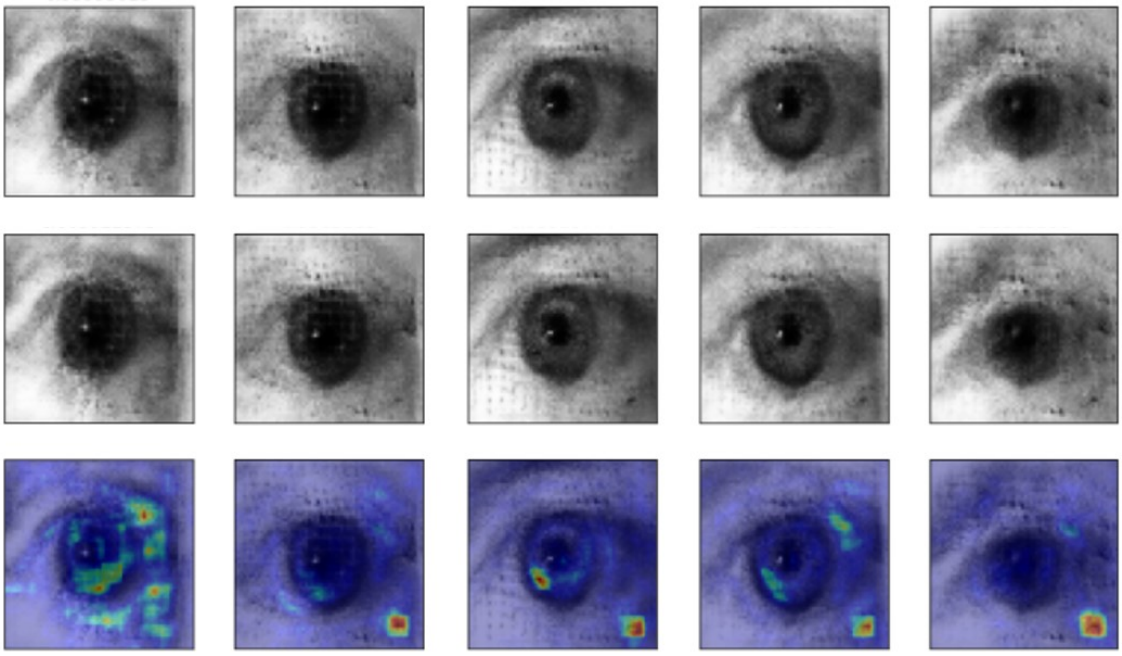


Figure 8.2: Results from counterfactual generation in the generative model with multi-class identity recognition. The first row represents the factuals, the second row represents the counterfactuals and the third row contains a saliency map that highlights the differences between the factuals and counterfactuals.

In the visual results, the differences between factuals and counterfactuals show that the image regions that are being altered are not located in the iris. Furthermore, there is a higher incidence of altered zones outside the iris in images where the factuals contain glaucoma and the counterfactuals do not, which is the case represented in the first column in Figure 8.2.

Since glaucoma-related features are expected to be located in the eye iris, the counterfactual explanations generated in this experiment lack quality and may be confusing to an explanation consumer. In order to obtain more plausible explanations, we used the Deep Taylor glaucoma

masks to promote changes in the glaucoma-related features and to preserve the eye structure, which should not contribute to the glaucoma recognition process. The resulting generator loss function that is used to update the counterfactuals decoder is represented in Equation 8.3.

$$L_C = E_{(I,M)\sim p_d(I,M)}[\lambda_1(F(I)\times(1-M)-C(I)\times(1-M))^2+ \\ \lambda_2 D^3(I)\log(1-D^3(C(I)))] \tag{8.3}$$

Using the glaucoma masks in the loss function, the network was trained for 560 epochs. We obtained 90.29% in glaucoma recognition accuracy when using the glaucoma recognition network to detect the original images' inverted glaucoma score in the counterfactuals. As can be seen in Figure 8.3, the differences between the counterfactuals and factuals are located mostly in the iris. However, there are still some parts outside the iris that are being changed.

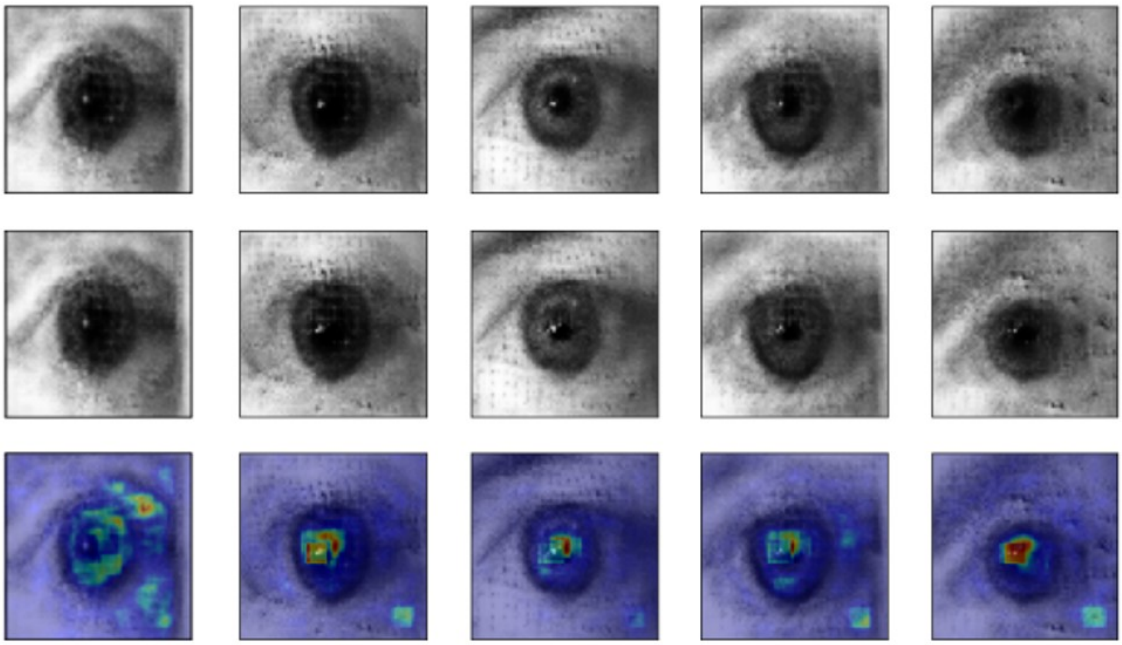


Figure 8.3: Results from counterfactual generation in the generative model with multi-class identity recognition, using glaucoma masks to guide the alteration of glaucoma-related features. The first row represents the factuals, the second row represents the counterfactuals and the third row contains a saliency map that highlights the differences between the factuals and counterfactuals.

To force the network to make changes more located on the iris, we can increase the weight we associate to the reconstruction of the eye structure in the loss function (λ_1). However, we verified that increasing this variable leads to a decrease in the model's capacity to invert the glaucoma classification in the counterfactuals. As can be seen in Figure 8.4, the higher the value in λ_1, the lower the accuracy of the glaucoma recognition network at identifying the original image's inverse glaucoma label in the counterfactuals. We can conclude from the analysis of this graph that solely

altering the features inside the glaucoma masks might not be enough to change the class of the factuals.

Figure 8.4: Graph that shows the results of changing the parameter λ_1, which promotes the similarity between the factual and the counterfactual explanations.

To conclude, in this experiment, we added a counterfactual generation module to the privacy-preserving model with multi-class identity recognition. Using glaucoma masks to guide the preservation of the eye structure allowed to obtain counterfactuals whose main differences from the factuals are located mostly in the iris.

8.2 Counterfactual Generation in Privacy-Preserving Model with Siamese Identity Recognition

In a second experiment, we applied the method for the generation of counterfactual explanations to the privacy-preserving model that uses a siamese identity recognition network, developed in Chapter 7. Similarly to the previous experiment, we add a decoder for the generation of counterfactuals, which is trained with the loss function mentioned in Equation 8.3. The counterfactuals decoder was capable of inverting the glaucoma classification assigned to the original image with 90.88% accuracy. Some examples of results can be seen in Figure 8.5. The differences between the counterfactuals and factuals are located in the iris.

Compared with the results from the previous section, these results seem to have higher incidence of changes inside the iris. Since there are no differences between the methods used in both networks to generate counterfactuals, we can infer that the quality of the counterfactual explanations may depend on the image quality, as the images from this model, which possess higher quality, provide better counterfactuals.

To conclude, in this experiment, we added a counterfactual generation module to the privacy-preserving module with siamese identity recognition. The counterfactual module was capable

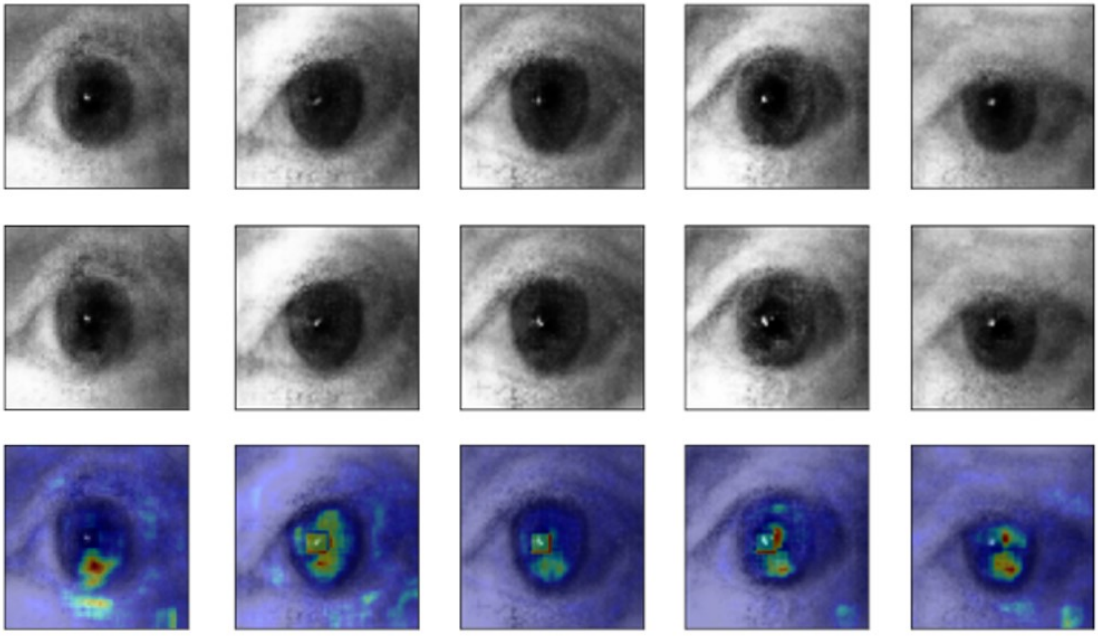


Figure 8.5: Results from counterfactual generation in the generative model with siamese identity recognition. The first row represents the factuals, the second row represents the counterfactuals and the third row contains a saliency map that highlights the differences between the factuals and counterfactuals. The first two columns represent eyes where the factual image presents glaucoma and the counterfactual does not. In the remaining columns, the counterfactuals have glaucoma and the factuals do not.

of generating high-quality counterfactuals that invert the glaucoma classification of the original image with high accuracy.

8.3 Main Conclusions

In this chapter, we added a module responsible for the generation of counterfactual explanations to our privacy-preserving models. This module is composed of a decoder that is trained to perform the least possible changes to the factual explanations that change their glaucoma-related class. Since the goal of the counterfactual explanations is to understand the changes in an image that would lead to a change in the glaucoma prediction, we provide saliency maps with the differences between the factual and counterfactual explanations, so that these changes are clear and easy to spot. We verified that we can obtain higher-quality counterfactuals by using Deep Taylor saliency maps containing glaucoma-related features to guide the feature alteration process. Using these saliency maps as masks, the changes to the images were mainly located in the eye iris.

In this experiment, we generated counterfactuals in the context of a binary classification problem. In this case, the counterfactual's target class is evident, as it is the class that was not predicted by the classification network. In multi-class problems, there may be several potential target

classes. As such, for multi-class classification tasks, we can either generate a counterfactual per class or choose a target class. To solve this problem, we thought of three possible solutions:

- **Multiple counterfactual decoders**: To generate a counterfactual per class, we could have multiple counterfactual decoders, one for each class. In this scenario and considering the training strategy that we used in the experimental work, we could train the factuals decoder as in the privacy-preserving models and then use the respective model's weights in each of the counterfactual decoders. On inference, the result from the counterfactual decoder responsible for the target class of the original image should be either ignored or interpreted as a factual explanation. This solution has various problems, as the resulting network could be significantly deeper, depending on the number of classes in the classification task, and take longer to train.

- **Conditional counterfactual decoder**: One other solution to generate a counterfactual per class could be to use a conditional counterfactual decoder. This decoder would receive as input the original image's latent representation and the target disease-related class, and generate an image similar to the factual explanation that the disease recognition network would classify as the target class. On inference, this model could be used to generate a counterfactual for every class or solely for one specific class, providing to the explanation consumer the ability to control the image's target class.

- **Selective counterfactual decoder**: We could also develop a network that automatically selects the counterfactual's target class as the class that would lead to the least differences between the factual and counterfactual explanations. Nonetheless, this last network would be limited in terms of controllability, as it would not allow the explanation consumer to control the counterfactual's target class. Compared with this decoder, the conditional counterfactual decoder offers greater controllability. The advantage in this network would be that the explanation consumer could obtain the class that would lead to the least changes in the original image.

Finally, the generation of counterfactual explanations allows to make the boundary between two classes clear in an image, allowing a medical expert to obtain more insights and gain more confidence in a Deep Learning model's predictions.

Chapter 9

Conclusions

Deep Learning has achieved great results in computer vision tasks. However, these models' usability in real-world contexts is hindered by the respective lack of interpretability. Among various interpretability techniques developed to garner trust in Deep Learning models' decisions, case-based interpretability stands out as it produces intuitive explanations by example. The problem in case-based explanations is that these may violate the privacy of subjects when used in contexts where the data exposes someone's identity, like in medical imaging. The main goal of this dissertation was to extend the use of case-based explanations to contexts with sensitive data by privatising case-based explanations, considering a medical scenario.

To achieve our goal, we reviewed the current literature in the topics of deep generative models, interpretability and visual privacy. Through the analysis of case-based interpretability methods and privacy-preserving methods, we arrived at three requirements that privacy-preserving case-based explanations must fulfill: anonymity, realism and explanatory evidence. In terms of anonymity, the images should not expose the identity in the original image nor any other identity present in the database. In terms of realism, the images should be realistic enough to be accepted and comprehended by the target explanation consumers, like medical experts in our case. Regarding explanatory evidence, the explanatory features should be preserved exactly as they are in the original images. Preserving the original image's task-related class does not guarantee the preservation of the original image's exact semantic features.

Current privacy-preserving models are not sufficiently developed to be applied to the domain of case-based explanations, as they do not ensure the preservation of explanatory evidence. Some models try to preserve general features, independently from a task while others use classification models to ensure that the generated image is classified as the original image. However, none of the methods guarantee the preservation of explanatory evidence as it is in the original image, which is a critical aspect of privacy-preserving explanations. Furthermore, some models present weaknesses regarding image quality and privacy, which are the other two characteristics needed to guarantee the usefulness of privacy-preserving explanations.

To fill the gaps in the literature, we developed two privacy-preserving models, bearing in mind the previously mentioned requirements. Our approach to the development of the privacy-preserving models consisted of making incremental changes to an existing privacy-preserving model: PPRL-VGAN [20]. We started by analysing the model's limitations. Then, we tackled these limitations from three perspectives: improving privacy, improving realism and improving preservation of explanatory evidence. This incremental approach allowed us to obtain several networks, at each step of the development process.

Since the PPRL-VGAN model uses a multi-class identity recognition network, we started by developing a privacy-preserving model that uses the same multi-class network. However, this privatisation model was not compatible with the characteristics of medical data, as the low number of images per patient hinders the training of a multi-class recognition network. To apply our model to a wider range of problems in the medical scene, we developed a second privacy-preserving model using a siamese identity recognition network.

Table 9.1 compares the privacy-preserving methods with the ones we developed. In this table, we consider that methods that directly use images from other patients in the privatisation process do not guarantee privacy for all subjects, even if these guarantee K-Anonymity. With our models, we were capable of preserving the explanatory evidence while also guaranteeing privacy. By preserving the exact semantic features of the original images, we guarantee the explanatory value of the resulting privatised explanations. The most significant limitation in our models is the lack of image quality, especially in the model that uses a multi-class recognition network, whose images contain a high degree of noise.

Table 9.1: Comparison between privacy-preserving methods.

Privacy-preserving model	**Preserves original image's class**	**Preserves original image's exact semantic features**	**Guarantees privacy for all data subjects**	**Generates high-quality images**	**Applicable to data with few images per subject**
CLEANIR [23]			×	×	
R^2VAE [33]				×	
PP-GAN [119]			×	×	×
SGAP [84]			×		×
PPRL-VGAN [20]	×			×	
DeepObfuscator [62]	×		×		
Ours (multiclass)	×	×	×		
Ours (siamese)	×	×	×	×	×

The most significant challenge we came across during the experimental work was to manage the trade-off between realism, privacy and explanatory evidence. We verified that enhancing one of these characteristics in the images usually meant that the other ones would be sacrificed. For instance, improving privacy would usually damage the image's intelligibility.

In addition to the privatisation of case-based explanations, we enhanced our privacy-preserving models with a counterfactual generation module, which provides counterfactual explanations similar to the privatised factual explanations. These counterfactuals help explain a model's decisions by highlighting the changes that should occur in an image so that its prediction changes. With this module, we were capable of not only privatising explanations, but also providing additional explanations that help the explanation consumer interpret a model's decisions.

As future work, the privacy-preserving models should be validated on more datasets, including medical data that does not serve as biometric data, to validate the results in data where identity recognition is more complex. Furthermore, the metrics used to evaluate the privacy-preserving models should be improved. Specifically, the models' evaluation should include metrics to objectively assess data quality and preservation of task-related features. In this work, these two dimensions have only been measured subjectively by analysing the visual results. Regarding the generation of counterfactual explanations, this module should be extended to multi-class classification tasks by developing multiple counterfactual decoders to obtain a counterfactual explanation for each class, a conditional counterfactual decoder, or a counterfactual decoder capable of selecting a target counterfactual class, as described in Chapter 8.3. Moreover, to optimise the quality of the privatised case-based explanations, privacy should be considered in the image retrieval process. In this work, privatisation is applied to case-based explanations after these have been retrieved. However, since the explanatory value of an image may differ from the explanatory value of its privatised version, the resulting privatised image may not be the best possible explanation for a case under analysis. By considering privacy in the image retrieval, we can optimise the case selected as an explanation. Finally, regarding the field of application of our work, although our main motivation for the development of a privacy-preserving model came from the medical field, there are other domains where privacy-preserving case-based explanations are relevant, such as presentation attack detection and forensics (e.g. sexual assault identification). As such, future work can also focus on adapting this work to other domains.

By merging three trending research fields in deep learning, this dissertation offers novel contributions to current research in the fields of deep generative models, interpretability, and privacy. We wrote a paper [76] which was published at the Interpretable Machine Learning in Healthcare workshop of ICML. In this paper, we highlight the need to privatise case-based explanations in the medical scene and compare deep learning and traditional privacy-preserving models (Chapter 5.3). Furthermore, we submitted a paper [75] presenting the novel privacy-preserving models developed in this dissertation to the Winter Conference on Applications of Computer Vision (WACV 2022). Finally, we are preparing a survey on case-based interpretability and visual privacy for a special issue on IEEE's Signal Processing Magazine [77], which covers the topics discussed in Chapters 3 and 4. These papers represent some of our contributions to the research community.

To conclude, this dissertation contributes towards improving the trust in deep learning algorithms for disease detection by providing privacy-preserving explanations by examples. Furthermore, the developed models enable the use of case-based explanations in the medical scene,

providing insights to support medical experts' decisions and enhance transparency in the decision-making process.

Appendix A

Visual Results from Privacy-preserving Models

In this appendix, we include more visual results obtained from the experimental work, to facilitate the evaluation of image quality.

A.1 Results from privacy-preserving model with multiclass identity recognition

The results in Figure A.1 correspond to the experiments in Chapter 6.1, which improves privacy preservation in the privacy-preserving model.

The results in Figure A.2 correspond to the Chapter 6.2.1, where we applied a WGAN-GP network to improve realism in the images and to solve the mode collapse problem. In Figure A.3, we expose results from Chapter 6.2.2, where we change the architecture of the generator.

Regarding the chapters about preservation of explanatory evidence, Figure A.4 contains the results from the experiments in Chapters 6.3.1 and 6.3.2, where we used iris segmentation masks and Deep Taylor masks in the generative model's loss function, to reconstruct relevant glaucoma-related features in the privatised images. Figure A.5 contains results from the experiments in Chapter 6.3.3, where Deep Taylor masks were given as input to the generative model to improve image quality and preservation of glaucoma-related features. Figure A.6 contains results from the experiments in Chapter 6.3.4 where we approximate the original image's glaucoma score in the privatised images.

A.2 Results from privacy-preserving model with siamese identity recognition

The results in Figure A.7 correspond to the experiments in Chapter 7.2, where we develop a privacy-preserving model that uses a siamese identity recognition network to privatise the images.

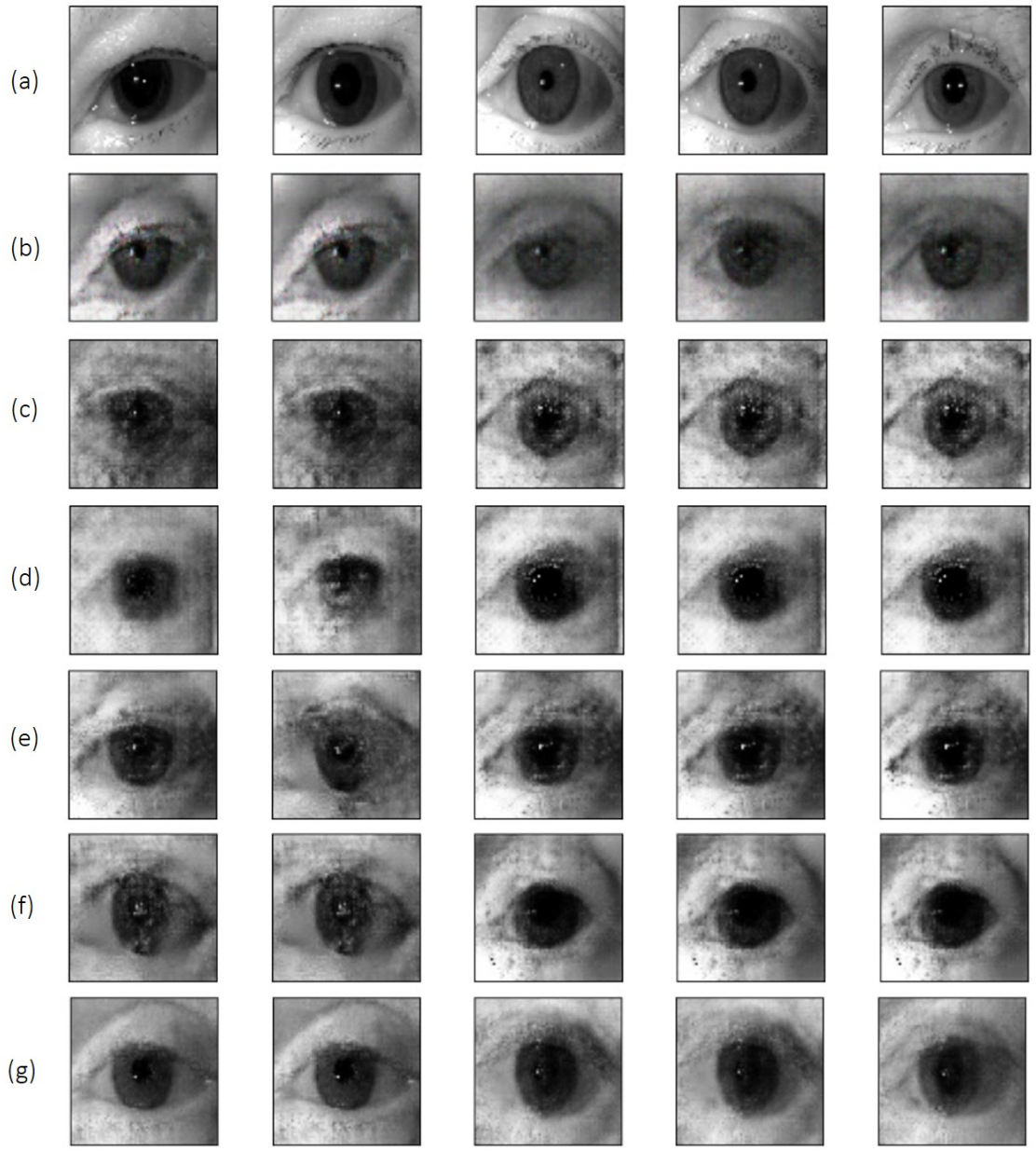


Figure A.1: Visual results obtained from the experiments that try to improve privacy in the privacy-preserving model. Each row exposes results from the privatised sets obtained in the experiment: (a) original image, (b) privatised set where we try to maximise the cross entropy between the identity distributions from the original and privatised images, (c) privatised set where we approximate a uniform identity distribution using cross entropy loss, (d) privatised set where we try to approximate a uniform identity distribution using KL loss, (e) privatised set with pre-trained recognition models obtained at 971 epochs, (f) privatised set with pre-trained recognition models obtained at 974 epochs, and (g) synthetic set that does not consider privacy (used as baseline).

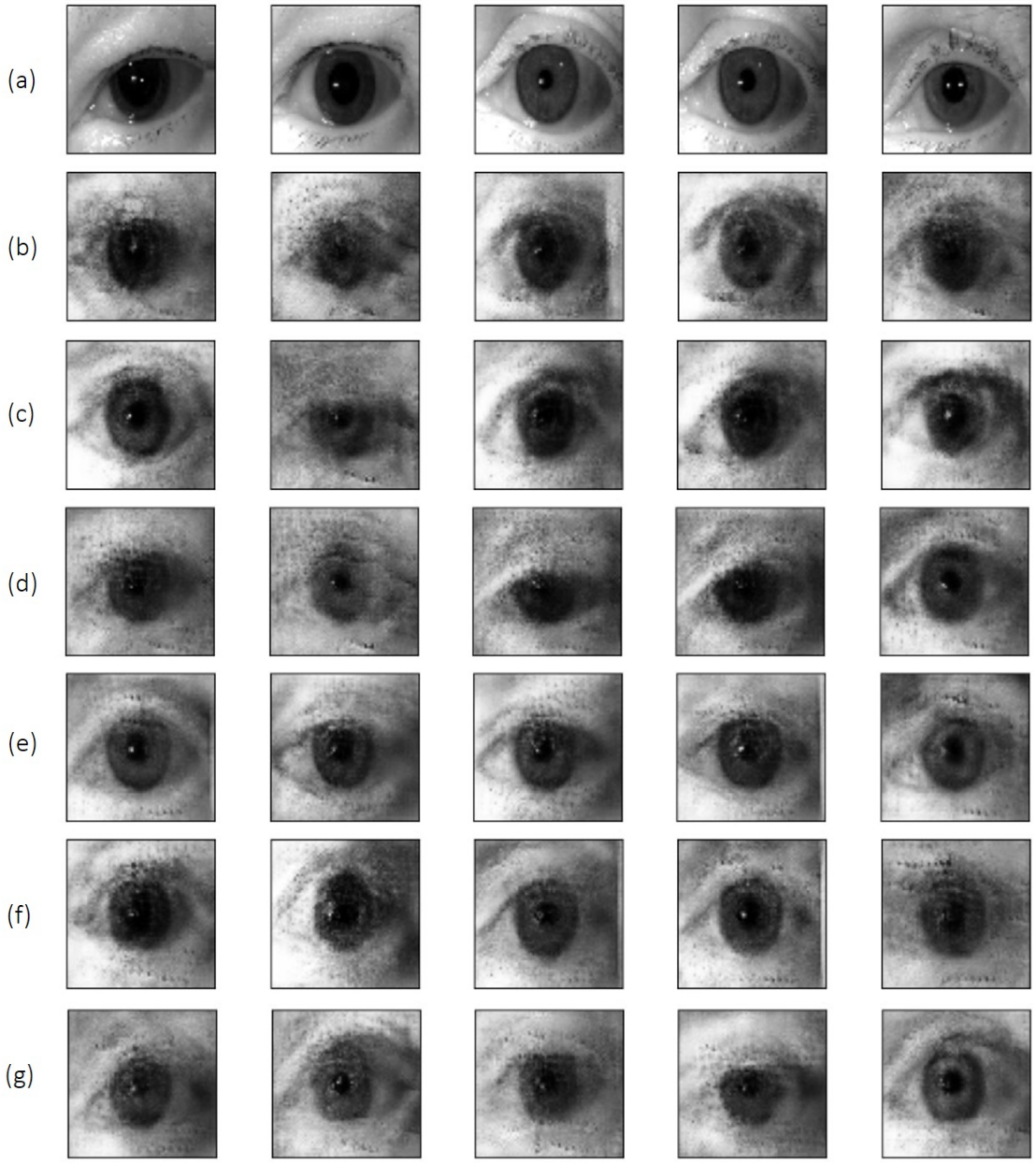


Figure A.2: Visual results obtained from the experiment using WGAN-GP. Each row exposes results from the privatised sets obtained in the experiment: (a) original image, (b) privatised set, (c) privatised set with noise, (d) privatised set with higher privacy degree, (e) privatised set with random uniform identity distribution, (f) privatised set with pre-trained generative network, and (g) privatised set with noise in latent representations and in uniform identity distribution.

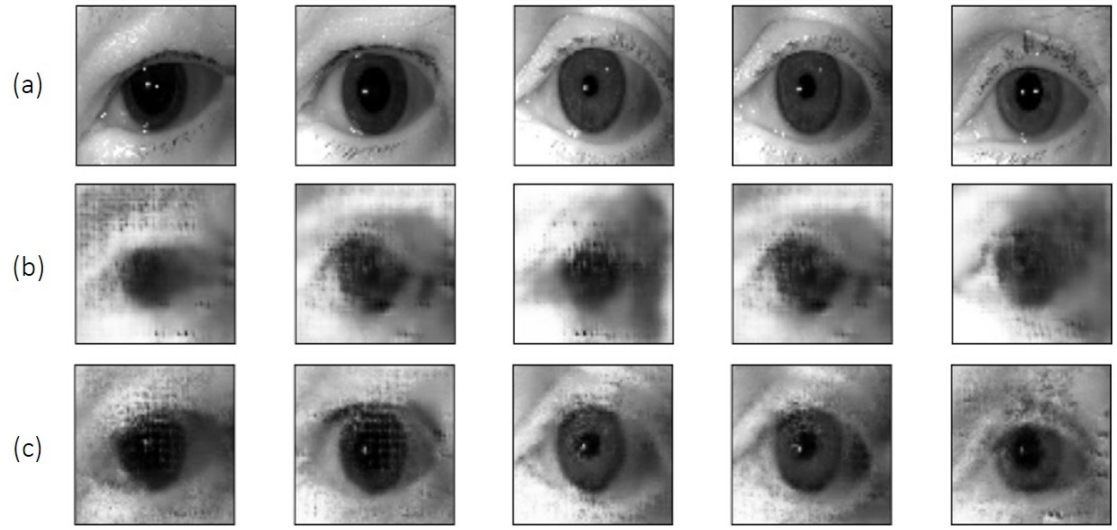


Figure A.3: Visual results obtained from the experiment using UNET and ResNet architectures in the generator. (a) corresponds to the original images. (b) and (c) are privatised versions of (a) using ResNet and UNET as generators, respectively.

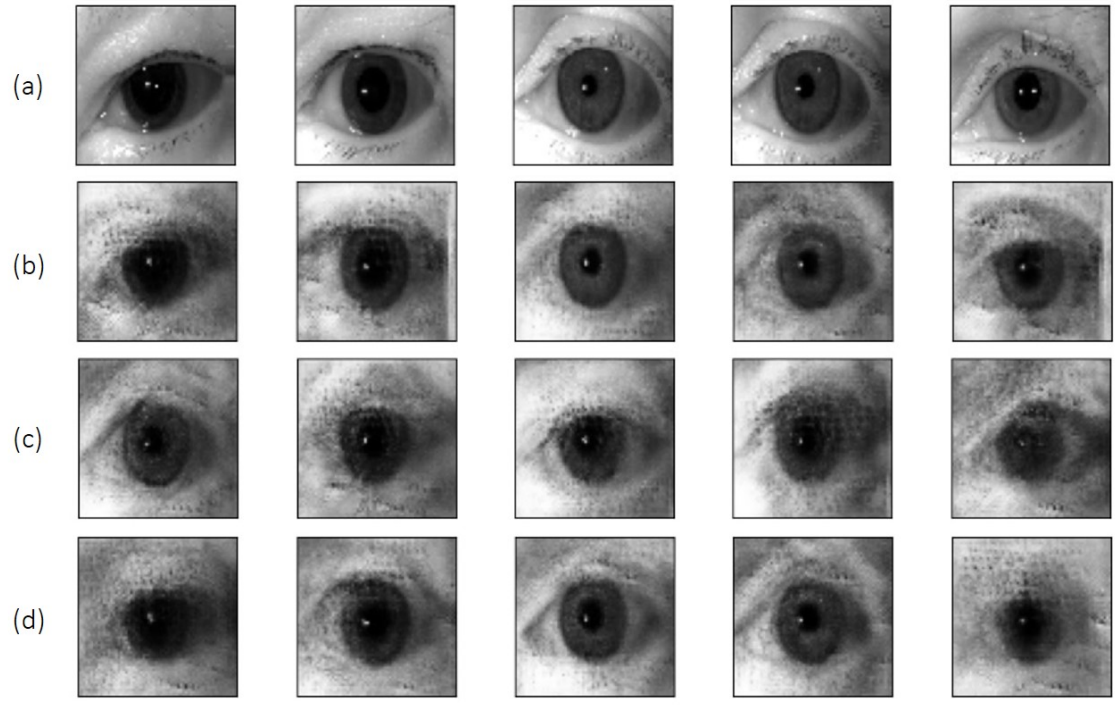


Figure A.4: Visual results obtained from the experiments to preserve explanatory evidence by reconstructing masks with glaucoma-related features, where the masks were only used in the loss function. (a) corresponds to the original image, (b) corresponds to the results of the experiment where we used iris segmentation masks, (c) are results from the experiment where we used non-binary Deep Taylor masks, and (d) are results from the experiment where we used binary Deep Taylor masks.

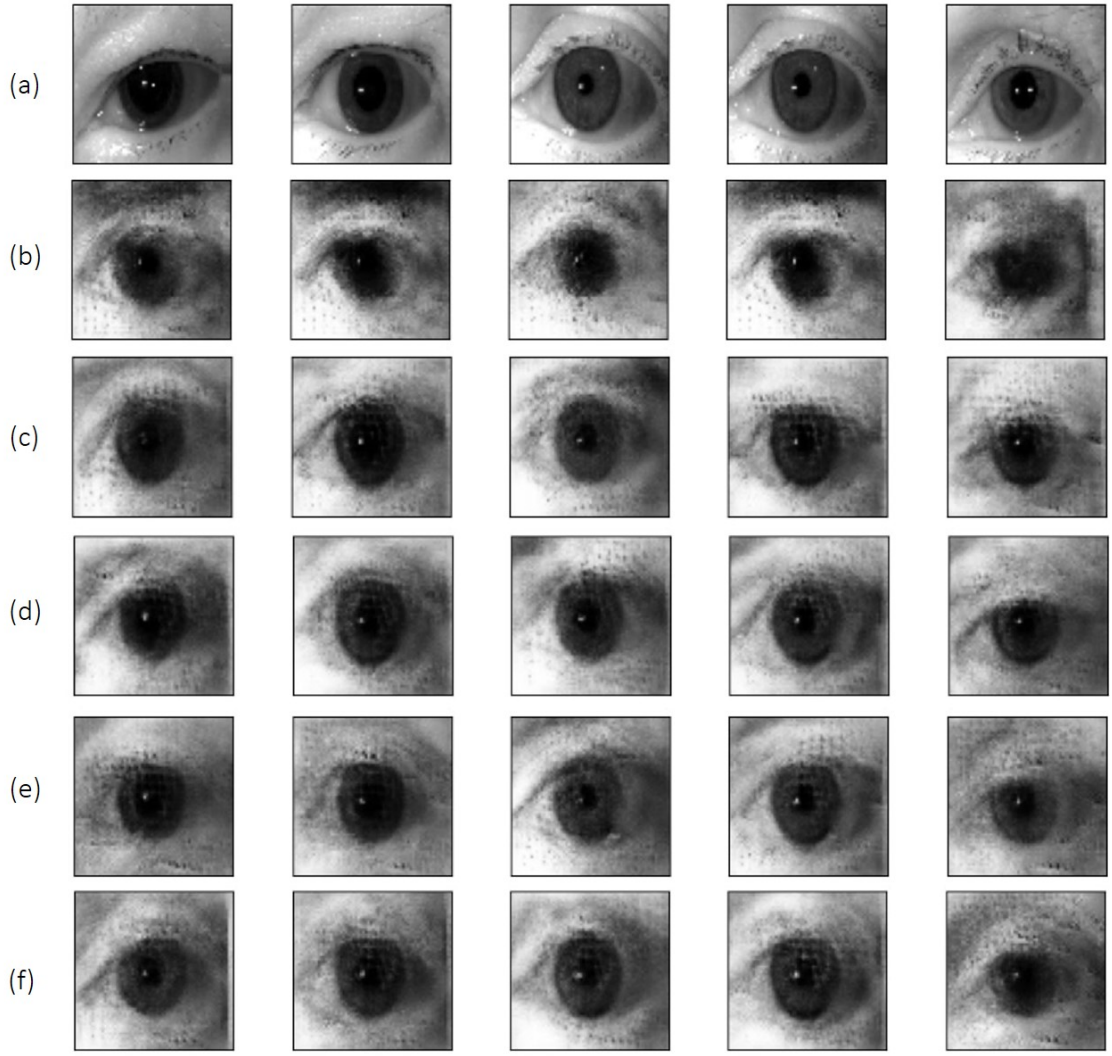


Figure A.5: Visual results obtained from the experiments to preserve explanatory evidence by reconstructing masks with glaucoma-related features, where the masks were introduced in the generative model. (a) corresponds to the original image. (b) are results from the generative model where concatenation between images and masks happened before these being introduced to the encoder. (c) are results from the generative model where concatenation between images and masks happened in the VAE's latent space. (d) are results from the generative model where concatenation between images and masks happened inside the encoder, after feature extraction. (e) and (f) are variations of (d) using as parameters $\lambda_3 = 2$ and $\lambda_5 = 0.0005$, respectively.

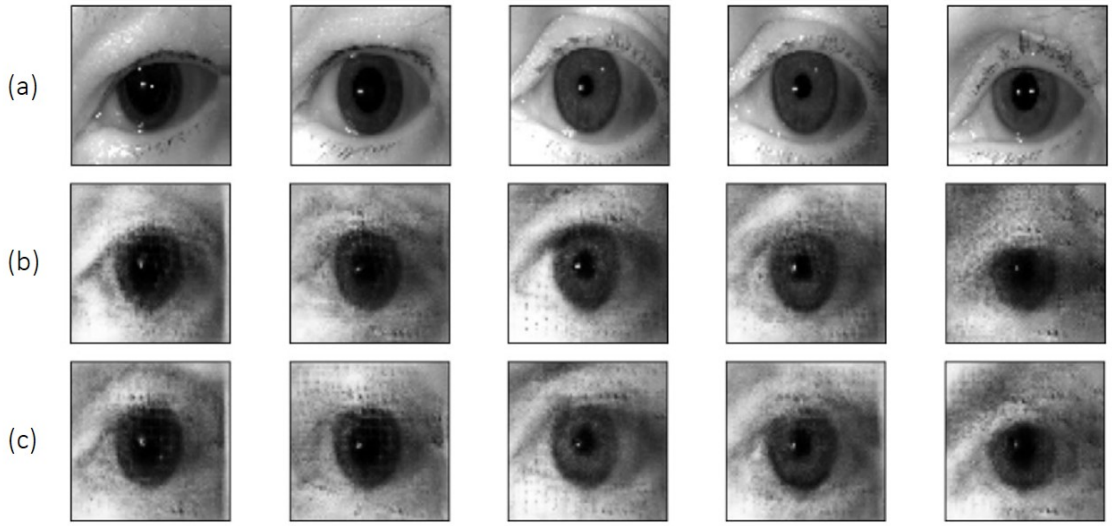


Figure A.6: Visual results obtained from the experiments to preserve explanatory evidence where we approximate the original image's glaucoma score. (a) corresponds to the original image. (b) and (c) are privatised versions of (a) where we use as parameters $\lambda_3 = 0.6$ and $\lambda_3 = 2$, respectively.

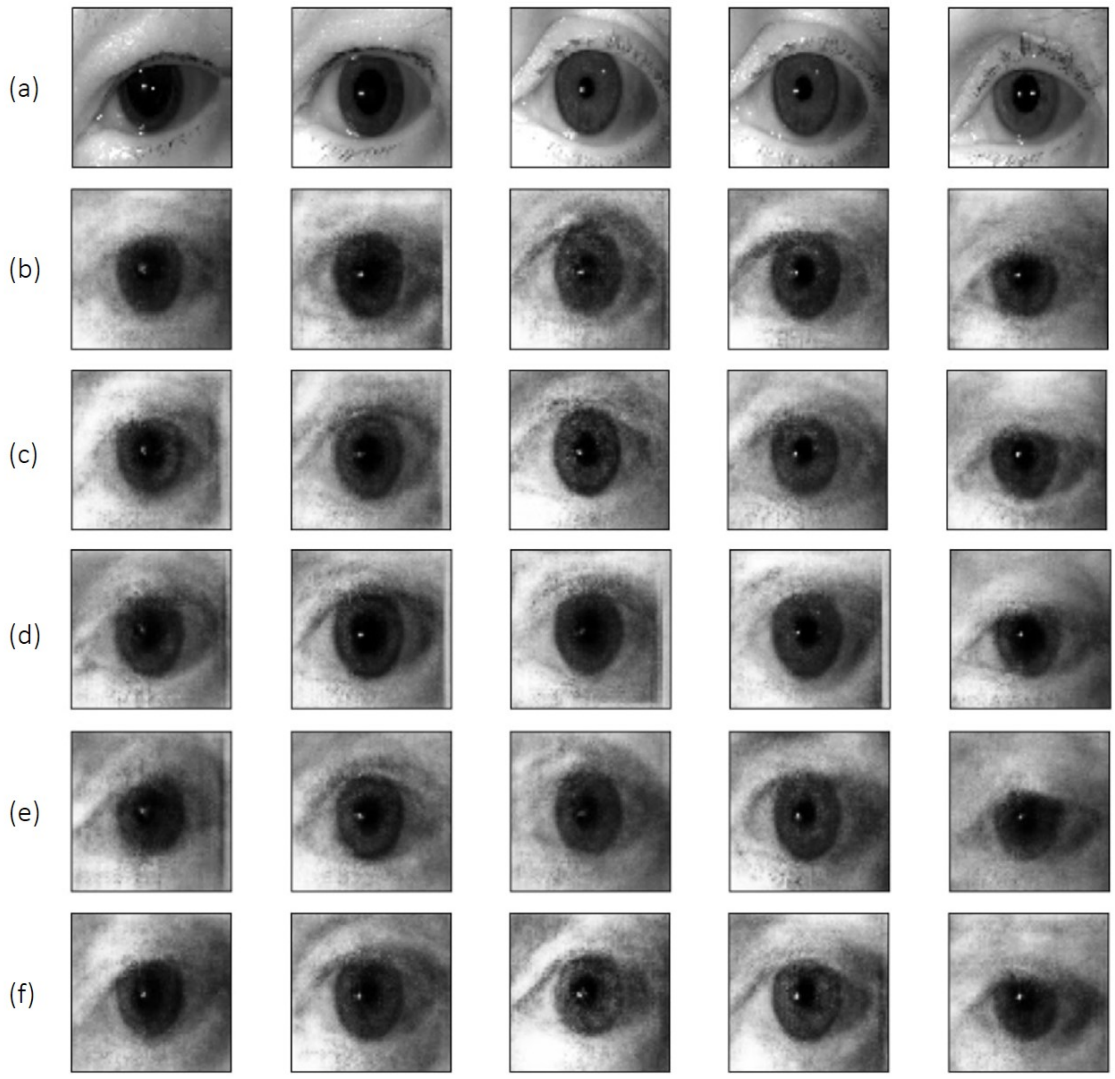


Figure A.7: Visual results obtained from the experiments with privacy-preserving model using siamese identity recognition. (a) corresponds to the original image. (b-d) are privatised versions of (a) where we do not guarantee the privacy of all subjects in the dataset, with $\lambda_2 = 1$, $\lambda_2 = 5$ and $\lambda_2 = 10$, respectively. (e) and (f) are privatised versions of (a) with the model where we guarantee privacy for all data subjects, with $\lambda_6 = 5$ and $\lambda_6 = 10$, respectively.

References

[1] Boston university: Privacy-preserving smart-room analytics, 2018. Last accessed January 2021. Available at vip.bu.edu/projects/vsns/privacy-smartroom/facial-expression-vgan.

[2] M. Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available at http://tensorflow.org/.

[3] N. Aifanti, C. Papachristou, and A. Delopoulos. The mug facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4, 2010.

[4] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K. Müller, S. Dähne, and P. Kindermans. innvestigate neural networks! *Journal of Machine Learning Research*, 20(93):1–8, 2019.

[5] G. An. The effects of adding noise during backpropagation training on a generalization performance. *Neural Computation*, 8(3):643–674, 1996.

[6] D. Aneja, A. Colburn, G. Faigin, L. Shapiro, and B. Mones. Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*, pages 136–153. Springer, 2016.

[7] P. Angelov and E. Soares. Towards deep machine reasoning: a prototype-based deep neural network with decision tree inference, 2020.

[8] P. Angelov and E. Soares. Towards explainable deep neural networks (xdnn). *Neural networks : the official journal of the International Neural Network Society*, 130:185—194, October 2020.

[9] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan, 2017.

[10] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar. Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graph.*, 27(3):1–8, August 2008.

[12] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[13] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, 2018.

[14] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019.

[15] 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. *International Journal of Pattern Recognition and Artificial Intelligence*, 7:25, 08 1993.

[16] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age, 2018.

[17] R. Caruana, H. Kangarloo, J. Dionisio, U. Sinha, and D. Johnson. Case-based explanation of non-case-based learning methods. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 212–5, 02 1999.

[18] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, Jul 2019.

[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 *Advances in Neural Information Processing Systems*, volume 32, pages 8930–8941. Curran Associates, Inc., 2019.

[20] J. Chen, J. Konrad, and P. Ishwar. Vgan-based image representation learning for privacy-preserving facial expression recognition, 2018.

[21] X. Chen, Y. Xu, D. W. Kee Wong, T. Y. Wong, and J. Liu. Glaucoma detection based on deep convolutional neural network. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 715–718, 2015.

[22] Z. Chen, Y. Bei, and C. Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, Dec 2020.

[23] D. Cho, J. H. Lee, and I. H. Suh. Cleanir: Controllable attribute-preserving natural identity remover. *Applied Sciences*, 10(3):1120, Feb 2020.

[24] F. Chollet et al. Keras, 2015. Software available at https://github.com/fchollet/keras.

[25] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[26] L. Dinh, D. Krueger, and Y. Bengio. Nice: Non-linear independent components estimation, 2015.

[27] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp, 2017.

[28] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning, 2017.

[29] L. Feng, S. Shu, Z. Lin, F. Lv, L. Li, and B. An. Can cross entropy loss be robust to label noise? In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2206–2212. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[30] K. Fernandes, J. Cardoso, and B. Astrup. A deep learning approach for the forensic evaluation of sexual assault. *Pattern Analysis and Applications*, 21:1–12, 08 2018.

[31] E. Fix and J. L. Hodges. *Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation Medicine, 1951.

[32] 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 *2009 IEEE 12th International Conference on Computer Vision*, pages 2373–2380, 2009.

[33] M. Gong, J. Liu, H. Li, Y. Xie, and Z. Tang. Disentangled representation learning for multiple attributes preserving face deidentification. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13, 2020.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.

[35] I. J. Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *CoRR*, 2017.

[36] R. Gross, E. Airoldi, B. Malin, and L. Sweeney. Integrating utility into face de-identification. In *Privacy Enhancing Technologies*, pages 227–242, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[37] X. Gu and W. Ding. A hierarchical prototype-based approach for classification. *Information Sciences*, 505:325 – 351, 2019.

[38] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), August 2018.

[39] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans, 2017.

[40] M. Gupta, A. Cotter, J. Pfeifer, K. Voevodski, K. Canini, A. Mangylov, W. Moczydlowski, and A. van Esbroeck. Monotonic calibrated interpolated look-up tables. *Journal of Machine Learning Research*, 17(109):1–47, 2016.

[41] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006.

[42] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[43] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[44] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, 2012.

[45] J. Illingworth and J. Kittler. The adaptive hough transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5):690–698, 1987.

[46] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

[47] S. Jadon. Introduction to different activation functions for deep learning. *Medium, Augmenting Humanity*, 16, 2018.

[48] W. Jin, J. Fan, D. Gromala, P. Pasquier, and G. Hamarneh. Euca: A practical prototyping framework towards end-user-centered explainable artificial intelligence, 2021.

[49] 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*, pages 155–171, Cham, 2019. Springer International Publishing.

[50] 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, 2019.

[51] E. M. Kenny and M. T. Keane. On generating plausible counterfactual and semi-factual explanations for deep learning, 2020.

[52] B. Kim, R. Khanna, and O. O. Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, volume 29, pages 2280–2288. Curran Associates, Inc., 2016.

[53] B. Kim, C. Rudin, and J. Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. NIPS'14, page 1952–1960, Cambridge, MA, USA, 2014. MIT Press.

[54] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions, 2018.

[55] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, volume 29, pages 4743–4751. Curran Associates, Inc., 2016.

[56] D. P. Kingma and M. Welling. Auto-encoding variational bayes. 2014.

[57] I. Kobyzev, S. Prince, and M. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[58] G. R. Koch. Siamese neural networks for one-shot image recognition. 2015.

[59] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951.

[60] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–44, 05 2015.

[61] D. A. Lee and E. J. Higginbotham. Glaucoma and its treatment: A review. *American Journal of Health-System Pharmacy*, 62(7):691–699, 04 2005.

[62] A. Li, J. Guo, H. Yang, and Y. Chen. Deepobfuscator: Adversarial training framework for privacy-preserving image classification, 2019.

[63] 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 *AAAI*, 2018.

[64] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[65] Y. Liu and S. O. Arik. Explaining deep neural networks using unsupervised clustering, 2020.

[66] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

[67] M. Marmot, D. Altman, D. Cameron, J. A. Dewar, S. G. Thompson, and M. Wilcox. The benefits and harms of breast cancer screening: an independent review. *British Journal of Cancer*, 108:2205–2240, 2013.

[68] S. M. McKinney, M. Sieniek, V. Godbole, and et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577:89–94, 2020.

[69] G. McLachlan and K. Basford. *Mixture Models: Inference and Applications to Clustering*, volume 38. 01 1988.

[70] G. McLachlan and D. Peel. *Finite mixture models*. Chichester: Wiley, 2000.

[71] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1 – 38, 2019.

[72] M. Mirza and S. Osindero. Conditional generative adversarial nets. 11 2014.

[73] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks, 2018.

[74] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[75] H. Montenegro, W. Silva, and J. S. Cardoso. Privacy-preserving generative adversarial network for case-based explainability in medical image analysis. In *submitted to WACV 2022*.

[76] H. Montenegro, W. Silva, and J. S. Cardoso. Towards privacy-preserving explanations in medical image analysis. In *submitted to ICML 2021 IMLH*.

[77] 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. In *Under Preparation*.

[78] C. Neustaedter and S. Greenberg. Balancing privacy and awareness in home media spaces 1. 2003.

[79] E. M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

[80] T. T. Nguyen and S. Sanner. Algorithms for direct 0-1 loss optimization in binary classification. *30th International Conference on Machine Learning, ICML 2013*, pages 2122–2130, 01 2013.

[81] H. Noh, T. You, J. Mun, and B. Han. Regularizing deep neural networks by noise: Its interpretation and optimization, 2017.

[82] C. Nugent, D. Doyle, and P. Cunningham. Gaining insight through case-based explanation. *J. Intell. Inf. Syst.*, 32:267–295, 06 2009.

[83] S. Odaibo. Tutorial: Deriving the standard variational autoencoder (vae) loss function, 2019.

[84] W. Oleszkiewicz, T. Wlodarczyk, K. Piczak, T. Trzcinski, P. Kairouz, and R. Rajagopal. Siamese generative adversarial privatizer for biometric data. 04 2018.

[85] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta. Visual privacy protection methods: A survey. *Expert Systems with Applications*, 42(9):4177 – 4195, 2015.

[86] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference, 2019.

[87] G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 30, pages 2338–2347. Curran Associates, Inc., 2017.

[88] N. Papernot and P. McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning, 2018.

[89] V. Petsiuk, A. Das, and K. Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018.

[90] B. Poole, J. Sohl-Dickstein, and S. Ganguli. Analyzing noise in autoencoders and deep networks, 2014.

[91] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[92] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

[93] A. Razavi, A. van den Oord, and O. Vinyals. Generating diverse high-fidelity images with vq-vae-2, 2019.

[94] D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1530–1538. JMLR.org, 2015.

[95] K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345, 2006.

[96] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[97] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408, 1958.

[98] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[99] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[100] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences, 2019.

[101] W. Silva, K. Fernandes, and J. S. Cardoso. How to produce complementary explanations using an ensemble model. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019.

[102] W. Silva, K. Fernandes, M. J. Cardoso, and J. S. Cardoso. Towards complementary explanations using deep neural networks. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 133–140, Cham, 2018. Springer International Publishing.

[103] W. Silva, A. Poellinger, J. S. Cardoso, and M. Reyes. Interpretability-guided content-based medical image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–314. Springer, 2020.

[104] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[105] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[106] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise, 2017.

[107] S. Srinivas and F. Fleuret. Full-gradient representation for neural network visualization, 2019.

[108] L. E. Sucar. *Probabilistic Graphical Models: Principles and Applications*. Springer Publishing Company, Incorporated, 2015.

[109] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning, 2016.

[110] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Assessment of iris recognition reliability for eyes affected by ocular pathologies. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6, 2015.

[111] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Implications of ocular pathologies for iris recognition reliability. *Image and Vision Computing*, 58:158–167, 2017.

[112] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks, 2016.

[113] A. van den Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6309–6318, Red Hook, NY, USA, 2017. Curran Associates Inc.

[114] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2, 6 2014.

[115] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.

[116] Z. Wang, A. C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[117] L. Weng. From gan to wgan. *lilianweng.github.io/lil-log*, 2017. Last accessed January 2021. Available at http://lilianweng.github.io/lil-log/2017/08/20/from-GAN-to-WGAN.html.

[118] L. Weng. Flow-based deep generative models. *lilianweng.github.io/lil-log*, 2018. Last accessed January 2021. Available at http://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html.

[119] Y. Wu, F. Yang, Y. Xu, and H. Ling. Privacy-protective-gan for privacy preserving face de-identification. *Journal of Computer Science and Technology*, 34:47–60, 01 2019.

[120] Y. Xiangli, Y. Deng, B. Dai, C. C. Loy, and D. Lin. Real or not real, that is the question. 2020.

[121] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, 2014.

[122] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization, 2015.

[123] R. M. Zur, Y. Jiang, L. L. Pesce, and K. Drukker. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Medical Physics*, 36(10):4810–4818, 2009.