applied sciences

MDPI

Article

Obfuscation Algorithm for Privacy-Preserving Deep Learning-Based Medical Image Analysis

Andreea Bianca Popescu [1,2,*], Ioana Antonia Taca [1,3], Anamaria Vizitiu [1,2], Cosmin Ioan Nita [2], Constantin Suciu [1,2], Lucian Mihai Itu [1,2] and Alexandru Scafa-Udriste [4,5]

[1] Advanta, Siemens SRL, 500097 Brasov, Romania; ioana_antonia29@yahoo.com (I.A.T.); anamaria.vizitiu@siemens.com (A.V.); suciu.constantin@siemens.com (C.S.); lucian.itu@siemens.com (L.M.I.)
[2] Department of Automation and Information Technology, Transilvania University of Brașov, 500174 Brasov, Romania; nita.cosmin.ioan@unitbv.ro
[3] Department of Mathematics and Computer Science, Transilvania University of Brașov, 500091 Brasov, Romania
[4] Department of Cardiology, Emergency Clinical Hospital, 8 Calea Floreasca, 014461 Bucharest, Romania; alexscafa@yahoo.com
[5] Department Cardio-Thoracic, University of Medicine and Pharmacy "Carol Davila", 8 Eroii Sanitari, 050474 Bucharest, Romania
* Correspondence: andreea.popescu.ext@siemens.com

Abstract: Deep learning (DL)-based algorithms have demonstrated remarkable results in potentially improving the performance and the efficiency of healthcare applications. Since the data typically needs to leave the healthcare facility for performing model training and inference, e.g., in a cloud based solution, privacy concerns have been raised. As a result, the demand for privacy-preserving techniques that enable DL model training and inference on secured data has significantly grown. We propose an image obfuscation algorithm that combines a variational autoencoder (VAE) with random non-bijective pixel intensity mapping to protect the content of medical images, which are subsequently employed in the development of DL-based solutions. A binary classifier is trained on secured coronary angiographic frames to evaluate the utility of obfuscated images in the context of model training. Two possible attack configurations are considered to assess the security level against artificial intelligence (AI)-based reconstruction attempts. Similarity metrics are employed to quantify the security against human perception (structural similarity index measure and peak signal-to-noise-ratio). Furthermore, expert readers performed a visual assessment to determine to what extent the reconstructed images are protected against human perception. The proposed algorithm successfully enables DL model training on obfuscated images with no significant computational overhead while ensuring protection against human eye perception and AI-based reconstruction attacks. Regardless of the threat actor's prior knowledge of the target content, the coronary vessels cannot be entirely recovered through an AI-based attack. Although a drop in accuracy can be observed when the classifier is trained on obfuscated images, the performance is deemed satisfactory in the context of a privacy–accuracy trade-off.

Keywords: image obfuscation; deep learning; medical imaging; privacy preserving classification

check for updates

Citation: Popescu, A.B.; Taca, I.A.; Vizitiu, A.; Nita, C.I.; Suciu, C.; Itu, L.M.; Scafa-Udriste, A. Obfuscation Algorithm for Privacy-Preserving Deep Learning-Based Medical Image Analysis. *Appl. Sci.* **2022**, *12*, 3997. https://doi.org/10.3390/app12083997

Academic Editor: Jan Egger

Received: 24 March 2022
Accepted: 13 April 2022
Published: 14 April 2022

1. Introduction

In the last decade, machine learning (ML) algorithms have demonstrated remarkable results in potentially improving the performance and the efficiency of healthcare applications. A recent study [1] provides an overview of the benefits that machine learning brings in healthcare, including aiding doctors in their decision making, and decreasing the cost and time it takes to reach a diagnosis. Even though such solutions allow for better resource allocation and treatment selection, they are challenging to implement in real-world circumstances due to several obstacles. The same study emphasizes that one of the most significant problems is the massive amount of high-quality data that are frequently necessary to create and evaluate machine learning models.

A related issue is the ethical aspect of data collection, which necessitates data sourcing for ML, to comply with personal information protection and privacy regulations [2]. The GDPR establishes precise permission standards for data uses in Europe, whereas the HIPAA regulates healthcare data from patient records in the United States. These laws are considerably more challenging to fulfill when clinical users prefer to delegate ML model development and deployment to third parties, and use them via cloud services, e.g., due to a lack of hardware capabilities. According to a recent survey [3], the Machine Learning as a Service (MLaaS) paradigm has appeared as a highly scalable approach for remotely running predictive models, raising at the same time increased security and privacy concerns. The same paper highlights that fully homomorphic encryption (HE) could be a straightforward approach that allows a third party to process encrypted data without knowing its content.

An early effort that combined HE with neural networks, involving the communication between the model owner and the data provider, is described in [4]. CryptoNets [5] eliminates this interaction, but it has the drawback that the encryption technique does not process real numbers. CryptoDL [6] approximates nonlinear functions with low-degree polynomials to overcome model complexity restrictions. However, the use of estimated activation functions reduces the prediction accuracy of the model. More recent studies propose different approaches to increase the classification accuracy at the inference phase in AI-based models employing homomorphic encryption. In [7], adopting a polynomial approximation of Google's Swish activation function, and applying batch normalization, enhanced classification performance on the MNIST and CIFAR-10 datasets. Additional optimizations are performed to reduce the consumption level. J.W. Lee et al. [8] emphasize that the most common activation functions are non-arithmetic functions (ReLU, sigmoid, leaky ReLU), which are not suited for homomorphic computing, because most HE schemes only enable addition and multiplication. They evaluate these non-arithmetic functions with adequate precision using approximation methods. In combination with multiple methods for reducing rescaling and relinearization errors, the bootstrapping strategy enables a deep learning model to be evaluated on encrypted data. According to the numerical verification, the ResNet-20 model produced equivalent results on the CIFAR-10 dataset for both encrypted and unencrypted data. The efficiency of MLaaS is drastically improved in [9], where GPUs acceleration is used to evaluate a pre-trained CNN on encrypted images from MNIST and CIFAR-10 datasets. None of the above-mentioned methods addresses the training phase of models on encrypted data due to the increased number of operations and the longer runtime, this being regarded as an open problem, especially in the case of image-based datasets. For privacy-preserving computations within deep learning models, we suggested a variant of a noise-free matrix-based homomorphic encryption method (MORE [10]) in our earlier work [11]. We validated the methodology using two medical data collections in addition to the MNIST dataset. The encryption step is employed during both training and inference. The experiments showed that the method provides comparable results to those obtained by unencrypted models, while having a low computational overhead. However, the changes made to the original HE scheme to allow computations on rational numbers come at a cost in terms of privacy, as it provides lower security than standard schemes. This method was further used in [12] to design a cloud-based platform for deploying ML algorithms for wearable sensor data, focused on data privacy. We have further addressed the security compromise in [13], where we combined a HE scheme based on modulo operations over integers [14], an encoding scheme that enables computations on rational numbers, and a numerical optimization strategy that facilitates training with a fixed number of operations. Nevertheless, the computational overhead introduced through encoding and encryption represents a significant drawback of the method.

The comprehensive survey [3] includes theoretical concepts, state-of-the-art capabilities, limits, and possible applications for more privacy-preserving machine learning (PPML) solutions based on HE. An overview of techniques based on other privacy-preserving primitives such as multi-party computation (MPC), differential privacy (DP) and federated learning (FL) is provided in [15]. The authors underline that a hybrid PPML system could feasibly imply a trade-off between ML performance and computational overhead.

Another privacy-preserving approach that has received increasing interest is image obfuscation. In the context of PPML, it entails modifying the image so that the content becomes unintelligible while retaining the underlying information to some extent. Obfuscation methods such as mosaicing, blurring and P3 are analyzed in [16]. Mosaicing is used to alter parts of an image inside a window whose size is inversely related to obfuscated image resolution. Blurring applies a Gaussian filter that removes details from images. Despite the fact that mosaicing and blurring make it impossible for the human eye to detect faces or digits in obfuscated images, the authors show that standard image recognition models can extract useful information from the transformed data. The strategy suggested in [17] uses Gaussian noise to obscure only a few images in the dataset (which are considered to have a sensitive content). The authors emphasize that this method could affect the model performance if too many frames require protection.

The obfuscation techniques described in [18] are variations on the mixup approach, which entails creating convex combinations of pairs of samples. The proposed approaches aim to improve the privacy of the training data, while optimizing the model accuracy without increasing the computational cost of the training process. The presented methods are variants of the mixup technique, which entails creating convex combinations of pairs of samples. After mixing, the newly created sample is further obfuscated through pixel grafting, pixel shuffling, noise addition or blurring. In the same research, authors demonstrate that metrics like SSIM (structural similarity index measure) and HaarPSI (Haar wavelet-based perceptual similarity index), which accord with human perception on picture degradation, may be used for privacy assessment. Two datasets that contain images depicting animals were used to validate the methods. The results highlight that a compromise between obfuscation and learning capabilities must always be considered. The Google Vision AI image classifier was queried with obfuscated images, and its recognition performance was lower than that of the human evaluators. Kim et al. [19] performed an interesting study focused on privacy-preservation for medical image analysis. They proposed a client-server system in which the client protects the patient identity by deforming the input image using a system comprising a transformation generator, a segmentation network, and a discriminator. The system is trained in an end-to-end adversarial manner to solve the task of MRI brain segmentation. Being focused on enabling protection against facial recognition, the approaches presented in [20,21] leverage generative adversarial networks to produce more visually pleasing outputs, while providing a solid defense against deep learning-based recognition systems. In [21], for the analyzed scenarios, the trade-off is formulated based on the privacy against face recognition versus the utility in terms of face detection.

Herein, we propose an obfuscation technique that combines variational autoencoders with non-bijective functions. The aim is to achieve a method that enables accurate model training, while ensuring privacy against human eye perception and AI-based reconstruction attacks. The experiments are constructed to reflect the perspective of a clinical user (e.g., hospital) in a specific use case (coronary angiography view classification), and the perspective of a threat actor. Because the hospital lacks the physical resources and the expertise to develop a DL classification model, the inference is performed by a third party, which is considered untrustworthy. In this scenario, this external party is a Machine Learning as a Service (MLaaS) provider who can train a DL model using the clinical data, and then make it available as a cloud service for inference. Since the patient data is considered to be sensitive and private, every angiographic frame used for training or inference is obfuscated to protect data privacy outside of the clinical environment. Conversely, a potential threat actor, that could be the MLaaS provider or an interceptor, may try to acquire illegal access to the clinical data. The considered attack strategy is based on the training of a reconstruction model on original-obfuscated pairs of samples from a public dataset. Because the obfuscation method is considered publicly available as a black-box tool for collaborative purposes, any external entity can use the tool to obfuscate images and obtain a dataset of corresponding image pairs. Two possible attack configurations are formulated. In the first one, the threat actor knows the data source (i.e., hospital) but is unaware of its specific type (coronary angiography, in our case), hence the training is performed on a public dataset containing medical-related samples. Another possibility is that the attacker is a collaborative hospital which knows that the target dataset consists of coronary angiographies, and which trains the reconstruction model on its own angiographic data.

All parties other than the hospital are regarded as untrustworthy in terms of data security, and, in consequence, every externalized angiographic frame is, in fact, an obfuscated image. Even the rightful receiver, in this case the MLaaS provider, is not considered honest regarding data confidentiality, which is why the proposed obfuscation method aims to be irreversible. The goal is to protect the medical images from a highly resourceful entity (in terms of both computer power and data), while allowing for the training of the desired deep learning model directly on the altered images.

The remainder of the paper is organized as follows. The obfuscation techniques, as well as the network architectures, datasets, and procedures for the suggested use case, are presented in Section 2. Section 3 describes the experiments performed from the perspectives of the clinical user and the threat actor, along with the findings. In Section 4, we iterate through the unique characteristics of the proposed technique, present remarks regarding its usefulness in deep learning-based applications, and finally draw the conclusions.

2. Methods and Materials

In the following, we propose a novel strategy that combines two obfuscation approaches to:

1. Hide the content of a sensitive image from the human eye;
2. Make AI-based image reconstruction challenging;
3. Facilitate DL model training using obfuscated images.

The first stage is to train a variational autoencoder, which uses the original (non-obfuscated) dataset as both input and target, and provides an obfuscated counterpart for each sample at the bottleneck. A detailed description of the VAE architecture, training and obfuscation process is presented in Section 2.1. The next step is also described as a stand-alone method in Section 2.2, where every pixel intensity value is randomly translated to another intensity value in a non-bijective manner, to alter the visual information. When the techniques are used in conjunction, the image encoded with the VAE is further obfuscated through pixel substitution, according to a non-bijective mapping function. The entire workflow is detailed in Section 2.3. The clinical usage scenario, the dataset, and the architecture used to solve the classification task are presented in Section 2.4. Section 2.5 describes the procedures employed to evaluate the privacy level provided by the proposed approach against human perception and against AI-based reconstruction attacks.

2.1. Obfuscation Method Based on a Variational Autoencoder

The Variational Autoencoder [22] considered herein is a generative model based on the work of Kingman et al. [23]. It consists of two models that support each other: an encoder (recognition model) and a decoder (generative model). The difference between VAEs and other AEs is that the input is not encoded as a single point, but as a distribution over the latent space, from which the decoder draws random samples. Due to the reparameterization trick, which allows for backpropagation through the layers, the two components of the VAE can be chosen to be (deep) neural networks.

The autoencoders, and by extension VAEs, generate an encoding of the inputs that allow for an accurate reconstruction. This property also ensures that the encoding contains useful information extracted from the input, and, hence, it can be employed in further DL-based analysis or model training, e.g., within an obfuscation method based on VAE.

From a probabilistic perspective, a VAE implies approximate inference in a latent Gaussian model, where the model likelihood and the approximate posterior are parameterized by neural networks. The recognition model compresses the input data x into a dimensionally reduced latent space χ, while the generative model reconstructs the data given the hidden representation $z \in \chi$. Let us denote the encoder $q_\theta(z|x)$ and the decoder $p_\phi(x|z)$, where θ and ϕ represent the neural network parameters.

The latent variables $z \in \chi$ are considered to be drawn from a simple distribution: $p(z) = \mathcal{N}(0, I)$, named prior (here, I denotes the identity matrix). The input data x have a likelihood $p(x|z)$ that is conditioned on z. As a result, a joint probability distribution over data and latent variables can be defined:

$$p(x, z) = p(x|z)p(z). \tag{1}$$

The aim is to calculate the posterior distribution $p(z|x)$. This can be achieved by applying Bayes' rule:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}, \tag{2}$$

where $p(x)$ can be obtained by marginalizing out z: $p(x) = \int p(x|z)p(z)dz$. Unfortunately, the integral is usually intractable [24]. As a consequence, an approximation of this posterior distribution is required.

There are two main ways for posterior approximation: applying Markov Chain Monte Carlo (MCMC) methods such as the Metropolis–Hastings algorithm [25] or Gibbs sampling [26], and variational inference (VI) [27]. VAE uses the latter because the sampling methods converge slower [28]. This approach implies approximating the posterior with a family of Gaussian distributions $q_\lambda(z|x)$, where parameters λ represent the mean and the variance of each hidden representation. As a result, the encoder parameterizes the approximate posterior $q_\theta(z|x, \lambda)$, taking x as input data, and parameters λ as outputs. On the other hand, the decoder parameterizes the likelihood $p(x|z)$, having the latent variables as input and the parameters to distribution $p_\phi(x|z)$ as output. The approximation is penalized by computing the Kullback–Leibler (KL) divergence that measures the distance between $q_\theta(z|x, \lambda)$ and $p(z)$.

Hereupon, the loss function which is minimized during training is composed of two terms: (i) the reconstruction error between input data x and output data x', and (ii) the KL divergence between the approximate posterior and $p(z)$, chosen to be a normal distribution:

$$Loss = \mathcal{L}(x, x') + KL(q_\theta(z|x, \lambda)||p(z)). \tag{3}$$

The first step of our method is to train a convolutional VAE on another dataset from the same domain as the working dataset. Additionally, one of the layers is used for noise addition. At the bottleneck, the information is divided between two channels to obtain an encoded version of the input. Those channels correspond to the mean (channel 1) and standard deviation (channel 2) of the normal distribution obtained from the encoder. Any of the channels can then be used for a subsequent DL model training on obfuscated images. From the trained VAE, only the encoder is retained as a black-box obfuscation tool. As there is no need for a reconstruction once an image has been obfuscated, the decoder is discarded. Figure 1 displays the workflow described for the obfuscation method based on a VAE.

For our experiments, the VAE is trained on the Medical MNIST dataset [29]. The dataset contains 6 classes of X-ray images, that are randomly distributed for training (30,000 images) and validation (12,000). More details about the Medical MNIST dataset are presented in Section 2.5.

During training, the 64 × 64 images are passed through three convolutional layers of 32, 8, and 4 filters, respectively, with a 3 × 3 receptive field. ReLU is the activation function chosen for each layer. The architecture of the decoder consists of three convolutional, ReLU activated layers of 4, 8, and 32 filters, followed by one dense layer. The VAE is trained for 10 epochs.

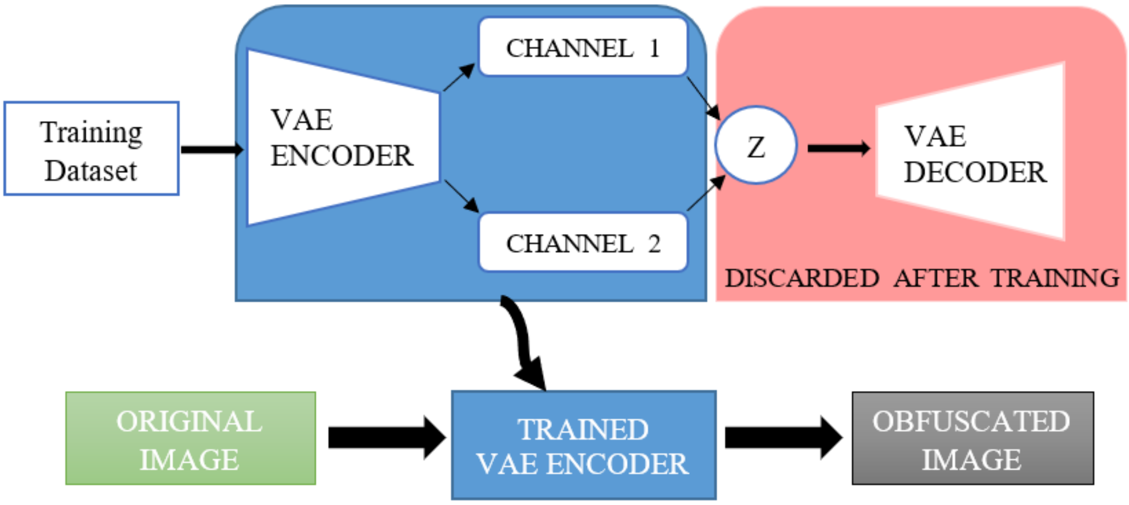

Figure 1. Schematic representation of the obfuscation technique based on a VAE.

The trained encoder can be used for obfuscating medical images. A channel option must be selected, depending on the desired result. The first channel, corresponding to the mean of the normal distribution, usually assures a better privacy level than the second channel, as it does not preserve as much detailed information from the input. This limits, though, its usefulness in further AI-algorithms. The channel corresponding to the standard deviation of the normal distribution tends to preserve more useful information from the original image. As a result, it is preferred in cases where the obfuscated images would be used in machine learning tasks. This channel, although depending on the initial structure of the original image, may or may not ensure the imposed or desired level of privacy. For example, in the encoding of an image with a monochromatic background, most probably sensitive details will be visible, which could uncover the nature of the original image. Such an example is shown in Figure 2, where the original image, representing a coronary angiography, has an almost monochromatic background. As a result, in the image obtained from channel 2, the main vessel can be seen.

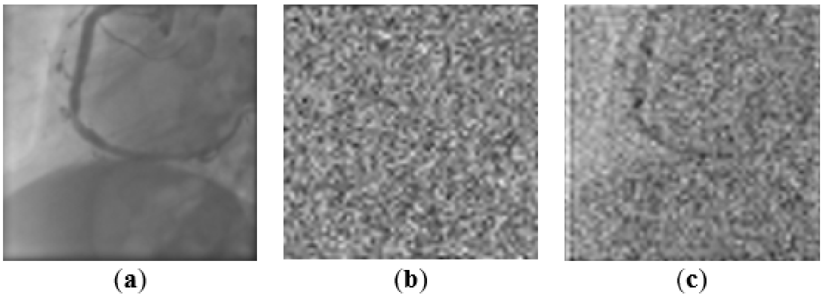

Figure 2. Comparison between the original frame (**a**) and the obfuscated counterparts when channel 1 (**b**) and channel 2 (**c**) are chosen.

2.2. Obfuscation Based on Non-Bijective Pixel Intensity Shuffling

This approach starts with a simple obfuscation technique—random pixel intensity shuffling. Every pixel intensity is randomly associated with another value from the same interval as described by Equation (4), where $range(a, b)$ is a function that returns all integer numbers between a and b, including the interval's endpoints, and $shuffle(x)$ is a function that randomly interchanges the positions of the elements of a list x inside the returned array. We call this array a map because it creates connections between each possible pixel intensity embodied in the list of indexes of the array and a new random value contained in the array at the corresponding position.

$$intensityMap = shuffle(range(0, 255)) \tag{4}$$

This association is a bijective function because for each domain component there is only one corresponding element in the codomain. Although this operation preserves the underlying information of the images, while making them unrecognizable for the human eye, the approach is still susceptible to AI-based attacks, statistical or even reverse engineering attacks. Presuming that an external party has access to the obfuscation algorithm in a black-box form, an unlimited number of new images can be obfuscated, and a statistical evaluation should reveal that a one-to-one mapping was used. By reversing this mapping, a potential attacker can obtain the original images with no information loss. Training a deep learning model to reconstruct the obfuscated images is another attack approach. In anticipation of this kind of attack, a second step is proposed for this obfuscation method. The bijective function is modified so that the injectivity property is lost. In other words, multiple elements of the domain will correspond to the same element of the codomain. This effect is achieved by applying the same *mod N* operation on each value of the previously obtained map. Hence, the obfuscation method can be defined by a function $f : A \to B$, where $A = [0, 255]$ and $B = [0, N)$. When obfuscating an image, an iteration across all pixels must be performed. In Equation (5), pv denotes the intensity of the pixel found at the (i, j) coordinates in the *image* matrix.

$$pv = image_{i,j} \tag{5}$$

This value is modified according to Equation (6), where the *mod* function represents the typical modulo operation and the pv value is used as an index.

$$image_{i,j} = intensityMap_{pv} \ mod \ N \tag{6}$$

Figure 3 synthesizes the steps proposed for this obfuscation technique. The key concept is that applying a *mod N* operation limits the range of possible values to N elements. However, this is not equivalent to filtering the highest intensities due to the previously performed random associations. Thus, more details are preserved in images by arbitrary but consistent replacement of $256 - N$ pixel intensities. Since the obfuscation function is represented by a many-to-one mapping, the task of reconstructing unseen images becomes more complex and more uncertain, even for an AI-based model trained on original-obfuscated image pairs.

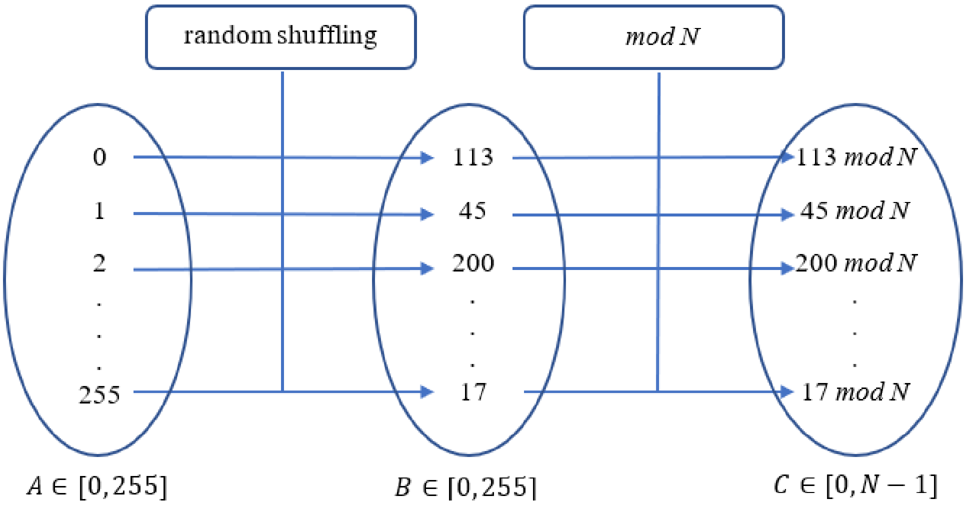

Figure 3. Schematic representation of the obfuscation technique based on pixel intensity shuffling.

The N value is an adjustable parameter that improves security when being set to lower values. As a function of this parameter, the underlying information is preserved in different degrees, presumably retaining enough details in the images for DL-based applications. Figure 4 displays a comparison between an angiographic frame Figure 4a and the obfuscated counterparts when bijective Figure 4b or non-bijectiveFigure 4c–e mapping is applied. The obfuscated samples are rescaled in $[0, 255]$ interval to allow a better visual comparison.

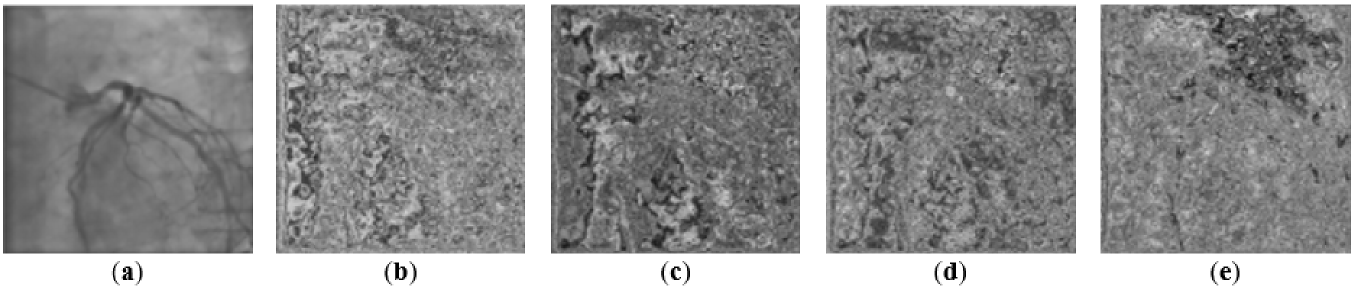

Figure 4. Comparison between the original frame (**a**) and the obfuscated counterparts when $N = 256$ (**b**), $N = 156$ (**c**), $N = 50$ (**d**) and $N = 45$ (**e**).

2.3. Secure Obfuscation Algorithm

As previously explained, the security of VAE obfuscation also depends on the image itself. For images with a uniform distribution of pixel intensities, the method will not only protect the content from the human eye perception but, due to the additional noise, also make it more difficult for an AI-based model to reconstruct the original image. In contrast, the human eye would be able to discern the environment from the main structures, or even details of the structures, in a dichromatic image where two predominant intensities describe the object and the background. The noise level can vary, but this would also affect the utility of the image. Using a non-bijective function to substitute the intensities makes the obfuscated images unrecognizable by the human eye. Although the modulo operation is meant to protect against more sophisticated attacks, the success rate of an AI-based reconstruction attack depends on the value of N. The smaller this parameter is, the more difficult the reconstruction becomes. However, this implies a trade-off between privacy and utility. We integrate the strengths of each method into a new obfuscation algorithm to maximize their effectiveness. The steps are as follows, in the order in which they should be performed :

1. The VAE model is trained on images similar to those that will be obfuscated in the clinical use case.
2. All pixel intensities are randomly shuffled, and a *modulo N* operation is performed on each resulting value leading to a non-bijective mapping between different intensities.
3. The original image is encoded using the VAE encoder.
4. Each pixel value of the encoded image is substituted with the corresponding value in the non-bijective map.

As a result, an obfuscated image is created, which retains the original image's underlying relevant information and can be used for further analysis and processing (e.g., image classification). Regardless of the initial structure of an image, combining the techniques improves privacy. First, the eye perception is affected by the intensity shuffling even if, after encoding, the sensitive content is still distinguishable. Then, the protection against AI-based reconstruction attacks is ensured by the conjunction of noise and non-bijectivity. The entire obfuscation workflow is schematically depicted in Figure 5.

Although the underlying information of an image is preserved using this technique, an essential requirement that must be met to use multiple images in the same application (e.g., training a classifier on obfuscated images) is that the same encoder and the same shuffling map should be applied on all images (both for training and inference). The trade-off between privacy and utility can be managed by tuning certain method-specific parameters according to the needs of the use case. For the technique based on non-bijective intensity mapping, the choice of parameter N may influence the image utility. Regarding confidentiality, a higher N implies less information retained in the obfuscated image and, thus, a more difficult to perform image reconstruction. Figure 6 displays an original angiographic frame and the obfuscated counterparts for each obfuscation approach. The chosen value for the modulo operator N in Figure 6c is 96. More examples are included in Appendix A, Figure A1.

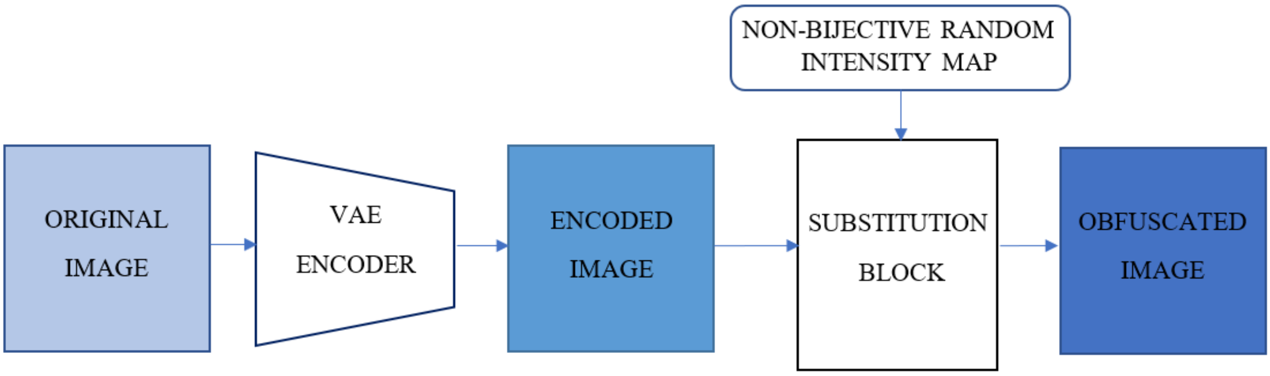

Figure 5. Schematic representation of the secure obfuscation algorithm.

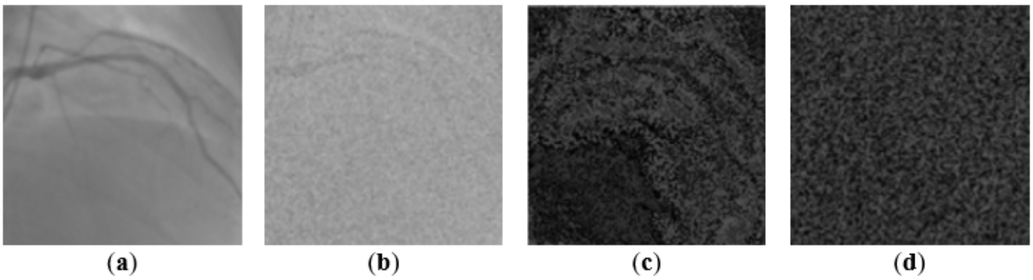

Figure 6. Comparison between an original angiographic frame (**a**) and the obfuscated counterparts when using (**b**) encoding, (**c**) non-bijective intensity mapping, and (**d**) combined algorithm.

2.4. Utility Level Evaluation

As the methods described above rely on reducing, to a certain degree, the information from the original images, their utility after the obfuscation must be evaluated. To perform this analysis, the same DL model is trained for multiple levels of obfuscation, including no obfuscation. The methods presented in Sections 2.1 and 2.2 are employed separately and in conjunction, as described in Section 2.3, to obfuscate an in-house dataset consisting of coronary angiography frames. The same experiment is run for multiple values of N, ranging between 1 and 255. The utility of obfuscated images is determined by comparing the accuracy achieved on a testing dataset for different degrees of obfuscation.

The task is to train a binary classifier to distinguish between RCA and LCA views in angiographic frames. Figure 7 depicts one sample of each category. The dataset contains 3280 coronary angiographies, balanced between the two classes. A subset of 600 images is used for validation, and another subset of 700 images is retained for evaluation purposes. The rest of the 1980 angiographic frames are used for training. Augmentation techniques such as shifting, flipping, zooming and rotation are applied. The original size of the frames is 512 × 512 pixels, but experiments with different input shapes have shown that a size of 128 × 128 ensures almost no loss in classification performance with a lower computational time. The pixels values are normalized through min-max scaling in the [0, 1] range.

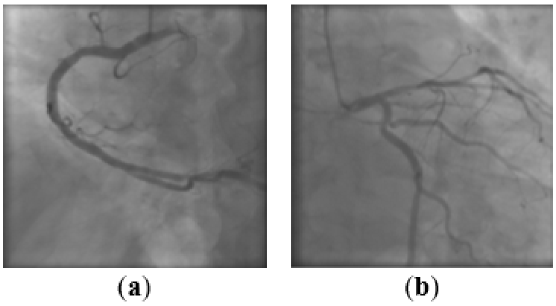

Figure 7. RCA—right coronary artery (**a**) and LCA—left coronary artery (**b**).

The images (obfuscated or not) are passed through four convolutional layers of 16 and 32 filters with a 3 × 3 receptive field during training. The pooling layers downsample the images by a factor of two by using the maximum value of a window. After the last convolutional layer, a flatting layer is added to convert the features matrix into a vector. The fully connected layers contain 512, 1024 and 2 nodes, respectively. The ReLU function is employed as an activation function for all layers, except for the last one where the softmax activation is used. Each convolutional layer is followed by a local normalization layer [30] to make the model more robust to image degradation. To limit the overfitting, between 25% and 50% of the connections of the neurons are dropped through dropout layers. Furthermore, although the maximum number of epochs is set to 30, early stopping is employed when the validation loss is not decreasing within 10 consecutive epochs. A learning rate scheduler is used to achieve good convergence, starting from 1×10^{-3}, and diminishing the value with every epoch. The workflow of an inference step using the obfuscation algorithm is depicted in Figure 8.

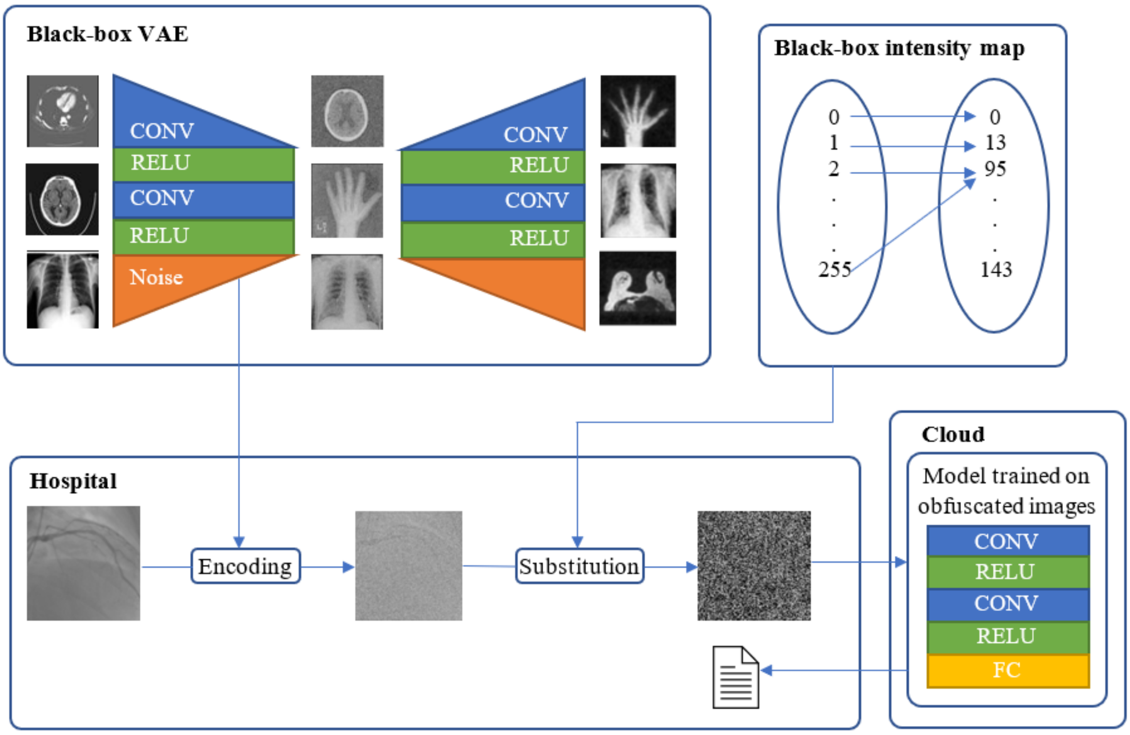

Figure 8. Detailed workflow of inference using the secure obfuscation algorithm.

The Keras framework [31] was used to build the convolutional neural network, and the local normalization layer is based on [30]. The experiments were run on a computer equipped with an Intel i7 CPU (Intel, Santa Clara, CA, USA) at 4.2 GHz, 32 GB RAM and an NVIDIA GeForce GTX 1050 Ti GPU (Nvidia, Santa Clara, CA, USA) with 4 GB of dedicated memory.

2.5. Privacy Level Evaluation

To compare the degree of privacy provided by each proposed technique, we employ similarity metrics such as SSIM and PSNR (peak signal-to-noise-ratio) assessed between the original and the corresponding obfuscated images. As stated in [18], SSIM is an image quality metric that can quantify image privacy. It considers perceptual phenomena like brightness and contrast, as well as structural information changes. SSIM can take values between 0 and 1, where 0 means no structural similarity, and 1 indicates identical images. Therefore, lower values correspond to an increased security. PSNR is expressed using the decibel scale, and typical values for good quality images (with a bit depth of 8) are between 30 and 50 dB. As a result, values below the lower threshold indicate that the image is protected against human perception. The entire testing subset owned by the hypothetical clinical user is employed for this evaluation. The averaged results are presented in Section 3.3.

Two possible attack configurations are considered to assess the level of security against AI-based reconstruction. The considered scenario is that of an external party willing to access the original data sent by the hospital or by a specific patient. The general assumption is that the obfuscation algorithm used by the hospital is publicly available as a black-box tool. The privacy parameter N is also presumed to be known. This means that another clinical user or an MLaaS provider, or even an external interceptor can use the tool to obfuscate images and obtain a dataset of corresponding image pairs. Moreover, because the data source is known, the threat actor might guess that the dataset consists of medical images. The workflow of an entity willing to gain unauthorized access to the data has the following steps: obfuscating a dataset of medical images using the same obfuscation tool as the hospital, training a deep learning model to reconstruct the original frames from the obfuscated images, intercepting obfuscated images, and reconstructing the original images using the previously trained model.

In the first attack configuration, the interceptor assumes that the targeted data contains medical images, but is unaware of their type (E_1); therefore, the malicious actor trains the reconstruction model using a publicly available dataset with different medical-related classes. In the following experiments (see Section 3.3), the reconstruction model is trained using the Medical MNIST dataset [29]. It contains six classes of X-ray images (abdomen CT, breast MRI, CXR, chest CT, hand radiography, head CT), each class totalling around 7000 samples. All 40,954 medical images are used for training, and the evaluation is performed on the intercepted obfuscated dataset. The Medical MNIST images have a size of 64 × 64 pixels, but they are resized to 128 × 128, the dimensions of the frames sent by the hospital. Figure 9 depicts a sample of each category of the Medical MNIST dataset.

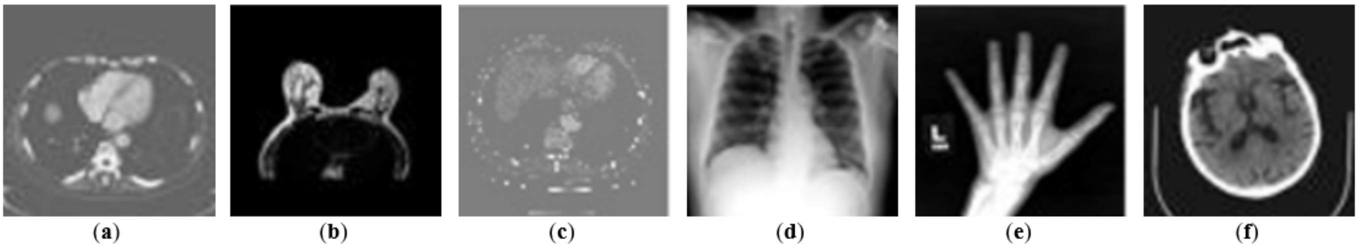

Figure 9. Medical MNIST samples: abdomen CT (**a**), breast MRI (**b**), CXR (**c**), chest CT (**d**), hand radiography (**e**), head CT (**f**).

Another possibility is that the type of the medical images is well known, so a similar dataset is used to train the reconstruction model (E_2). For example, two clinical partners want to create an aggregated dataset containing coronary angiographies for training a view classification model, but they both wish to keep their data confidential. However, one of the partners is willing to obtain the content provided by the other. As they both use the same obfuscation tool, the threat actor obfuscates his angiographic dataset, and uses it to train a reconstruction model. Then, the malicious actor intercepts the obfuscated frames of the victim, and tries to undo the obfuscation. The (in-house) dataset used in these experiments contains 8365 angiographies (5779 LCA and 2586 RCA), all employed for training. Their original size (512 × 512) is modified to 128 × 128.

Before training, both the inputs (obfuscated images) and the targets (original images) are normalized through min-max scaling in the [0, 1] interval. The U-Net architecture introduced in [32] is employed for reconstruction. The first half of the network, which behaves like an encoder, consists of convolutional and pooling layers that perform downsampling. Each decoder block combines its input with information from the corresponding encoder block, and performs convolutional and upsampling operations. The same activation function, number of filters, kernel size, pooling window and stride as in the original paper were used. The batch size and momentum values were set to 1 and 0.99, respectively. The model was trained for 30 epochs with a learning rate of 0.001. The architecture was implemented in the PyTorch framework [33], and the models were trained on a machine equipped with 128 GB RAM and NVIDIA GeForce GTX 1080 Ti GPU with 11 GB of dedicated memory.

The reconstruction network was trained on images obfuscated using the methods described in Sections 2.1 and 2.2, and the algorithm presented in Section 2.3 for multiple values of the parameter N. To determine the degree of similarity between the reconstructed images and the original counterparts, SSIM and PSNR are computed across all frames sent by the victim (the training dataset of the classifier). Considering the threshold values of SSIM, in the results presented in Section 3.3, a lower SSIM value denotes a poor reconstruction performance and a high privacy level. Regarding the interpretation of PSNR, in the following experiments, values under 30 indicate inaccurate reconstruction and high security. The scikit-image library [34] was employed for computing the similarity metrics.

Expert readers manually performed a visual assessment to determine to what extent the reconstructed images are protected against human perception. The assessment was performed on 50 frames (25 LCA, 25 RCA). Since in most cases the background was reconstructed more accurately than the arteries, two separate scores were assigned for each image. A scale from 1 to 5 was chosen, where 1 indicates that the object was not reconstructed at all and 5 denotes a visual similarity larger than 95%. Some scoring guidelines were formulated to limit the evaluation bias. Tables 1 and 2 synthesize the links between scores and image descriptions.

Figures 10 and 11 display for each score an evaluation example corresponding to the scoring guidelines. The mean scores are computed for all evaluations of all frames. The LCA and RCA frames were also considered separately to determine if reconstruction performs better on a specific class.

Table 1. Scoring guidelines concerning the vessels' accurateness.

Score	Vessel Tree Description
1	No vessel is visible in the image.
2	There are some fine lines in the background, but it is hard to distinguish whether they are blood vessels or to identify the angiographic view.
3	The main vessel is visible, but there are many missing details, and additional artifacts are present.
4	All branches are visible but not with the same clarity as in the original image. Enough details are present to be able to distinguish the angiographic view.
5	The reconstruction is more than 95% similar to the original image. Some portions might be unclear, or some additional artifacts might be present, but the main arteries are well visible.

Table 2. Scoring guidelines concerning the background accurateness.

Score	Background Description
1	The background is almost monochromatic.
2	The prominent shadows are vaguely captured.
3	More accurate intensities are captured, but the background is still diffused overall.
4	The background is close to the original one in shape and pixel intensities. Some diffused areas or additional artifacts might be present.
5	The reconstructed background is more than 95% similar to the original one. The same shapes and shadows are depicted, but might differ in pixel intensity in specific regions.

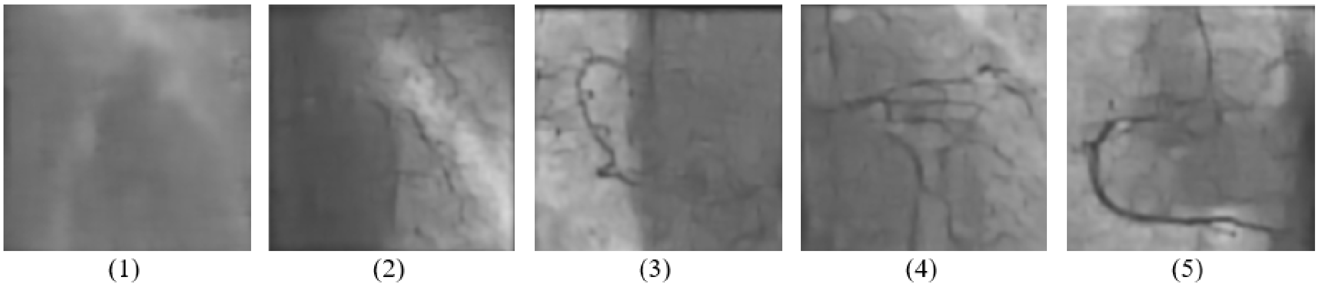

Figure 10. Examples of reconstructed angiographies and the scores assigned concerning the vessels' accurateness: (**1**) no visible vessels; (**2**–**4**) intermediate scores; (**5**) accurate vessels reconstruction.

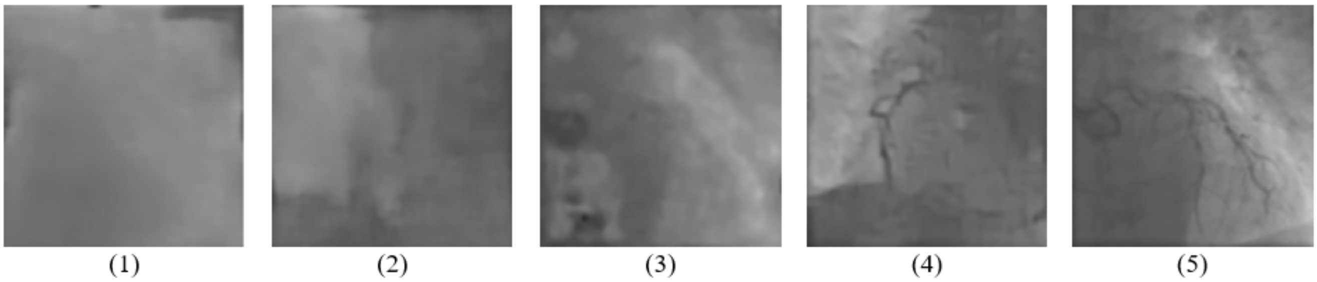

Figure 11. Examples of reconstructed angiographies and the scores assigned concerning the background accurateness: (**1**) monochromatic background; (**2**–**4**) intermediate scores; (5) accurate background reconstruction.

3. Experiments and Results

3.1. Angiographic View Classification

To evaluate the utility of angiographic frames after obfuscation, we formulate four experiments in which convolutional neural networks are trained to solve the angiographic view classification task:

- C_1—original images are used (no obfuscation);
- C_2—images are obfuscated using only the VAE encoder;
- C_3—images are obfuscated only through intensity substitution according to a non-bijective map;
- C_4—images are obfuscated using both methods, as described in Section 2.3.

The details regarding these experiments are presented in Section 2.4. The accuracy obtained by the DL model for each configuration on a testing subset is reported in Table 3.

Table 3. Comparison between DL-model performance when trained on original and obfuscated images, respectively.

	C_1	C_2	C_3	C_4	[11]
Test Accuracy	97.57%	93.71%	88.57%	82.71%	96.20%

After altering the angiographic frames using the VAE encoder, the performance drops by approximately 4%. The method based on a non-bijective map applied to pixel intensities ($N = 96$) leads to a decrease in accuracy of 9%. Although using both techniques causes a significant performance drop compared to the model trained on original images, the accuracy value remains above 80% and may be considered satisfactory in the context of a privacy--accuracy trade-off. The purpose of these experiments is not to achieve state-of-the-art performance on obfuscated images but to compare the results when the same architecture and different obfuscation techniques are employed.

The last column of the table displays the performance previously achieved on the same dataset, using a different DL model and employing the MORE [10] homomorphic encryption scheme as a privacy-preserving technique. The accuracy was identical for the encrypted and the unencrypted model, but the computational time was around 32 times larger when encrypted data was used. In the experiments C_1–C_4 both training and inference were performed with the same runtime since the complexity of the data is not increased by the obfuscation method. Although the MORE encryption scheme provides some advantages in terms of simplicity, clarity, and practicability, when adapted for PPML its linear structure can raise security concerns [11]. By having access to a large enough number of pairings of encrypted and unencrypted data, and by formulating the key search attack as an optimization problem, this linearity may allow one to find the secret key. Furthermore, the fact that the message to be encrypted will always be found among the eigenvalues of the ciphertext matrix is a benefit in terms of utility but also represents privacy-related disadvantages. The obfuscation method overcomes these limitations, since it is highly non-linear and no decryption key is involved.

The C_4 experiment was run multiple times for different values of parameter N. The results achieved on a testing subset are depicted in Figure 12. As expected, for $N = 1$, the accuracy drops to 50% (random guess) because all images become monochromatic. For the other values of N, no monotonous tendency can be observed, suggesting that even for smaller values, enough details are preserved for the classification to be successfully performed.

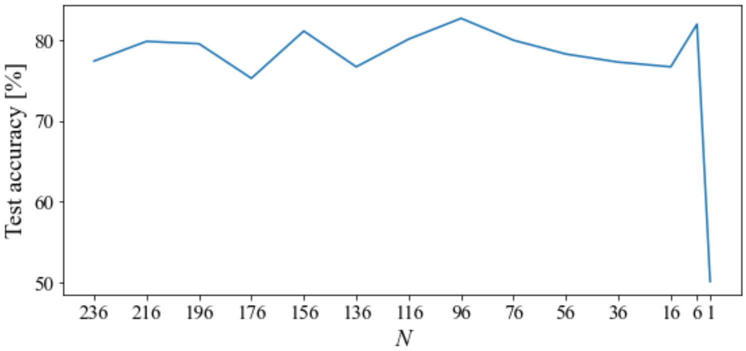

Figure 12. Influence of parameter N on the test accuracy in C_4 configuration experiments.

3.2. Privacy Level of Obfuscated Images

A comparison of the similarity metrics derived for all angiographic frames in the test subset for each of the three procedures is presented in Table 4.

Table 4. Similarity between the original frames and the obfuscated images.

	Encoding	Non-Bijective Mapping	Combined Techniques
SSIM	0.1999	0.0602	0.0512
PSNR [dB]	19.18	9.96	9.92

When employing the VAE encoder to obfuscate images, the results show low similarity and poor quality compared to the original ones. However, applying a non-bijective random mapping leads to an SSIM value below 0.1, corresponding to almost no structural similarity. The PSNR also indicates that applying non-bijective function features increases privacy. Nevertheless, the metrics decrease even further when the methods are simultaneously employed. Thus, the results support the initial hypothesis and motivate the usage in conjunction with the proposed techniques.

3.3. AI-Based Reconstruction Attack

To evaluate the security of obfuscated frames against AI-based reconstruction attacks, we formulate six experiments:

- A_1—E_1 attack configuration when images are obfuscated using only the VAE encoder;
- A_2—E_1 attack configuration when images are obfuscated using only the non-bijective mapping;
- A_3—E_1 attack configuration when images are obfuscated using both encoding and non-bijective mapping;
- A_4—E_2 attack configuration when images are obfuscated using only the VAE encoder;
- A_5—E_2 attack configuration when images are obfuscated using only the non-bijective mapping;
- A_6—E_2 attack configuration when images are obfuscated using both encoding and non-bijective mapping.

Figures 13 and 14 display an example of a reconstructed angiography for each attack configuration. More angiographic samples and their recovered counterparts are presented in Appendix A, Figures A2 and A3. A visual comparison provides the first intuition on the reconstruction capabilities of the AI model in different scenarios. As expected, the performance of the reconstruction model is improved when the training dataset is similar to the targeted dataset, but, even so, all it can restore is the background of the angiographic frames. Because it is typically not possible to identify a patient based on the background of an angiography, this information is not considered sensitive. Even if the background can be recreated through AI-based methods, the obfuscation techniques are deemed secure against AI-based attacks as long as the object of interest (in this case, the coronary vessels) remains unrecognizable after the reconstruction.

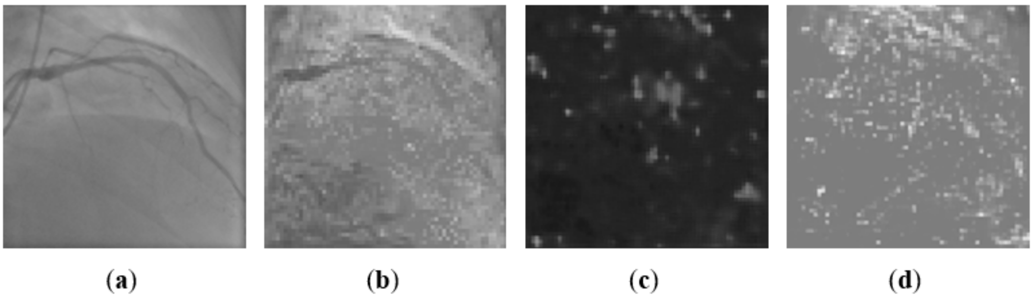

Figure 13. Comparison between (**a**) an original angiographic frame and the reconstructions obtained with the attack configurations (**b**) A_1, (**c**) A_2, and (**d**) A_3.

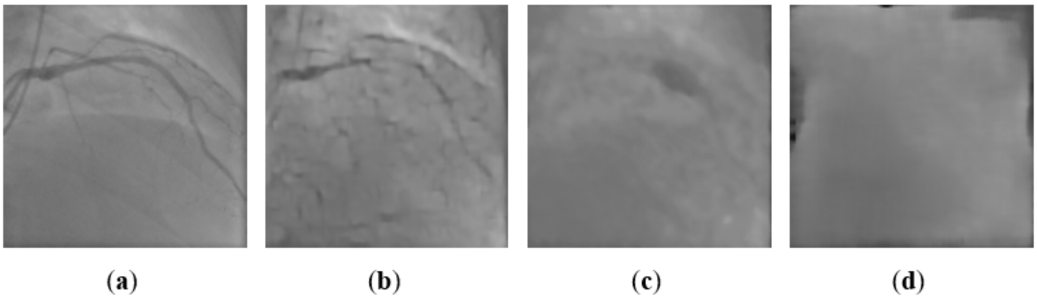

Figure 14. Comparison between (**a**) an original angiographic frame and the reconstructions obtained with the attack configurations (**b**) A_4, (**c**) A_5, and (**d**) A_6.

For a quantitative analysis, the similarity metrics discussed in Section 2.5 are computed. Higher scores for both SSIM and PSNR indicate better reconstructions. The average values for each attack configuration are presented in Table 5.

Table 5. Similarity between the original frames and the reconstructed images.

Attack Configuration	Experiment	SSIM	PSNR [dB]
E_1	A_1	0.6120	25.23
	A_2	0.3079	9.96
	A_3	0.5100	22.30
E_2	A_4	0.8173	29.54
	A_5	0.7593	26.91
	A_6	0.6855	23.63

These results support the conclusions drawn from the visual inspection. The VAE-based technique enables a certain degree of reconstruction but applying the non-bijective intensity mapping eliminates this shortcoming. Although for the E_1 attack configuration, the best privacy is achieved in experiment A_2 (only non-bijective intensity mapping), this result is not confirmed when a similar dataset is used for training the reconstruction model. For the E_2 setup, the results from A_4, A_5 and A_6 experiments show that, when the methods are used in conjunction, the quality of recreated frames is significantly affected: the vessels are no longer visible, and the background is diffuse.

Experiments A_3 and A_6 were repeatedly run for different values of N. Figures 15 and 16 show how SSIM and PSNR vary as a function of parameter N for the attack configurations E_1 and E_2. As expected, the reconstruction is impossible for $N = 1$ (all information is removed) and is slightly better in E_2 than in E_1. However, there is no monotonous tendency in any configuration. Conversely, in the case of PSNR, although the metrics for $N = 1$ are higher due to the background similarity, there is an oscillating downward trend suggesting that a smaller N implies an increased security level.

The results of the manual evaluation for all three obfuscation approaches are depicted in Figure 17. The mean scores regarding vessels and background reconstruction quality are displayed for each evaluator, alongside the average value. While similar scores were attributed to both vessels and background reconstructions when only encoding was used, for the other two obfuscation approaches, the recovered background presents a higher quality compared to the reconstructed vessels. However, overall, we observe a decreasing trend when comparing the three employed techniques. Even if applying the non-bijective intensity mapping results in a significant privacy improvement, a further decrease in reconstruction quality is noticed when the techniques are used in conjunction.

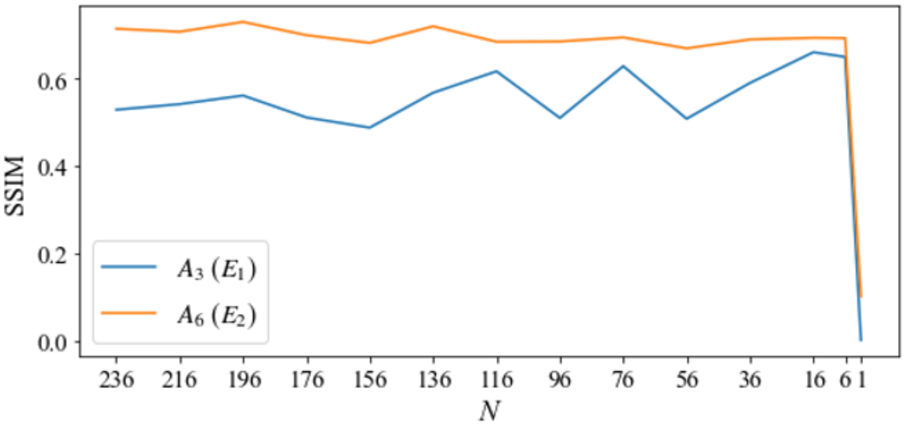

Figure 15. Influence of parameter N on SSIM in A_3 (E_1) and A_6 (E_2), respectively, configuration experiments.

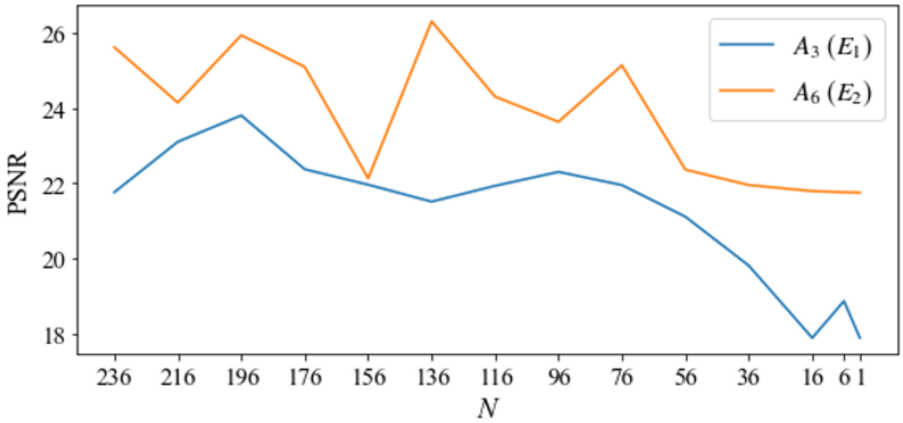

Figure 16. Influence of parameter N on PSNR in A_3 (E_1) and A_6 (E_2), respectively, configuration experiments.

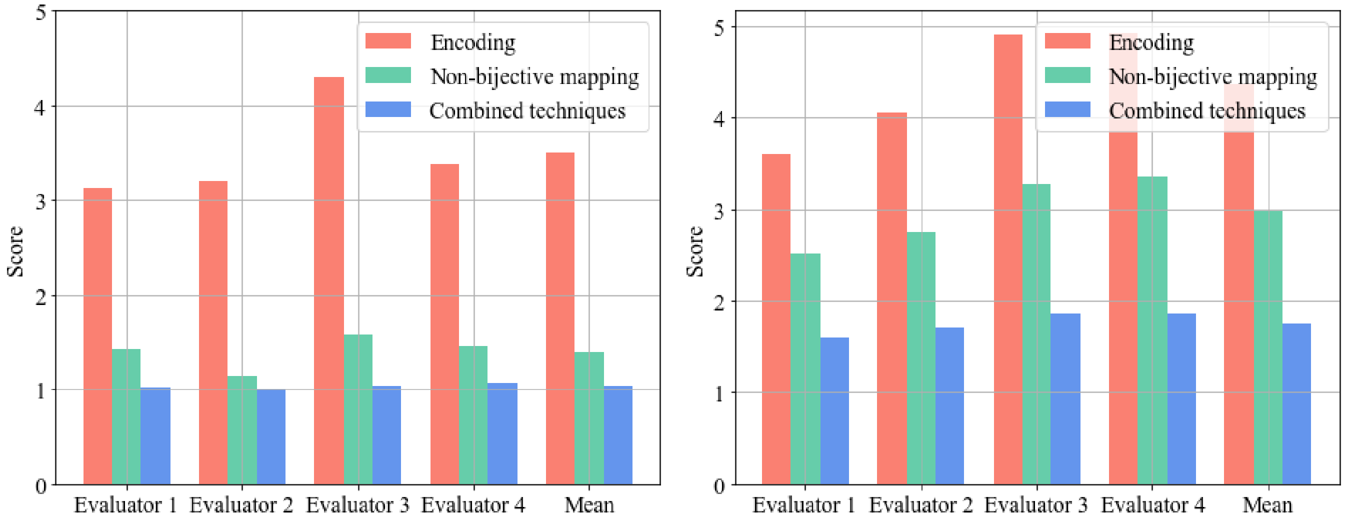

Figure 17. Mean scores of manual evaluation for vessels (**left**) and background (**right**) reconstructions.

Table 6 presents a numerical synthesis of the results. The mean and standard deviation of vessels and background evaluations are displayed for each obfuscation method. The standard deviation is smaller than 1 for each evaluation case, indicating low inter-user variability. The fact that the vessels were mainly evaluated with a score of 1 for the combined procedures strengthens the idea that this strategy provides robust security against recovery attempts.

Table 6. Mean scores regarding the quality of the reconstructed images.

	Encoding		**Non-Bijective Mapping**		**Combined Techniques**	
	Vessels	**Background**	**Vessels**	**Background**	**Vessels**	**Background**
Mean score	3.50 ± 0.91	4.37 ± 0.70	1.40 ± 0.51	2.97 ± 0.79	1.03 ± 0.17	1.75 ± 0.61

4. Discussion and Conclusions

4.1. Advantageous Properties and Limitations

A first key feature of the proposed obfuscation algorithm is its irreversibility. It is impossible to undo the encoding stage, since performing the decoding without having access to the trained decoder is impossible. Furthermore, reversing the non-bijective mapping is also impossible since it establishes many-to-one relationships, resulting in a large number of alternative substitutes. When the data is sensitive but the partners are untrustworthy, this property addresses the challenge of encryption key management generally involved in collaborative model training. Because there is no inverse function, this obfuscation algorithm can be used by multiple entities to create a common dataset and train a more robust DL model without exposing the original data to each other. Even if the receiver is considered trustworthy, the simple fact of externalizing the data exposes it to the risk of being accessed by an unauthorized party. Thus, an essential requirement that the approach must meet is the preservation of the image utility in the altered state. Once this is achieved, the collaborative work can be carried out based on a zero-trust architecture where the original data can be accessed only by authorized personnel inside the hospital environment.

Another benefit of this strategy is that the VAE does not have to be trained on the same or similar dataset as the one being obfuscated. Even if the encoder is not trained on the target dataset, the underlying information is preserved, and the privacy level is unaffected. As a result, there is no need to disclose sensitive data when training the encoder because any publicly available dataset can be used. Furthermore, it does not need to be trained in the clinical environment; it may be provided as a black-box tool. Furthermore, an outside party can not exploit the decoder to reverse the encoding if it is combined with non-bijective intensity shuffling.

The main drawback of the method is the drop in accuracy for the model that uses obfuscated images to perform a specific task. Such solutions, particularly in the medical domain, should provide performance comparable to that of an expert, to be adopted in the clinical decision-making process. Thus, a path for further development of the method would consist of integrating different strategies for enhancing the performance and robustness of models trained on secured images. An interesting research idea in this direction is to assess how training a DL-based model on a mixed dataset (containing both obfuscated and non-obfuscated synthetic images) would improve the performance. The usage of denoising modules directly in the classification network to increase accuracy without compromising privacy should also be investigated.

Another shortcoming is the lack of a precise security level quantification that would allow a clinical user to choose a particular algorithm configuration for a specific use case. To achieve a rigorous separation of the privacy and accuracy levels according to the obfuscation technique specifications, we intend to conduct within future work additional experiments which solve other medical tasks and employ different datasets and DL solutions.

4.2. Privacy-Utility Trade-off Considerations

While parameter N influences the confidentiality level of the obfuscation method that uses non-bijective functions, the degree of privacy provided by the VAE-based approach is dependent on the level of noise added during encoding, and the number of channels obtained at the bottleneck. When several channels are employed, the information is shared between them, resulting in fewer details being preserved in one channel and in more robust security. Furthermore, the valuable information is not evenly dispersed across the different channels, and one may select a particular representation to fulfill a specific requirement. Hence, the clinical user may select between different options for the trade-off between accuracy and privacy (e.g., categorical choice: very high accuracy, high accuracy, balanced, high privacy, very high privacy). For example, a very high privacy requirement may be chosen if easily recognizable patient features are present in the images (MRI data [35]).

Regarding the classification accuracy, although its value is still above 80% when the obfuscation approaches are combined, which is acceptable in the context of a privacy-accuracy trade-off, there is still room for improvement. The purpose of the classification experiments is not to achieve state-of-the-art performance on obfuscated images but to compare the results when the same architecture and different obfuscation techniques are employed. Because the classification task can be successfully performed even when using small values of N, we sought to explore the existence of structural dissimilarities between the LCA and the RCA, which could allow for a superior reconstruction for one of the classes. Figure 18 depicts two different samples (one LCA and one RCA angiographic acquisition) and their dichromatic obfuscated counterparts. Although assigning the proper category solely by visually inspecting the binary images is difficult, it is clear that some characteristics are preserved even when $N = 2$, which allows for a relatively accurate DL-based classification. The scores given by the expert readers for LCA and RCA reconstructions are recorded separately, to see if such differentiating details allow for a more qualitative recovery for specific samples. In terms of vessel scores, there is no substantial difference between the two categories, according to the visual inspection. However, it appears that the background can be better reconstructed for RCA views.

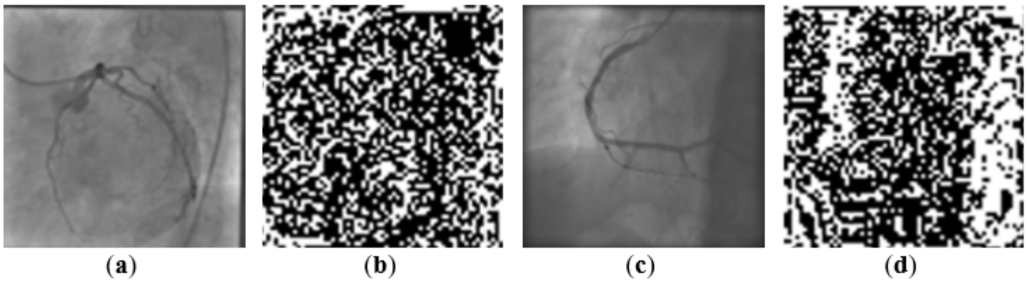

Figure 18. Comparison between LCA (**a**) and RCA (**c**) samples and their obfuscated counterparts when $N = 2$: LCA (**b**) and RCA (**d**).

To demonstrate that there is no statistically significant difference between the two groups, we compute the p-value. The scores are first standardized into the t-score. The p-value is calculated by considering a two-tailed hypothesis. A comparison between the results obtained for the three obfuscation approaches, where the vessels and the background are separately assessed, is presented in Table 7. As the significance level is set to 0.05 and all computed p-values exceed this threshold, we can confirm that the difference between reconstructed RCA and LCA frames is not statistically significant.

Table 7. Statistical significance assessment regarding the reconstruction difference between LCA and RCA views.

	Encoding		Non-Bijective Mapping		Combined Techniques	
	Vessels	**Background**	**Vessels**	**Background**	**Vessels**	**Background**
p-value	1	0.765	0.337	0.183	0.678	0.076

4.3. Final Conclusions

In this paper, we present an obfuscation approach that protects the privacy of medical images while allowing for DL model training. Although obfuscation techniques have been previously researched, integrating them into medical applications might be challenging due to the strict privacy and performance requirements. Mosaicing and blurring can be used to make faces and digits unrecognizable to the human eye, as shown in [16]. According to the authors, the obfuscation methods that were evaluated preserve enough information correlated to the original images. Thus, an accurate reconstruction is possible using AI-based models. The approaches proposed in [17] assume that only a part of the images from the dataset contains sensitive information, and these will be obfuscated. However, this is not the case when training models in medical DL-based applications, where the same level of confidentiality is required for all employed data. The method also implies the risk of affecting model accuracy if too many samples need to be secured, which again is not acceptable in a medical application where both privacy and accuracy are crucial.

A promising technique is presented in [18], where images are obfuscated by mixing their pixels with the pixels of another image. Other obfuscation methods were combined with the proposed technique to enhance security, and the experiments showed that the images are protected both from human perception and artificial recognition systems. The performance of models trained on obfuscated images varies with the level of privacy. The loss in accuracy significantly increases when methods are combined, and when privacy parameters are tuned for better security. Another aspect to consider is that the model is trained to perform the cat vs dog classification task, hence the properties of the classes are well defined, and there are many training samples available. However, since the differences between images are very subtle in specific medical imaging applications, and the available data is limited, it is unlikely that training a model on mixed images would achieve high accuracy. In the approach presented by Kim et al. [19], the patient identity is protected by transforming the brain MRI into a proxy image that is sent to the server for segmentation. The altered segmentation mask is then sent to the client, who restores it to the useful version. Compared to our method, this approach differs from the initial requirements perspective, as it is designed to allow for an accurate reconstruction of the processed image. To achieve this, an identity obfuscation loss and a transformation invertibility loss based on SSIM are minimized. The mean average precision and the F1-score are used to assess the re-identification accuracy in the case of an attacker attempting to match an encoded image or segmentation against an existing database. In [20,21], generative models (GANs) were used to create visually appealing images similar to the original ones in terms of basic shape, but distinct in terms of details. Applying this method to X-ray coronary angiographies, for example, might result in synthetic angiographic frames with characteristics which are significantly different from those in the original images (possible stenoses might be excluded, vessel ramifications might be modified, etc). This method is particularly challenging to apply in personalized medicine since the details of each image are required for a proper assessment, but the entire content is confidential. Furthermore, unlike the techniques discussed above, GAN-based methods do not secure information regarding the target objects or the objective of model training (in our use case, the MLaaS provider, or an interceptor who visualizes the obfuscated images, could tell that they are angiographic frames).

The proposed obfuscation algorithm was created with the requirements of a medical use case in mind. Only the computational overhead associated with the obfuscation phase is introduced. Once the data have been secured, training and inference are carried out as if plain data were used. Because the result of the obfuscation is still an image, there is no need for special deep learning libraries or frameworks. Although the privacy-accuracy trade-off must be considered, applying the obfuscation algorithm on medical images successfully hides the sensitive content from human perception and protects it against AI-based reconstruction attacks, while allowing for DL model training with satisfactory performance.

Author Contributions: Conceptualization, A.B.P. and C.I.N.; methodology, A.B.P., I.A.T., A.V., C.S. and L.M.I.; software, A.B.P. and I.A.T.; validation, A.V., C.I.N., A.S.-U. and L.M.I.; formal analysis, A.S.-U., C.I.N. and A.V.; investigation, A.B.P., I.A.T., C.I.N., A.V. and L.M.I.; resources, L.M.I. and A.S.-U.; data curation, A.B.P. and I.A.T.; writing—original draft preparation, A.B.P. and I.A.T.; writing—review and editing, C.I.N., A.V., C.S. and L.M.I.; visualization, A.B.P. and I.A.T.; supervision, C.S. and L.M.I.; project administration, L.M.I.; funding acquisition, L.M.I. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a grant of the Romanian Ministry of Education and Research, CCCDI—UEFISCDI, project number PN-III-P2-2.1-PED-2019-2415, within PNCDI III. This work was also supported by a grant of the Ministry of Education and Research, CNCS/CCCDI— UEFISCDI, project number PN-III-P3-3.6-H2020-2020-0145/2021. This project received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 875351.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Medical MNIST dataset is available at https://github.com/apolanco3225/Medical-MNIST-Classification (accessed on 23 March 2022). The coronary angiography data used to support the findings of this study have not been made available because they were acquired in a research grant (heart.unitbv.ro) which does not allow the publication of the data.

Conflicts of Interest: Andreea Bianca Popescu, Anamaria Vizitiu, Constantin Suciu, and Lucian Mihai Itu are employees of Siemens SRL, Brasov, Romania. Ioana Antonia Taca received scholarships from Siemens SRL, Brasov, Romania. The other authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

DL	Deep Learning
VAE	Variational Autoencoder
AI	Artificial Intelligence
ML	Machine Learning
MLaaS	Machine Learning as a Service
SSIM	Structural Similarity Index Measure
PSNR	Peak Signal-to-Noise-Ratio
ReLU	Rectified Linear Unit
LCA	Left Coronary Artery
RCA	Right Coronary Artery
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
CXR	Chest X-Ray
CNN	Convolutional Neural Network

Appendix A

Multiple examples of angiographic frames and the corresponding obfuscated or reconstructed counterparts are presented in this appendix.

Figure A1 displays for each original sample the obfuscated version obtained when encoding and the non-bijective map are used independently and in conjunction. The value of the parameter N used to attain the images displayed under (c) and (d) is 96.

Figure A2 presents reconstructed images in the E_1 attack configuration, when the malicious actor is aware that the target data are medical images but does not know their specific type. The original angiographies are shown in the first column.

The same frames are displayed in Figure A3 along with the recovered images in the E_2 attack configuration, where the threat actor knows that the targeted dataset contains coronary angiographies, and the reconstruction model is trained on a similar dataset. We observe that the more knowledgable the attacker is, the better the reconstruction performance is when only encoding is employed as a security measure. However, the coronary vessels are difficult to recover in both attack configurations, when the second step of the obfuscation algorithm is included.

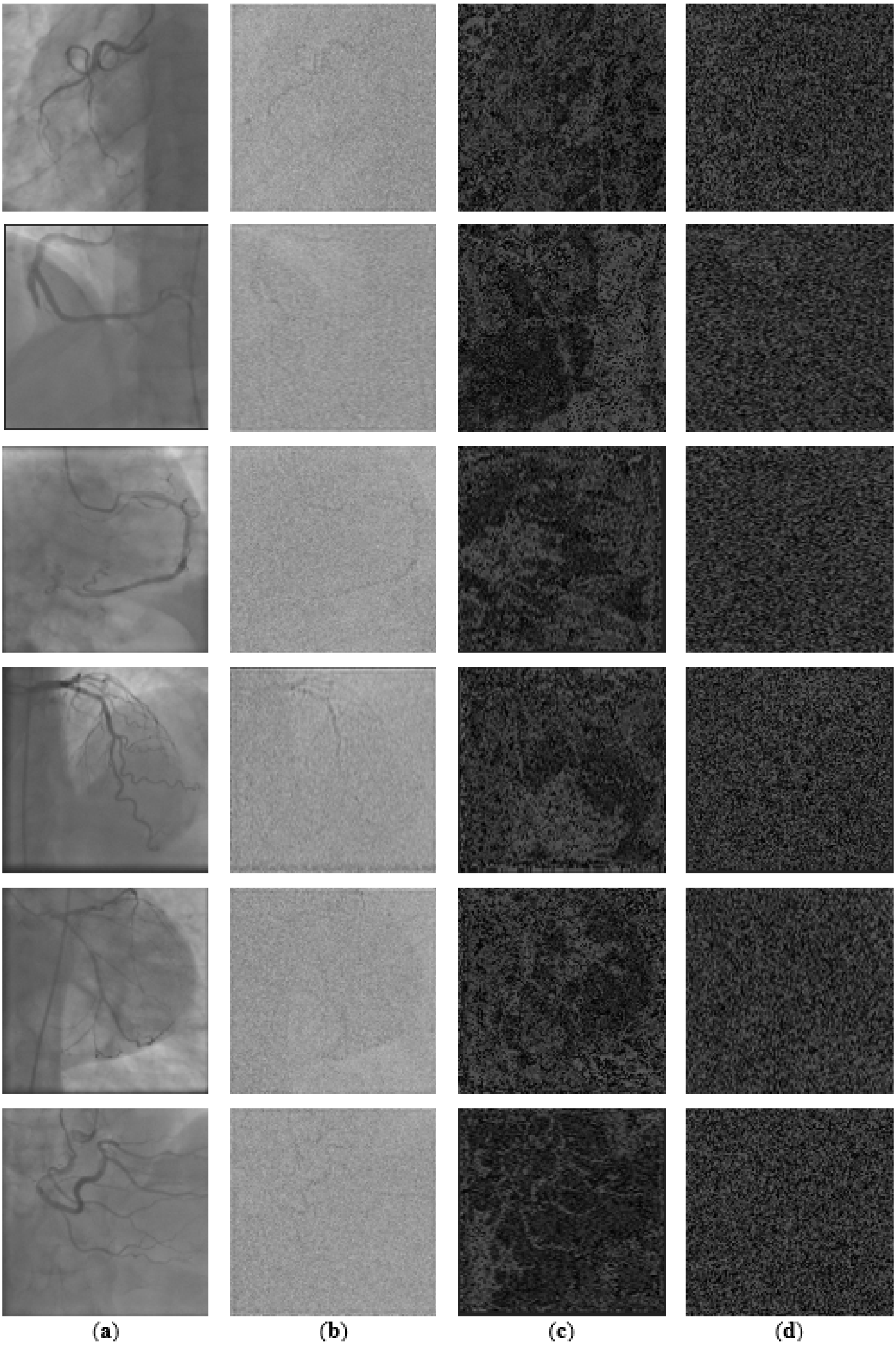

Figure A1. Comparison between (**a**) original and corresponding obfuscated angiographic frames using (**b**) encoding, (**c**) non-bijective intensity mapping, and (**d**) combined algorithm.

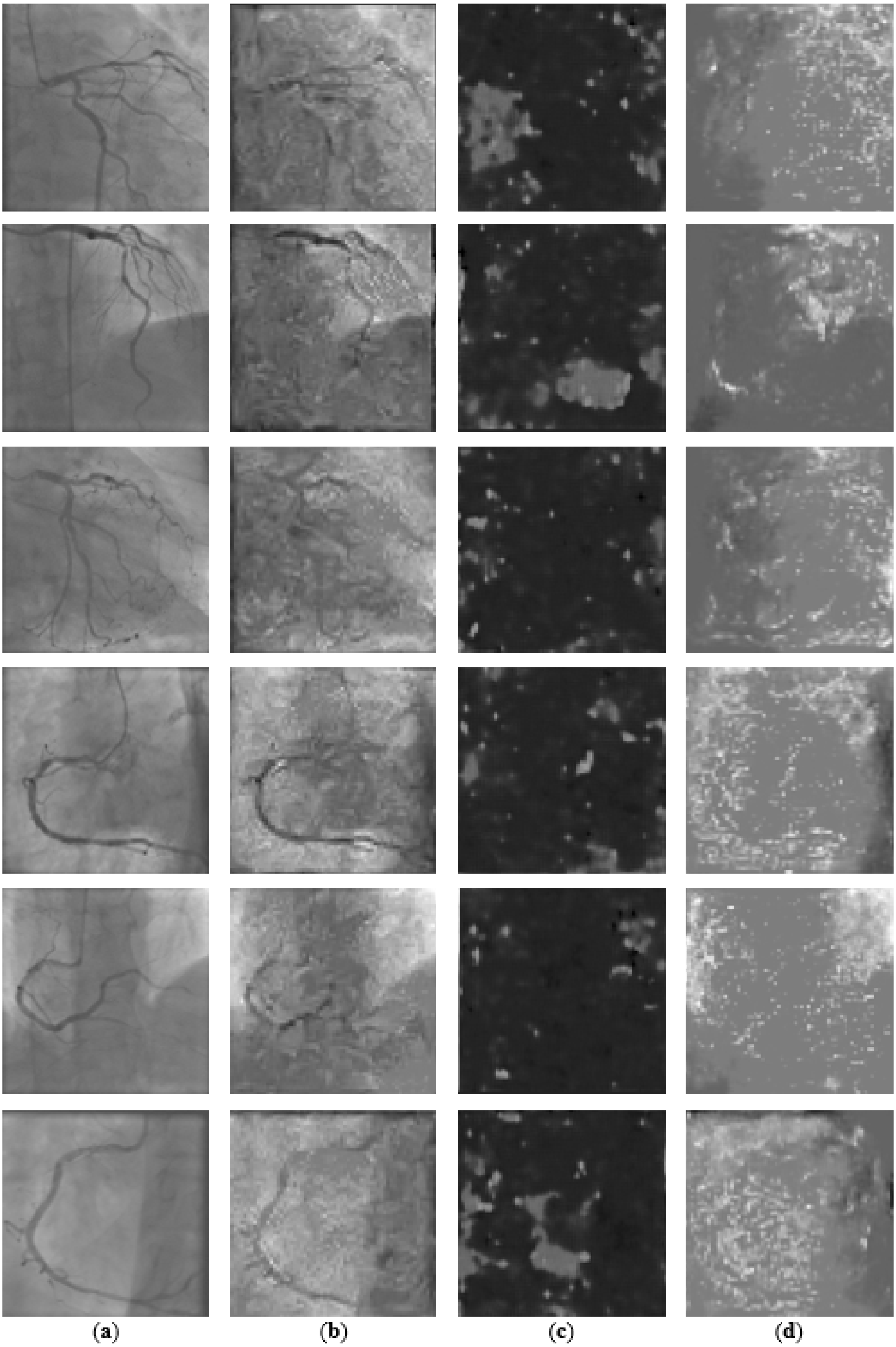

Figure A2. Comparison between (**a**) original angiographic frames and the reconstructions obtained with the attack configurations (**b**) A_1, (**c**) A_2, and (**d**) A_3.

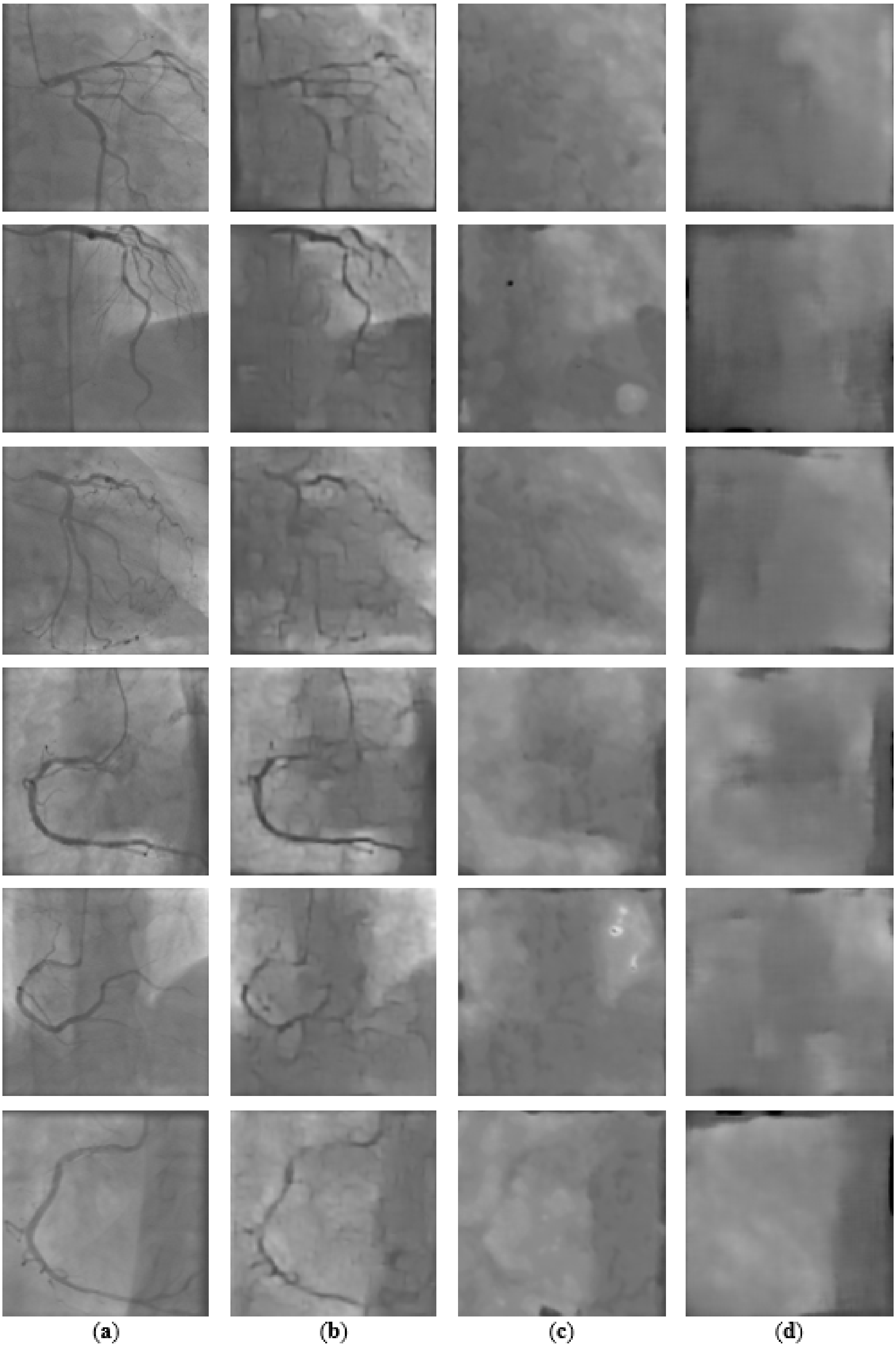

Figure A3. Comparison between (**a**) original angiographic frames and the reconstructions obtained with the attack configurations (**b**) A_4, (**c**) A_5, and (**d**) A_6.

References

1. Gui, C.; Chan, V. Machine learning in medicine. *Univ. West. Ont. Med. J.* **2017**, *86*, 76–78. [CrossRef]
2. Vayena, E.; Blasimme, A.; Cohen, I.G. Machine learning in medicine: Addressing ethical challenges. *PLoS Med.* **2018**, *15*, e1002689. [CrossRef]
3. Pulido-Gaytan, L.B.; Tchernykh, A.; Cortés-Mendoza, J.M.; Babenko, M.; Radchenko, G. A Survey on Privacy-Preserving Machine Learning with Fully Homomorphic Encryption. In *Latin American High Performance Computing Conference*; Springer: Cham, Switzerland, 2020; pp. 115–129.
4. Orlandi, C.; Piva, A.; Barni, M. Oblivious neural network computing via homomorphic encryption. *EURASIP J. Inf. Secur.* **2007**, *2007*, 37343. [CrossRef]
5. Gilad-Bachrach, R.; Dowlin, N.; Laine, K.; Lauter, K.; Naehrig, M.; Wernsing, J. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 201–210.
6. Hesamifard, E.; Takabi, H.; Ghasemi, M. Cryptodl: Deep neural networks over encrypted data. *arXiv* **2017**, arXiv:1711.05189.
7. Ishiyama, T.; Suzuki, T.; Yamana, H. Highly accurate CNN inference using approximate activation functions over homomorphic encryption. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 3989–3995.
8. Lee, J.W.; Kang, H.; Lee, Y.; Choi, W.; Eom, J.; Deryabin, M.; Lee, E.; Lee, J.; Yoo, D.; Kim, Y.S.; et al. Privacy-preserving machine learning with fully homomorphic encryption for deep neural network. *arXiv* **2021**, arXiv:2106.07229.
9. Al Badawi, A.; Jin, C.; Lin, J.; Mun, C.F.; Jie, S.J.; Tan, B.H.M.; Nan, X.; Aung, K.M.M.; Chandrasekhar, V.R. Towards the alexnet moment for homomorphic encryption: Hcnn, the first homomorphic cnn on encrypted data with gpus. *IEEE Trans. Emerg. Top. Comput.* **2020**, *9*, 1330–1343. [CrossRef]
10. Kipnis, A.; Hibshoosh, E. Efficient methods for practical fully homomorphic symmetric-key encrypton, randomization and verification. *Cryptology ePrint Archive*. 2012. Available online: https://ia.cr/2012/637 (accessed on 23 March 2022).
11. Vizitiu, A.; Niţă, C.I.; Puiu, A.; Suciu, C.; Itu, L.M. Applying Deep Neural Networks over Homomorphic Encrypted Medical Data. *Comput. Math. Methods Med.* **2020**, *2020*, 3910250. [CrossRef]
12. Vizitiu, A.; Nita, C.I.; Toev, R.M.; Suditu, T.; Suciu, C.; Itu, L.M. Framework for Privacy-Preserving Wearable Health Data Analysis: Proof-of-Concept Study for Atrial Fibrillation Detection. *Appl. Sci.* **2021**, *11*, 9049. [CrossRef]
13. Popescu, A.B.; Taca, I.A.; Nita, C.I.; Vizitiu, A.; Demeter, R.; Suciu, C.; Itu, L.M. Privacy preserving classification of eeg data using machine learning and homomorphic encryption. *Appl. Sci.* **2021**, *11*, 7360. [CrossRef]
14. Guang-Li, X.; Xin-Meng, C.; Ping, Z.; Jie, M. A method of homomorphic encryption. *Wuhan Univ. J. Nat. Sci.* **2006**, *11*, 181–184. [CrossRef]
15. Chen, H.; Hussain, S.U.; Boemer, F.; Stapf, E.; Sadeghi, A.R.; Koushanfar, F.; Cammarota, R. Developing privacy-preserving AI systems: The lessons learned. In Proceedings of the 2020 57th ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA , 20–24 July 2020; pp. 1–4.
16. McPherson, R.; Shokri, R.; Shmatikov, V. Defeating image obfuscation with deep learning. *arXiv* **2016**, arXiv:1609.00408.
17. Zhang, T.; He, Z.; Lee, R.B. Privacy-preserving machine learning through data obfuscation. *arXiv* **2018**, arXiv:1807.01860.
18. Raynal, M.; Achanta, R.; Humbert, M. Image obfuscation for privacy-preserving machine learning. *arXiv* **2020**, arXiv:2010.10139.
19. Kim, B.N.; Dolz, J.; Desrosiers, C.; Jodoin, P.M. Privacy Preserving for Medical Image Analysis via Non-Linear Deformation Proxy. *arXiv* **2020**, arXiv:2011.12835.
20. Li, T.; Choi, M.S. DeepBlur: A simple and effective method for natural image obfuscation. *arXiv* **2021**, arXiv:2104.02655.
21. Chen, J.W.; Chen, L.J.; Yu, C.M.; Lu, C.S. Perceptual Indistinguishability-Net (PI-Net): Facial image obfuscation with manipulable semantics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6478–6487.
22. Kingma, D.P.; Welling, M. An introduction to variational autoencoders. *arXiv* **2019**, arXiv:1906.02691.
23. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
24. Barber, D. *Bayesian Reasoning and Machine Learning*; Cambridge University Press: Cambridge, UK, 2012.
25. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092. [CrossRef]
26. Geman, S.; Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *PAMI-6*, 721–741. [CrossRef]
27. Bernardo, J.; Bayarri, M.; Berger, J.; Dawid, A.; Heckerman, D.; Smith, A.; West, M.; Beal, M.; Ghahramani, Z. The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. *Bayesian Stat.* **2003**, *7*, 210.
28. Ganguly, A.; Earp, S.W. An Introduction to Variational Inference. *arXiv* **2021**, arXiv:2108.13083.
29. Apolanco3225. Medical MNIST Classification. 2017. Available online: https://github.com/apolanco3225/Medical-MNIST-Classification (accessed on 23 March 2022).
30. Yin, B.; Scholte, H.S.; Bohté, S. LocalNorm: Robust Image Classification Through Dynamically Regularized Normalization. In *International Conference on Artificial Neural Networks*; Springer: Cham, Switzerland, 2021; pp. 240–252.
31. Chollet, F. keras. 2015. Available online: https://github.com/fchollet/keras (accessed on 23 March 2022).

32. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015; pp. 234–241.
33. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
34. Van der Walt, S.; Schönberger, J.L.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J.D.; Yager, N.; Gouillart, E.; Yu, T. Scikit-image: Image processing in Python. *PeerJ* **2014**, *2*, e453. [CrossRef] [PubMed]
35. Schwarz, C.G.; Kremers, W.K.; Therneau, T.M.; Sharp, R.R.; Gunter, J.L.; Vemuri, P.; Arani, A.; Spychalla, A.J.; Kantarci, K.; Knopman, D.S.; et al. Identification of anonymous MRI research participants with face-recognition software. *N. Engl. J. Med.* **2019**, *381*, 1684–1686. [CrossRef] [PubMed]