Non-bijectivity-based image obfuscation method for deep learning based medical applications

Andreea Bianca Popescu, Ioana Antonia Taca, Anamaria Vizitiu, Lucian Mihai Itu
Siemens SRL
Brasov, Romania
andreea.popescu.ext@siemens.com, nita.cosmin.ioan@unitbv.ro, ioana_antonia29@yahoo.com

Andreea Bianca Popescu, Cosmin Ioan Nita, Ioana Antonia Taca, Anamaria Vizitiu, Lucian Mihai Itu
Transilvania University of Brasov
Brasov, Romania
anamaria.vizitiu@siemens.com, lucian.itu@siemens.com

Abstract— **As more and more deep learning (DL) solutions are employed in the healthcare domain using the Machine Learning as a Service (MLaaS) paradigm, concerns regarding personal data privacy have been raised. In this context, especially in medical imaging, the demand for privacy-preserving techniques, that allow for DL model development, has recently increased significantly. Herein, we propose a medical image obfuscation algorithm based on pixel intensity shuffling and non-bijective functions. The proposed algorithm is evaluated on a medical use case based on coronary angiographic images. Multiple convolutional neural networks are trained to measure the utility of the obfuscated images. An attack configuration based on artificial intelligence (AI) is evaluated to validate the level of privacy. The classification performance on the obfuscated images is satisfactory, while the computational time is not affected significantly. Visual and metrics-based analyses show that the data is protected from human perception and from AI-based image reconstruction approaches.**

Keywords—image obfuscation, non-bijective functions, deep learning, medical imaging, coronary angiography

I. INTRODUCTION

Deep learning (DL) based solutions have proved their utility in multiple areas in the past years. A significant accomplishment is the use of DL applications in healthcare, where such approaches have shown remarkable results in assisting clinicians in diagnosis, treatment, and prevention [1]. Conversely, significant data amounts are required to ensure that DL models achieve high accuracy. Health data typically contains sensitive and personal patient information, hence, the sharing of data outside the clinical center is conditioned by performing a proper anonymization [2]. To overcome the concerns regarding data confidentiality, privacy-preserving techniques that allow for neural network training have been developed (homomorphic encryption, secure multiparty computation, differential privacy). Although publications demonstrate the possibility of integrating homomorphic encryption (HE) in artificial intelligence methods [3-5], most proposed schemes have limitations that hinder their real-world utility. For instance, the increasing noise affects the number of correct consecutive operations in the BFV scheme [6], while other solutions allow only for addition and multiplication on small integers [7], [8]. Moreover, their mathematical complexity influences also the computational time. These drawbacks make homomorphic encryption unsuitable for DL-based medical applications, where both time and accuracy are crucial. Furthermore, medical data are acquired and stored in a large variety of formats. Besides tabular or time-series data (e.g., EKG), that are easier to process when privacy-preserving techniques are employed, more complex acquisitions are represented by medical images. A single image sample contains significantly more information than tabular or time series data. Thus, applying DL methods on images protected through HE is infeasible, as it implies a more substantial computational overhead.

Another approach is to hide the image content through obfuscation while allowing for DL model training, with no computational overhead. McPherson et al. [9] demonstrated that DL can still achieve high performance in face, number, or object recognition, even if the images are obfuscated. The authors showed that "mosaicing" and blurring could transform faces and digits and make them unrecognizable by the human eye, but an artificial intelligence standard model can still extract useful information from the obfuscated images. The approaches proposed in [10] assume that only some of the images from the dataset contain sensitive information, and these will be obfuscated. There is though the risk of affecting model accuracy if too many samples need to be secured. To ensure protection against attacks based on statistical methods, the authors have proposed dataset augmentation with fake synthetic generated samples that do not influence the model performance. A promising technique is presented in [11], where images are obfuscated by mixing the pixels of two images. Multiple obfuscation methods were combined with the proposed technique to enhance security, and the experiments indicated that the images are protected both from human perception and artificial recognition systems. A different approach was considered in [12] and [13], where generative models were used to create visually pleasing images similar to the original ones in terms of general shape but different concerning the details. According to the authors of [13], this technique could be helpful in the context of training a model for face detection while ensuring privacy against face recognition. This approach would be challenging to adapt for

978-1-6654-8162-5/22/$31.00 ©2022 IEEE

medical imaging applications, where the details are essential for classification, but the entire content needs to be protected.

The application scenario is formulated in the context of developing AI algorithms that can automatically analyze images. As these solutions may not be owned or deployed by the entity that uses them, confidentiality concerns dictate the demand for privacy preserving techniques that allow for AI model training on secure data. Although the images are altered through obfuscation, this does not affect the physicians' analyses since they can access the original images from inside the secure environment of the hospital. The obfuscation method aims to protect the images while they are analyzed by artificial intelligence models deployed by an external party.

In this paper we present an image obfuscation method based on pixel intensity shuffling, designed for meeting the following requirements: (i) hiding the content of an image from the human eye, (ii) making AI-based image reconstruction difficult, and (iii) allowing for model training which leads to high accuracy. The rest of the manuscript is structured as follows. Section II describes the methods and materials used in our experiments: obfuscation algorithm, datasets, workflows, and network architectures. The experiments conducted from the clinical user and the threat actor perspectives, along with the corresponding results, are presented in Section III. Section IV highlights the advantages of the obfuscation techniques for image-based DL analyses, concluding our work.

II. METHODS

A. Image obfuscation algorithm

The first step of the proposed obfuscation algorithm consists of randomly shuffling the pixel intensities. For this, every potential pixel intensity (integer values in the range [0, 255]) is paired with a value from the same interval. Because each domain member has only one corresponding element in the codomain, this correlation represents a bijective function. Although the substitutions are arbitrarily chosen, which makes images unrecognizable, this method is vulnerable to DL-based or reverse-engineering attacks. Because third parties may have a black-box version of the obfuscation technique, new images may be obfuscated with the same tool, and statistical analysis could be able to indicate that a one-to-one mapping was utilized. Inverting this mapping would allow a possible threat actor to recover the original frames with zero loss. Another attack technique could rely on a deep learning model to recover the obfuscated images. Because such methods are considered powerful enough to learn a bijective function, we presume that this is the approach chosen by the attacker in the experiments that will follow. To prevent this type of attack, the obfuscation algorithm's second step is to change the mapping such that the injectivity attribute is lost. Consequently, several domain elements will be associated with the same codomain element. A modulo operation is used on each value of the bijective map to accomplish this result. The two steps of the proposed algorithm are depicted in Fig. 1. An important observation is that, even if the use of different non-bijective functions for distinct images would improve the security, such obfuscated frames could not be successfully employed in DL applications.

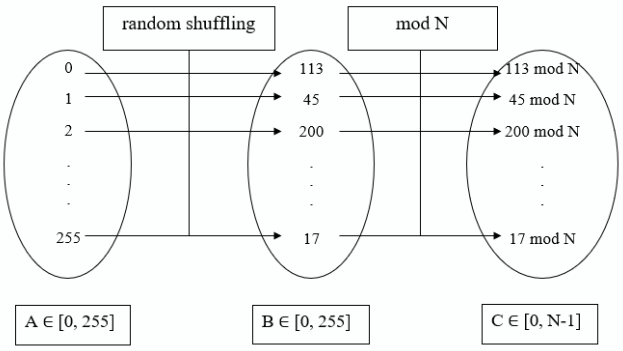

Fig. 1. Schematic representation of the obfuscation algorithm.

To train a classifier, for example, all the images (used either for training or inference) should be part of the same distribution, so the same bijective function must be applied. The structural similarity metric (SSIM) and peak signal-to-noise ratio (PSNR) are computed between original and obfuscated images to evaluate the security against human eye recognition. Since these similarity metrics agree with the human perception, they can be regarded as obfuscation level measures.

B. Clinical user perspective and use case description

A first perspective considered when analyzing the usage scenario of an obfuscation technique is that of the clinical user (e.g., hospital, patient) who regards data as sensitive and private. However, to reduce the diagnosis time in certain use cases, there is a need for developing a DL based model that could solve an easy but tedious task before a doctor performs the actual evaluation. An exemplary task is the classification of the view of an X-ray coronary angiography. Considering that the hospital does not have the hardware resources and especially the expertise to develop and deploy a deep learning classification model, the solution is to use the services of a third party, which may have an impact on data confidentiality. This external party is in this case a Machine Learning as a Service (MLaaS) provider that can train a DL model with the data provided by the clinical user, and then make it accessible as a service in the cloud for inference. The workflow is as follows: patients' consent that their data will be used in model training, the hospital collects data and creates a dataset; this dataset is sent to an MLaaS provider that trains a model; the hospital uses this model for remote inference by sending a sample to the MLaaS platform and receiving the classification result. Every angiographic frame used for training or inference is obfuscated to preserve the privacy of the data outside the hospital environment.

In our experiments, we use an in-house dataset of frames depicting either the right coronary artery (RCA) or the left coronary artery (LCA). It contains 3280 coronary angiographies and is balanced between the two classes. A subset of 680 images is used for validation, and another subset of 702 images is kept for evaluation purposes. The remaining frames are augmented (e.g., through shifting, zooming, rotation), resulting in 9980 images used in training. The size of these frames is 512x512 pixels, but experiments with different

input shapes have shown that 128x128 pixels is a setting which ensures satisfactory classification performance while requiring less computational time. Min-max scaling is applied to normalize the pixel values in the [0, 1] interval. The architecture of the classifier trained to distinguish between RCA and LCA in angiographic frames consists of four convolutional layers followed by two fully connected layers. The first and the third convolutional layers use kernels with a size of 5x5 for an increased receptive field, while the other two convolutional layers use 3x3 filters. Average-pooling is used for down-sampling, and the SELU activation function is chosen. The activation function on the final layer is the sigmoid function, and binary cross-entropy is employed as a loss function. Training is performed for 30 epochs, with a batch size of 64 and a learning rate equal to 0.001. The classification accuracy is used as evaluation metric.

C. *Threat actor perspective*

In the following, we also analyze the perspective of an external party (e.g., the MLaaS provider, an interceptor) willing to gain access to the non-obfuscated version of the data sent by the hospital for inference. Since the obfuscation algorithm is publicly released as a black-box tool, we also assume that the threat actor can use this tool to obfuscate any dataset. Moreover, because the data source is known, the attacker can estimate that the dataset consists of medical images, but does not know their specific type (coronary angiographies in our case). The workflow of an entity willing to gain unauthorized access to the data sent by the hospital 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
- intercept obfuscated images sent by the hospital and reconstruct the original ones using the previously trained model.

In the following experiments, we simulate an attack based on the U-net architecture proposed for the first time in [14]. The objective is to develop a model that takes as input an obfuscated image and outputs an image ideally identical, or at least very similar, with the original one. The architecture consists of an encoder and a decoder. The encoder contains two convolution operations (a convolution block) that preserve the size of the image, followed by a max-pooling operation that down-samples the activation map. On this map, another convolution block is applied. Then, as a part of the decoder, a transpose convolution is performed to up-sample the activation map, and the result is concatenated with specific intermediate values from the encoder. For every pixel of the obfuscated image, a corresponding pixel in the reconstructed image is predicted. As it is presumed that the attacker knows that the target data consists of medical images but cannot precisely determine what those images contain, the Medical MNIST dataset, publicly available [15], is used to train the reconstruction model. It contains six classes of X-ray images, each of them totaling around 7,000 samples. A subset of 70% is used in training, and the rest is kept for validation and testing. Image size is 64x64 pixels, allowing for a larger batch size (128). The model is trained for 20 epochs with a learning rate of 0.001. SSIM and PSNR are computed between original and reconstructed images to evaluate image similarity. The obfuscation algorithm is implemented in Python, and all deep learning models are developed using the PyTorch [16] framework.

III. EXPERIMENTS AND RESULTS

A. *Classification experiments*

The first set of experiments consists of training multiple classifiers on original and obfuscated data for different values of the parameter N. The results are synthesized in Table I. Fig. 2 represents a visual comparison between the original frame and the four different levels of obfuscation.

Although the model performs well on the original data, the obfuscation step (in the bijective form, when N=256) decreases the accuracy for the test set by more than 10%. However, as the dataset is balanced and the two classes are predicted equally, this performance may still be considered satisfactory. The subsequent three experiments show that, up to a certain point, reducing the parameter N, and, thus, applying a non-bijective obfuscation does not have a significant influence on the classification performance. For this specific use case, the threshold seems to be around N=50: the accuracy drops more when a smaller value of N is used. Although the use case application is represented by a binary classification which is not complex, the experiments demonstrate that the non-bijective-based obfuscation preserves enough information for an AI model to learn and achieve decent results. The intent is not to present a state-of-the-art AI-based solution for clinical applications but to determine how an AI model would perform when the training images are secured through different levels of obfuscation.

The corresponding metrics that numerically describe the similarity between the original and obfuscated images presented in Fig. 2 are also displayed in Table I. All SSIM values are under 0.1, and the PSNR is below 30 dB for all four cases, indicating almost no structural similarity with respect to the original angiographic image.

B. *Reconstruction experiments*

Two reconstruction experiments are considered to measure the impact of non-injectivity in the obfuscation method. In the first one, the images are obfuscated through a bijective function. Fig. 3 displays three samples along with the obfuscated and the reconstructed versions. A similar comparison, but for a non-bijective obfuscation, is depicted in Fig. 4.

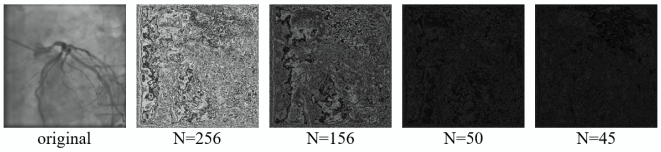

Fig. 2. Comparison between the original frame and four levels of obfuscation.

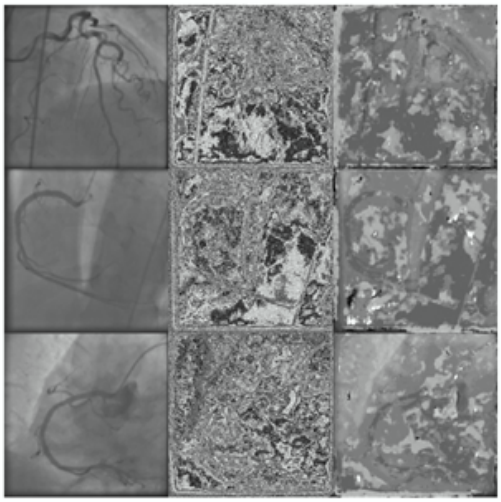

Fig. 3. Comparison between original (left), bijective obfuscated (middle) and reconstructed images (right).

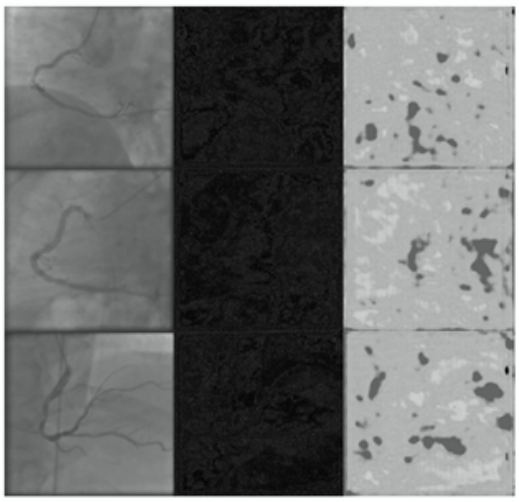

Fig. 4. Comparison between original (left), non-bijective obfuscated (middle) and reconstructed images (right).

Even in the bijective case, the reconstruction is significantly different from the target image. However, the curvatures of the aortic vessels are still visible both in the obfuscated and the reconstructed images. Conversely, in the non-bijective case, any information suggesting what is depicted in the medical frames is hidden. The similarity metrics confirm the difference between these approaches formulated after the visual evaluation. Table II includes the average SSIM and PSNR values computed over the entire testing angiographic subset.

IV. CONCLUSIONS

A related work that is focused on protecting medical data is presented in [17]. A client-server system is proposed in which the client protects the patient's identity by deforming the input image using an end-to-end adversarial system. The brain MRI is converted into a proxy image by the client and sent to the server for segmentation. The client receives the distorted segmentation mask and returns it to its original state. This methodology differs from our method in terms of initial needs, as it is meant to allow for an accurate reconstruction of the processed image. The attack vector consists of matching an encoded image or segmentation to an existing database, the re-identification accuracy being evaluated using the mean average precision and the F1-score.

TABLE I. CLASSIFICATION PERFORMANCE AND SIMILARITY METRICS FOR DIFFERENT LEVELS OF OBFUSCATION

Data description	Test accuracy	SSIM	PSNR [dB]
Original	94.73 %	-	-
Obfuscated (bijective; N=256)	83.48 %	0.0109	9.780
Obfuscated (non-bijective; N=150)	84.47 %	0.0368	11.1784
Obfuscated (non-bijective; N=50)	84.05 %	0.0940	8.230
Obfuscated (non-bijective; N=45)	79.77 %	0.0147	8.129

Compared to other privacy-preserving techniques that increase the complexity of data representation, image obfuscation has the advantage of conserving the shape and the size of input data. Thus, the computational time is not significantly influenced, the only overhead being introduced by the obfuscation process itself, which is negligible. Because the data format is also preserved (the output of the obfuscation algorithm is still an image), applying an already developed classification model is possible without any change in the implementation. The non-bijective function proposed as obfuscation technique transforms images, making them unrecognizable by the human eye, and impossible to reconstruct with a DL model. Although using obfuscated images as input to the classifier implies a trade-off between accuracy and privacy, the method can successfully hide sensitive information while allowing for DL-based analyses.

TABLE II. AVERAGED SIMILARITY METRICS BETWEEN THE ORIGINAL FRAMES AND THE RECONSTRUCTED IMAGES

Data description	Average SSIM	Average PSNR [dB]
Obfuscated (bijective; N=256)	0.676	18.11
Obfuscated (non-bijective; N=50)	0.608	12.45

ACKNOWLEDGMENT

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 project received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 875351.

REFERENCES

[1] R. Miotto, F. Wang, S. Wang, X. Jiang, and J.T. Dudley, "Deep learning for healthcare: review, opportunities and challenges", in Briefings in bioinformatics, vol. 19 (6), pp. 1236-1246, 2018.

[2] R. Shokri and V. Shmatikov, “Privacy-preserving deep learning”, in Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, pp. 1310-1321, 2015.

[3] C. Orlandi, A. Piva, and M. Barni, “Oblivious neural network computing via homomorphic encryption”, EURASIP Journal on Information Security, pp. 1-11, 2007.

[4] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, “Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy”, in International conference on machine learning, pp. 201-210, 2016.

[5] E. Hesamifard, H. Takabi, and M. Ghasemi, “Cryptodl: Deep neural networks over encrypted data”, arXiv preprint arXiv:1711.05189, 2017.

[6] J. Fan and F. Vercauteren, “Somewhat practical fully homomorphic encryption”, Cryptology ePrint Archive, 2012.

[7] P. Paillier, “Public-key cryptosystems based on composite degree residuosity classes”, in International conference on the theory and applications of cryptographic techniques, Springer, Berlin, Heidelberg, pp. 223-238, 1999.

[8] T. ElGamal, “A public key cryptosystem and a signature scheme based on discrete logarithms”, IEEE transactions on information theory, vol. 31 (4), pp. 469-472, 1985.

[9] R. McPherson, R. Shokri, and V. Shmatikov, “Defeating image obfuscation with deep learning”, arXiv preprint arXiv:1609.00408, 2016.

[10] T. Zhang, Z. He, R. B. Lee, “Privacy-preserving machine learning through data obfuscation”, arXiv preprint arXiv:1807.01860, 2018.

[11] M. Raynal, R. Achanta, and M. Humbert, “Image Obfuscation for Privacy-Preserving Machine Learning”, arXiv preprint arXiv:2010.10139, 2020.

[12] T. Li and M. S. Choi, “DeepBlur: A simple and effective method for natural image obfuscation”, arXiv preprint arXiv:2104.02655 1, 2021.

[13] J. W. Chen, L. J. Chen, C. M. Yu, and C. S. Lu, “Perceptual Indistinguishability-Net (PI-Net): Facial Image Obfuscation with Manipulable Semantics”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6478-6487, 2021.

[14] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation”, in International Conference on Medical image computing and computer-assisted intervention, Springer, Cham, pp. 234-241, 2015.

[15] Medical MNIST dataset: https://www.kaggle.com/andrewmvd/medical-mnist. Accessed: 17.09.2021.

[16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library”, in Advances in Neural Information Processing Systems 32 [Internet]. Curran Associates, Inc., pp. 8024–35, 2019.

[17] B. N. Kim, J. Dolz, C. Desrosiers, and P. M. Jodoin, “Privacy Preserving for Medical Image Analysis via Non-Linear Deformation Proxy”, arXiv preprint arXiv:2011.12835