De-identification and Obfuscation of Gender Attributes From Retinal Scans

Anonymous[1 [0000−1111−2222−3333]]

Anonymous Organization
xxx@xxx.xxx

Abstract. Retina images are considered to be important biomarkers and have been used as clinical diagnostic tools to detect multiple diseases. We examine multiple techniques for de-identifying retina images while maintaining their clinical ability for detecting diabetic retinopathy (DR), using gender as a proxy for identifiability. We apply two differential privacy algorithms, Snow and VS-Snow, on the entire image (globally) and on blood vessels only (locally) to obfuscate important image features that can predict a patient's sex. We evaluate the level of privacy and retained clinical predictive power of these de-identified images by using attacking gender classifier models and downstream disease classifiers. We show empirically that our proposed VS-Snow framework achieves strong privacy while preserving a meaningful clinical predictive power across different patient populations.

Keywords: Fundus images · Data Privacy · De-identification.

1 Introduction

Automated machine learning (ML) systems have shown great potential to enhance medical decisions and improve healthcare access. However, public medical datasets sharing has always been a bottleneck for the development of equitable clinical AI [18]. The release of such datasets is strictly regulated by the Health Insurance Portability and Accountability Act (HIPAA) to control the risk of leakage of private attributes. This risk is particularly notable in ophthalmology, as retinal fundus photos are considered accurate and identifiable biomarkers. Rapid advancements in artificial intelligence also introduce novel privacy risks. Previous research has demonstrated that deep learning models could accurately predict private attributes like gender from retinal fundus images [7], raising concerns for potential malicious uses of this information.

More importantly, datasets that contain biases in private attributes like gender tend to produce biased AI algorithms [8]. Consequently, biased algorithms tend to produce stereotypical diagnoses and under-perform on minority patient groups, which is extremely dangerous in the context of healthcare as this can yield discriminatory outcomes.

We propose a framework to mitigate the privacy and fairness risks from the root: datasets. We aim not only at promoting AI fairness, but we envision making

private data sharing a viable future. Specifically, our work investigates privacy concerns arising from the public release of two retina fundus image datasets, using gender as a proxy for privacy. We show that our clinically-inspired de-identification algorithms significantly reduce the ability of an adversary to distinguish a patient's gender, while retaining most utility for downstream tasks like the identification of diabetic retinopathy (DR).

1.1 Differential privacy for image obfuscation

Originally designed for statistical databases (i.e., the US Census), differential privacy algorithms have also been reinvented for medical image obfuscation. The purpose of image obfuscation is to modify an image such that sensitive information is no longer discernible in the image. Recently, differential privacy has been applied to iris images, which are also human biomarkers. For example, the Snow method [5] employs pixel-level noise by arbitrarily re-assigning pixel intensities to a constant value, i.e., 127 for grayscale images, based on hyperparameter probability of p. Snow achieves $(0, \delta)$-differential privacy with $\delta = 1 - p$ and protects individual pixels in the input image. More formally, $Pr[snow(x) \in S] \leq Pr[snow(x') \in S] + \delta$, where x and x' are neighboring pixels.

1.2 Deep learning for diabetic retinopathy and sex classification

In recent years, researchers have applied various computer vision models for diabetic retinopathy detection on diverse data and population [15, 17]. Some models achieve high performance in detecting referable diabetic retinopathy, which is comparable to the performance of ophthalmologists. [3, 11]

We are also interested in using gender classifiers to simulate real-life privacy attacks. Several recent studies have attempted to perform sex prediction on color fundus images and achieved reliable results [6, 7, 12, 14]: Munk et al. [12] built a ResNet-152-based model in predicting patient information such as age and sex from three distinct retinal imaging modalities with accuracy of 0.73; Korot et al. [7] achieved around 0.84 accuracy using retinal fundus images. The aforementioned studies generally identify regions within images that are important for sex classification. Kim et al. [6] conducted an experiment to examine the sex prediction results after erasing fovea and blood vessels from the fundus images. The results show that erasing both anatomical regions decreases the model performance, indicating that both regions are helpful for sex prediction. Despite these results, retina specialists have not reached a consensus on fundus structures that are distinct for different sexes. [7].

2 Materials and Methods

2.1 Dataset

This work explores two distinct datasets, one publicly available and a private one. The first one is the publicly available Brazilian Open-Access Ophthalmological

Dataset (BRSET) [2, 13]. The BRSET consists of retinal fundus images and clinical diagnosis and assessments from three Brazilian ophthalmological centers in Sao Paulo with a total of 16,266 images from 8,108 patients seen from 2010 to 2020. Sensitive demographic variables are also included, such as patient sex and age. In terms of diabetic retinopathy label distribution, the dataset is highly imbalanced. Only 4.2% of the images have positive DR, which makes this a difficult classification task. The images were captured by a Nikon NF505 and a Canon CR-2 professional retinal camera, resulting in images with a resolution of about 900x1000 pixels and a centered composition.

The second dataset is a private Diabetes Center Dataset with 19,224 ultrawide field retinal images collected through a diabetic retinopathy screening program in the US. The dataset includes information about the presence and severity of diabetic retinopathy in each patient eye. This dataset is more balanced, as 21% of the images have positive DR. The images were captured with a high-resolution ultra-wide camera of 4000x4000 pixels resolution. The Diabetes Center images do not all have a centered composition, and the orientation of the images varies significantly. In addition, the private Diabetes Center dataset images contain some extraneous noise, such as eyelashes and fingers obscuring parts of the retina.

2.2 Pre-processing

Our images were normalized based on the dataset statistics (mean, std) and resized to 256x256. In training, we employed a combination of horizontal flip and Shift-Scale-Rotate both with probabilities of 0.25 as data augmentations to add model generalizability.

The diabetic retinopathy severity in both datasets is rated following the International Clinical Diabetic Retinopathy (ICDR) Severity Scale. Because we have highly-skewed data, where most patients do not have diabetic retinopathy, we regrouped patients with mild, moderate, and severe non-proliferative diabetic retinopathy (NPDR) into one diabetic retinopathy group. We kept patients with proliferative diabetic retinopathy (PDR) as the other group, leaving us three final disease categories – healthy, NPDR, PDR. In this way, we have sufficient samples for both NPDR and PDR groups.

The images from the BRSET and Diabetes Center data are split into a training set, a validation set, and a test set of ratios 70%, 10%, and 20%, respectively, by patient stratified sampling to prevent data leakage.

2.3 De-identification Framework

Figure 1 shows our general de-identification and evaluation workflow with some example retinal scans and de-identified images.

Full Image Obfuscation: Snow To determine the effectiveness of differentially private image obfuscation techniques on retinal fundus images, we first applied

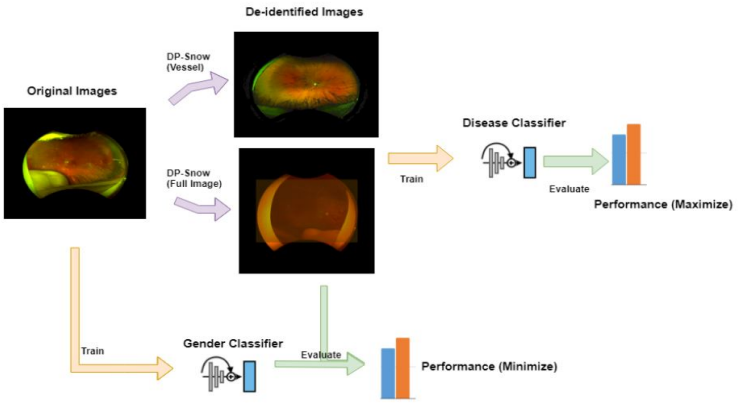


Fig. 1: De-identification & evaluation workflow

the Snow algorithm to the full image region. Snow is a differentially private de-identification technique first introduced by John et al. [5]. It randomly adds noise to the image to obfuscate sensitive demographic information. Here we modified it to fit the RGB retina images de-identification task. Specifically, the Snow method first computes the whole dataset pixel intensity average for each RGB channel. Next, for each RGB channel, the method randomly selects a proportion, p, of image pixels. It assigns the corresponding channel average pixel intensity to the selected pixels. Consider a retinal fundus image as $I(x)$ where x refers to the index of each pixel in the image. For a 256x256 image, $x \in [1, 2, ..., 65536]$. The intensity of the pixel x is represented by $I(x) \in [0, 255]$. A subset S of size $p * I_{rows} * I_{cols}$ indices are randomly selected, and a new image $I'(x)$ is created such that

$$I'(x) = \begin{cases} \mu_c & x \in S \\ I(x) & x \notin S \end{cases}, \tag{1}$$

where μ_c is the average pixel intensity of channel c of all images in the given dataset.

Vascular Region Obfuscation: VS-Snow Our observation of the Snow method (in results section) is that a full image de-identification technique may be sacrificing too much utility for privacy, and the whole image pixel noising approach would lose local details that are clinically useful. Inspired by studies showing that blood vessels are related to sex identifications [6], we propose VS-Snow, a local de-identification method to only apply obfuscation to the vessel regions. VS-Snow is composed of two parts: First, the blood vessel region mask was segmented using a modified Full Resolution network (FR-UNet) [10], demonstrating state-of-the-art retinal vessel segmentation performance. Next, we randomly selects a proportion, p, of the vessel mask pixels, (as opposed to full image in Snow) and change their values to the average of their neighboring vessel pixels

(as opposed to the average of full pixels in Snow). We modified the original FR-UNet as it was designed for gray-scale images and took a lot of disk space at runtime. To better fit our use case, we changed the aggregation module of the U-Net to perform depth-wise convolution across RGB channels and incorporated a random sliding window sampler in the dataloading module. Consider a retinal fundus image as $I(x)$ where x refers to the index of each pixel in the image. For a 256x256 image, $x \in [1, 2, ..., 65536]$. Denote the vessel mask as $m \in [1, 2, ..., 100]$, a subset of $I(x)$, and the $n_i \in m$ as the set of neighboring vessel pixels for each given x_i Then a subset S of size $p * Size(m)$ indices are randomly selected, and a new image $I'(x)$ is created such that

$$I'(x_i) = \begin{cases} \mu_{ni} & x_i \in S \\ I(x_i) & x_i \notin S \end{cases} \tag{2}$$

For both Snow and VS-Snow we generated images with p = 0.1, p = 0.3, p = 0.5 on both BR-Set and private Diabetes Center dataset.

2.4 Evaluation Framework

To evaluate how robust our de-identified images are against attacking models and how much clinical utility these images retain compared to the original image, we designed the two following pipelines.

Gender Classifier (Attacker) The gender classifier's (Attacker) objective is to re-identify patients' identity, in our settings, to recognize the gender of a patient. In a realistic scenario, a hacker would train the gender classifier on available datasets with private information and test on our released dataset to infer gender. Thus we trained our attacker models on the original images with gender labels and tested on 1) original images, 2) Snow de-identified images and 3) VS-Snow de-identified images, to remove gender-related information as much as possible via de-identification. We quantify the level of privacy gain by computing the performance drop of the attacker model successfully recognizing gender on de-identified images compared to original images.

Diabetic Retinopathy (DR) Classifier (Downstream Task) The diabetic retinopathy classifier's (Downstream Task) objective is to classify the disease categories from the images. In a realistic setting, a researcher would take our de-identified images and train models to predict diabetic retinopathy. Thus we trained our downstream models on all three types of images with DR labels, (original, Snow de-identified images, and VS-Snow images) and tested on respective test sets. We quantify the level of clinical utility by computing the performance drop of the downstream model on original images compared to de-identified images.

Implementation Details Separate models were trained for sex and diabetic retinopathy classification. We used PyTorch 2.0.1 as the deep-learning library and Nvidia V100 GPUs to train the models. Our backbone network is ResNet-200D [4], a modification of the ResNet architecture that utilizes an average pooling tweak for downsampling. We chose ResNet-200D after comparing its performance with various other architectures like EfficientNet[16] and XceptionNet[1]. Furthermore, in the unmodified ResNet, the 1x1 convolution for the downsampling block ignores 3/4 of the input feature maps, whereas ResNet-D takes the whole feature maps as inputs, and no information will be ignored. Our ResNet-200d was pre-trained in general image classification from the ImageNet database, as transfer learning is generally faster with better performance than training from scratch [19]. We modified the fully connected last layer (changed output dimension from 1000 to 2 for sex classification or 3 for diabetic retinopathy classification) to tailor the CNN to desired outputs. The cross-entropy loss function was adopted in the sex prediction model for binary classification; the class-weighted focal loss (gamma = 0.6) [9] was adopted for the 3-class diabetic retinopathy classification because it has the appropriate properties to handle the class imbalance and overfitting issues. We used Adam as the optimization scheme (learning rate = 1e-4, weight decay = 1e-5) and trained through 50 epochs with early stopping mechanism. The CNNs were validated for each epoch, and the model with highest F1, accounting for data imbalance, were selected as the final predictor.

3 Results

3.1 Full Image Snow Results

We tested our attacker and downstream models on the original and de-identified sets of images (Table 1). As expected, the model performance decreases as we increase the Snow method's parameter, p, since more information is obfuscated. The gender classifier performed as high as 0.75 and 0.83 on the two datasets, proving itself to be a good attacker on par with other gender prediction work [7]. To balance patient privacy and the images' clinical utility, we obfuscated the images with the goal of reducing the sex classification accuracy to approximately 50% while maintaining an F1 score for DR classification as close to the original images as possible. For the Brazilian dataset (BRSET), the sex classification accuracy dropped to 52.15% when we set the parameter p to 0.5. In this case, the F1 score of the DR classification is 64.5%, which reflects a 18.1% drop with respect to the baseline (82.6%). On the Diabetes Center dataset, we reduced the sex classifier's accuracy to 51% by only setting parameter p to 0.3. This led to a significant decrease in the DR classifier's performance from 79.5% on the original images to 58.4% on obfuscated images, with p equal to 0.3. These results show that the Snow approach successfully reduces privacy, but may obfuscate images too much at the expense of losing clinical utility.

Table 1: Snow De-identification

Data	Classification	Original	$p = .1$	$p = .3$	$p = .5$
BRSET	Sex (Acc)	75.12	64.99	61.35	52.15
	DR (F1)	82.6	72.3	69.6	64.5
Diabetes Center	Sex (Acc)	83.33	57.43	51.00	42.59
	DR (F1)	79.5	65.8	58.4	55.9

3.2 VS-Snow Results

A more targeted de-identification approach –only applied to the vessel regions and neighboring vessel pixels– improves the clinical usefulness of the fundus images. Similar to Snow experiments, the classifiers' performance decreases as we increase the value of the p parameter. For the BRSET, at p parameter of 0.5, the sex classifier's attacking ability was reduced to 58.1% accuracy, higher than a random classifier, while the diabetic retinopathy classifier experienced an 10.6% F1 score drop from training on the original images (72% vs. 82.6%). For the diabetes center dataset, we successfully reduced the sex classification accuracy to 50.7% with only 0.1 p parameter. We maintained a diabetic retinopathy classification F1 score very close to the training results on the original images (75.9% vs. 79.5%). These results shows that VS-Snow is capable of preserving privacy and utility at a good standing across patient populations.

Table 2: VS-Snow De-identification

Data	Classification	Original	$p = .1$	$p = .3$	$p = .5$
BRSET	Sex (Acc)	75.12	67.7	60.1	58.1
	DR (F1)	82.6	78.1	75	72
Diabetes Center	Sex (Acc)	83.33	50.7	47.9	47.5
	DR (F1)	79.5	75.9	73.8	74.6

4 Discussion

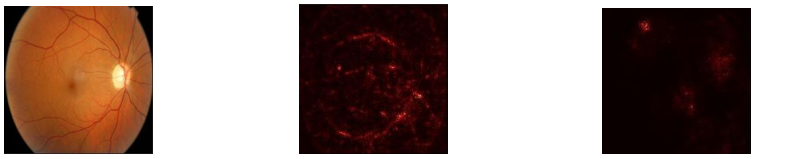
4.1 Privacy-utility tradeoff

Our results show a clear privacy-utility tradeoff, meaning there is an inverse relationship between privacy protection (obfuscating the patient sex) and statistical utility (identifying DR) of the images. In order to reach almost perfect privacy (indicated when gender classification accuracy ≤ 50), the F1 score of the DR classifier reduces at least 11.4 and 4.9 points on the BRSET and Diabetes Center datasets, respectively. Still, our de-identification framework is able to strike a reasonable balance between privacy and utility. For both datasets, we are able

to achieve almost perfect privacy while maintaining an F1 score > 70. The only experiment that yielded unsatisfactory results was applying Snow to the entire image on the Diabetes Center dataset, implying that local obfuscation may be better than global obfuscation.

4.2 Importance of vasculature

Comparing the results from applying Snow to the entire image and VS-Snow to only the blood vessels demonstrates the importance of the vasculature for both sex and DR classification. However, it appears that the vasculature is much more important for sex classification than DR classification. This is because, compared to sex classifier's performance, the DR classifier performance is significantly better when applying Snow to only the vessels (p=0.5) than when it's applied to the entire image. This is contrary to the performance of the sex classifier, as the performance is comparable regardless of the de-identification method. In figure 2 we shows an example of saliency maps and it is clear that gender information is much more prevalent in the vessel regions than the disease information, supporting our hypothesis that a local obfuscation method might be more optimal for retinal de-identification.



(a) Original Images (b) Gender Classification (c) DR Classification

Fig. 2: Example of original retina image vs saliency maps

4.3 Limitations and future work

In the future, we would like to extend this work to examine and improve VS-Snow's ability to obfuscate other sensitive demographic features, such as income, race and age. Moreover, other advanced image models, such as large pretrained vision transformers, would be tested to determine whether our de-identification method is robust against different attackers.

References

1. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
2. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P.C., Mark, R., et al.: Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation [Online] **101**(23), E215 – E220 (2000). https://doi.org/10.1161/01.cir.101.23.e215
3. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., et al.: Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA **316**(22), 2402–2410 (12 2016). https://doi.org/10.1001/jama.2016.17216
4. He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M.: Bag of tricks for image classification with convolutional neural networks (2018)
5. John, B., Liu, A., Xia, L., Koppal, S., Jain, E.: Let it snow: Adding pixel noise to protect the user's identity. In: ACM Symposium on Eye Tracking Research and Applications. ETRA '20 Adjunct, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3379157.3390512, https://doi.org/10.1145/3379157.3390512
6. Kim, Y.D., Noh, K.J., Byun, S.J., Lee, S., Kim, T., Sunwoo, L., et al.: Effects of hypertension, diabetes, and smoking on age and sex prediction from retinal fundus images. Scientific Reports **10**, 4623 (2020). https://doi.org/10.1038/s41598-020-61519-9
7. Korot, E., Pontikos, N., Liu, X., Wagner, S.K., Faes, L., Huemer, J., et al.: Predicting sex from retinal fundus photographs using automated deep learning. Scientific Reports **11**, 10286 (2021). https://doi.org/10.1038/s41598-021-89743-x
8. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences **117**(23), 12592–12594 (2020)
9. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection (2018)
10. Liu, W., Yang, H., Tian, T., Cao, Z., Pan, X., Xu, W., et al.: Full-resolution network and dual-threshold iteration for retinal vessel and coronary angiograph segmentation. IEEE Journal of Biomedical and Health Informatics **26**(9), 4623–4634 (Sep 2022). https://doi.org/10.1109/JBHI.2022.3188710
11. Liu, X., Ali, T.K., Singh, P., Shah, A., McKinney, S.M., Ruamviboonsuk, P., et al.: Deep learning to detect oct-derived diabetic macular edema from color retinal photographs: A multicenter validation study. Ophthalmology Retina **6**(5), 398–410 (2022). https://doi.org/https://doi.org/10.1016/j.oret.2021.12.021
12. Munk, M.R., Kurmann, T., Márquez-Neila, P., Zinkernagel, M.S., Wolf, S., Sznitman, R.: Assessment of patient specific information in the wild on fundus photography and optical coherence tomography. Scientific Reports **11**, 8621 (2021). https://doi.org/10.1038/s41598-021-86577-5
13. Nakayama, L.F., Goncalves, M., Zago Ribeiro, L., Santos, H., Ferraz, D., Malerbi, F., et al.: A brazilian multilabel ophthalmological dataset (brset) (2023). https://doi.org/10.13026/xcxw-8198
14. Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., et al.: Prediction of cardiovascular risk factors from retinal fundus pho-

tographs via deep learning. Nature Biomedical Engineering **2**, 158–164 (2018). https://doi.org/10.1038/s41551-018-0195-0
15. Ruamviboonsuk, P., Krause, J., Chotcomwongse, P., Sayres, R., Raman, R., Widner, K., et al.: Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. npj Digital Medicine **2**, 25 (2019). https://doi.org/10.1038/s41746-019-0099-8
16. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
17. Ting, D.S.W., Cheung, C.Y.L., Lim, G., Tan, G.S.W., Quang, N.D., Gan, A., et al.: Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. JAMA **318**(22), 2211–2223 (12 2017). https://doi.org/10.1001/jama.2017.18152
18. Yala, A., Quach, V., Esfahanizadeh, H., D'Oliveira, R.G., Duffy, K.R., Médard, M., Jaakkola, T.S., Barzilay, R.: Syfer: Neural obfuscation for private data release. arXiv preprint arXiv:2201.12406 (2022)
19. Yu, Y., Lin, H., Meng, J., Wei, X., Guo, H., Zhao, Z.: Deep transfer learning for modality classification of medical images. Information **8**(3) (2017). https://doi.org/10.3390/info8030091, https://www.mdpi.com/2078-2489/8/3/91