Journal of Imaging Informatics in Medicine
https://doi.org/10.1007/s10278-025-01776-0

ORIGINAL PAPER

Pixel Tampering: Does Face Redaction Harm Medical AI Performance?

Eduardo M. J. M. Farina[1,2] · Felipe A. Matsuoka[1,2] · Gustavo Corradi[1] · Yosuke Yamagishi[4] · Masatoshi Abe[5] · Maximilian Pfeiffer[6] · Andrea S. Souza[7] · Raquel Moreno[7] · Ivanei Bramati[7] · Fernanda Moll[7] · Almir Bitencourt[8] · Carlos Sacomani[8] · Soraia Quaranta Damião[8] · Rubens Chojniak[8] · Nitamar Abdala[2] · Rodrigo Ragazzini[2] · Henrique Carrete Jr.[2] · Paulo E. A. Kuriki[3] · Marcelo Straus Takahashi[12] · Nelson Caserta[9] · Cesar H. Nomura[10] · Felipe C. Kitamura[2,11]

Received: 15 September 2025 / Revised: 6 November 2025 / Accepted: 24 November 2025

Abstract

Balancing data sharing and patient privacy is essential in medical imaging. Face redaction tools anonymize head CTs by removing identifiable features, but their impact on deep learning (DL) model performance remains a concern. We present an open-source face redaction tool designed to enhance data-sharing security while preserving DL performance, validated through a Kaggle competition on age prediction from brain CTs. This study aims to evaluate whether models trained on redacted images perform comparably on both redacted and non-redacted test sets, and how they compare to similar models trained on non-redacted images reported in the literature. A Kaggle challenge was conducted between March 2 and April 30, 2024, to crowdsource age prediction models. The dataset comprised 2,377 redacted head CT studies for training and 148 for testing, sourced from multiple institutions. In a post-hoc analysis, the top-performing models were evaluated on both redacted and non-redacted formats of the test set, with performance measured using mean absolute error (MAE). The two best models achieved MAEs of 2.8 and 3.4 years on redacted test data. On the non-redacted format, MAEs increased to 3.2 and 3.8, respectively. Paired *t*-tests showed a significant performance drop for one model ($p = 0.038$) but not the other ($p = 0.051$). There was no significant difference between the models on the redacted test set ($p = 0.610$). Models trained on redacted data may show minimal performance decline when applied to non-redacted images, yet still outperform existing benchmarks. Our tool enables secure data sharing with a limited impact on DL accuracy.

Keywords Medical imaging · Head CT · Face redaction · Deep learning · Model performance · Age prediction

Eduardo M. J. M. Farina and Felipe A. Matsuoka contributed equally to this work and are and share first authorship.

✉ Felipe A. Matsuoka
felipe.matsu.2003@gmail.com

Felipe C. Kitamura
felipe.matsu.2003@gmail.com

1 Dasa, São Paulo, Brazil

2 Universidade Federal de São Paulo, Unifesp, São Paulo, Brazil

3 UT Southwestern Medical Center, Dallas, TX, USA

4 Division of Radiology and Biomedical Engineering, University of Tokyo, Tokyo, Japan

5 Department of Nephrology, Osaka University, Osaka, Japan

6 Siegen, Germany, North Rhine-Westphalia

7 IDOR, São Paulo, Brazil

8 AC Camargo, São Paulo, Brazil

9 Universidade Estadual de Campinas, São Paulo, Brazil

10 Instituto do Coração - InCor, São Paulo, Brazil

11 Bunkerhill Health, San Francisco, CA, USA

12 Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, USA

Published online: 16 December 2025

Introduction

The increasing use of medical imaging data for machine learning (ML) research has raised important concerns regarding patient privacy and data security [1]. Head CT and MRI scans, while critical for diagnostic and research purposes, contain identifiable facial structures that can potentially be reconstructed and recognized by both human observers and advanced facial recognition systems [2, 3]. Consequently, ensuring data anonymization is crucial to comply with regulatory standards while enabling secure data sharing for research collaboration [4, 5].

Previous studies have addressed the importance of anonymization in medical imaging. For example, Collins et al. [6] proposed a Gaussian smoothing-based method to obscure facial features in head CT images, effectively preventing re-identification. Our method was inspired by the one created by Collins et al. due to its simplicity, ease of implementation, redaction of the entire head (including hair and ears), high redaction rate, and low tampering with central nervous system structures. Similarly, Selfridge et al. [7] developed a face redaction workflow for total-body PET/CT images, demonstrating that facial identifiability could be significantly reduced after redaction. Uchida et al. [8] further explored deformation-based redaction techniques to preserve internal anatomical structures while rendering external features unrecognizable. Although these approaches successfully reduced identifiability, they did not systematically evaluate the impact of redaction on downstream ML tasks, particularly deep learning (DL) models.

This gap is critical, as face redaction techniques alter pixel values, introduce visual artifacts, or disrupt anatomical structures, potentially affecting the performance of ML models trained on such data.

In this study, we tested the impact of an open-source face redaction tool on DL models that predict patient age from head CTs. Specifically, we leveraged the data from a Kaggle competition that used face-redaced exams in the training and test sets. We studied two of the winning models (trained on redacted images) by doing inference on the redacted and non-redacted formats of the test set to compare if the performance differs.

Methodology

Face Redaction Tool

The de-identification algorithm was designed to obscure superficial facial features while preserving deeper anatomical structures relevant for analysis. The method operates directly on DICOM image stacks. First, voxel intensities were converted to Hounsfield units (HU) using the RescaleSlope and RescaleIntercept tags. We then binarized the volume using a threshold of −800 HU to distinguish air from tissue, creating an initial air–body mask. From this binary volume, only the largest connected component was retained, ensuring that the head and body were isolated from external artifacts.

Next, to guarantee full coverage of skin and subcutaneous regions, the boundary of this component was dilated with a 35×35 morphological kernel, expanding the mask slightly beyond the skin surface. The region corresponding to the expanded air volume—essentially the "space around the head"—was then filled with randomly sampled HU values between −125 and +50 HU, corresponding to typical skin and subcutaneous fat densities. These values were obtained from the subject's own tissue distribution to maintain realistic texture and contrast. Random sampling avoided uniform patterns, giving the de-identified region a natural, noise-like appearance.

Finally, the modified voxel intensities were written back into new DICOM files using the same acquisition metadata, preserving spatial orientation, slice thickness, and windowing behavior. The resulting images effectively remove all identifiable facial surfaces while maintaining internal anatomy integrity for downstream medical AI analysis (Figure 1).

Kaggle Competition and Dataset

To assess the impact of a face redaction tool on model performance, we launched an international Kaggle AI competition focused on age prediction from head CTs (https://www.kaggle.com/competitions/spr-head-ct-age-prediction-challenge), using a redacted dataset. That competition was hosted by Sociedade Paulista de Radiologia (SPR) (https://www.spr.org.br/). The retrospective dataset, compiled from four diverse Brazilian institutions with IRB approval, underwent de-identification: identifiable metadata was removed using the RSNA anonymizer [9], and a face redaction algorithm was applied with the "face" parameter to obscure facial features in the images (https://github.com/kitamura-felipe/face_deid_ct). Metadata anonymization happened before the data left each institution. Face redaction was performed by the competition organizers, except for one donating institution that performed it before transferring the data to the organizer. All institutions signed a Data Use Agreement. The age labels were extracted from the DICOM metadata.

For the competition, the dataset was split to maintain a clear separation between training and testing sets: the training set included data from four institutions, while the test set consisted of data from a single institution but from a different time period. The age labels, used as ground truth, were extracted from the DICOM metadata.

The training dataset contained 2377 brain CT scans with subject ages ranging from 18 to 89 years. The mean age in the training set was 41.0 years (SD = 14.3), with a median age of 39.0

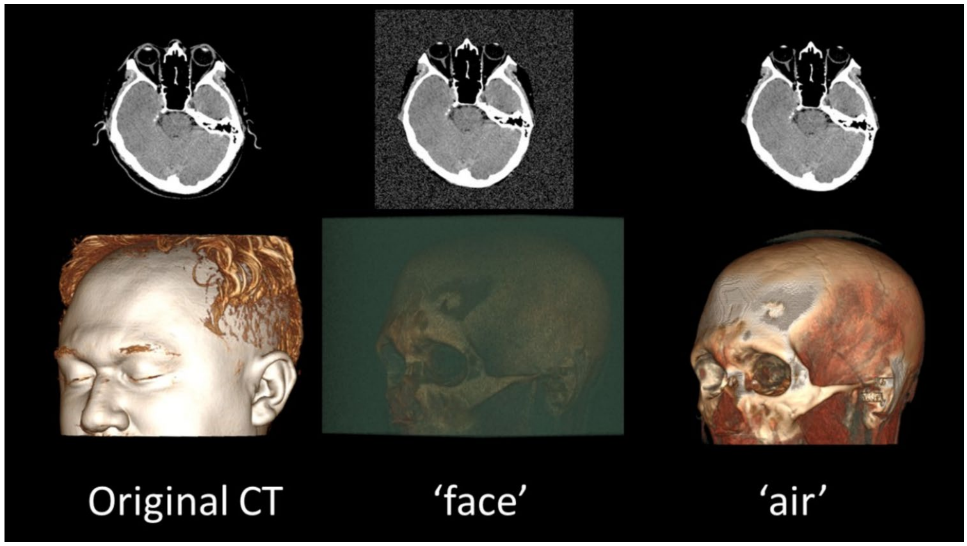

Fig. 1 Example of the results from the defacing algorithm using the “air” and “face” parameters (head CT images used with patient consent)

years and an interquartile range (IQR) of 19.0. The test dataset included 148 brain CT scans, with ages ranging from 18 to 85 years. The mean age in the test set was 38.1 years (SD = 15.4), with a median age of 36.0 years and an IQR of 24.5 (Figure 2).

Impact on Model Performance Analysis

To evaluate the redaction algorithm’s effect, we analyzed the inference results of the first and third-place models by running them on the original, non-redacted test set and on the redacted format (the one made available to participants in the competition). We were not able to rerun the model from the second place due to technical issues. We compared the mean average error (MAE) scores of each model before and after redaction using a paired t-test. We also compared model 1 to model 3 without redaction using a paired t-test. Statistical significance was defined as $p < 0.05$.

To assess agreement between predictions on redacted and non-redacted datasets, we generated Bland-Altman plots for each model. For each test case, we calculated the difference between the predicted and chronological age (ground truth obtained from the DICOM tag) and compared these differences across conditions. Limits of agreement were computed as the mean difference ±1.96 standard deviations, allowing us to visualize potential systematic bias and variability introduced by the redaction process.

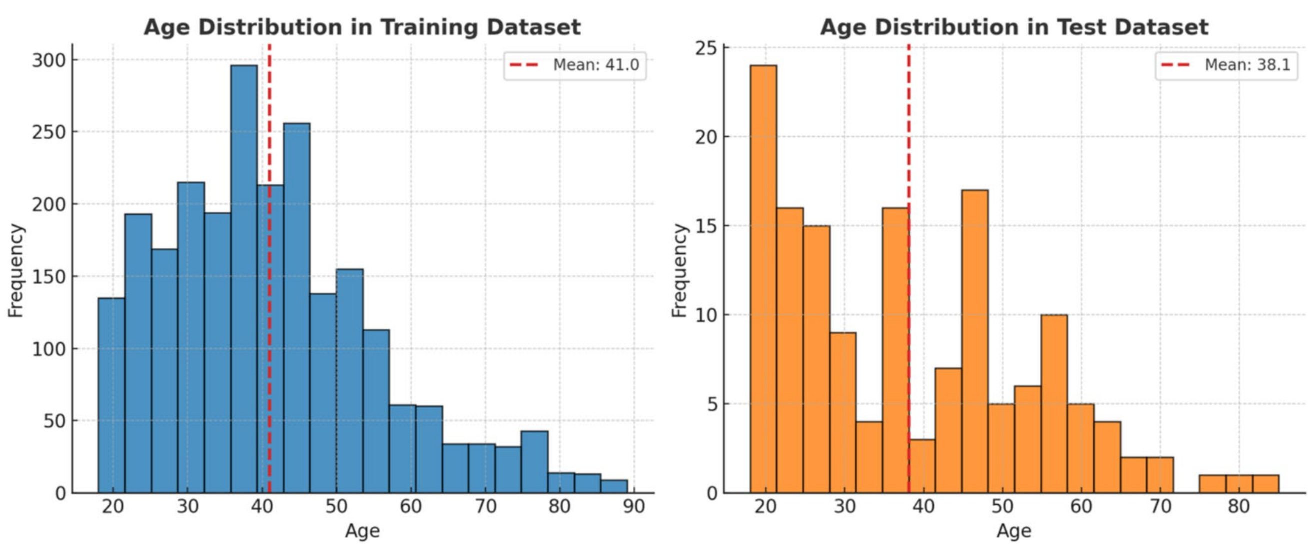

Fig. 2 Age distribution of subjects in the training and test datasets. The histograms display the frequency of ages in each dataset, with the training set shown in blue and the test set in orange. The red dashed lines indicate the mean age for each dataset

Models Evaluated

First Place Solution

The team used a single EfficientNetV2-S model [10] and aggregated slice predictions for final study-level predictions. Data splitting utilized a 9:1 holdout method. Images were preprocessed using brain parenchyma, bone, and soft tissue windows, stacked in the channel direction to create 3-channel images, and removed noisy slices without bone. Most images were at 512 × 512 size, resizing only when necessary. The model was trained for 5 epochs with basic augmentations, aligning all slice labels in the same study to the study age. Predictions were aggregated using the mean of the caudal 40th percentile, improving performance. The code is available at https://github.com/yamagishi0824/spr_head_ct_age_1st.

Second Place Solution

After processing each slice using a window function with WL = 40 and WW = 80, three adjacent slices were stacked to form a 3-channel image. Using the 3-channel image, the head region was detected using morphological transformation, and images that did not include the head were excluded based on a threshold. Using the detected region, 25 cropped images in the z-axis direction were selected and input into a 2.5D convoluted neural network (CNN). The backbone of the 2.5D CNN was EfficientNet-V2-L. The target was divided by 100 for age prediction, and the model was trained using Binary Cross-Entropy. The code is available at https://github.com/abebe9849/SPRHeadCTage/tree/main.

Third Place Solution

This solution introduces a novel 3D CNN that diverges from existing EfficientNet architectures, with around 2.3 million parameters and featuring both channel and spatial attention mechanisms, inspired by CBAM [11] and adapted to the 3D domain. The activation function [12], a per-channel-parameterized version of APTx that emulates MISH performance with less computation, is introduced [13]. The model processes single-window (WL = 40 and WW = 80) 3D head CT scans. An AdaBelief optimizer [14] was initially set at a learning rate of 3e−4, then adjusted to 3e−5 with a significant weight decay to enhance regularization. The architecture optimally manages multiple scans by averaging predictions, efficiently parallelizing scans across the batch dimension. This strategy ensures robust performance, leveraging full 3D spatial context for improved age prediction accuracy. The code is available at https://gitlab.com/Skelp/headct3dtinynet.

Results

The Kaggle competition attracted 146 registered entrants, 26 active participants, and 15 teams, collectively submitting 316 entries. On the private leaderboard of the SPR Head CT Age Prediction Challenge—which was calculated using approximately 60% of the hidden test set—the mean absolute error (MAE) across all submissions ranged from 2.596 to 18.337. This leaderboard represented the final post-competition standings and was designed to assess model generalization and discourage overfitting to the public subset. In our own evaluation on the complete test set (including both public and private portions), two of the top-performing models achieved MAEs of 2.8 (model 1) and 3.4 years (model 3) on the redacted dataset. When running inference on the non-redacted test set, MAEs increased to 3.2 and 3.8 years, respectively. A paired t-test (Table 1) indicated a significant difference for model 1 (redacted vs. non-redacted, $p = 0.038$) but not for model 3 ($p = 0.051$). No significant difference was observed between the two models when evaluated on the non-redacted data ($p = 0.610$).

To further analyze the agreement between model predictions and ground truth, we conducted a Bland-Altman [15] analysis (Figure 3). The redacted and non-redacted datasets had similar error distribution across the reference standard age range for both models 1 and 3.

Discussion

Our findings demonstrate that the consistent application of the proposed redaction algorithm on the training set preserves deep learning model performance on a redacted test set, leading to comparable performance on the non-redacted test set on the age prediction task.

In comparison to previous works, our study addresses a critical gap by systematically evaluating the impact of redaction on downstream model performance. Collins et al. proposed a Gaussian smoothing-based method for facial anonymization in head CT scans [6]. While effective at reducing facial identifiability, their method did not include validation on machine learning tasks. Similarly, Selfridge

Table 1 Comparison of MAE values for the top-performing models on redacted and non-redacted datasets along with p-values from paired t-tests

Model	MAE (redacted dataset	MAE (non-redacted dataset)	p-value
Model 1	2.8	3.2	0.038
Model 3	3.4	3.8	0.051
Comparison			0.610

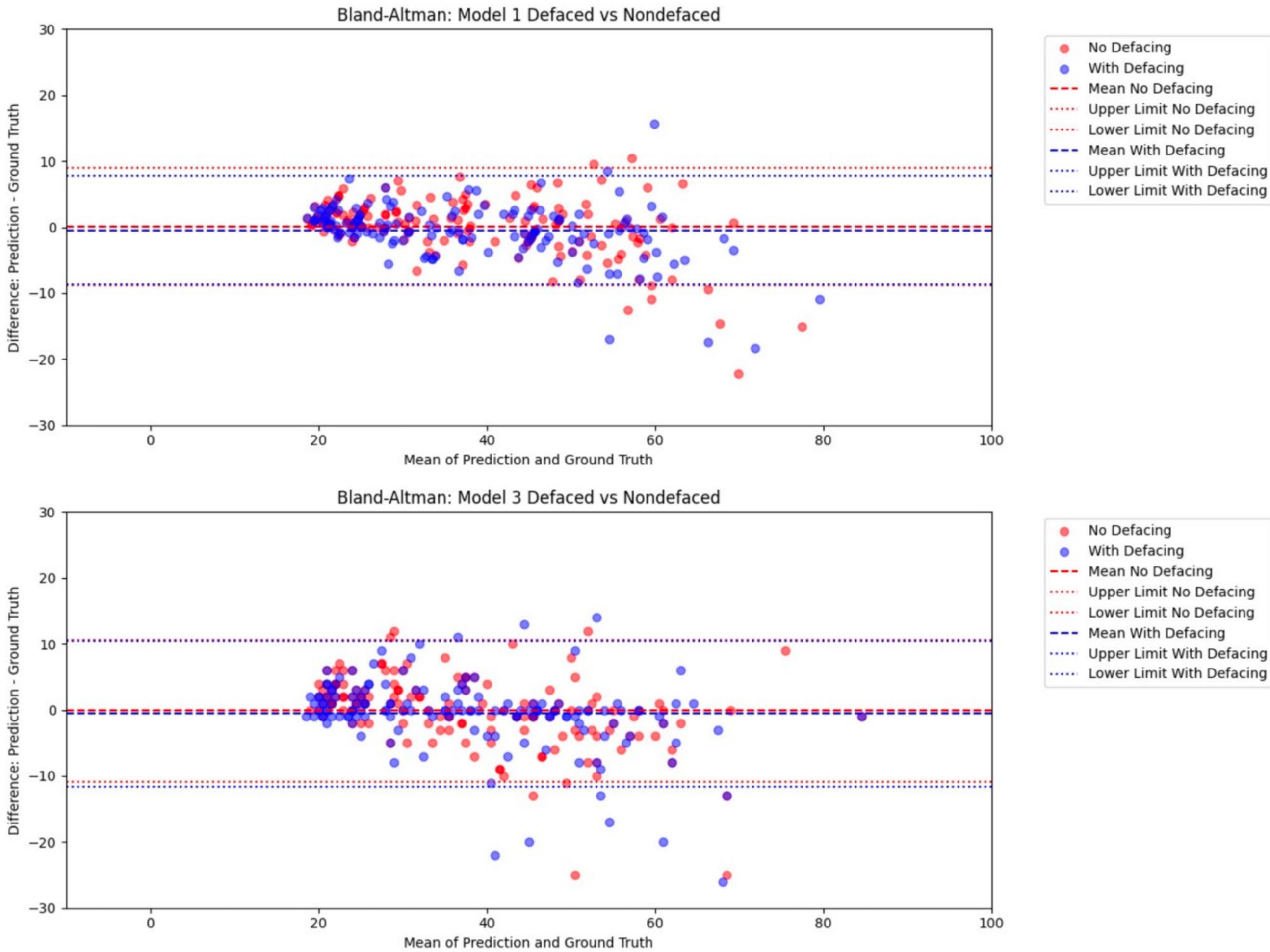

Fig. 3 Bland-Altman plot comparing each prediction with and without the proposed redacting method. The *x*-axis represents the ground truth values, and the *y*-axis shows the difference between the ground truth and each model's prediction. Blue points indicate predictions from the model with the proposed redating method, and red points represent predictions from the model without the redacting method. Dashed lines denote the limits of agreement (±1.96 SD), indicating the range within which 95% of the differences are expected to lie

et al. [7] demonstrated that redaction reduces identifiability in total-body PET/CT images, but they reported biases in PET attenuation correction near redacted regions. Uchida et al. [8] employed deformation-based redaction techniques that preserve internal anatomical structures while obscuring facial features. However, their evaluation relied primarily on correlation metrics and did not assess the impact on predictive tasks. In contrast, our study systematically evaluates the performance of deep learning models on redacted and non-redacted datasets using a competitive Kaggle-based framework, providing direct evidence for the safe use of the proposed redaction tool.

To further contextualize our results, we compared with the work of Bermudez et al. [16], who reported an MAE of 11.02 years for brain age prediction using head CT scans with convolutional neural networks. Models trained on brain CT exams redacted with our tool and validated on normal exams achieved lower MAEs, ranging from 2.8 to 3.4 years. This result highlights the robustness of deep learning models when our redaction tool was applied and demonstrates that it effectively preserves critical information necessary for accurate predictions.

Our findings also underscore the importance of consistent application of redaction algorithms across the entire AI pipeline. Inconsistent use, such as applying redaction only during training while leaving test or deployment data untouched, introduces visual discrepancies that could compromise model generalization. The slight performance degradation observed on non-redacted test data reinforces this point.

Although the absolute difference in MAE between redacted and non-redacted predictions was relatively small (≈0.4 years, or about 4.8 months), this variation remains statistically significant and highlights the sensitivity of deep learning models in capturing subtle age-related morphological cues on head CTs. While such a difference may not hold direct clinical implications, since chronological age prediction itself is not a diagnostic endpoint, it reinforces the precision and robustness of the models. In future research, divergences between predicted and chronological age could serve as quantitative biomarkers of accelerated or delayed craniofacial or brain aging. Additionally, these age-prediction models could support operational quality control by identifying discrepancies between predicted and recorded patient ages, helping detect mislabeled or mismatched imaging data in large-scale datasets.

One limitation of our study is that we did not test the performance of the same models trained on redacted versus non-redacted formats of the training set. Since part of the training set was redacted before the data was transferred to the competition organizing team, we do not have access to the entire non-redacted training set. Hence, this study can not infer if training on redacted exams would degrade performance when compared to training on non-redacted exams. Even then, prior studies trained on non-redacted exams resulted in worse performance than the models we presented. Another limitation of our study is that we tested a single task (age prediction). We can not extrapolate the results to other tasks, like segmentation or classification of other diseases.

Overall, our study demonstrates that models trained on exams redacted with our tool perform equally or slightly worse on non-redacted images. We recommend that inference done during model deployment follows exactly the same preprocessing as used during training, including face redaction, to guarantee the best performance possible. By ensuring consistency in the application of redaction algorithms, both patient privacy and model reliability can be preserved across real-world deployment scenarios.

Acknowledgements The authors would like to thank Errison dos Santos Alves and Thiago Matheus Bruno Da Silva for their assistance with the promotion of the competition.

Author Contribution All authors contributed to the study conception and design. The models evaluated were developed by Y. Y. and M.P.

Data Availability Defaced data analyzed during the study are available at (https://www.kaggle.com/competitions/spr-head-ct-age-prediction-challenge), and non-defaced data is not available for sharing.

Declarations

Ethics Approval As this work used a public dataset, no ethical approval is required.

Consent to Participate The patient provided informed consent for the use and display of their facial CT images.

Competing interests F. K. is a consultant for Bunkerhill Health, a consultant for GE Healthcare, a consultant for MD.ai, a speaker for Sharing Progress in Cancer Care, an early career consultant to the Editor of Radiology, an associate editor of Radiology Artificial Intelligence, a vice chair of the SIIM ML Committee, a member of the RSNA AI Committee, and a member of the RSNA Radiology Informatics Council.

References

1. Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. *When machine learning meets privacy: A survey and outlook.* ACM Comput. Surv., 54(2), March 2021.
2. Joseph Jen-Sho Chen, Krishna Juluru, Tara Morgan, Ryan Moffitt, Khan M. Siddiqui, and Eliot L. Siegel. *Implications of surface-rendered facial CT images in patient privacy.* American Journal of Roentgenology, 202(6):1267–1271, 2014. PMID: 24848824.
3. C. L. Parks and K. L. Monson. *Automated facial recognition of computed tomography-derived facial images: Patient privacy implications.* Journal of Digital Imaging, 30:204–214, 2017.
4. Francesco Sardanelli, Mariangela Alì, M. G. Myriam Hunink, et al. *To share or not to share? Expected pros and cons of data sharing in radiological research.* European Radiology, 28:2328–2335, 2018.
5. M. Bosserdt, B. Hamm, and M. Dewey. *Clinical trials in radiology and data sharing: Results from a survey of the European Society of Radiology (ESR) Research Committee.* European Radiology, 29:4794–4802, 2019.
6. Scott A. Collins, Jing Wu, and Harrison X. Bai. *Facial de-identification of head CT scans.* Radiology, 296(1):22, 2020.
7. Aaron R. Selfridge, Benjamin A. Spencer, Yasser G. Abdelhafez, Keisuke Nakagawa, John D. Tupin, and Ramsey D. Badawi. *Facial anonymization and privacy concerns in total-body PET/CT.* Journal of Nuclear Medicine, 64(8):1304–1309, 2023.
8. Masato Uchida et al. *Facial defacing methods in medical imaging.* Neuroinformatics, 21:123–135, 2023.
9. RSNA MircWiki. (2017, December 10). *The CTP DICOM Anonymizer*. MircWiki. Retrieved *[Jan 2024]*, from https://mircwiki.rsna.org/index.php?title=The_CTP_DICOM_Anonymizer
10. Mingxing Tan and Quoc V. Le. *EfficientNet: Rethinking model scaling for convolutional neural networks.* CoRR, abs/1905.11946, 2019.
11. S. Woo, J. Park, J. Y. Lee, and I. S. Kweon. CBAM: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision (ECCV), 2018.
12. Ravin Kumar. *APTX: Better activation function than Mish, Swish, and ReLU's variants used in deep learning.*International Journal of Artificial Intelligence and Machine Learning, 2(2):56–61, July 2022.
13. D. Misra. Mish: A self-regularized non-monotonic neural activation function. *CoRR*, abs/1908.08681, 2019.
14. Dian Huang, Dawei Yin, Lijun Zhang, Dezhong Peng, and QiLin Li. *AdaDiff: Adaptive gradient descent with the differential of gradient.* Journal of Physics: Conference Series, 2010:012027, September 2021.
15. Davide Giavarina. *Understanding Bland-Altman analysis.* Biochemia Medica, 25(2):141–151, June 5, 2015.
16. Camilo Bermudez, Andrew J. Plassard, Shikha Chaganti, Yuankai Huo, Katherine E. Aboud, Laurie E. Cutting, Susan M. Resnick, and Bennett A. Landman. *Anatomical context improves deep learning on the brain age estimation task.* Neuroinformatics, 18(4):553–567, 2020.