





Proceedings Article

Privacy Risks in the Anonymization of Medical Image Data

Lukas Schmahl ^{a,*} · Mattias P. Heinrich ^b · Malte Maria Sieren ^{c,d} · Lennart Berkel ^c

^aStudy program Medical Engineering Science, Universität zu Lübeck, Lübeck, Germany

^bInstitute of Medical Informatics, Universität zu Lübeck, Lübeck, Germany

^cInstitute for Radiology and Nuclear Medicine, Universitätsklinikum Schleswig-Holstein, Lübeck, Germany

^dInstitute for Interventional Radiology, Universitätsklinikum Schleswig-Holstein, Lübeck, Germany

*Corresponding author, email: lukas.schmahl@student.uni-luebeck.de; mattias.heinrich@uni-luebeck.de

Received 05 February 2025; Accepted 21 November 2025; Published online 05 December 2025

© 2025 Lukas Schmahl *et al.*; licensee Infinite Science Publishing

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Publicly available medical image datasets are essential for the progress of supporting diagnostic algorithms in radiology. In terms of patient data protection, it is necessary to anonymize images before publication. However, there is a risk that these anonymization procedures are insufficient. In this study, we employ a specially developed siamese neural network to assess the ability of re-identifying additional X-ray images of specific patients within supposedly anonymized datasets. This is analyzed using different neural network constellations and images from two variable datasets, CheXpert and KI-Rad-MSK, which include chest and wrist radiographs. Our results show that conventional anonymization approaches cannot withstand attacks using modern deep learning methods: One image of a patient is sufficient to re-identify other images of the same person within large datasets. Instead, we recommend innovative variants, such as latent diffusion models, to ensure data protection without compromising progress in medical imaging.

1. Introduction

In recent years, the diagnosis of various diseases based on medical image data has increasingly developed towards the use of machine learning support in radiology. In addition to improvements in early detection of diseases, the time saved by these methods is proving to be a major advantage. However, a huge obstacle in developing automated diagnostic algorithms at expert level is the large amount of data required to train the underlying artificial neural networks. This data demand has led to an increase in the availability of public medical image datasets, including CT, MRI and X-ray images. Nevertheless, a significant portion of patient scans generated worldwide remain inaccessible to the research community due to patient privacy concerns.

The current process of anonymization or pseudonymization prior to the publication of patient image data is only partially effective. In the majority of cases, only the most concise information, such as name or age, is removed from the associated metadata. However, this procedure does not prevent a sufficiently trained deep learning network from being able to re-identify images of a known patient from a supposedly anonymized public dataset, as Packhäuser *et al.* showed in [1]. Figure 1 illustrates a conceivable problem scenario in connection with re-identification: A potentially compromised X-ray of a known patient is compared with the individual scans within a publicly available, anonymized image dataset by a powerful deep learning network. Based on the similarity, a ranking list of the most similar images in the dataset can then be created, with possible other images of the person landing in the top ranks. This illus-

trates how sensitive information, such as a reference to a medical report, can be obtained by attackers from the supposedly anonymized metadata through inadequate anonymization procedures.

In [2], Reddy et al. analyzed several prominent incidents from the past, in which data leaks led to the leakage of large amounts of image data along with medical reports. These are often caused by internal vulnerabilities such as insufficient protected networks or human error. This data could, as described, be linked to public datasets, further increasing the risk of attacks related to re-identification methods. Therefore, our goal is to demonstrate the problem of re-identification with a siamese neural network using two different datasets. In particular, we aim to draw attention to the need for improved anonymization techniques before publishing large amounts of medical image data.

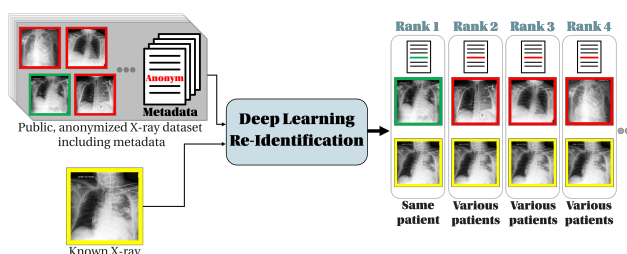


Figure 1: Problem scenario of re-identification: A known or compromised image of a patient (framed yellow) is used to find other images of this person in a public anonymized dataset. The scans can be ranked according to their similarity, with the actual second image of the person (framed green) being placed first. This demonstrates the risk of exposing sensitive patient data through insufficient anonymization processes.

II. Methods and materials

In the following section, the datasets used, the architecture of the implemented neural network and the evaluation methodology are described in detail. Our aim is to clearly present the basics of the re-identification task and to make the methods used comprehensible.

II.I. Datasets

For this study, we utilize two datasets: CheXpert [3] and KI-Rad-MSK. Both serve as a comprehensive foundation for developing and analyzing the proposed re-identification network.

The first dataset, CheXpert, was published by Stanford University in 2019 and comprises a total of 224,316 chest X-ray images from 65,240 patients. In addition to the scans, the dataset contains annotations for 14 different pathologies. For this work, we only use frontal

images, while lateral images are excluded to ensure uniformity of the data. Before applying the deep learning model, the images are also subjected to extensive pre-processing. Accordingly, the images are first cropped to the central thoracic region while retaining essential adjacent anatomical information. To achieve this, the lungs are segmented using a pre-trained nnU-Net presented in [4]. Based on these segmentations, bounding boxes are generated that allow us to crop the images. After this transformation, the images are scaled down to 256×256 pixels. These preprocessing steps ensure high anatomical consistency across the entire dataset.

Exactly one image pair is created per patient, whereby the two images are selected at random if more than two scans of this person are available. This results in a total of 31,746 image pairs. A detailed analysis of the diverse images in CheXpert reveals that even X-rays of the same patient often exhibit considerable variations in image quality. For this reason, an additional homogeneous subset of the dataset is created, which has a higher consistency with regard to aspects such as image quality and perspective. This selection process significantly reduces the number of patients included to 944, which increases the informative value of the analysis for specific questions. Moreover, the effect of using groups of four images per patient is investigated. For this purpose, the homogeneous subset is further adjusted so that only patients with at least four X-rays are considered. The remaining 458 patients allow the observation of the network behavior with more than two X-rays per patient, which is explained in more detail in the following section.

The second dataset, KI-Rad-MSK, originates from the UKSH Lübeck and includes X-ray images of various anatomical regions. Only the available wrist radiographs are selected for this study. This results in a total of 10,247 scans from 6,061 patients. The images in KI-Rad-MSK are available in two digital X-ray techniques: digital radiography (DX) and computed radiography (CR). Only the DX images are used for this research. After downsampling to 512×512 pixels and the formation of image pairs analogous to CheXpert, 2,071 patient scans are obtained.

II.II. Network architecture

The implemented siamese neural network is oriented on the approach in [5] and specially adapted for the intended re-identification of image pairs. The architecture of the model is visualized in Figure 2. The basic framework consists of two identical ResNet34 streams, each of which processes one image of a pair from a batch of size N . The output configuration of the fully connected layer is chosen so that both streams generate a compact 256-dimensional feature embedding for each input image. These compressed representations form the fundamental basis for the efficiency and accuracy of the intended re-identification. To quantify the image simi-

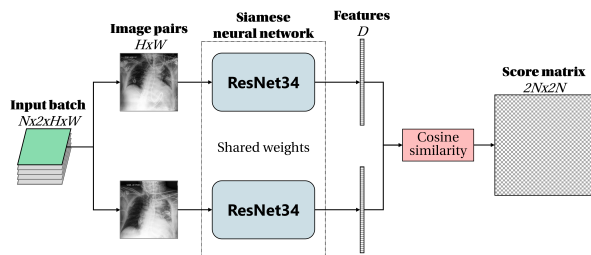


Figure 2: Schematic representation of the network architecture. The image pairs from the batch are each passed through the siamese ResNet34 streams, whereby the feature representations are extracted. The cosine similarity of the feature vectors is then calculated and the similarities in the batch are written into the score matrix.

larities, the cosine similarity between the feature representations is calculated. The similarity values are then used to create a $2N \times 2N$ score matrix, which contains all similarities between the individual images within the batch under consideration. The network is trained with a batch size of 32 over 8000 epochs. In addition, experiments are carried out with a batch size of 64 to investigate the influence of larger batch sizes. The network is adapted and optimized with the InfoNCE-Loss from [6] and the Adam optimizer with a learning rate of 1×10^{-3} . A learning rate scheduler reduces the learning rate by a factor of 0.2 after 4000 epochs.

During training, augmentations such as random horizontal flipping, photometric distortions and random deletion of image areas are used to increase the robustness of the model. As mentioned in the previous section, a quadruplet method is added to the network configuration. In this strategy, image groups of four images per patient are processed instead of image pairs. The aim of this extension is to investigate whether the network learns more efficiently due to the greater variability and range of features within the images per patient. Validation during training is performed with untransformed data to objectively evaluate the performance.

II.III. Evaluation

The evaluation procedure is based on the work of Heinrich and Hansen in [7] and is also visible in Figure 1. The fundamental idea is to evaluate the re-identification ability of the model by creating ranking lists. Potential image pairs are sorted according to their cosine similarities in the score matrix. The rank of the actual partner in this list is then determined for each image and used to analyze the network performance.

The test images are processed analogously by the siamese neural network. The 256-dimensional feature representations of the images from the batch are compared by cosine similarity and written into the score matrix. A ranking list is then created for each image based

on the matrix entries, in which the most similar images are sorted according to the similarity values. The rank of the actual image partners is then used as the central metric for evaluating recognition performance: Based on this ranking, the top- k accuracies are calculated, which measure the proportion of cases in which the actual partner is below the k highest results. When evaluating the network, geometric and photometric transformations are applied or omitted to assess the robustness of the model to variations in the input data.

III. Results and discussion

The re-identification results reveal some significant differences in performance between the different network configurations and datasets, as shown in Table 1. The corresponding plot in Figure 3 visually illustrates the relationship between re-identification accuracy and rank k , with overall accuracy consistently improving with increasing k in all cases examined.

At evaluating the entire CheXpert dataset, the re-identification network achieves a top-1 accuracy of 47.25%. This means that the actual image partner can correctly be ranked first in almost half of the test cases using the generated feature representations and the subsequent cosine similarity. The performance of the model increases significantly with higher ranks: top-5 accuracy is 73.00%, ranking the partner in top-10 81.00% and top-15 even 94.50%. These results demonstrate that the model effectively enables us to place the correct partner images predominantly among the top positions. The plot in Figure 3 therefore shows a steady increase, which indicates robust feature generation and similarity calculation.

In comparison, the use of the qualitatively enhanced homogeneous subset shows a better performance. The top-1 accuracy in Table 1 is 51.67%, while a placement in the top-10 already reaches 83.33%. This improvement can be explained by the lower image errors and the higher consistency within the subgroup, which allow the model to process more accurate feature representations.

A further increase in performance is achieved by enlarging the batch size to 64. In this configuration, the top-1 precision is 53.33% and the value for placing the correct partner image within the top-15 is 88.33%. Con-

Table 1: Top- k re-identification accuracies for different network configurations on CheXpert and KI-Rad-MSK

Model setup	Top-1	Top-5	Top-10	Top-15
Whole CheXpert	0.4725	0.7300	0.8100	0.9450
Homog. subset	0.5167	0.7667	0.8333	0.8667
Batch 64	0.5333	0.7500	0.8333	0.8833
No test aug.	0.4417	0.6833	0.7333	0.8333
Group of 4	0.8929	0.9821	1.0000	1.0000
KI-Rad-MSK	0.4200	0.6775	0.7875	0.8375

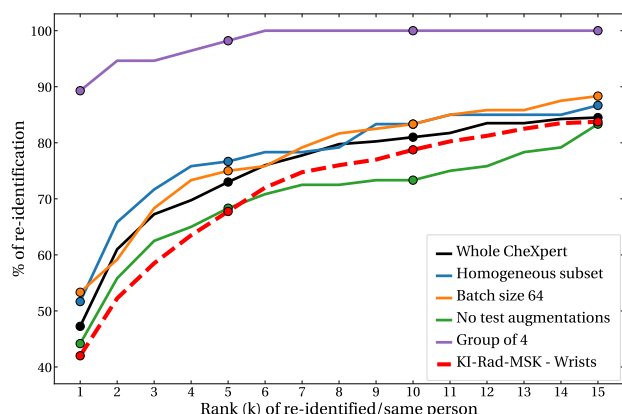


Figure 3: Graphical representation of the re-identification accuracies from different network constellations on CheXpert and KI-Rad-MSK. The accuracy percentage is shown for the configurations and datasets as a function of rank k .

sequently, a larger batch size during training promotes the stability and convergence of the model, leading to better generalization.

Results without test augmentations highlight the importance of image transformations for network generalizability. In absence of additional synthetic image modifications, the top- k accuracies drop. Accordingly, the graph in Figure 3 is significantly lower than in the previously discussed model configurations. This underlines that augmentations help account for image variability and improve the robustness of the model.

The best results can be obtained after a training process with image groups. Using this method, the model achieves a top-1 performance of 89.29%. An accuracy of 100.00% is already visible from a rank of $k=6$, so that the true partner is guaranteed to be among the top- k candidates from this rank. Increasing the feature set by using several different, qualitatively usable images per patient significantly improves the learning process of the network and enables a more precise re-identification. This indicates that a higher quantity of diversified image data per patient generally leads to optimized feature processing and consequently better network performance.

Compared to CheXpert, the KI-Rad-MSK dataset shows a slightly lower re-identification performance. Accordingly, Table 1 indicates that the top- k accuracies generally achieve a little lower percentages. Nevertheless, the graph in Figure 3 illustrates that the network performance also improves with increasing rank and is only marginally below the average CheXpert results. In general, the worse outcome can be attributed to the specific characteristics of the wrist radiographs, which provide less distinct biometric information compared to chest X-ray images. In addition, differences in image quality and acquisition conditions could further affect the re-identification.

IV. Conclusion

Our research highlights the capability of the implemented siamese neural network to re-identify additional medical radiographs of specific patients in anonymized datasets with high precision. In most network configurations used for CheXpert, the true partner of an image can be placed at the first rank with an accuracy of at least 50%. Particularly good results are achieved if the highest possible variation within individual patient data is used for training. Even with the wrist scans from KI-Rad-MSK, the performance is only slightly below the CheXpert average despite less distinct biometric information compared to the thorax, demonstrating the robustness of the model. Our findings show that conventional anonymization methods, such as metadata removal, are inadequate against modern deep learning technologies. In the future, innovative procedures for image synthesis, including latent diffusion models, could provide a solution to the problem addressed. These methods may minimize the risk of re-identification without restricting the availability of data, which remains essential for medical progress.

Acknowledgments

The work has been carried out at and supervised by the Institute of Medical Informatics, Universität zu Lübeck.

Author's statement

Conflict of interest: Authors state no conflict of interest. DeepL was used for linguistic refinement of this paper.

References

- [1] K. Packhäuser *et al.* Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest x-ray data. *Scientific Reports*, 12(1):14851, 2022, doi:[10.1038/s41598-022-19045-3](https://doi.org/10.1038/s41598-022-19045-3).
- [2] J. Reddy *et al.* A review on data breaches in healthcare security systems. *International Journal of Computer Applications*, 184(45):1–7, 2023, doi:[10.5120/ijca2023922333](https://doi.org/10.5120/ijca2023922333).
- [3] J. Irvin *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, 2019, doi:[10.1609/aaai.v33i01.3301590](https://doi.org/10.1609/aaai.v33i01.3301590).
- [4] F. Isensee *et al.* nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):1–9, 2021, doi:[10.1038/s41592-020-01008-z](https://doi.org/10.1038/s41592-020-01008-z).
- [5] Y. Taigman *et al.*, Deepface: Closing the gap to human-level performance in face verification, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1701–1708, 2014. doi:[10.1109/CVPR.2014.220](https://doi.org/10.1109/CVPR.2014.220).
- [6] A. van den Oord *et al.* Representation learning with contrastive predictive coding. *CoRR*, 2018, doi:[10.48550/arXiv.1807.03748](https://doi.org/10.48550/arXiv.1807.03748).
- [7] M. P. Heinrich and L. Hansen, Implicit neural compression for privacy preserving medical image sharing, in *Medical Imaging with Deep Learning*, 2024. URL: <https://proceedings.mlr.press/v250/heinrich24a.html>.