Hide Identity, Preserve Pathology: Diffusion-Based Anonymization for Chest X-rays

Yasmeena Akhter[*], **Muskan Dosi**[*], **Mayank Vatsa**, **Richa Singh**

Indian Institute of Technology Jodhpur, India
{akhter.1,dosi.1,mvatsa,richa}@iitj.ac.in

Abstract

Chest X-rays are a widely-used, cost-effective imaging modality for medical investigations; however, they encode distinctive biometric signatures that enable identification attacks. We introduce *PrivDiff-Net*, a novel diffusion-based framework that addresses critical privacy vulnerabilities in chest X-rays through anatomical biometric features while preserving essential diagnostic utility for clinical applications. Our approach introduces two modules in a latent diffusion framework: (1) a Selective Attribute Suppression (SAS) module that removes sensitive identity cues using orthogonal projection in cross-attention, and (2) a Selective Privacy Guidance (SPG) loss that discourages identity features while preserving diagnostic information during diffusion. Quantitative results show that PrivDiff-Net achieves near-random identification (AUC: 47%) while maintaining high diagnostic accuracy (AUC: 78%). It effectively suppresses sensitive attributes and produces high-quality anonymized CXRs, validated by clinicians for diagnostic utility. These results establish *PrivDiff-Net* as a new benchmark for privacy-preserving Chest X-rays, providing a practical solution for secure data sharing in collaborative research environments while enabling ethical deployment of AI systems in healthcare where transparency and patient privacy are critical.

Introduction

Chest X-rays (CXRs) are the most widely performed radiological examination for assessing and monitoring lung and thoracic disorders. With over 1.5 billion procedures conducted annually worldwide, they remain the cornerstone of modern diagnostic imaging (Smith-Bindman et al. 2019; Abhisheka et al. 2024). Their affordability, accessibility, and diagnostic reliability make them indispensable in both advanced and resource-limited healthcare systems. As the most frequently used imaging modality for pulmonary diseases, CXRs offer vast potential for automated screening, diagnosis, and monitoring of respiratory conditions (Akhter, Singh, and Vatsa 2023; Qin et al. 2018; Agrawal and Choudhary 2022; Thirukrishna 2022; Çallı et al. 2021; Akhter et al. 2025). Beyond routine screening, they play a crucial role in emergency medicine, critical care, and global health initiatives, where rapid and accurate diagnosis directly impacts outcomes. The integration of Artificial Intelligence (AI) has further amplified their clinical value, enabling automated detection of subtle radiographic patterns and supporting rapid, data-driven decision-making (Khalifa and Albadawy 2024; Akhter et al. 2023; Zhou et al. 2021; Akhter, Singh, and Vatsa 2023; Akhter 2025). This synergy between AI and CXR analysis holds particular promise for resource-limited regions where AI-powered diagnostic tools can deliver scalable healthcare solutions (Ram and Bodduluri 2023).

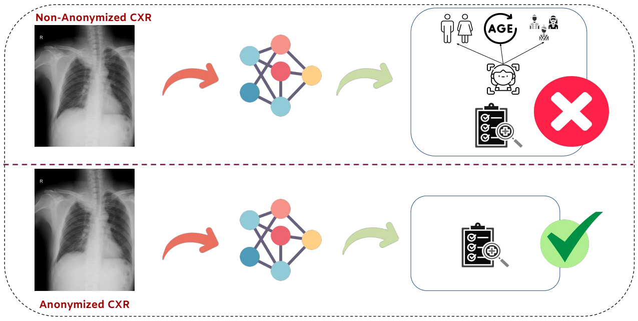

Figure 1: Highlights the challenge of possible privacy leakage in CXR-AI diagnostic systems.

The rapid adoption of AI in chest radiography has revealed an unexpected vulnerability: *chest biometrics*, referring to person-specific anatomical and physiological signatures within the thoracic cavity that persist across imaging sessions. These biometric cues, such as rib cage geometry, cardiac silhouette, and spinal curvature, can enable automated identification by linking anonymized scans to individuals (Packhäuser et al. 2022; Vayena, Blasimme, and Cohen 2018). While such patterns enhance diagnostic accuracy, they also expose patients to privacy risks, creating a tension between clinical utility and confidentiality. Inadequate protection of these features can lead to unauthorized tracking, demographic profiling, and breaches of medical confidentiality. The privacy threat is not theoretical: under HIPAA Privacy Rules, over 176 million patients in the United States have already been affected by Protected Health Information (PHI) breaches (K Pool et al. 2019; Nass, Levit, and Gostin 2009; Abbasi and Smith 2024; Isibor 2024). With the global medical X-ray market projected to grow from USD 14.99

[*] These authors contributed equally.

billion in 2024 to USD 23.93 billion by 2032[1], the scale of potential exposure continues to rise. As healthcare increasingly relies on cloud-based AI systems and cross-institutional data sharing, secure anonymization of diagnostically useful CXRs becomes essential. Moreover, emerging paradigms such as federated learning and foundation models for chest imaging demand stronger, modern privacy safeguards to prevent identification while supporting large-scale, data-driven medical innovation (Glocker et al. 2023; Paschali et al. 2025).

Developing robust anonymization techniques that preserve diagnostic integrity while removing identity-revealing features is essential for ethical AI deployment and regulatory compliance in medical imaging (Vizitiu et al. 2021; Popescu et al. 2021; Kim et al. 2020). Existing privacy-preserving approaches mainly follow two paradigms: perturbation-based and adversarial methods. Perturbation-based techniques, such as differential privacy (DP) (Dwork et al. 2006; Fan 2018), protect biometric information by adding randomized noise to maintain overall data statistics (Abadi et al. 2016; Bu et al. 2020; Dong, Roth, and Su 2019; Dwork et al. 2006; Ziller et al. 2021). DP-Pix (Fan 2018, 2019) extended these principles to medical images through pixel block averaging and calibrated noise injection, though often at the cost of image quality. More recent deep learning approaches focus on privacy-guaranteed synthetic image generation and adversarial frameworks like Privacy-Net (Popescu et al. 2021), which better balance biometric concealment with diagnostic utility. The core challenge remains achieving selective suppression of identity-specific cues without degrading the pathological features vital for accurate diagnosis.

We present *PrivDiff-Net*, a diffusion-based framework for selective attribute suppression in chest X-rays, enabling effective de-identification while preserving diagnostic utility. To our knowledge, this is the first approach to use diffusion models specifically for CXR anonymization. Unlike conventional methods that apply uniform transformations, *PrivDiff-Net* disentangles identity-related cues from diagnostic features in the latent space. Through a latent diffusion process, it selectively removes biometric markers such as identity, age, and gender, while retaining pathology-relevant patterns essential for disease detection. The proposed **Selective Attribute Suppression (SAS)** module modifies cross-attention using orthogonal projection to eliminate sensitive components, and the **Selective Privacy Guidance (SPG)** loss penalizes identity-related information during generation. Together, these components enable secure data sharing for AI-driven medical research and diagnostics without compromising privacy. Our main contributions include:

- **Diffusion-Based De-Identification:** We propose a diffusion-based approach to anonymize CXR images by selectively removing sensitive markers while preserving diagnostic information. Our method integrates anonymization within the denoising process, ensuring sensitive features are suppressed without compromising medical structures.
- **Cross-Attention Feature Filtering with Orthogonal Projection:** The Selective Attribute Suppression (*SAS*) module modifies cross-attention in diffusion models using orthogonal projection cleaning to remove identity, age, and gender-related components while preserving diagnostic signals, ensuring AI models focus on clinically relevant features.
- **Selective Privacy Guidance (*SPG*) loss**: We introduce SPG Loss to penalize identity-related features in generated images, enforcing stronger disentanglement between personal attributes and diagnostic information. This loss minimizes identification risks while preserving disease-related patterns.
- **Experimental Validation:** To rigorously validate *PrivDiff-Net*'s effectiveness, we conducted extensive experiments across multiple datasets for Privacy-Utility performance.
- **Clinical Validation:** To assess the practical utility of *PrivDiff-Net* in real-world clinical settings, we conducted an evaluation with board-certified radiologists. The results show that anonymized CXRs generated by *PrivDiff-Net* retain sufficient diagnostic quality for accurate interpretation, achieving an inter-rater agreement of $\kappa = 0.89$ between original and anonymized images.

Proposed Methodology

Problem Formulation

We formulate the anonymization task as a conditional image generation problem within the latent diffusion framework. Let the dataset be defined as $\mathcal{D} = (x_i, y_i^{\text{id}}, y_i^{\text{age}}, y_i^{\text{gender}}, y_i^{\text{diag}})_{i=1}^{N}$, where each chest X-ray image x_i is associated with four categorical labels: identity, age, gender, and diagnosis. Specifically, the identity label $y_i^{\text{id}} \in \{1, 2, \ldots, I\}$ corresponds to the subject identifier, the age label $y_i^{\text{age}} \in \{1, 2, \ldots, A\}$ represents the patient's age group, the gender label $y_i^{\text{gender}} \in \{1, 2\}$ indicates male or female, and the diagnosis label $y_i^{\text{diag}} \in \{1, 2, \ldots, D\}$ denotes the clinical condition associated with the image. The identity, age, and gender attributes are obtained directly from patient metadata in the dataset and serve as sensitive factors to be suppressed during anonymization, while the diagnosis label is used to preserve clinically relevant information.

The objective is to transform x_i into a new image $\hat{x}_i$ such that the sensitive attributes are effectively suppressed, while the diagnostic information remains intact. We consider privacy risks where an adversary may try to identify a patient or infer sensitive attributes such as identity, age, or gender from medical images. Such attacks can occur when a recognition or embedding model is used on generated or shared images. Our goal is to prevent these identification or attribute inference attempts by removing sensitive information, while ensuring that diagnostic cues important for clinical use is preserved.

- *Privacy Preservation*: The sensitive attributes $y_i^s = (y_i^{id}, y_i^{age}, y_i^{gender})$ become unrecoverable from $\hat{x}_i$
- *Diagnostic Utility*: The diagnostic information y_i^{diag} remains intact and accurately predictable from $\hat{x}_i$

[1] https://www.fortunebusinessinsights.com/medical-x-ray-market-104862.

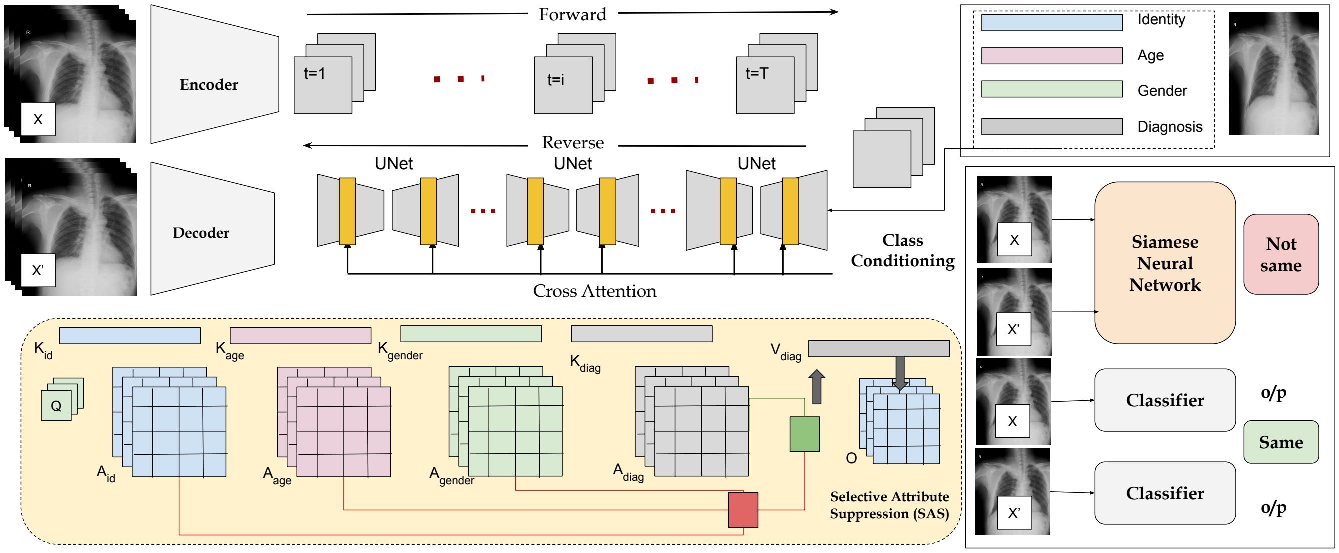

Figure 2: Overview of the proposed *PrivDiff-Net*, a diffusion-based anonymization framework. The model uses forward and reverse diffusion processes conditioned on class information, with the Selective Attribute Suppression (*SAS*) module applying cross-attention and orthogonal projection to remove sensitive features while preserving diagnostic relevance. A Siamese network and classifiers evaluate privacy preservation and diagnostic utility.

Mathematically, this can be expressed as an optimization problem:

$$\min_{\mathcal{T}} E_{x,y^{diag}}[\mathcal{L}_{diag}(f_{diag}(\mathcal{T}(x)), y^{diag})]+$$
$$\lambda \max_{\mathcal{T}} E_{x,y^s}[\mathcal{L}_s(f_s(\mathcal{T}(x)), y^s)]$$

where f_{diag} and f_s are pretrained classifiers for diagnostic and sensitive attributes, respectively. $\mathcal{L}_{diag}$ and $\mathcal{L}_s$ are corresponding loss functions, and λ is a hyperparameter balancing the privacy-utility trade-off. The choice of λ reflects the desired emphasis and optimal balance; too high a value may harm diagnostic fidelity, while too low a value may yield weaker privacy.

PrivDiff-Net Framework

To achieve this, we employ a Latent Diffusion Model (LDM) (Sohl-Dickstein et al. 2015) framework, where the input image is first encoded into a compressed latent space using a pretrained Variational Autoencoder (VAE) and then progressively modified through a controlled denoising process rather than directly on pixel values.

Given an input image x_i, we first encode it into a latent representation $z_0 = \mathcal{E}(x_i)$ using a pre-trained encoder $\mathcal{E}$, where $z_0 \in R^{h\times w\times c}$ with $h < H, w < W$, and c representing the channel dimension. Given a noisy latent representation z_t at timestep t, the reverse denoising process is conditioned to preserve diagnosis-related information while suppressing sensitive attributes. The forward diffusion process gradually adds Gaussian noise to the latent representation over T timesteps:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I})$$

where $\{\beta_t\}_{t=1}^T$ is a predefined noise schedule. The forward process can be expressed in closed form:

$$q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t}z_0, (1-\bar{\alpha}_t)\mathbf{I})$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. The reverse diffusion process learns to denoise the latent representation through a U-Net architecture ϵ_θ:

$$p_\theta(z_{t-1}|z_t, c) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t, c), \sigma_t^2\mathbf{I})$$

where c represents the conditioning information and μ_θ is parameterized as:

$$\mu_\theta(z_t, t, c) = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t, t, c)\right)$$

We employ two key techniques to achieve our goal: **Selective Attribute Suppression (SAS)**, leveraging the cross-attention mechanism to filter out unwanted features, and **Selective Privacy Guidance Loss (SPG)** during each reverse diffusion step to explicitly penalize the presence of such information in the generated images. Together, these methods work synergistically to suppress unwanted attributes while preserving the clinical utility of the CXR images.

Selective Attribute Suppression

We introduce the Selective Attribute Suppression (*SAS*) module, to ensure the effective removal of sensitive information from CXR images while retaining diagnostic relevance. This module modifies the cross-attention mechanism within the *LDM* framework by using orthogonal projection cleaning. This operation systematically filters out sensitive features while preserving disease-related information during the denoising process. At each timestep t, let $\mathbf{z}_t \in R^{H\times W\times D}$ denote the latent representation of the image, where $H \times W$

represents the spatial dimensions and D is the feature dimension. For each category $c \in \{id, age, gender, diag\}$, we compute the query Q, key K_c, and value V_c matrices as follows:

$$\mathbf{Q} = \mathbf{W}_Q \cdot z_t \in R^{N \times d_k}$$

$$\mathbf{K}_c = \mathbf{W}_{K_c} \cdot \mathbf{C}_c \in R^{L_c \times d_k}$$

$$\mathbf{V}_c = \mathbf{W}_{V_c} \cdot \mathbf{C}_c \in R^{L_c \times d_v}$$

where $\mathbf{W}_Q \in R^{d_k \times D_f}, \mathbf{W}_{K_c} \in R^{d_k \times d_c}$ and $\mathbf{W}_{V_c} \in R^{d_v \times d_c}$ are learnable projection matrices for queries, $\mathbf{C}_c \in R^{L_c \times d_c}$ is the embedding matrix for category c, $N = H_f \times W_f$ is the number of spatial locations, and L_c is the number of classes in category c. For each category, the attention map is generated using scaled dot-product attention mechanism:

$$\mathbf{A}_c = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}_c^T}{\sqrt{d_k}}\right) \in R^{N \times L_c}$$

The softmax operation normalizes the attention scores, ensuring that the focus is distributed appropriately across the feature space corresponding to each category. To remove sensitive features, we first aggregate the attention maps corresponding to identity, age, and gender attributes.

$$\mathbf{F}_s = [\mathbf{A}_{id}, \mathbf{A}_{age}, \mathbf{A}_{gender}] \in R^{N \times (L_{id} + L_{age} + L_{gender})}$$

This concatenated matrix captures all sensitive information embedded within the latent space. To project out these sensitive components from the diagnosis-related attention map, we construct an orthogonal projection matrix:

$$\mathbf{P} = \mathbf{I} - F_s(F_s^\top F_s + \epsilon I)^{-1} F_s^\top$$

Here $\mathbf{I}$ is the identity matrix that ensures dimensional consistency. Applying this projection removes the influence of sensitive features from the diagnostic attention map to compute a clean representation, ensuring diagnostic features remain intact while systematically removing contributions from sensitive attributes. The cleaned diagnostic attention map is obtained by applying the projection:

$$\mathbf{A}_{\text{diag}}^{'} = \mathbf{P} \cdot \mathbf{A}_{\text{diag}}$$

which satisfies $\mathbf{P}F_s \approx \mathbf{0}$ and hence removes contributions aligned with the sensitive subspace. Finally, the output features are computed, which are further normalized for use:

$$\mathbf{O} = \mathbf{A}_{\text{diag}}^{'} \cdot \mathbf{V}_{\text{diag}}$$

We aim to remove cues that enable re-identification or demographic inference, while preserving radiographic pathology; correlations may exist, but our goal is to prevent recovery of the sensitive labels. Hence, it is reasonable to approximate diagnostic and sensitive features as lying in distinct, linearly separable subspaces within the latent representation.

The orthogonal projection employed in the Selective Attribute Suppression (SAS) module removes components of the diagnostic attention map that lie within the subspace spanned by sensitive attribute features. The projection matrix effectively nullifies any component aligned with these sensitive features, enforcing orthogonality between preserved diagnostic information and removed sensitive cues and disentangles identity-related signals from diagnostic representations.

Selective Privacy Guidance Loss

We define our privacy-focused noise ϵ_p prediction as:

$$\begin{aligned}\epsilon_{\text{p}} = \epsilon_\theta(\mathbf{z}_t, \phi) + w_{\text{diag}}\left[\epsilon_\theta(\mathbf{z}_t, \text{diag}) - \epsilon_\theta(\mathbf{z}_t, \phi)\right] \\ -|w_{\text{s}}|\left[\epsilon_\theta(\mathbf{z}_t, \text{s}) - \epsilon\theta(\mathbf{z}t, \phi)\right]\end{aligned}$$

Here, $\epsilon_\theta(\mathbf{z}_t, \phi)$ represents the unconditional noise prediction. In contrast, $\epsilon_\theta(\mathbf{z}_t, \text{diag})$ and $\epsilon_\theta(\mathbf{z}_t, \text{s})$ are the noise predictions conditioned on diagnostic information and sensitive attributes respectively. The parameter $w_{\text{diag}} > 1$ amplifies diagnostic features, ensuring strong retention of medically relevant details. While, negative guidance $w_{\text{s}} < 0$ actively suppresses the influence of sensitive features by subtracting their contribution. Adjusting w_{s} controls the privacy–utility trade-off: a larger magnitude enhances anonymization but may risk partial loss of diagnostic fidelity. ϵ_p is computed for a single attribute (s) at a iteration. We define the Selective Privacy Guidance Loss (*SPG*) function as:

$$\mathcal{L}_{\text{SPG}} = \|\epsilon_{\text{p}} - \epsilon_{\text{true}}\|_2^2$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is the true noise added during the forward process. It ensures that the overall noise prediction matches the true noise while enforcing the removal of unwanted attributes. By balancing the diagnostic amplification and sensitive suppression, the *SPG* loss effectively integrates the dual objectives of accurate denoising and privacy preservation. The complete framework is illustrated in Figure 2.

Experimental Setup

Dataset Composition: To evaluate the proposed anonymization framework, we used two large-scale CXR datasets: ChestX-ray14 (CXR14) (Wang et al. 2017) and CheXpert (Irvin et al. 2019). The **ChestX-ray14 (CXR14) dataset** comprises 112,120 frontal chest X-rays from 30,805 patients, standardized as 8 bit grayscale images with 1024×1024 resolution, typically downsampled to 256×256 for computational efficiency while maintaining diagnostic quality. It covers 14 thoracic pathologies, including Atelectasis, Pneumonia, Pneumothorax, Emphysema, Cardiomegaly, Consolidation, Fibrosis, Pleural Thickening, Edema, Effusion, Hernia, Mass, Nodule, and Infiltration, enabling a broad evaluation of diagnostic utility preservation. Following the *PriCheXy-Net* (Packhäuser et al. 2023) protocol, we used 10K, 2K, and 5K image pairs for training, validation, and testing, respectively. To assess generalization beyond CXR14, we also evaluated on the **CheXpert dataset**, which contains 224,316 chest X-rays from 65,240 patients collected at Stanford Hospital. CheXpert provides both frontal and lateral views with varied imaging protocols and demographics, covering 14 pathological observations, including No Finding, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, and Support Devices. The dataset was split at the patient level into 10K, 2K, and 5K image pairs for training, validation, and testing, ensuring no overlap across subsets.

Framework Evaluation: We evaluate our proposed *PrivDiff-Net* on Image generation, disease classification, and patient identification. To assess the quality of the generated CXR sample, we used the Fréchet Inception Distance (FID) to measure the distributional similarity between original and anonymized images.

$$\text{FID} = ||\mu_r - \mu_g||^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$$

and Structural Similarity Index Measure ($SSIM$) for measuring perceptual image quality,

$$\text{SSIM}(x, \hat{x}) = \frac{(2\mu_x\mu_{\hat{x}} + c_1)(2\sigma_{x\hat{x}} + c_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + c_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2)}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of real and generated feature distributions. Disease classification and patient identification tasks are assessed using the Area under the ROC Curve (AUC) score.

Implementation Details The complete pipeline is developed using the PyTorch framework (Paszke et al. 2019). The core architecture employs a U-Net-based VAE diffusion model, incorporating multi-head self-attention layers at 16×16 and 8×8 resolutions to capture both local and global anatomical dependencies. Each cross-attention module within the U-Net integrates our novel *SAS* component, utilizing eight attention heads with key and value dimensions of 64, learnable category embeddings of size 256 for each attribute type, and a projection dimension of 256 to ensure rich feature representation. To train the proposed model, we set the batch size, learning rate and optimizer to 32, 1e-5, and Adam (Kingma and Ba 2015), respectively, for 300 epochs. To report the results, we used the evaluation metrics from the Scikit-learn library (Pedregosa et al. 2011). The training is performed on NVIDIA V100-DGXS GPUs.

Experimental Validation

We rigorously assess the efficacy of our proposed privacy preservation framework for various tasks: image generation, diagnosis utility and privacy leakage. We also compare our performance with existing baseline models.

Baseline methods: We evaluated our proposed *PrivDiff-Net* against established anonymization frameworks, including DP-Pix (Fan 2018, 2019), Privacy-Net (Kim et al. 2021), and PriCheXy-Net (Packhäuser et al. 2023). These baselines represent the current state-of-the-art privacy-preserving CXR imaging. Our comprehensive evaluation examined disease classification performance and patient anonymization efficacy. Table 1 presents the quantitative results across all methods. The following sections examine these results in detail, highlighting the trade-offs between privacy preservation and diagnostic accuracy across different approaches.

Disease Diagnostic Task: To evaluate the diagnostic task on the generated anonymized test set, we used the pretrained ChexNet(Rajpurkar et al. 2017), which is a 121-layered DenseNet (Huang et al. 2017). This evaluation ensures that the data utility is preserved, without affecting the underlying abnormalities in the CXR samples. Results (Table 1) demonstrate that *PrivDiff-Net* achieves a mean AUC score of 0.78, maintaining higher diagnostic accuracy. It validates our method's ability to preserve essential pathological features while suppressing identifying characteristics in CXRs.

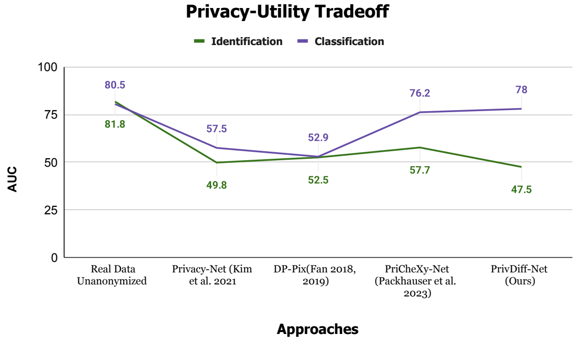

Figure 3: Comparing patient identification (lower is better) and disease classification performance (higher is better) across different anonymization approaches.

Patient Identification Task: Our experimental results as shown in Table 1 demonstrate that our proposed *PrivDiff-Net* significantly advances the state of privacy preservation in CXR imaging. The achieved patient verification performance of 47.5% represents a substantial improvement over baseline approaches (Kim et al. 2021; Packhäuser et al. 2023; Fan 2018, 2019). PrivDiff-Net effectively pushes the identification capability closer to random chance (50%). This performance metric is particularly noteworthy because it indicates that our latent diffusion-based approach successfully disrupts privacy leakage from CXRs.

The effectiveness of *PrivDiff-Net* lies in its ability to operate in the latent space, where it learns to remove subtle anatomical cues that could reveal patient identity. Unlike traditional obfuscation methods based on masking or pixel-level alterations, it captures a deeper understanding of anatomical structures, allowing selective suppression of identity-related features without affecting diagnostic content. Beyond privacy protection, *PrivDiff-Net* addresses a major challenge in medical data sharing by eliminating the need to store or distribute original, unanonymized data. This enables secure collaboration across institutions while preserving data utility, thereby accelerating medical imaging research and AI development. Moreover, the framework can be extended to other medical imaging modalities where privacy is critical, and it establishes a foundation for creating privacy-preserving methods that retain high performance in downstream tasks such as disease classification and abnormality detection.

Analysis of Privacy-Utility Trade-off in CXR Images: Our experimental results, summarized in Table 1, demonstrate that *PrivDiff-Net* achieves an optimal balance between patient privacy protection and preservation of diagnostic utility. The comparative analysis reveals several significant findings across the evaluated anonymization techniques. The non-anonymized medical images present substantial privacy risks, with patient identification possible at an AUC of 81.8%. This confirms the critical need for effective anonymization

Table 1: Quantitative comparison of anonymization techniques measuring privacy protection (verification AUC, mean ± std from 10 independent runs) and diagnostic utility on ChestXray14 dataset. Non-anonymized data serves as the baseline.

Task	Real Data Unanonymized	Privacy-Net (Kim et al. 2021)	DP-Pix (b=8) (Fan 2018, 2019)	PriCheXy-Net (μ =0.01) (Packhäuser et al. 2023)	*PrivDiff-Net* (**Ours**)
Identification↓	81.8 ±0.6	49.8 ±2.2	52.5 ±3.2	57.7 ±4.0	47.5 ±0.8
Classification↑	80.5 ±0.4	57.5 ±1.3	52.9 ±0.6	76.2 ±0.4	78.0 ±1.0

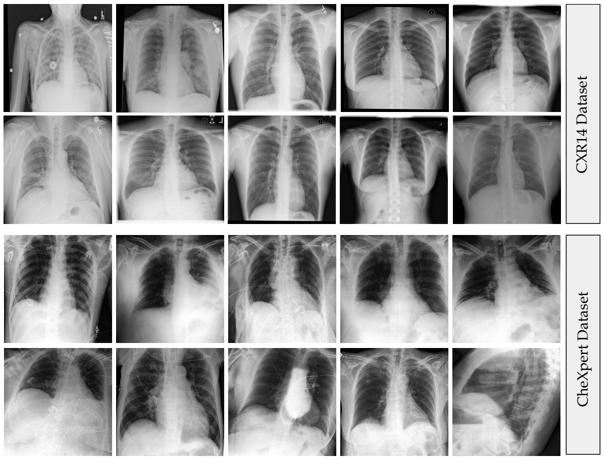

Figure 4: Illustrates the examples of CXR samples generated using *PrivDiff-Net*.

methods in medical imaging. When examining existing approaches, we observe that DP-Pix reduces identification to 50.0% – 52.5% AUC. At the same time, Privacy-Net achieves 49.8% AUC. Our proposed *PrivDiff-Net* framework further improves privacy protection with a identification AUC of 47.5%, which approaches random classification performance. This indicates that *PrivDiff-Net* effectively obfuscates patient-specific biometric information within the CXRs.

Generally, the critical challenge in medical image anonymization lies not only in protecting privacy but also in maintaining diagnostic value On this front, in our case, the non-anonymized data demonstrates high utility with a disease classification AUC of 80.5% However, existing methods significantly compromise this utility: DP-Pix shows severe degradation (50.0%–52.9% AUC), Privacy-Net provides moderate performance (57.5% AUC), and PriCheXy-Net (at μ=0.01) achieves similar results (76.2% AUC) In contrast, *PrivDiff-Net* preserves high diagnostic utility with a classification AUC of 78%, representing a modest margin of 2% improvement over PriCheXy-Net and providing substantially stronger privacy preservation. The improvement on the Privacy-Utility Trade-off with *PrivDiff-Net* over the existing baseline approaches is also highlighted in Figure 3.

These results demonstrate that *PrivDiff-Net* addresses the fundamental tension between privacy protection and diagnostic utility preservation in CXR images. Unlike previous approaches where enhanced privacy often incurs considerable utility loss, our method effectively removes identity-revealing features while preserving the disease-relevant information necessary for accurate clinical analysis and research.

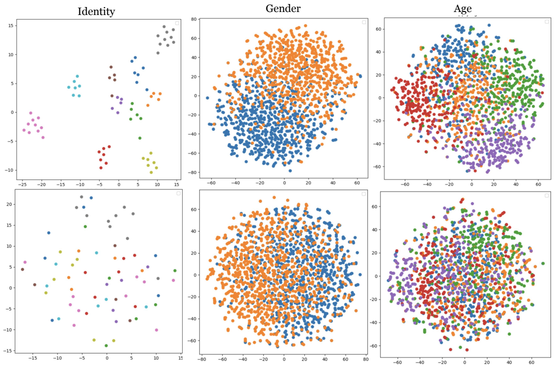

Figure 5: t-SNE visualization of feature representations before (top) and after (bottom) applying anonymization. Initially, clear separability is observed for identity, gender, and age attributes. After applying our method, the clusters for sensitive attributes become significantly entangled, effectively suppressing identity, age, and gender information while retaining diagnostic relevance.

Image Generation Task: *PrivDiff-Net* is evaluated for image generation quality using the standard metrics of FID and $SSIM$, achieving scores of 1.37 and 0.85, respectively. These quantitative results demonstrate that our generated CXRs closely align with the original data distribution while maintaining privacy. The high fidelity of the generated images is further validated by successful disease classification performance on the anonymized samples.

The visual quality of our results is shown in Figure 4, which presents generated CXR samples. Strong performance in image quality and diagnostic tasks confirms that *PrivDiff-Net* produces realistic, clinically useful CXRs without requiring access to unanonymized data. This enables secure medical image sharing while maintaining visual fidelity and diagnostic utility. Our findings highlight key implications for:

1. Privacy-preserving dataset sharing,
2. Secure clinical AI development,
3. Data augmentation in medical imaging, and
4. Collaborative research

The effectiveness of our anonymization approach is further supported by feature visualization through t-SNE plots in Figure 5. The generated data will be made publicly available for research purposes.

Table 2: Ablation study demonstrating the incremental contribution of each component in *PrivDiff-Net*. Results show patient identification accuracy (AUC (%)) and disease classification AUC (%) on ChestX-ray14 test set.

Model	Identification ↓	Classification ↑
Vanilla Diffusion	75.2	72.1
Diffusion + SAS	61.3	74.0
Diffusion + SPG	56.6	76.8
Diffusion + SAS + SPG	**47.5**	**78.0**

Ablation Study: To assess the effectiveness of the individual components of the proposed anonymization framework, we performed the ablation study. We compared the effectiveness of proposed with three settings: vanilla diffusion, diffusion with SAS, and diffusion with SPG. The results are provided in Table 2. The ablation study reveals the critical contribution of each component in achieving PrivDiff-Net's privacy-utility balance.

- *Baseline Vanilla Diffusion Model*: The vanilla diffusion model alone provides limited privacy protection with 75.2% identification accuracy, only marginally better than non-anonymized images (81.8%). Its diagnostic performance of 72.1% also falls short, indicating that undirected diffusion processes cannot adequately balance privacy and utility.
- *Impact of Selective Attribute Suppression (SAS)*: Adding the SAS module yields a substantial 13.9% reduction in identification (from 75.2% to 61.3%) while improving diagnostic accuracy by 1.9%. This demonstrates that orthogonal projection in cross-attention effectively filters identity-related features from diagnostic representations without degrading clinical information. The simultaneous improvement in both metrics validates our hypothesis that sensitive and diagnostic features occupy separable subspaces.
- *Impact of Selective Privacy Guidance (SPG)*: The SPG loss alone achieves even stronger privacy protection, reducing identification to 56.6% (18.6% improvement) while boosting diagnostic accuracy to 76.8%. This significant dual improvement confirms that explicit guidance during the diffusion process can effectively steer generation away from identity-revealing features while reinforcing pathological diagnostic patterns.
- *Synergistic Effect of SAS and SPG*: The complete framework (Diffusion + SAS + SPG) achieves optimal performance with 47.5% identification, approaching random chance (50%) while maintaining the highest diagnostic accuracy at 78.0%. The combined approach outperforms the sum of individual components, indicating synergistic interaction. The SAS module's feature-level filtering complements SPG's generation-level guidance, creating a robust dual-mechanism privacy preservation system.

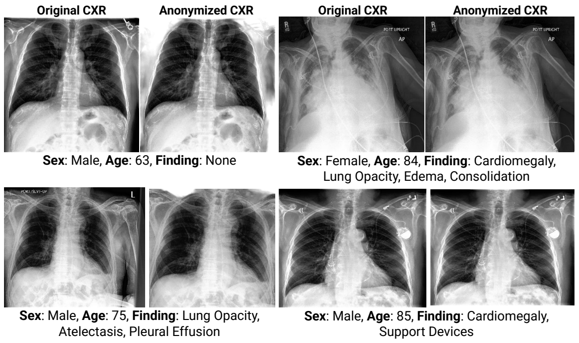
Original CXR
Anonymized CXR
Original CXR
Anonymized CXR
Sex: Male, Age: 63, Finding: None
Sex: Female, Age: 84, Finding: Cardiomegaly, Lung Opacity, Edema, Consolidation
Sex: Male, Age: 75, Finding: Lung Opacity, Atelectasis, Pleural Effusion
Sex: Male, Age: 85, Finding: Cardiomegaly, Support Devices

Figure 6: Samples of original versus anonymized CXRs validated by five radiologists, confirming 98% diagnostic adequacy enabling secure multi-institutional collaboration.

Clinical Validation for Real-Time Application

To validate the clinical applicability and visual quality of *PrivDiff-Net*-generated anonymized CXRs, we conducted a comprehensive user study with a cohort of five board-certified radiologists, including three chest radiologists and two general radiologists with extensive experience in thoracic imaging. The work has been approved by the Institutional Review Board (IRB) (Ref No. IEC/IITJ/2022-23/03). The study was designed to assess whether *PrivDiff-Net*-generated images maintain sufficient diagnostic quality for real-world clinical interpretation while effectively concealing patient identity. The study protocol involved blind assessment of 50 randomly selected image pairs (original and anonymized) across two key dimensions: diagnostic quality, and pathology detection accuracy. Each radiologist reviewed the complete set of images individually to provide a strong inter-expert consensus. Few samples of the reviewed CXRs are provided in Figure 6. Experts evaluated the diagnostic adequacy of anonymized images for 14 common thoracic conditions, comparing their findings against original CXRs. The cohort reported that 98% of anonymized images maintained sufficient diagnostic quality for clinical interpretation, with a diagnostic agreement of κ=0.89 between original and anonymized images. All participating doctors agreed that the generated anonymized CXRs preserve essential diagnostic features including lung parenchyma detail, cardiac borders, and osseous (bony) structures. This validation provides practical evidence that our approach can facilitate secure cross-institutional collaboration while maintaining the clinical standards necessary for accurate diagnosis and patient care.

Generalization Performance and External Validation: To evaluate the clinical robustness and generalization of *PrivDiff-Net*, we tested it on the CheXpert dataset, which differs from the ChestX-ray14 training data in geography and institution. The cross-dataset experiments covered diagnostic classification, privacy preservation, age and gender prediction, and image generation quality. Results show that *PrivDiff-Net* maintains consistent performance across both datasets, confirming its strong generalization for real-world medical use. Specifically, *PrivDiff-Net* maintains strong privacy protection on CheXpert with a patient identification AUC of 43.95% (unanonymized original data: 71.24%), closely mirroring the 47.5% performance observed on ChestX-ray14, indicating that our selective attribute suppression mechanism effectively

Table 3: Gender and age attribute suppression performance on CheXpert Dataset. The result demonstrates the effectiveness of *PrivDiff-Net* in reducing sensitive attribute classification accuracy to near-random levels while preserving diagnostic utility.

Attribute	**Real Data (%)**	**PrivDiff-Net (%)**	**Reduction (%)**
Age Classification	77.5	43.2	34.3
Gender Classification	81.91	50.11	31.8

disrupts biometric patterns regardless of imaging protocols or patient demographics.

The anonymized CheXpert images achieve an average AUC of 81.82% across 14 pathologies, with only a 5.67% drop from the unanonymized original images. This consistency across diverse imaging conditions and patient populations shows that *PrivDiff-Net* learns generalizable anonymization principles rather than dataset-specific patterns. The strong cross-institutional results validate its readiness for real-world clinical use, enabling secure multi-institutional research and AI model development without compromising privacy or diagnostic accuracy. For the CheXpert dataset, the generated CXRs achieve an FID of 3.28 and SSIM of 0.76, demonstrating high visual quality.

Gender and Age Attribute Suppression Performance Analysis on CheXpert Dataset: We evaluated the proposed *PrivDiff-Net* for the attribute suppression capabilities on the CheXpert dataset and demonstrated effectiveness in removing sensitive demographic information while maintaining diagnostic utility. Table 3 presents the quantitative results for sensitive attribute classification. The age classification performance demonstrates effective suppression, with accuracy dropping from 77.5% on original non-anonymized CXRs to 43.2% on *PrivDiff-Net* anonymized images with a substantial reduction of 34.3%. This performance approaches random classification for age prediction, indicating that our Selective Attribute Suppression (SAS) module successfully obfuscates the age-related anatomical markers typically encoded in CXRs. Gender classification accuracy indicates equally impressive suppression, decreasing from 81.91% on original CXRs to 50.11% on anonymized images with a reduction of 31.8% . The final performance of 50.11% is significantly close to random chance (50% for binary classification), indicating near-complete suppression of gender-revealing anatomical features. The near-random gender classification performance represents a critical privacy achievement to restrict gender based tasks such as demographic profiling and gender bias for CXR-AI diagnostics systems.

The similar suppression performance on CheXpert validates that *PrivDiff-Net* learns generalizable anonymization principles rather than dataset-specific artefacts. Dataset-independent performance is crucial for real-world deployment, where AI systems must handle diverse patient populations across healthcare institutions. The consistent attribute suppression across different imaging protocols, patient demographics, and institutional practices confirms that our proposed anonymization approach maintains its privacy-preserving capabilities with diagnostic utility.

Computational Complexity and Runtime Considerations: *PrivDiff-Net* employs Latent Diffusion Models (LDMs), which offer a significant computational advantage over conventional diffusion approaches. LDMs operate in a compressed latent space, typically 8× to 16× smaller than the original image resolution rather than directly on full-resolution images. This compression reduces GPU memory usage by approximately 75–85% and accelerates both training and inference by nearly 10×. Compared to pixel-space diffusion methods, LDMs lower overall computational cost by one to two orders of magnitude, making them practical for deployment in medical imaging settings. In terms of inference speed, *PrivDiff-Net* can generate high-quality anonymized chest X-rays (256×256 resolution) in about 2.5–3.2 seconds per image on a single NVIDIA V100 (32 GB) GPU, fully compatible with clinical workflow requirements. Model training over 300 epochs using the Adam optimizer requires roughly 72 GPU hours, representing a balanced trade-off between computational cost and model performance typical of advanced deep learning systems in healthcare.

Discussion

In this work, we presented a novel latent diffusion-based framework for generating anonymous chest X-rays that maintain diagnostic utility while protecting patient privacy. Our approach demonstrates significant advantages over traditional anonymization methods by learning deformation fields that effectively obscure biometric identifiers while preserving clinically relevant features. Quantitative evaluation shows that the generated images retain high visual quality and radiological consistency, enabling secure visualization of model decisions without compromising patient data. These results establish a promising direction for privacy-preserving medical image analysis, particularly in scenarios requiring interpretable AI systems. Although performance is close to random, some residual biometric cues may persist in anonymized samples. Thus, clinical deployment should include safeguards like policy controls, encryption, and monitoring. Distinct anatomical traits—such as skeletal abnormalities, unique rib patterns, congenital variants, or surgical implants—can still serve as identifiable markers. Future work will assess cohorts with surgical hardware, skeletal deformities, and rare conditions to strengthen privacy protection, and extend the framework to other imaging modalities with enhanced privacy mechanisms. Our method represents an essential step toward deploying trustworthy AI systems in clinical settings where transparency and patient privacy are paramount.

Acknowledgment

The authors acknowledge the doctors who contributed to the clinical validation of this work, as well as the support provided by IndiaAI and Meta through the Srijan: Centre of Excellence for Generative AI.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC, Conference on Computer and Communications Security*, 308–318.

Abbasi, N.; and Smith, D. A. 2024. Cybersecurity in Healthcare: Securing Patient Health Information (PHI), HIPPA Compliance Framework and the Responsibilities of Healthcare Providers. *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online)*, 3(3): 278–287.

Abhisheka, B.; Biswas, S. K.; Purkayastha, B.; Das, D.; and Escargueil, A. 2024. Recent trend in medical imaging modalities and their applications in disease diagnosis: a review. *Multimedia Tools and Applications*, 83(14): 43035–43070.

Agrawal, T.; and Choudhary, P. 2022. Segmentation and classification on chest radiography: a systematic survey. *The Visual Computer*, 1–39.

Akhter, Y. 2025. Data-centric ai for chest x-ray analysis in resource-constrained settings. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25. ISBN 978-1-956792-06-5.

Akhter, Y.; Ranjan, R.; Singh, R.; and Vatsa, M. 2025. SHIELD: a self-supervised, silicosis-focused hierarchical imaging framework for occupational lung disease diagnosis. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25. ISBN 978-1-956792-06-5.

Akhter, Y.; Ranjan, R.; Singh, R.; Vatsa, M.; and Chaudhury, S. 2023. On AI-assisted pneumoconiosis detection from chest x-rays. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23. ISBN 978-1-956792-03-4.

Akhter, Y.; Singh, R.; and Vatsa, M. 2023. AI-based radiodiagnosis using chest X-rays: A review. *Frontiers in big data*, 6: 1120989.

Bu, Z.; Dong, J.; Long, Q.; and Su, W. J. 2020. Deep learning with gaussian differential privacy. *Harvard Data Science Review*, 2020(23): 10–1162.

Çallı, E.; Sogancioglu, E.; van Ginneken, B.; van Leeuwen, K. G.; and Murphy, K. 2021. Deep learning for chest X-ray analysis: A survey. *Medical Image Analysis*, 72: 102125.

Dong, J.; Roth, A.; and Su, W. J. 2019. Gaussian Differential Privacy. *CoRR*, abs/1905.02383.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of Third Theory of Cryptography Conference, TCC New York, NY, USA, March 4-7*, 265–284. Springer.

Fan, L. 2018. Image pixelization with differential privacy. In *IFIP Annual Conference on Data and Applications Security and Privacy*, 148–162. Springer.

Fan, L. 2019. Differential privacy for image publication. In *Theory and Practice of Differential Privacy (TPDP) Workshop*, volume 1, 6.

Glocker, B.; Jones, C.; Roschewitz, M.; and Winzeck, S. 2023. Risk of Bias in Chest Radiography Deep Learning Foundation Models. *Radiology: Artificial Intelligence*, 5(6): e230060.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269. Los Alamitos, CA, USA: IEEE Computer Society.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence,*, volume 33, 590–597.

Isibor, E. 2024. Regulation of Healthcare Data Security: Legal Obligations in a Digital Age. *Available at SSRN 4957244*.

K Pool, J.; Akhlaghpour, S.; Fatehi, F.; and Burton-Jones, A. 2019. Causes and Impacts of Personal Health Information (PHI) Breaches: A Scoping Review and Thematic Analysis. In *Twenty-Third pacific Asia conference on information systems, China July*.

Khalifa, M.; and Albadawy, M. 2024. AI in diagnostic imaging: Revolutionising accuracy and efficiency. *Computer Methods and Programs in Biomedicine Update*, 5: 100146.

Kim, B. N.; Dolz, J.; Desrosiers, C.; and Jodoin, P.-M. 2020. Privacy preserving for medical image analysis via non-linear deformation proxy. *arXiv preprint arXiv:2011.12835*.

Kim, B. N.; Dolz, J.; Jodoin, P.-M.; and Desrosiers, C. 2021. Privacy-Net: An Adversarial Approach for Identity-Obfuscated Segmentation of Medical Images. *IEEE Transactions on Medical Imaging*, 40(7): 1737–1749.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9*.

Nass, S.; Levit, L.; and Gostin, L. 2009. *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health through Research*. National Academies Press. ISBN 978-0-309-12499-7.

Packhäuser, K.; Gündel, S.; Münster, N.; Syben, C.; Christlein, V.; and Maier, A. 2022. Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data. *Scientific Reports*, 12(1): 14851.

Packhäuser, K.; Gündel, S.; Thamm, F.; Denzinger, F.; and Maier, A. 2023. Deep Learning-Based Anonymization of Chest Radiographs: A Utility-Preserving Measure for Patient Privacy. In Greenspan, H.; Madabhushi, A.; Mousavi, P.; Salcudean, S.; Duncan, J.; Syeda-Mahmood, T.; and Taylor, R., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, 262–272. Cham: Springer Nature Switzerland. ISBN 978-3-031-43898-1.

Paschali, M.; Chen, Z.; Blankemeier, L.; Varma, M.; Youssef, A.; Bluethgen, C.; Langlotz, C.; Gatidis, S.; and Chaudhari, A. 2025. Foundation Models in Radiology: What, How, Why, and Why not. *Radiology*, 314(2): e240597.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E. Z.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, 8024–8035.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85): 2825–2830. https://scikit-learn.org/stable/.

Popescu, A. B.; Taca, I. A.; Nita, C. I.; Vizitiu, A.; Demeter, R.; Suciu, C.; and Itu, L. M. 2021. Privacy preserving classification of EEG data using machine learning and homomorphic encryption. *Applied Sciences*, 11(16): 7360.

Qin, C.; Yao, D.; Shi, Y.; and Song, Z. 2018. Computer-aided detection in chest radiography based on Artificial Intelligence: a survey. *Biomedical engineering online*, 17(1): 1–23.

Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Ram, S.; and Bodduluri, S. 2023. Implementation of Artificial Intelligence–assisted chest X-ray interpretation: it is about time.

Smith-Bindman, R.; Kwan, M. L.; Marlow, E. C.; Theis, M. K.; Bolch, W.; Cheng, S. Y.; Bowles, E. J.; Duncan, J. R.; Greenlee, R. T.; Kushi, L. H.; et al. 2019. Trends in use of medical imaging in US health care systems and in Ontario, Canada, 2000-2016. *JAMA*, 322(9): 843–856.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *CoRR*, abs/1503.03585.

Thirukrishna, K. S. S. P. e. a., J.T. 2022. Survey on Diagnosing CORONA VIRUS from Radiography Chest X-ray Images Using Convolutional Neural Networks. *Wireless Pers Commun 124, 2261–2270 (2022).*

Vayena, E.; Blasimme, A.; and Cohen, I. G. 2018. Machine Learning in Medicine: Addressing Ethical Challenges. *PLoS medicine*, 15(11): e1002689.

Vizitiu, A.; Nita, C.-I.; Toev, R. M.; Suditu, T.; Suciu, C.; and Itu, L. M. 2021. Framework for privacy-preserving wearable health data analysis: proof-of-concept study for atrial fibrillation detection. *Applied Sciences*, 11(19): 9049.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2097–2106.

Zhou, S. K.; Greenspan, H.; Davatzikos, C.; Duncan, J. S.; Van Ginneken, B.; Madabhushi, A.; Prince, J. L.; Rueckert, D.; and Summers, R. M. 2021. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5): 820–838.

Ziller, A.; Usynin, D.; Braren, R.; Makowski, M.; Rueckert, D.; and Kaissis, G. 2021. Medical imaging deep learning with differential privacy. *Scientific Reports*, 11(1): 13524.