FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Privacy-preserving Case-based Explanations for Federated Learning Models

Inês de Sousa Cardoso

Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Luís Filipe Teixeira

Co-Supervisor: Prof. Wilson dos Santos Silva

September 30, 2025

Privacy-preserving Case-based Explanations for Federated Learning Models

Inês de Sousa Cardoso

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Rui Rodrigues
External examiner: Prof. Tiago Gonçalves
Supervisor: Prof. Luís Filipe Teixeira
Co-Supervisor: Prof. Wilson dos Santos Silva

September 30, 2025

Resumo

A crescente utilização de inteligência artificial (IA) no diagnóstico médico tem vindo a transformar a prática clínica, permitindo maior eficiência e precisão. Contudo, para que as pessoas confiem em sistemas baseados em IA, é fundamental garantir a interpretabilidade destes sistemas através de, por exemplo, explicações baseadas em casos, que justificam decisões recorrendo a exemplos semelhantes.

Tradicionalmente, este tipo de explicações era construído a partir de conjuntos de dados centralizados. No entanto, devido a novas leis e questões de privacidade, a partilha de dados médicos entre instituições torna-se limitada ou mesmo inviável, reduzindo a diversidade dos dados de treino e, consequentemente, comprometendo a capacidade de generalização e robustez dos modelos de IA.

A Aprendizagem Federada (*Federated Learning*, FL) surgiu como uma resposta a este desafio, permitindo treinar modelos de forma colaborativa sem necessidade de centralizar dados sensíveis. Ainda assim, a geração de explicações baseadas em casos em cenários de FL continua a ser difícil: como os dados dos pacientes não podem ser partilhados, o acesso direto a exemplos representativos fica condicionado. Uma alternativa é recorrer a imagens sintéticas geradas localmente nos clientes, mas esta solução também acarreta riscos de privacidade, dado que os modelos generativos podem memorizar e reproduzir informação identificável dos pacientes (PII, *Personally Identifiable Information*).

Neste contexto, esta dissertação explora como o machine unlearning (MU) – um conjunto de técnicas destinadas a remover a influência de dados específicos dos modelos – pode ser aplicado a modelos generativos baseados em difusão, por forma a remover PII memorizada de pacientes, e consequentemente, gerar imagens sintéticas anónimas mas clinicamente relevantes.

Esta dissertação, para além de ser o primeiro trabalho a abordar a anonimização de imagens médicas sintéticas através de machine unlearning, e um dos poucos a explorar o *data-point unlearning* em modelos de generativos, apresenta dois contributos principais: em primeiro lugar, propõe uma nova técnica de *unlearning* adaptada a modelos de difusão *Image-to-Image*; em segundo lugar, introduz uma *pipeline* para identificar de forma sistemática conjuntos de pacientes memorizados em modelos de difusão latente e reduzir a sua influência, garantindo simultaneamente a conformidade com as normas de privacidade e a preservação da utilidade clínica das imagens geradas.

Embora as abordagens propostas não tenham conseguido esquecer por completo todos os pacientes identificados, os resultados demonstram que o unlearning pode reduzir substancialmente os riscos de reidentificação em conjuntos de dados sintéticos, preservando em parte a fidelidade diagnóstica das explicações baseadas em casos. Estes resultados reforçam a necessidade de desenvolver novas técnicas de *data-point unlearning* adaptadas especificamente a modelos de difusão *image-to-image*, de modo a melhorar tanto a proteção da privacidade como a usabilidade clínica.

Em suma, esta dissertação contribui para aproximar as áreas da interpretabilidade em IA e da proteção de dados na síntese de imagens médicas, através da aplicação de *machine unlearning* em modelos de difusão. Tal permite abrir novas perspetivas para a investigação deste tipo de técnicas em modelos generativos e para a implementação de sistemas de aprendizagem federada em saúde, promovendo uma adoção de IA mais ética, segura e confiável em contextos clínicos.

Palavras-chave: Machine Learning, Inteligência Artificial Explicável, Interpretabilidade, Explicações Baseadas em Casos, Machine Learning com Preservação de Privacidade, Machine Unlearning, Modelos

de Difusão, Modelos Generativos, Visão por Computador, Imagem Médica, Anonimização de Dados, Direito ao Esquecimento, Regulamento Geral sobre a Proteção de Dados.

Abstract

The increasing integration of artificial intelligence (AI) into medical imaging has transformed diagnostic practice, enabling higher efficiency, accuracy, and the development of new clinical support tools. However, for people to trust AI-driven systems, interpretability must be ensured through approaches such as case-based explanations, which provide decision rationales by retrieving similar examples.

Traditionally, the models responsible for performing the diagnostics have been trained in a centralized setting. However, due to privacy concerns and regulations, sharing medical data between different institutions can be very difficult or even impossible, resulting in limited diversity of training data and affecting the generalization and robustness of AI models.

Federated Learning (FL) has emerged as a solution to this challenge, enabling collaborative model training without requiring the centralization of sensitive data. However, producing case-based explanations in FL settings remains particularly difficult. Although synthetic samples generated in the clients can be used as substitutes, they still carry the risk of leaking sensitive information since generative models tend to memorize and reproduce patients' personally identifiable information (PII).

This dissertation addresses these challenges by exploring how machine unlearning (MU), a family of techniques designed to erase the influence of specific training data, can be integrated with diffusion-based generative models to create anonymous but clinically meaningful synthetic examples. MU not only mitigates risks of memorization and data leakage in generative models but also provides a pathway to comply with legal requirements such as the European General Data Protection Regulation (GDPR) and its "Right to Be Forgotten" (RTBF).

Besides being the first work to address synthetic medical image anonymization with machine unlearning and one of the few to explore data-point unlearning in generative models, this dissertation makes two key contributions: First, it proposes a new unlearning technique tailored to Image-to-Image diffusion models. Second, it introduces a pipeline for systematically identifying sets of memorized patients within latent diffusion models and reducing their influence, thereby complying with privacy regulations while preserving the clinical utility of the generated images.

Although the proposed unlearning approaches could not completely forget all identified patients and led to a reduction in clinical utility, the results demonstrate that unlearning can substantially reduce the risks of patient re-identification in synthetic datasets. At the same time, these findings highlight the need for developing new unlearning techniques specifically tailored to Image-to-Image diffusion models, such as latent diffusion models, and also to data-point unlearning in diffusion models, to further improve both privacy protection and clinical usability.

Overall, this dissertation contributes to bridging the gap between explainable AI and data protection in medical imaging with the use of the growing field of MU. By proposing a method for generating anonymous case-based explanations through machine unlearning in diffusion models, it advances both the scientific understanding of unlearning in generative AI and the practical implementation of federated learning systems in healthcare, ultimately fostering more ethical, secure, and trustworthy AI adoption in clinical environments.

Keywords: Federated Learning, Explainable Artificial Intelligence, Case-based Explanations, Privacy-preserving Machine Learning, Machine Unlearning, Diffusion Models, Generative Models, Computer

Vision, Medical Imaging, Data Anonymization, Right to Be Forgotten, General Data Protection Regulation.

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework to achieve a better and more sustainable future for all. They include 17 goals to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice. This thesis aligns with four of these goals: SDG 3, SDG 8, SDG 9, and SDG 12.

SDG 3 Ensure healthy lives and promote well-being for all at all ages

SDG 8 Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation

SDG 12 Ensure sustainable consumption and production patterns

Table 1 presents an overview of the specific targets addressed, the contributions of this thesis, and performance indicators.

Table 1: Overview of the specific Sustainable Development Goals (SDGs), their targets, contributions of this thesis, and performance indicators.

SDG	**Target**	**Contribution**	**Performance Indicators and Metrics**
3	3.8	Developing privacy-preserving AI methods increases trust in medical systems, improving access to safe diagnostic tools.	Adoption of AI-based diagnostic support systems with privacy guarantees.
	3.d	Strengthening health data protection contributes to more resilient healthcare infrastructures.	Reduction of patient re-identification risks in synthetic datasets.
8	8.2	Promoting innovation in privacy-preserving AI fosters productivity and inclusive participation across institutions.	Number of institutions able to collaborate in federated learning without data sharing.
	8.3	Supporting secure and ethical digital health technologies contributes to sustainable economic growth in healthcare.	Growth of privacy-preserving AI research and adoption in medical imaging.
9	9.1	Strengthening digital infrastructures in healthcare by developing federated learning systems.	Deployment of federated AI frameworks in clinical research settings.
	9.5	Fostering research and innovation in machine unlearning and diffusion models.	Methods proposed and evaluated for privacy-preserving AI.
12	12.2	Reducing the need for full retraining to remove patient information through machine unlearning lowers computational costs.	Decrease in GPU hours required for compliance with RTBF requests.
	12.6	Encouraging sustainable AI practices by integrating privacy compliance mechanisms.	Efficiency of unlearning

Acknowledgements

This dissertation was developed in collaboration with the AIT4LIFE group at Utrecht University (UU). I am grateful to everyone I met there for making me feel welcome from the very first day, for the stimulating discussions, and, of course, for the Friday cake meetings and borrels at NKI, with the Trustworthy AI Group. A special thanks goes to Soufyan and Aniek, who were always willing to help and provided support with technical issues.

I also want to express my gratitude to my supervisors, Professor Wilson Silva and Professor Luís Teixeira, for their guidance, support, and valuable discussions throughout the development of this thesis.

To the friends I met in Utrecht, thank you for making the city feel like home and for sharing so many unforgettable moments with me. From lighthearted conversations to near-bike accidents in Dutch traffic, you made my stay both meaningful and fun.

Finally, I could not go without thanking my family, my boyfriend, and my long-time friends, who, even from afar, never stopped giving me their endless love, patience, and encouragement throughout this semester.

Inês de Sousa Cardoso

'If you are not failing every now and again, it's a sign you're not doing anything very innovative.'

Woody Allen

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Problem and Motivation . . . 2
- 1.3 Research Questions . . . 2
- 1.4 Main Contributions . . . 2
- 1.5 Dissertation Structure . . . 3

2 Background **4**
- 2.1 Overview . . . 4
- 2.2 Data Protection Regulations . . . 4
- 2.3 Federated Learning . . . 5
- 2.4 Case-Based Explanations . . . 5
  - 2.4.1 Case-Based Explanation Retrieval in Centralized Setting . . . 6
  - 2.4.2 Case-Based Explanation Retrieval in Federative Setting . . . 6
  - 2.4.3 Privacy-Aware Case-Based Explanations . . . 8
- 2.5 Machine Unlearning . . . 8
  - 2.5.1 Categories of Machine Unlearning . . . 8
  - 2.5.2 Common Machine Unlearning Baselines . . . 9
- 2.6 Diffusion Models . . . 10
  - 2.6.1 Diffusion Probabilistic Models (DPM) . . . 10
  - 2.6.2 Denoising Diffusion Implicit Models (DDIMs) . . . 11
  - 2.6.3 Latent Diffusion Models (LDMs) . . . 11
  - 2.6.4 Medical Image Generation with Diffusion Models . . . 12
- 2.7 Summary . . . 13

3 Literature Review: Machine Unlearning as a Data Removal Tool in Generative Models **14**
- 3.1 Overview . . . 14
- 3.2 Document Retrieval Methodology . . . 14
  - 3.2.1 Research Domain . . . 14
  - 3.2.2 Identification Phase . . . 16
  - 3.2.3 Screening Phase . . . 16
  - 3.2.4 Screening Phase Results . . . 17
  - 3.2.5 Eligibility Phase . . . 17
  - 3.2.6 Included Documents . . . 17
- 3.3 Machine Unlearning in Generative Models . . . 18
  - 3.3.1 Unlearning Scenarios . . . 18
  - 3.3.2 Type of Models used in Unlearning in Generative Models . . . 18
  - 3.3.3 Common Unlearning Evaluation Metrics . . . 19
- 3.4 Applications of Machine Unlearning in Federated Learning . . . 21
  - 3.4.1 Horizontal Federated Learning (HFL) . . . 21
  - 3.4.2 Vertical Federated Learning (VFL) . . . 21

3.5 Applications of Machine Unlearning in Medical Imaging . . . 22
3.5.1 Image Classification . . . 22
3.5.2 Image Harmonization and Bias Removal . . . 22
3.5.3 Image Generation and Reconstruction . . . 23
3.6 Related Work . . . 23
3.6.1 Saliency Unlearing (SalUn) . . . 23
3.6.2 KL–L_2 . . . 25
3.6.3 Subtracted Importance Sampled Scores (SISS) . . . 26
3.6.4 Erase to Enhance (MRI Unlearning) . . . 29
3.7 Summary . . . 31

4 Literature Review: Anonymization Techniques in Medical Imaging 32
4.1 Overview . . . 32
4.2 Document Retrieval Methodology . . . 32
4.3 Image Anonymization Techniques . . . 32
4.3.1 Classic methods . . . 33
4.3.2 Deep Learning Based Methods . . . 33
4.3.3 In-Model Methods . . . 34
4.3.4 Post-Model Methods . . . 35
4.4 Image De-Anonymization Techniques . . . 37
4.4.1 Re-Identification Attacks (Explanation Access) . . . 37
4.4.2 Model Exploitation Attacks (Black-Box and White-Box Access) . . . 38
4.4.3 Adversarial Queries (Black-Box Access) . . . 38
4.5 Summary . . . 38

5 Machine Unlearning Experimental Setup 40
5.1 Overview . . . 40
5.2 Machine Unlearning Scenario . . . 40
5.2.1 Difficulty Assessment of Unlearning Scenario . . . 40
5.3 Model Choice . . . 41
5.4 Dataset and Task Choice . . . 41
5.4.1 Data Preprocessing: Eliminating Mislabeled Samples . . . 42
5.5 Machine Unlearning Methods . . . 44
5.5.1 Gradient Ascent (GA) . . . 44
5.5.2 KL-Divergence based Unlearning (KL-away) . . . 45
5.5.3 Subtracted Importance Sampled Scores (SISS) . . . 46
5.5.4 Saliency Unlearning (SalUn) . . . 46
5.6 Evaluation Metrics . . . 48
5.7 Summary . . . 50

6 Generate anonymous case-based explanations with Machine Unlearning 51
6.1 Overview . . . 51
6.2 Method . . . 51
6.3 Experiments . . . 52
6.4 Results and Discussion . . . 53
6.5 Summary . . . 57

7 Improvement-Oriented Experiments 60
7.1 Overview . . . 60
7.2 Defining Different Forget and Remain Sets . . . 60
7.2.1 Method . . . 60
7.2.2 Results and Discussion . . . 61

7.3 Forgetting a Single Patient . . . 62
7.3.1 Method . . . 62
7.3.2 Results and Discussion . . . 62
7.4 Summary . . . 64

8 Conclusion and Future Work **65**

References **67**

A Appendix A **76**
A.1 Declarations . . . 76
A.2 Derivation of KL Simplification under Gaussian Assumption . . . 76

List of Figures

2.1 Pipeline for the DP-FL model [80]. 7
2.2 Visual representation of Localized Unlearning by Torkzadehmahani et al. [135]. 9
2.3 Illustrative image of Denoising Diffusion Probabilistic Model Process in *Denoising Diffusion Probabilistic Models* [81] 10
2.4 Illustration of the Medfusion model's structure by Müller-Franzes et al. [106]. 12

3.1 PRISMA flowchart. 15
3.2 Examples of images generated by the unlearned DDPM under different sparsity levels when forgetting the class "airplane". Lower sparsity (e.g., 10%) achieves forgetting but reduces generation quality, while higher sparsity (e.g., 90%) preserves quality but fails to forget. A balanced sparsity of 50% provides the best trade-off [39]. 25
3.3 Visual Results of KL–L_2 unlearning method by Li et al. [83]. The images in the retain set remain almost unaffected before and after unlearning, while the images in the Forget set are nearly noise after unlearning. 27
3.4 Visualization of celebrity unlearning with SISS method over fine-tuning steps [3]. . . . 30

4.1 Anonymization Pipeline proposed by Campos et al. [11]: A generative model generates synthetic samples based on the real training data. From these samples, those closely resembling patients in the training set are removed based on a patient retrieval and a patient verification model. 36

5.1 Comparison between a healthy X-ray (a) and one that presents Cardiomegaly (b). Both images were obtained from the MIMIC-CXR-JPG [73] dataset. 42
5.2 Examples of synthetic chest X-rays generated after applying unlearning techniques. Instead of producing postero-anterior (PA) views as intended, the model outputs resemble lateral views because of mislabeled samples in the MIMIC-CXR-JPG [73] dataset. . . . 43
5.3 Examples of found mislabeled chest X-ray images in the MIMIC-CXR-JPG [74] dataset. 43

6.1 Pipeline to allow deletion of memorized patients' PII. Synthetic images from the original model are first evaluated using a patient re-identification attack with a Retrieval and a Verification Network. If the similarity between a synthetic image and its Rank-1 real match exceeds 50%, the corresponding patient is marked as memorized, and all its images used in training are added to the Forget set. Otherwise, the synthetic image is added to the Remain set. After unlearning, new synthetic samples are generated with the same seeds and re-evaluated to assess anonymization and unlearning effectiveness. 52
6.2 Images produced by the mixup of the SalUn localization strategy with Gradient Ascent (On the top row) and by Gradient Ascent Unlearning Algorithm (On bottom row). 54
6.3 Comparison of anonymization quality. KL-Away effectively removes patient-specific features, while KL-Away + SalUn approaches fail to anonymize. 55
6.4 Examples of the images produced with the SISS unlearning technique, showcasing the image quality degradation. 56

6.5 Examples where all identified patients could be anonymized by the different unlearning methods. In cases where all KL-Away variants succeed, the resulting images appear visually similar, although KL-Away without a saliency mask introduces more noticeable differences compared to the original synthetic image. The figure also highlights quality and utility degradation with SISS: in the middle row, SISS produces an image with cardiomegaly, reducing clinical utility, whereas the other methods preserve the absence of cardiomegaly. 59

7.1 Redefined pipeline for detecting memorized patients' PII. Unlike the initial approach, which only considered the rank-1 retrieved image, this method evaluates similarity scores with the top-3 retrieved images. Patients associated with any retrieval above the similarity threshold are added to the Forget set, capturing cases where synthetic images embed features from multiple patients simultaneously. 61

7.2 Visual example of the successful example in the single-patient unlearning experiment. The top row shows the real image of the target patient; the middle row presents the most similar synthetic images (non-anonymous) before unlearning, and the bottom row shows the anonymized images obtained after unlearning with KL-Away. 64

List of Tables

1 Overview of the specific Sustainable Development Goals (SDGs), their targets, contributions of this thesis, and performance indicators. . . . vi

3.1 Distribution of Papers by Data Source and Query. . . . 16
3.2 Types of generative models used in unlearning studies and the corresponding papers. . . . 19
3.3 Summary of state-of-the-art HFL federated machine unlearning methods. . . . 22
3.4 Comparison of unlearning methods most similar to this dissertation goals. . . . 23

4.1 Overview of de-anonymization attacks proposed and used in the literature, grouped by attack type and required access level. . . . 37

6.1 Quantitative evaluation of the images generated using the Medfusion model before unlearning. . . . 53
6.2 Utility evaluation of the images generated using the Medfusion model before unlearning. 53
6.3 Quantitative evaluation of patient retrieval model. . . . 54
6.4 Quantitative evaluation of patient verification model. . . . 54
6.5 Evaluation of the datasets' images by the different unlearning metrics across the various metrics. . . . 58

7.1 Evaluation of the datasets' images by the different unlearning metrics across the various metrics. . . . 63

List of Acronyms

AI	Artificial Intelligence
DDIM	Denoising Diffusion Implicit Model
DDPM	Denoising Diffusion Probabilistic Model
DL	Deep Learning
DM	Diffusion Model
DPM	Diffusion Probabilistic Model
DP-FL	Differentially Private Federated Learning
ELBO	Evidence Lower Bound
FedAvg	Federated Averaging
FedSGD	Federated Stochastic Gradient Descent
FID	Fréchet Inception Distance
FL	Federated Learning
GA	Gradient Ascent
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
HFL	Horizontal Federated Learning
I2I	Image-to-Image
KL	Kullback–Leibler (Divergence)
LDM	Latent Diffusion Model
MAE	Masked Autoencoder
MIA	Membership Inference Attack
MU	Machine Unlearning
NLL	Negative Log-Likelihood
PII	Personally Identifiable Information
RTBF	Right To Be Forgotten
SalUn	Saliency Unlearning
SISS	Subtracted Importance Sampled Scores
SSIM	Structural Similarity Index Measure
UA	Unlearning Accuracy
U-Net	U-Net (convolutional) architecture
VAE	Variational Autoencoder
VFL	Vertical Federated Learning
ViT	Vision Transformer
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

1.1 Context

The integration of artificial intelligence (AI) into medical image analysis has transformed diagnostics, improving efficiency and accuracy while leading to significant progress in computer-aided diagnosis systems. As these systems become more complex, the need for transparency and interpretability grows. Explainable AI (XAI) has emerged as a field dedicated to making AI model decisions comprehensible to human users [5]. Among the various XAI techniques, case-based explanations have gained relevance in clinical domains, such as radiology, as they provide interpretable, patient-specific insights that align closely with the clinical reasoning process.

Traditionally, the models responsible for performing these diagnostics were trained in a centralized setting. However, due to privacy concerns and regulations, sharing medical data between different institutions can be very difficult or even impossible, resulting in limited diversity of training data and affecting the generalization and robustness of AI models [76].

To address this challenge, Federated Learning (FL) has been proposed as a solution, enabling collaborative model training without centralizing data. Instead, institutions train models locally and share only model updates, preserving data privacy. However, although FL addresses data-sharing constraints, it introduces new challenges for explainability, particularly in generating case-based explanations when institutions lack access to a global dataset.

In a FL setting, case-based explanations require a decentralized approach. Since each institution holds only a subset of the data and cannot directly share patient records, one proposed solution is for the jointly trained retrieval model to rely on synthetic catalogs, locally generated by the participating institutions and shared, instead of real patient data. These synthetic datasets are typically produced using generative models, such as Generative Adversarial Networks (GANs) or Diffusion Models (DMs), allowing a distributed explanation system without direct data exchange. However, recent studies reveal that these deep learning models can memorize training samples, potentially leaking private data in generated outputs [14]. This poses a significant privacy risk, both in the centralized and federated learning environment, as synthetic images can violate legal and ethical standards and privacy regulations, such as the European General Data Protection Regulation (GDPR) [25] or the Health Insurance Portability and Accountability Act (HIPAA) [139] if they contain recoverable patient information.

Furthermore, in addition to anonymization, these privacy regulations, such as GDPR [25], mandate that systems must guarantee the "Right to Be Forgotten" (RTBF), which extends to medical settings and

their patients. This means that, in addition to generating privacy-preserving case-based explanations, healthcare AI systems must also provide mechanisms to completely remove the influence of a patient's data when requested.

1.2 Problem and Motivation

The increasing reliance on generative models to synthesize medical images introduces privacy concerns. Despite efforts to generate anonymized datasets, studies [14, 56, 99] have shown that generative models, such as diffusion models, can memorize training data and potentially leak sensitive information [14]. At the same time, privacy regulations such as GDPR [25] mandate mechanisms such as the "Right to be Forgotten", requiring the complete removal of an individual's data from any learned system or database. Recent studies have shown that both the memorization risks in generative models and the regulatory requirement for data deletion can be addressed through machine unlearning techniques, as these methods allow the selective removal of specific data influences from trained machine learning models without requiring complete retraining [10].

This dissertation addresses the feasibility of applying machine unlearning techniques to generative models, particularly diffusion models, to produce case-based explanations in medical imaging that enhance interpretability while preserving patient privacy. This means determining whether unlearning methods can reliably erase memorized patient information from diffusion models while maintaining their performance and clinical utility after data removal.

1.3 Research Questions

Formally, this dissertation seeks to answer the following research questions:

RQ1. How can a generative model unlearn specific data to comply with privacy regulations?

RQ2. Can unlearning enable diffusion models to produce fully anonymous synthetic images that remain clinically meaningful?

RQ3. Can a diffusion model unlearn information contributed by a specific patient or set of patients?

RQ4. How can we define and identify the set of patients memorized by a generative model to determine which patients must be forgotten, to allow complete anonymization?

1.4 Main Contributions

The main contributions of this dissertation are the following:

- We investigate and adapt different machine unlearning techniques, evaluating their effectiveness and utility to produce anonymous medical images.
- We propose the first systematic attempt at applying machine unlearning to diffusion models for patient-level forgetting in medical imaging, addressing the GDPR's anonymization and RTBF requirements.

- We introduce a new unlearning technique tailored to diffusion models, which allows the simultaneous removal of multiple patient identities while minimizing the loss of generative utility.
- We demonstrate that machine unlearning can reduce re-identification risks in synthetic datasets, highlighting both the potential and the current limitations of unlearning for clinical usability.
- We conduct one of the first experimental studies on machine unlearning for medical image anonymization, highlighting future directions.

1.5 Dissertation Structure

In addition to this introductory chapter, the dissertation is organized into seven other chapters. Chapter 2 presents the background and foundational concepts necessary to understand this work, including data protection regulations, federated learning, case-based explanations, machine unlearning, and diffusion models in medical imaging. Chapter 3 provides a systematic literature review on the use of machine unlearning as a data removal tool in generative models, analyzing unlearning scenarios, methodologies, and evaluation metrics. Chapter 4 reviews medical image anonymization techniques, with a focus on both traditional and deep learning-based approaches, as well as de-anonymization attacks. Chapter 5 introduces the experimental setting, detailing the dataset, tasks, model choices, unlearning methods, and evaluation metrics. Chapter 6 describes the proposed pipeline for generating anonymous case-based explanations with machine unlearning and reports the main experimental results. Chapter 7 presents additional experiments oriented towards understanding and improving the obtained results in the previous chapter, such as a different definition of the forget and Remain sets and experimenting with single-patient forgetting. Finally, Chapter 8 concludes the dissertation by summarizing the contributions, discussing limitations, and suggesting directions for future work.

Chapter 2

Background

2.1 Overview

This chapter presents the background knowledge required to understand the contributions of this dissertation. It begins in Section 2.2 with a review of data protection regulations, focusing on the General Data Protection Regulation (GDPR) and on the Health Insurance Portability and Accountability Act (HIPAA), introducing the key concepts of pseudonymization, anonymization, and the Right to Be Forgotten, which motivate the need for unlearning in medical imaging. Section 2.3 then presents the foundations of Federated Learning (FL), a paradigm that enables collaborative training across institutions without sharing raw data, making it especially relevant for privacy-preserving case-based explanations. Following this, Section 2.4 discusses case-based explanations, an interpretability approach that justifies model predictions through examples. The section introduces categories of case-based methods, their role in centralized and federated settings, and the associated privacy concerns. Next, Section 2.5 introduces the field of Machine Unlearning, presenting its categories, challenges, and commonly used baselines. This section highlights how unlearning enables removing specific training data influences from machine learning models, supporting compliance with privacy regulations. Section 2.6 presents an overview of current diffusion models, focusing on latent diffusion and its applications in medical imaging. This includes the Medfusion architecture, which serves as the generative backbone in the experiments conducted in this dissertation. The chapter concludes in Section 2.7 with a summary and closing remarks.

2.2 Data Protection Regulations

With the increase of AI use in many fields, data protection laws have emerged to guarantee the right of the citizen to privacy, such as the European General Data Protection Regulation (GDPR) [25] and the Health Insurance Portability and Accountability Act (HIPAA) [139], which defines how data can be used, shared, and stored.

From GDPR, there are three essential concepts to understand: Pseudonymization, Anonymization, and the Right To Be forgotten (RTBF).

Pseudonymization, as defined in Article 4(5) of GDPR, refers to the processing of personal data in a way that personal data can no longer be attributed or linked to a specific data subject without the use of additional information kept separately.

In contrast, anonymization, while not explicitly defined in GDPR, is referenced in Recital 26 as the processing of personal data in such a way that re-identification is no longer possible by any means likely to be used by attackers or data processors, even when considering additional contextual data. Under HIPAA, there is a similar concept in the form of "de-identified health information", which is data from which all personal identifiers have been removed using either expert determination or a strict "safe harbor" method that excludes 18 specific identifiers.

However, in medical imaging, achieving true anonymization is particularly difficult. For example, diagnostic scans such as MRI or PET often contain intrinsic biometric features, which can be used to identify individuals even when metadata is removed. Studies, such as Packhäuser et al. [112] or Montenegro et al. [104], have demonstrated the feasibility of re-identifying individuals based solely on medical scans, highlighting the inadequacy of conventional anonymization techniques in high-stakes domains, such as the medical image one.

This is especially relevant when sharing or visualizing case-based explanations derived from AI models. Although these examples are critical for interpretability, they should not compromise patient confidentiality. Anonymization methods in this context must go beyond simple data masking and consider adversarial scenarios, such as patient re-identification attacks.

2.3 Federated Learning

Federated learning (FL) is an approach that allows collaborative model training among multiple data owners without requiring them to share their raw data [98]. In contrast to traditional centralized learning, where data from all sources must be aggregated into a single location, FL allows each client to train the model locally using their private data. The locally trained updates are then sent to a central server, where they are aggregated to update a global model.

The simplest approach in federated learning is Federated Stochastic Gradient Descent (FedSGD) [98]. In FedSGD, each client performs only a single step of local gradient descent before sending its updates to the central server for aggregation. While this method is conceptually straightforward, it requires frequent communication between clients and the server, leading to high communication overhead and often slow convergence.

To address these limitations, Federated Averaging (FedAvg) [98] was introduced and has since become the most widely used federated learning algorithm. In FedAvg, each client performs multiple steps of local training, often across several epochs, before sending its model weight updates. The server then aggregates these updates using a weighted average based on each client's local dataset size. By reducing synchronization frequency, FedAvg offers a better trade-off between computation and communication, making it the standard choice in many federated learning applications.

2.4 Case-Based Explanations

Case-based explanations are a class of interpretability methods that aim to justify a model's decision by retrieving examples from the training data. This technique mimics human reasoning, where decisions are often made by comparing new situations with ones that have been encountered before. Therefore,

case-based explanations provide straightforward and human-understandable rationales, particularly in complicated fields such as medical imaging.

In the field of medical imaging, case-based explanations offer several important benefits. Firstly, they provide clinical alignment by presenting reasoning in a familiar form to healthcare professionals, who often rely on case comparisons when making diagnostic and treatment decisions. Second, they promote transparency, allowing verification of whether the reasoning of a model is grounded in relevant clinical patterns or is instead influenced by false correlations. Finally, by offering concrete and relatable examples, it increases clinicians' confidence and trust in AI-assisted decision-making.

Case-based explanations can be grouped into four main categories:

- **Prototypes**, which are representative samples that provide an overview of the dataset or a specific class.
- **Critics**, which highlight data points that deviate from typical model predictions, providing insight into areas where prototypes are insufficient. Critics can be thought of as points that are poorly represented by Prototypes [114].
- **Semi-factuals**, which show how outcomes remain consistent despite potential changes ("even if" scenarios). One can think of a semi-factual as the point that is furthest from the query while still maintaining the same label [114].
- **Counter-factuals**, which illustrate what could have changed the decision of the model ("if only" scenarios). One can think of a counter-factual as the point closest to the query yielding a different label [114].

2.4.1 Case-Based Explanation Retrieval in Centralized Setting

In centralized medical AI systems, all training and reference data are stored in a single location, allowing the direct retrieval of similar cases from a global dataset. While counterfactual explanations can be generated from a single reference image, similarity-based explanations require searching the entire data set for comparable examples. Traditional retrieval methods often rely on global similarity measures such as Euclidean distance in the image space or feature space [77]. However, in medical imaging, diagnostically relevant patterns are often small and localized, making such metrics suboptimal.

To address this, task-specific retrieval methods have been proposed. One example is the study of Holzinger et al. [61], which focuses the similarity computation on diagnostically relevant image regions identified via Deep Taylor decomposition [103]. This approach has been shown to improve retrieval quality compared to global similarity measures.

2.4.2 Case-Based Explanation Retrieval in Federative Setting

In a federated learning setting, direct retrieval from a global dataset is impossible since the raw data remain at local institutions. However, several strategies enable case-based explanation retrieval in FL:

- **Local retrieval with aggregated results:** Each institution searches its local dataset and shares only the embeddings of the top results or anonymized representations, preserving privacy by never transmitting raw images [58].

- **Federated prototype learning**: Methods such as FedProto by Tan et al. [129] and AproMFL by Gai et al. [42] allow clients to learn and share abstract class prototypes, which can act as representative cases without exposing patient data.
- **Synthetic cross-site sharing:** Institutions generate synthetic but representative examples for their top retrievals using generative models, then share these examples for explanatory purposes. A notable example is the **DP-FL** model by Latorre et al. [80], a FedAvg-based framework for detecting pleural effusion from chest radiographs (X-rays). In DP-FL, hospitals collaboratively train a model and retrieve explanatory cases from a synthetic dataset jointly generated by all participating sites. While effective, this approach suffers from the limitation that insufficiently anonymized synthetic data may still leak private and sensitive medical information. An image representing this model is shown in Figure 2.1.

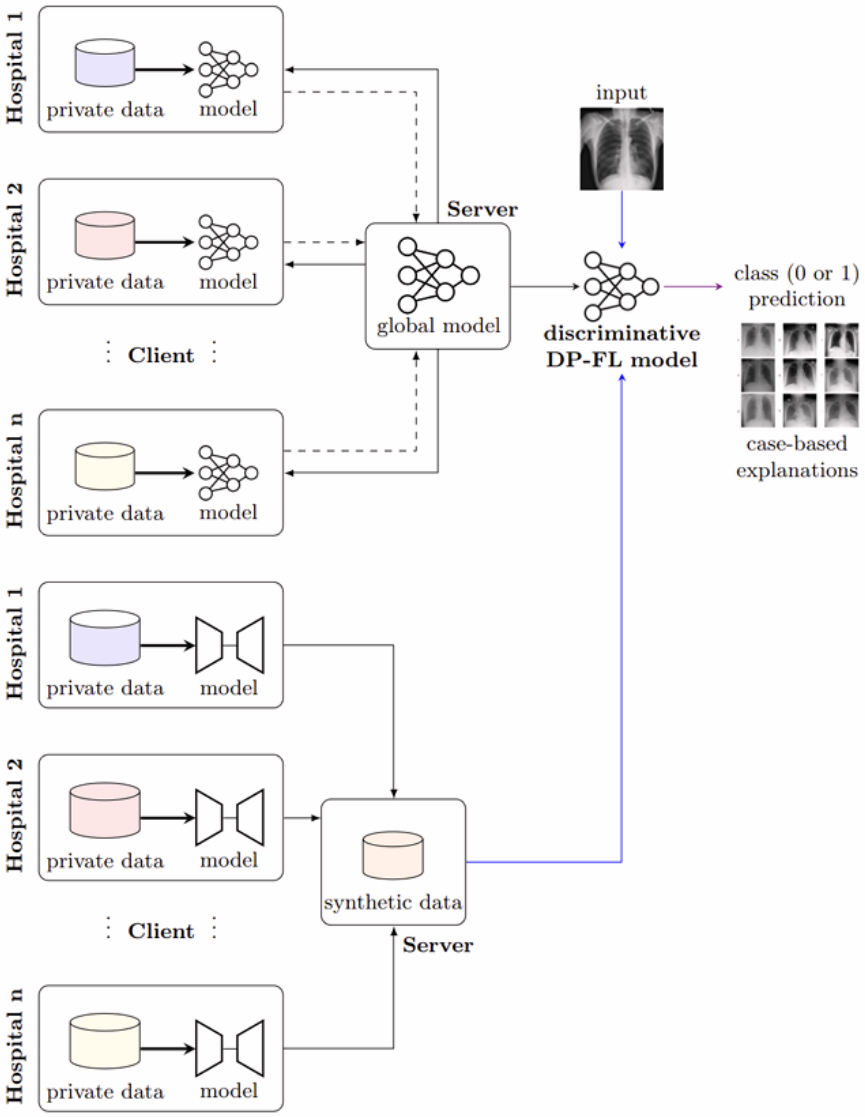


Figure 2.1: Pipeline for the DP-FL model [80].

- **Secure Multi-Party Computation (SMPC)**: SMPC enables institutions to jointly compute similarity scores or rankings without revealing their underlying data [62, 90].

2.4.3 Privacy-Aware Case-Based Explanations

The deployment of case-based explanation methods in healthcare raises significant privacy concerns. Retrieved examples are often derived from real patient data and may still contain identifiable information, even when explicit identifiers have been removed. Consequently, sharing case-based explanations, even within a federated learning framework, can still result in violations of data protection laws if the retrieved images are not properly anonymized. Mitigating this risk requires the adoption of robust medical imaging anonymization techniques, which allow the secure sharing of explanatory examples without compromising patient confidentiality.

2.5 Machine Unlearning

Machine Unlearning (MU) refers to a set of techniques that aim to remove the influence of specific data points from a trained machine learning model as if those data points had never been included in training. This concept has gained increasing attention in the context of privacy regulations such as the General Data Protection Regulation (GDPR) [25], which grants individuals the "Right To Be Forgotten".

Applying MU in modern deep learning systems is particularly challenging, especially in sensitive domains, such as medical AI, where memorization of identifiable patient information can lead to the leakage of patient privacy information. Kaissis et al. [76] highlight this risk in federated learning settings, noting that memorized patient data could be extracted through model attacks such as model inversion or membership inference.

To formalize the unlearning process, there are typically two disjoint subsets to be defined: the *Forget set* D_f and the *Remain set* D_r. The Forget set contains the samples whose influence must be erased from the model, while the Remain set contains all other data that should continue to inform the model's predictions. An effective unlearning method must therefore ensure that the model behaves as if D_f had never been seen, while preserving performance and generalization on D_r.

In this dissertation, machine unlearning is investigated as a mechanism to ensure that generative models, such as diffusion models used to create anonymized case-based explanations, align with both ethical responsibilities and legal mandates for data privacy and control.

2.5.1 Categories of Machine Unlearning

Machine unlearning approaches can be categorized into two main families: exact (or certified) MU and approximate MU.

The exact MU consists of removing the target data from the dataset and retraining the whole model. Although this is an intuitive and understandable approach, it requires a significant computational cost, consuming considerable GPU hours, which not only is financially expensive but also carries a substantial carbon footprint [119]. The approximate MU does not require retraining the model and is focused on developing methods with provable error guarantees or unlearning certifications [39]. These methods attempt to approximate the behavior of a re-trained model by applying techniques such as gradient reversal or influence function updates.

There are two types of approximate MU: Full-parameter Unlearning and Localized Unlearning.

Full-parameter unlearning methods involve modifying all model parameters to reduce or remove the influence of specific data. Localized unlearning focuses on operating on a small identified subset of parameters and is composed of two steps: the first is the localization strategy that identifies the subset of parameters in the model that are most relevant to the data to be unlearned, and the second is the unlearning algorithm that modifies the determined parameters to eliminate the influence of the target data while trying to maintain the model's overall performance on other tasks. A visual representation of localized unlearning is provided in Figure 2.2.

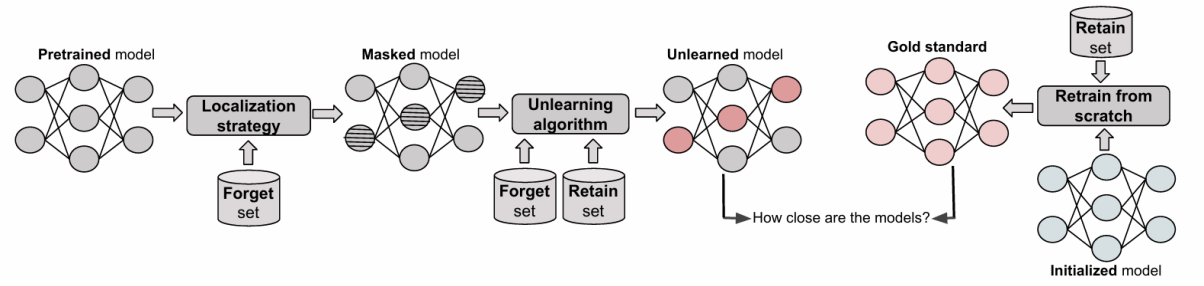


Figure 2.2: Visual representation of Localized Unlearning by Torkzadehmahani et al. [135].

2.5.2 Common Machine Unlearning Baselines

Machine Unlearning has some established baselines that provide simple reference points, widely adopted in the literature for comparison with new methods. These common baselines can be grouped into three categories: retraining-based methods, which serve as the gold standard but are computationally costly; label manipulation methods, which alter the labels of forget samples to disrupt learned associations; and loss-based methods, which directly modify the training objective to erase information from the Forget set. The main representatives of each category are summarized below.

- **Retraining-based Methods:**
  - **Retraining from Scratch (RFS):** The model is completely retrained using only the remaining data. This is the gold standard for unlearning, but computationally expensive [13].
  - **Fine-Tuning (FT):** The model is fine-tuned on the remaining data after removing the Forget set. Simple but may not fully eliminate memorized patterns [47].
- **Label Manipulation Methods:**
  - **Random Relabeling (RL):** Labels of the Forget set D_f are randomly reassigned to incorrect classes during training. For a forget sample (x_f, y_f), the new label becomes $\tilde{y}_f \sim \text{Uniform}(\mathscr{Y} \setminus \{y_f\})$, forcing the model to dissociate forgotten data from original labels [130].
  - **Noisy Labeling (NL):** Noise is added to Forget set labels: $\tilde{y}_f = y_f + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. This disrupts the input-output mapping, degrading the model's ability to reconstruct forgotten data [130].
- **Loss-based Methods:**

– **Confusion Loss:** Maximizes prediction entropy on forget data: $\mathscr{L}_{\text{conf}} = -\sum_{x_f \in D_f} H(p(y|x_f))$, where $H(\cdot)$ is the entropy function. This encourages uncertain predictions, reducing model confidence on forgotten data [7].

– **Gradient Ascent (GA):** Performs gradient ascent on the standard loss over forget data: $\mathscr{L}_{\text{GA}} = -\mathscr{L}_{\text{standard}}(D_f)$. This directly pushes model parameters away from forgotten data influence but may cause training instability, and often leads to overforgetting [72].

2.6 Diffusion Models

Generative AI has become a central research area in modern machine learning, enabling the synthesis of realistic data samples from learned probability distributions. While earlier approaches such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) have achieved remarkable results, recent developments in diffusion models have established a new benchmark for generative tasks such as image generation or super-resolution [27].

2.6.1 Diffusion Probabilistic Models (DPM)

Among various diffusion-based approaches, Diffusion Probabilistic Models (DPMs) provide the theoretical backbone for this family. These models work by learning to reverse a gradual noise corruption process, enabling them to generate high-quality and high-fidelity samples from complex data distributions. This process is illustrated in Figure 2.3.

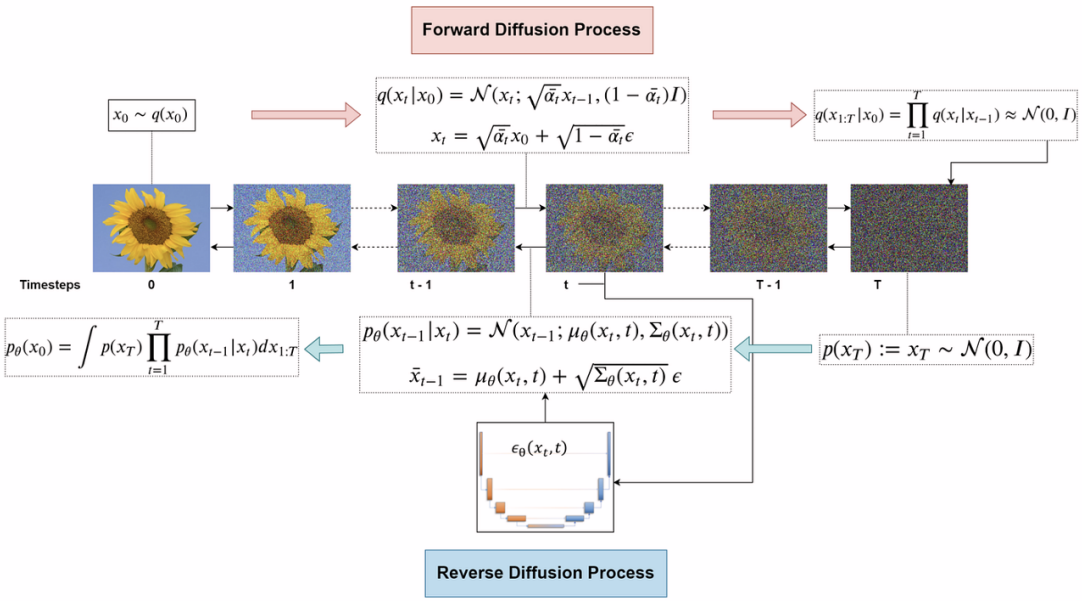


Figure 2.3: Illustrative image of Denoising Diffusion Probabilistic Model Process in *Denoising Diffusion Probabilistic Models* [81]

2.6.1.1 Forward diffusion process

The forward process is defined by a Markov chain that gradually adds Gaussian noise over a series of T time steps. At each step t, a small amount of noise is added to the data from the previous step $t-1$. This

process can be mathematically described as:

$$q(\mathbf{x}t \mid \mathbf{x}t-1) = \mathcal{N}(\mathbf{x}t; \sqrt{1-\beta_t}\mathbf{x}t-1, \beta_t\mathbf{I}) \tag{2.1}$$

Where β_t is a small positive constant that controls the noise schedule. A key insight from Ho et al. [59] is that this entire sequential process can be expressed as a single, direct sampling step for any arbitrary timestep t:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2.2}$$

where $\bar{\alpha}t = \prod i = 1^t(1-\beta_i)$. This reparameterization enables efficient training, as any timestep t can be sampled independently during training. Modern research focuses on optimizing the noise schedule β_t to improve both generation quality and sampling efficiency.

2.6.1.2 Reverse Diffusion Process

The reverse process employs a neural network, typically based on the U-Net architecture, to learn the inverse of the forward diffusion process. This network is trained to predict the noise component added at each step, enabling iterative denoising from pure Gaussian noise back to clean data. The reverse process can be modeled as:

$$p_\theta(\mathbf{x}t-1 \mid \mathbf{x}t) = \mathcal{N}(\mathbf{x}t-1; \mu\theta(\mathbf{x}t,t), \Sigma\theta(\mathbf{x}_t,t)) \tag{2.3}$$

The model's training objective is to minimize the difference between the true reverse distribution and the predicted one, which is typically formulated as a noise prediction task. The U-Net architecture proposed by Ronneberger et al. [116] has become the standard choice due to its effectiveness in handling multi-scale features through skip connections, which is crucial for the denoising task across different levels of corruption.

2.6.2 Denoising Diffusion Implicit Models (DDIMs)

A significant limitation of DDPMs is their slow sampling process, which requires a lot of denoising steps to produce high-quality samples. This bottleneck was addressed by Song et al. [125], which introduced this new family of models: Denoising Diffusion Implicit Model (DDIMs). The key innovation of DDIMs is in treating the reverse process as a non-Markovian process, meaning that the sampling process does not need to follow the exact reverse of the forward chain. Instead of stochastically sampling a distribution at each step, DDIMs provide a deterministic mapping from noise to data, creating "shortcuts" through the denoising trajectory. This approach enables the same high-quality generation but with significantly less number of denoising steps, making diffusion models a more practical choice for generative tasks.

2.6.3 Latent Diffusion Models (LDMs)

Latent Diffusion Models (LDMs) are a class of generative models that combine the strengths of diffusion-based generative processes with latent-space representation learning. The central idea is to first compress high-dimensional data, such as images, into a lower-dimensional latent space using an autoencoder, thereby reducing computational cost during the generative process.

In the latent space, a denoising diffusion probabilistic model (DDPM), or one of its variants, is trained to iteratively remove noise from a corrupted latent vector, effectively learning the reverse of a

predefined forward diffusion process. By operating in the latent domain rather than pixel space, LDMs achieve significant efficiency gains while maintaining high visual fidelity.

Formally, given an input x, the encoder E_θ maps the image into a latent code

$$z = E_\theta(x).$$

The diffusion process is then applied to z, progressively adding Gaussian noise over a fixed number of timesteps. The reverse process, parameterized by a neural network (often a U-Net), learns to denoise and reconstruct z step-by-step. The final output is decoded back into the image space via a decoder D_ϕ, producing:

$$\hat{x} = D_\phi(z).$$

This design enables scalable, high-quality generation while allowing the incorporation of conditioning information (such as text, labels, or medical metadata) directly into the diffusion process.

2.6.4 Medical Image Generation with Diffusion Models

Recently, diffusion models have been applied to the medical domain for multiple tasks and have been demonstrated to be a superior alternative to GANs for image synthesis in the medical domain [106]. A remarkable solution that has been used in medical applications, including in the production of case-based explanations [11, 80], is the Medfusion model proposed by Müller-Franzes et al. [106].

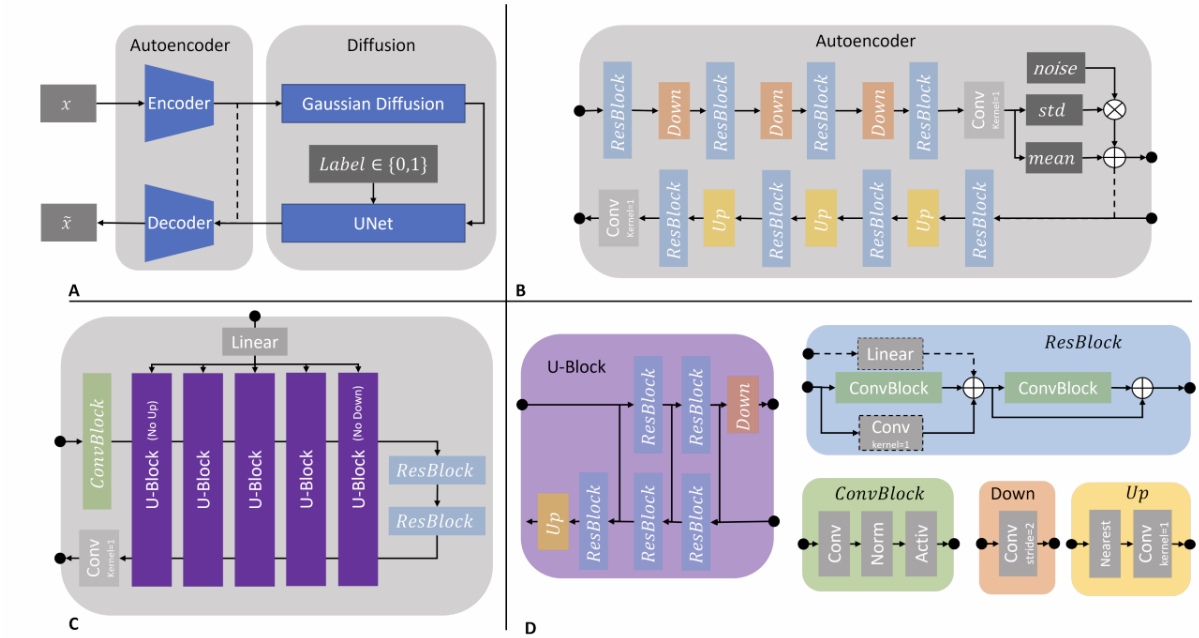


Figure 2.4: Illustration of the Medfusion model's structure by Müller-Franzes et al. [106].

The Medfusion model adapts the latent diffusion framework for medical image generation by combining a convolutional autoencoder with a conditional diffusion process operating in the latent space. The autoencoder encodes an input image into a lower-dimensional representation through a series of residual and downsampling blocks and reconstructs it using the corresponding upsampling and residual blocks. At the bottleneck, the encoder outputs mean and standard deviation maps, which are combined

with Gaussian noise to produce latent variables via the reparameterization trick. A U-Net processes these noisy latent variables during the reverse diffusion process, integrating both timestep and class-label embeddings to enable conditional generation. The network architecture is composed of residual blocks with convolution, normalization, and Swish activation, as well as up- and down-sampling layers that adjust spatial resolution. Operating entirely in latent space allows the model to synthesize high-quality medical images efficiently while conditioning on relevant clinical labels.

A key modification compared to the original Latent Diffusion Model (LDM) lies in the variational autoencoder (VAE) design. While the original LDM employed 4 latent channels, Medfusion uses 8 channels, which were found to significantly reduce visual artifacts and improve reconstruction fidelity across diverse medical imaging domains.

2.7 Summary

This chapter established the theoretical background for this dissertation by connecting privacy regulations, case-based explanations, unlearning, and medical image generation with synthetic models.

The discussion began with data protection frameworks such as GDPR and HIPAA, which emphasize the importance of anonymization and motivate the adoption of federated learning as a paradigm for collaborative yet privacy-preserving training. Case-based explanations were then presented as a powerful interpretability tool. However, their use, even in federated settings, raises privacy concerns, as patient-identifiable information (PII) can still be revealed in the images if not properly anonymized. To address these challenges, machine unlearning was introduced as a mechanism to selectively remove the influence of specific data points, for example, to eliminate memorized patient information or to comply with requests aligned with the “Right To Be Forgotten” (RTBF). Finally, diffusion models were described as the generative backbone of this work, with particular emphasis on latent diffusion and the MedFusion model, which provide efficient and effective means for generating high-quality synthetic medical images.

Chapter 3

Literature Review: Machine Unlearning as a Data Removal Tool in Generative Models

3.1 Overview

This chapter reviews the existing literature on machine unlearning techniques with a particular focus on their application as tools to remove data from generative models. Section 3.2 details the PRISMA methodology employed to identify and select the most relevant documents. The subsequent sections analyze these works to provide insight into the current trends in this area. Section 3.3 discusses the use of machine unlearning in generative models by presenting common unlearning scenarios, types of generative models in which unlearning has been employed, and the evaluation metrics typically used. Section 3.4 explores applications of machine unlearning within federated learning frameworks, while Section 3.5 examines its role in the domain of medical imaging. Finally, Section 3.6 highlights prior works that are most aligned with the goals of this dissertation, providing foundational knowledge and methodological context for the experiments conducted in later chapters. This chapter ends with Section 3.7 with a summary and closing remarks on the content presented.

3.2 Document Retrieval Methodology

To retrieve the documents for this investigation, a systematic strategy was adopted employing the PRISMA framework *(Preferred Reporting Items for Systematic Reviews and Meta-Analyses)* [113] as it provides a rigorous methodology to ensure the transparent selection, inclusion, and evaluation of relevant studies. Figure 3.1 shows a visual representation of all the steps performed, which will be explained further in the following subsections.

3.2.1 Research Domain

This research aims to generate privacy-preserving case-based explanations for diffusion models by applying techniques of machine unlearning. Consequently, the documents considered for review should

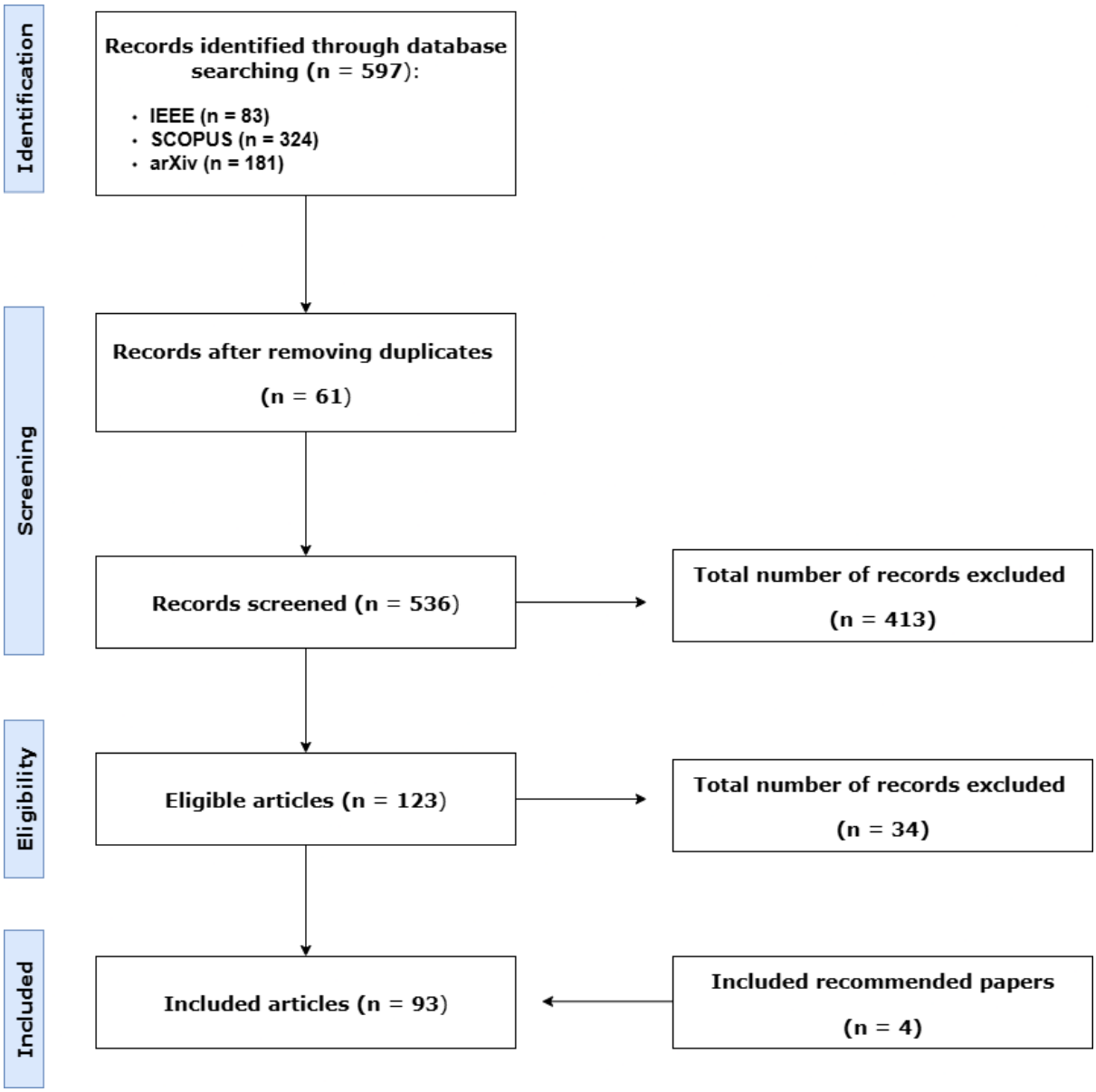


Figure 3.1: PRISMA flowchart.

align with topics related to machine unlearning as a tool to remove specific content from a trained generative model.

3.2.2 Identification Phase

To ensure that the documents were obtained in a systematic and replicable manner, identical queries were applied across all selected data sources.

Data Sources Three digital libraries were chosen as the primary sources for this research: IEEE Xplore [67], Scopus [36], and arXiv [24]. For arXiv, the queries were applied to all fields; in IEEE Xplore, it was restricted to the abstract; and in Scopus, it was applied to the abstract, title, and keywords.

Search Queries To retrieve the most important documents in the context of this investigation, the following queries were performed:

> **Q1. Search about Unlearning Mechanisms in Generative Models.**
> unlearn* AND (genera* OR diffusion OR "deep learn*") AND imag*
>
> **Q2. Exploring unlearning mechanisms in federated learning for image generation tasks.**
> unlearn* AND (federat* OR distributed OR decentralized) AND ("image generation" OR genera* OR image)
>
> **Q3. Exploring unlearning mechanisms in medical imaging.**
> unlearn* AND (medical OR health* OR clinical) AND imag*

Search Queries Results Using the set of search queries presented, a total of 597 documents were retrieved in all digital libraries. Table 3.1 presents the number of results per query on each data source.

Table 3.1: Distribution of Papers by Data Source and Query.

Source	Q1	Q2	Q3
Scopus	243	47	43
arXiv	162	4	15
IEEE Xplore	68	4	11

3.2.3 Screening Phase

The goal of the Screening Phase is to eliminate duplicate records and those that do not respect a set of rules related to their metadata, title, and abstract. The chosen criteria for a document to be considered are defined as follows:

- The document must be written in English.
- The publication date must be 2019 or later.

- ArXiv documents published more than a year ago must also have been formally published or peer-reviewed.
- The document must be available in full text, either as open access or through the institutional account.

3.2.4 Screening Phase Results

After removing 61 duplicate records, the total number of records was reduced to 536. Then these documents were evaluated against the eligibility criteria. The number of documents deleted according to each specific criterion is detailed below.

- Documents not published in English: 8
- Documents published before 2019: 58
- Documents from arXiv with more than 1 year, that haven't been published to any journal or conference: 71
- Not-Accessible documents: 3

The 396 resulting records were analyzed for their abstracts and titles to assess their relevance to the context of the research. Of these, only 123 advanced to the eligibility phase. Examples of documents discarded in this process included those addressing unlearning in recommender systems or retrieval-based models, papers focusing on optimizations for unlearning in cloud ML systems without relevance to computer vision, papers that addressed backdoor attacks using machine unlearning techniques, and studies where unlearning was only tangential (such as fairness, domain adaptation, continual learning) rather than the main contribution. Documents limited to classification models were also removed, except works in medical imaging and federated learning, where classification studies were retained in order to better understand the applications of machine unlearning in these domains.

3.2.5 Eligibility Phase

The eligibility phase involves individually analyzing the full texts of the remaining 123 records to determine their relevance to the research objectives. After careful evaluation, 89 records were considered pertinent to the study and retained as the final selection. Papers recommended by the supervisors, as well as other relevant works suggested during the process, were also included, resulting in a final collection of 93 documents.

3.2.6 Included Documents

The final set of included documents is listed in the References section: [2–4, 8, 9, 12, 15–23, 29–32, 35, 38, 41, 43–46, 48–55, 64–66, 69–71, 78, 82–84, 86–89, 91–96, 100–102, 105, 107, 108, 110, 115, 117, 118, 123, 126–128, 132–134, 138, 140–146, 148–151, 153–157]

3.3 Machine Unlearning in Generative Models

3.3.1 Unlearning Scenarios

In the context of machine unlearning, the unlearning scenario defines the type of data to be removed and at what granularity unlearning must occur. Common unlearning scenarios in generative models are:

- **Class-wise or Attribute-wise Labels Forgetting:** Involves removing the ability of a generative model to produce images belonging to a specific class or label attribute used during training, such as "dogs" or "cars." Methods for class-wise forgetting are discussed in: [39, 86, 143, 145]
- **Concept-wise Forgetting:** Ensures that the model cannot generate content related to a particular concept, such as material containing nudity or specific artistic styles, such as "Van Gogh". In the documents retrieved, this approach is always applied to remove sensitive or undesired concepts in Text-to-Image models, where generation is guided through text prompts. Methods for concept-wise forgetting are discussed in: [2, 9, 15–18, 21, 39, 41, 64, 69, 78, 82, 84, 92, 94, 101, 102, 110, 133, 138, 140–142, 144–146, 149, 154]
- **Data-point Unlearning:** Focuses on removing the influence of specific training samples, such as memorized images or images associated with individual users. Although this type of forgetting has been explored in the context of image classification, it has received less attention in generative modeling. Recent works that address data-point unlearning in generative settings include: [3, 83].
- **Pattern erasure:** Focus on deleting a specific response pattern, for example, deleting a watermark. This unlearning scenario is explored in the following studies: [95, 100, 134]

3.3.1.1 Challenging Forgets

Fan et al. [38] explore the notion of worst-case Forget sets in machine unlearning, highlighting that not all unlearning requests are equally difficult and that identifying and evaluating worst-case Forget sets is crucial to developing robust unlearning methods. These sets are defined as those that include samples with either:

(i) High entanglement score (ES), meaning the forget samples are embedded very close to retained data in the feature space;

(ii) Examples with strong memorization, where the model assigns overly confident predictions to the forget samples.

These properties make certain Forget sets particularly resistant to unlearning. Experiments explored in this study show that worst-case sets are significantly harder to erase than random sets, leading to substantial drops in unlearning accuracy (UA) and reduced effectiveness against membership inference attacks (MIA).

3.3.2 Type of Models used in Unlearning in Generative Models

By analyzing the papers retrieved, it is possible to see that most studies that explore unlearning in generative models do it on text-to-image models, with concept-wise forgetting as unlearning scenario. An overview of the generative models used is presented in Table 3.2

Table 3.2: Types of generative models used in unlearning studies and the corresponding papers.

Model Type	Papers
Image-to-Image Models	
GANs	[83]
Diffusion Models (DDIM / DDPM)	[3, 39, 83]
ViT (Vision Transformers)	[23, 83]
MAEs (Masked Autoencoders)	[83]
Text-to-Image Models	
Stable Diffusion (v1.4 / v2.1)	[2, 9, 15, 16, 18, 21, 39, 41, 64, 69, 78, 82, 84, 92, 94, 100–102, 110, 133, 138, 140–146, 149]
Rectified Flow Transformers (SD 3 / Flux)	[16, 43]

3.3.3 Common Unlearning Evaluation Metrics

The goals of unlearning vary depending on the unlearning scenario, and consequently, different metrics have been adopted in the literature to measure forgetting success, knowledge retention, and overall model quality. Below are presented some of the common metrics for each scenario.

Class-wise or Attribute-wise Forgetting When forgetting is performed at the class or attribute level, the evaluation typically relies on:

- **Unlearning Accuracy (UA)** that corresponds to the proportion of generated samples from the forgotten class that are no longer classified correctly, measured using an external classifier.
- **Fréchet Inception Distance (FID)** used to quantify the distributional quality of generated images for the retain classes, ensuring that forgetting does not degrade overall generative performance.
- **Inception Score (IS)** occasionally used to assess image diversity and sharpness.

Concept-wise Forgetting Concept-wise Forgetting was always seen linked to text-to-image models, and as the goal is to erase an entire concept, the evaluation usually combines perceptual quality with semantic alignment. Some common metrics for this scenario are:

- **CLIP Similarity / CLIP-based Scores** to measure semantic alignment between prompts and generated images, and to confirm that forgotten concepts are no longer rendered.
- **LPIPS (Learned Perceptual Image Patch Similarity)** is a perceptual distance metric to assess reconstruction quality after unlearning.
- **FID / IS** to evaluate fidelity and diversity for remaining concepts.
- **Robustness to Recovery Attacks** consists of assessing the robustness to recovery attacks such as UnlearnDiff Attack proposed by Zhang et al. [153], which attempts to revive forgotten concepts using adversarial prompts.

Data-point Unlearning For data-point-level forgetting (erasing individual images or identities), metrics focus on memorization and similarity between the Forget set and the model's output. Some metrics used include:

- **Negative Log-Likelihood (NLL):** In diffusion models, the data likelihood $p_\theta(x)$ is defined implicitly through the denoising process, where the model learns conditional distributions $p_\theta(x_{t-1} \mid x_t)$ across timesteps. The marginal probability of a datapoint can be expressed as:

$$p_\theta(x) = \int p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t)\, dx_{1:T},$$

with $x_T \sim \mathcal{N}(0, I)$. Since this integral is intractable, a variational bound (the ELBO) is used, leading to an approximation of the NLL by the diffusion training loss:

$$\mathrm{NLL}(x;\theta) \approx \mathbb{E}_{t,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2\right].$$

If the model has memorized a datapoint x, the predicted noise ε_θ matches the true noise ε closely, yielding a low NLL. If unlearning was successfull, the model assigns lower likelihood to forgotten samples, resulting in a higher NLL. This property makes NLL a useful indicator of forgetting strength, as employed in the SISS framework for data unlearning in diffusion models [3].

- **Self-Supervised Copy Detection (SSCD):** Given a forgotten sample x_f and a generated sample $\hat{x}$, similarity in embedding space is measured via:

$$\mathrm{SSCD}(\hat{x}, x_f) = \frac{f(\hat{x}) \cdot f(x_f)}{\|f(\hat{x})\|\, \|f(x_f)\|},$$

where $f(\cdot)$ denotes a self-supervised embedding extractor. Lower SSCD values indicate effective forgetting, as generated outputs no longer resemble the forgotten examples. This metric was adapted by Alberti et al. [3] to quantify unlearning strength in diffusion models.

Pattern Erasure For unlearning specific patterns, such as watermarks or response templates, some metrics used are:

- **Pattern Detection Accuracy:** probability that a forgotten watermark or pattern can still be identified in generated samples.
- **Downstream Utility Metrics:** task-specific measures (e.g., CLIP alignment or classification accuracy) to ensure that erasure does not degrade overall generative quality.

Emerging Metrics Focusing on text-to-image generative models, Rusanovsky et al. [117] criticize how unlearning is commonly evaluated. Although most methods rely on metrics, such as CLIP scores, FID, or unlearning accuracy, the authors demonstrate that forgotten concepts can still be recovered through carefully chosen latent seeds, even when prompt-based tests suggest successful erasure. To better assess unlearning and robustness, the authors propose additional metrics:

- **Relative Latent Likelihood Distance (RLLD):** measures the change in the likelihood of latent seeds associated with forgotten versus retained concepts, quantifying whether the model still assigns a high probability to forgotten representations.

- **PSNR (Peak Signal-to-Noise Ratio):** evaluates reconstruction quality when attempting to regenerate forgotten concepts from latent seeds, indicating how much structural information remains embedded in the model.
- **Membership Inference Attack (MIA) Success Rate:** assesses whether an adversary can still detect if a datapoint or concept was part of the training set, despite being targeted for forgetting. High MIA accuracy suggests incomplete erasure and persistent privacy risk.

3.4 Applications of Machine Unlearning in Federated Learning

As described in Section 2.3, Federated Learning (FL) is a decentralized machine learning paradigm where multiple clients collaborate to train a shared model without exchanging raw data, allowing privacy-preserving collaboration between institutions. To comply with privacy regulations such as the RTBF, machine unlearning has been used in FL environments to tackle the challenge of when data or client contributions must be removed from a trained model.

Two primary FL paradigms are relevant in the context of unlearning: Horizontal Federated Learning (HFL) and Vertical Federated Learning (VFL).

3.4.1 Horizontal Federated Learning (HFL)

In HFL, clients hold datasets that share the same feature space but correspond to different subsets of the sample space. Unlearning in this setting typically targets the removal of entire clients, specific data samples, or whole classes. Some methods that address these issues include:

- **Sequential Informed Federated Unlearning (SIFU)** performs client-level unlearning by reconstructing and sequentially subtracting the historical updates of the forgotten client from the global model, achieving theoretical guarantees without extra overhead for other participants [87].
- **Federated Unlearning During Learning (FedAU)** embeds unlearning directly into the training process through adjusted gradient updates, enabling sample-, class-, or client-level removal with minimal cost [157].
- **Federated Unlearning with Class-aware Representation Transformation (FUCRT)** applies a class-aware projection in the representation space to suppress information about the forgotten class, which is particularly effective in non-IID client distributions [12].
- **Fast Federated Machine Unlearning (FFMU)** approximates the effect of data removal using nonlinear functional theory, providing certified guarantees of forgetting while avoiding costly retraining [91].

Table 3.4.1 summarizes the main characteristics of these methods.

3.4.2 Vertical Federated Learning (VFL)

In VFL, by contrast, all clients share the same set of samples but hold complementary feature spaces. Labels are often controlled by a single party, making label privacy and protection a critical concern.

Table 3.3: Summary of state-of-the-art HFL federated machine unlearning methods.

FL Type	Method	Paper	Target	Model Type
HFL	SIFU	[87]	Client-level	Classification
HFL	FedAU	[157]	Sample/class/client-level	Classification
HFL	FUCRT	[12]	Class-level	Classification
HFL	FFMU	[91]	Client/sample-level	Classification

Unlearning in this setting may involve erasing specific labels or removing the contribution of sensitive feature partitions. A notable example is the study by Gu et al. [50], which combines manifold mixup with gradient ascent to erase labels efficiently, requiring only a small number of labeled samples.

3.5 Applications of Machine Unlearning in Medical Imaging

Machine unlearning is an emerging area of research in the field of medical imaging, driven by increasing concerns over data privacy, bias mitigation, and regulatory compliance, such as the "Right To Be Forgotten" (RTBF). This section surveys recent studies that explore the practical use of unlearning in medical imaging tasks, including classification, harmonization, and image reconstruction.

3.5.1 Image Classification

A common application of machine unlearning in the medical field and medical imaging is to remove the learned influence of specific patient data from a trained classifier model. In such cases, those particular images' effects are erased from the model's parameters, so the classifier behaves as if it had never seen them, while still retaining knowledge from the remaining dataset.

For example, in a federated learning context, Deng et al. [29] proposed a collaborative retinal disease identification framework across multiple medical institutions, integrating unlearning to enable the global model to forget withdrawn or erroneous data used in training. ElBedoui et al. [35] applied the same principle to an ECG classification model, and Gao et al. [45] extended this idea by introducing a framework that not only removes a client's data influence efficiently but also produces cryptographic proofs to verify that unlearning was performed correctly. Similarly, other studies [30, 46, 156] also explored client-level unlearning in federated learning for medical image classification tasks, ensuring that models can adapt to data removal requests while maintaining diagnostic performance.

In a centralized setting, Aliyeva et al. [4] developed a deep unlearning approach to classification of histopathological images of breast cancer, using adaptive gradient ascent to selectively remove the effect of specific samples from a MobileNet classifier. Nasirigerdeh et al. [108] use saliency unlearning for classification models, proposed by Fan et al. [39], to remove the influence of specific clients from a disease classification model trained on the CheXpert dataset.

3.5.2 Image Harmonization and Bias Removal

Machine unlearning has also been explored in the context of image harmonization tasks, where the goal is to remove non-biological variance caused by different scanners or acquisition protocols. Dinsdale et al. [31] address the challenge of variability in cardiac MRI images caused by differences between scanner manufacturers. Such vendor-specific characteristics can lead segmentation models to learn and

rely on acquisition-related features rather than purely anatomical information, reducing their ability to generalize across vendors. To mitigate this, the study incorporates an unlearning strategy within a U-Net–based segmentation framework. The model is first trained to perform both cardiac segmentation and vendor classification, enabling it to capture vendor-specific cues. An unlearning stage is then introduced in which a confusion loss penalizes the encoder for retaining vendor-discriminative features, effectively forcing it to "forget" scanner identity information. This results in a vendor-neutral segmentation model that maintains high segmentation accuracy while demonstrating improved robustness and consistency across data from multiple MRI manufacturers. FedHarmony, proposed by Dinsdale et al. [32], extends this concept to the federated learning paradigm. The framework uses a domain adaptation approach with three loss functions to train a model that performs a primary task while simultaneously unlearning scanner-specific effects.

3.5.3 Image Generation and Reconstruction

While machine unlearning has been predominantly applied to classification and de-biasing tasks in he medical field, Xue et al. [148], proposes a method to unlearn the influence of undesired training data from a reconstruction model, without full retraining. Specifically, they address the issue of hallucinations (false anatomical structures) that appear when models are trained on mixed-organ datasets. In the specific case of this paper, brain MRI and knee MRI. To achieve unlearning, the study adapts techniques originally designed for classification, and achieves best results and performance with Noisy Labeling with Fine-Tuning (NL-FT). This approach will be further explained in subsection 3.6.4.

3.6 Related Work

This section presents the literature works whose proposed methods are most aligned with the main goal of this dissertation. An overview of these methods is provided in Table 3.4.

Table 3.4: Comparison of unlearning methods most similar to this dissertation goals.

Method	Paper	I2I Model	DM	Data-Point Unlearning	Applied to Medical Data
SalUn	[39]	✓	✓	—	—
KL–L_2	[83]	✓	✓	✓	—
SISS	[3]	✓	✓	✓	—
MRI Unlearning	[148]	✓	—	✓	✓
This dissertation	—	✓	✓	✓	✓

3.6.1 Saliency Unlearing (SalUn)

Saliency-based Unlearning by Fan et al. [39] is an unlearning technique suitable for both image classification and generation tasks. In generation tasks, it tackles concept-wise forgetting in text-to-image models, such as Stable Diffusion, and also class-wise forgetting from denoising diffusion probabilistic models (DDPMs) with classifier-free guidance. Although it has not been applied to medical data in the author's paper, it has been successfully applied to medical data in classification models by Nasirigerdeh et al. [108].

Unlearning scenario For the unlearning experiments in the image-to-image setting, the experiments are carried out on the CIFAR-10 dataset, where DDPM is trained to generate images conditioned on class labels. The unlearning scenario is class-wise forgetting, as the model must forget a specific CIFAR-10 class (e.g., “airplane”) by no longer generating realistic samples of that class, while retaining the ability to synthesize images from all other classes.

Forget Set and Remain Set: In the image-to-image unlearning task, the Forget Set consists of images and conditions from the class to be erased (e.g., “airplane”), while the Remain Set includes all other CIFAR-10 classes. In the experiments with Stable Diffusion, to delete the concept of nudity, 800 images were generated using the prompt "a photo of a nude person", and 800 images using the prompt "a photo of a person wearing clothes". The Forget set and the Remain set have the same number of samples.

Method Overview SalUn introduces a gradient-based weight saliency map to localize which model parameters are most responsible for representing the forgetting set. The process works as follows:

1. **Saliency Masking:** Compute the gradient of the forgetting loss by calculating the MSE diffusion loss restricted to the forgetting set, with respect to model parameters. Parameters with saliency above a threshold are marked as salient.
2. **Selective Updating:** During unlearning, only the salient parameters are updated, while all others remain fixed, preventing unnecessary changes and maintaining the overall performance of the model.
3. **Forgetting Loss on the Forget Set:** Forgetting is enforced through Random Relabeling (RL), where the original class condition is replaced with an incorrect one, forcing the model to decouple the forgotten concept from its original label.
4. **Regularization on the Remain Set:** A standard diffusion loss is simultaneously applied to the Remain set, ensuring that generative quality for non-forgetting classes is preserved.

It is important to note that, although in the experiments addressed in the author’s paper, Random Relabeling was the loss function that led to the best results, the framework is flexible and can incorporate other loss functions, such as Noisy Labeling or Gradient Ascent.

Formally, the unlearning objective for SalUn can be expressed as:

$$\min_{\Delta\theta} L_{\text{SalUn}}(\theta_u) = \mathbb{E}_{(x,c)\sim D_f, t, \varepsilon\sim\mathcal{N}(0,1), c'\neq c}\left[\left\|\varepsilon_{\theta_u}(x_t \mid c') - \varepsilon_{\theta_u}(x_t \mid c)\right\|^2\right] + \beta\, \ell_{\text{MSE}}(\theta_u; D_r), \tag{3.1}$$

where D_f is the forgetting set, D_r the Remain set, and β balances forgetting and retention.

Effect of Thresholding on Saliency Masks It is important to note that the threshold used in constructing the saliency mask plays a key role in determining which parameters are updated during unlearning. By varying the threshold, different saliency masks, with different sparsity levels, are obtained. According to the authors, a lower threshold marks more parameters as salient, potentially leading to underforgetting and degraded image quality, while a higher threshold may result in the image quality. In

practice, setting the treshold to the median of the gradient magnitudes provides a balanced mask, yielding stable unlearning performance. As illustrated in Figure 3.2, varying the saliency sparsity strongly affects the trade-off between forgetting effectiveness and image generation quality.

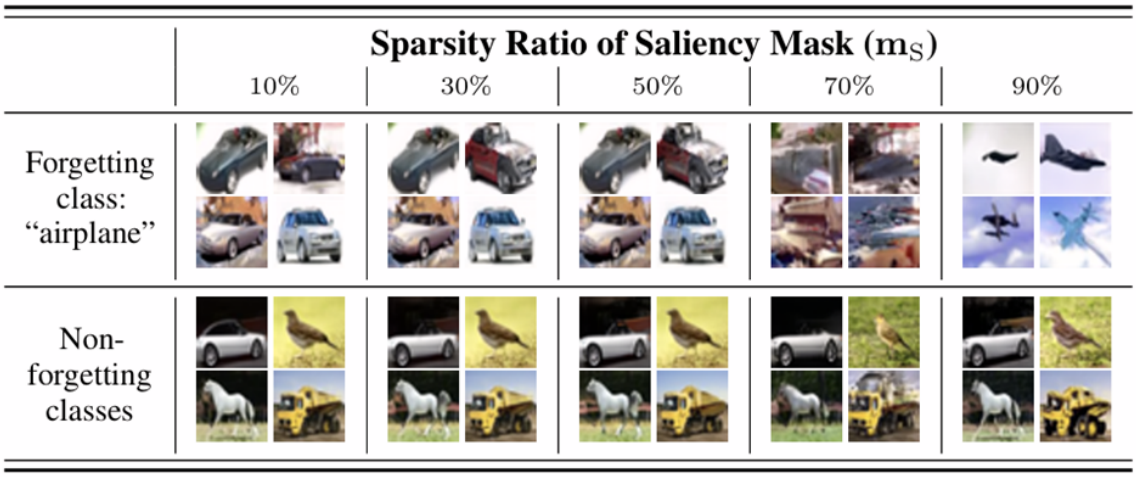


Figure 3.2: Examples of images generated by the unlearned DDPM under different sparsity levels when forgetting the class "airplane". Lower sparsity (e.g., 10%) achieves forgetting but reduces generation quality, while higher sparsity (e.g., 90%) preserves quality but fails to forget. A balanced sparsity of 50% provides the best trade-off [39].

Evaluation Metrics: To evaluate unlearning effectiveness in generative models, two complementary metrics are employed:

- **Unlearning Accuracy (UA)**: the percentage of generated images (conditioned on the forgotten class) that are correctly identified as not belonging to that class, measured by an external classifier (ResNet-34 trained on CIFAR-10).
- **Fréchet Inception Distance (FID)**: quantifies the quality of generated images for the Remain set; lower values indicate better fidelity.

Results On CIFAR-10 with classifier-free guided DDPM, SalUn was evaluated for class-wise forgetting. The method achieved Unlearning Accuracy (UA) close to 100% and a FID nearly identical to retraining from scratch. These results indicate that SalUn can completely suppress the forgotten class while preserving the generation quality for the remaining set, nearly matching the gold standard of full retraining.

3.6.2 KL–L_2

Li et al. [83] explore machine unlearning in the context of image-to-image (I2I) generative models, such as VQ-GANs, Diffusion Models, and masked autoencoders (MAEs). According to the authors, data unlearning settings remain a gap in the diffusion model literature, especially in image-to-image models. *For simplicity, and since the authors didn't gave a name to their method they propose, we will refer to their approach as KL–L_2 throughout this dissertation.*

Unlearning scenario The unlearning scenario of this study falls into data-point unlearning, as the goal of this study is to forget specific training samples so that the model cannot reconstruct them, even when provided with partial input, while maintaining reconstruction quality on all other samples.

Forget Set and Remain Set For the experiments with the ImageNet-1K dataset, they randomly select 100 classes as Forget Set and another 100 classes as Remain Set. For Places-365, from 365 classes, they randomly select 50 classes as Remain Set and another 50 classes as Forget Set.

Method Overview In this study, unlearning is performed by using the Kullback–Leibler divergence (KL-divergence) [26], a standard way to measure how different two probability distributions are. A small KL value means two distributions are similar, while a large KL value means they are very different. In this setting, generated retain samples are encouraged to be statistically close to their ground-truth counterparts (small KL), while generated forget samples are pushed to look completely different (large KL).

To avoid instability from unbounded KL values, the authors prove that the safest and most effective way to maximize divergence is to push forget outputs toward Gaussian noise with the same covariance as the original data.

Finally, since computing KL directly on high-dimensional images is often intractable, the paper shows how to approximate it with a simple L_2 loss in the encoder space. For retain samples, the encoder of the unlearned model is kept close to the original encoder, preserving fidelity. For forget samples, embeddings are aligned with Gaussian noise, ensuring the decoder reconstructs meaningless outputs.

Evaluation Metrics Evaluation focuses on forgetting success (whether the forget samples degrade into noise) and retention fidelity (PSNR, SSIM on remain samples).

Results Across ImageNet-1K and Places-365 with multiple backbones (Diffusion Models, VQ-GAN, MAE), the method effectively removes the ability to reconstruct forget samples, while retaining nearly identical reconstruction quality for remain samples. An example of the results obtained in the paper is presented in Figure 3.3.

Limitations A limitation of this approach is that the unlearned reconstruction goal is to produce noise in a part of the image that contains undesirable content, rather than to reconstruct it into a different object. This differs from many other unlearning scenarios, where the goal is not to produce pure noise in the image, but rather to selectively remove sensitive or undesired information while keeping the overall image coherent.

3.6.3 Subtracted Importance Sampled Scores (SISS)

Subtracted Importance Sampled Scores (SISS), proposed by Alberti et al. [3], similarly to KL–L_2 [83], performs data-point unlearning in I2I models. Unlike techniques tailored for concept or class unlearning, like SalUn, SISS is designed to remove specific data points from the training set while preserving overall model quality. SISS is applied to both unconditional DDPMs and text-to-image Stable Diffusion, and

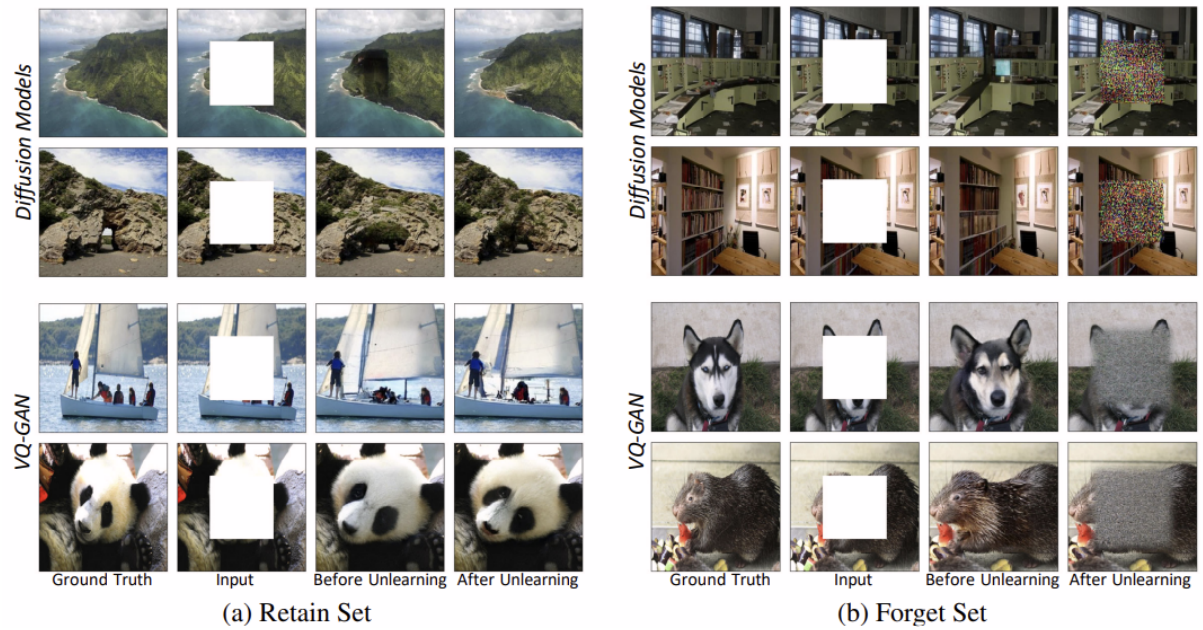


Figure 3.3: Visual Results of KL–L_2 unlearning method by Li et al. [83]. The images in the retain set remain almost unaffected before and after unlearning, while the images in the Forget set are nearly noise after unlearning.

it has been evaluated on CelebA-HQ, MNIST augmented with Fashion-MNIST T-shirts, and Stable Diffusion v1.4 trained on LAION-5B.

Unlearning Scenario The unlearning scenario of this paper is data unlearning: removing the influence of individual training datapoints, such as a specific celebrity face or the inserted T-shirt example. This differs from concept unlearning, since there is no guiding class or prompt; the unlearning must directly suppress a desired datapoint. On CelebA-HQ, six celebrity faces were randomly selected, and each was unlearned individually. For MNIST augmented with Fashion-MNIST, the Forget set is a single T-shirt image introduced into the dataset, whereas the Remain set corresponds to the remaining handwritten digits.

Method Overview This technique combines naive deletion with gradient ascent through importance sampling, addressing an important trade-off in data unlearning: naive deletion preserves model quality but unlearns slowly, while gradient ascent unlearns effectively but degrades overall performance.

Before diving into the formulation of this method, it is important to understand the concept of importance sampling. Importance sampling is a classical technique in statistics and machine learning that allows expectations under one distribution to be estimated using samples from another. This is achieved by reweighting samples with an importance factor, correcting for the difference between the target and sampling distributions [3].

In the context of unlearning, importance sampling enables SISS to place greater weight on forget samples during training updates, ensuring they are effectively suppressed, while down-weighting their influence on the Remain set. This balances the trade-off between efficient unlearning (as in gradient ascent) and preserving model quality (as in naive deletion).

In this approach, the unlearning loss is derived by decomposing the naive deletion objective into two terms: (1) a retention term that preserves knowledge of the retain set and (2) a forgetting term that suppresses the influence of the forget set. SISS leverages importance sampling to compute both terms with a single model forward pass, rather than requiring separate evaluations. Given a retain sample x_r and forget sample x_f at diffusion timestep t, the method samples from a mixture distribution $q_\lambda(m_t|x_r,x_f)$ parameterized by λ:

$$q_\lambda(m_t|x_r,x_f) = (1-\lambda)q(m_t|x_r) + \lambda q(m_t|x_f), \tag{3.2}$$

where $q(m_t|x)$ is the standard noising distribution for sample x.

The SISS unlearning loss is then defined as:

$$\mathscr{L}_{\text{SISS}} = \mathbb{E}_{m_t\sim q_\lambda}\left[w_r \cdot \|\varepsilon - \hat{\varepsilon}(m_t,t)\|^2 - (1+s)\, w_f \cdot \|\varepsilon - \hat{\varepsilon}(m_t,t)\|^2\right], \tag{3.3}$$

with importance weights

$$w_r = \frac{n}{n-k}\frac{q(m_t|x_r)}{q_\lambda(m_t|x_r,x_f)}, \quad w_f = \frac{k}{n-k}\frac{q(m_t|x_f)}{q_\lambda(m_t|x_r,x_f)}. \tag{3.4}$$

Here, the first term (weighted by w_r) preserves model performance on retain data, while the second term (weighted by w_f) drives forgetting through gradient ascent. Note that n represents the total dataset size (Forget + Remain set), k the Forget set size, and $(n-k)$ the Remain set size. The additional factor $(1+s)$, called the *superfactor*, controls the strength of the forgetting signal to improve stability.

The parameter λ balances between the two extremes: $\lambda = 0$ reduces to naive deletion, while $\lambda = 1$ approximates gradient ascent. In practice, $\lambda = 0.5$ provides an effective trade-off, combining strong unlearning with good quality preservation.

Forget Set and Remain Set The Forget set consists of one or more images to be erased from the model's knowledge, related to a single entity. The Remain set includes all other images, whose reconstructions should remain close to the original model's outputs. Therefore, in the CelebA-HQ experiments, the Forget set is formed by images of a specific celebrity, while the Remain set is formed by all the other faces in the dataset. In the MNIST + T-shirt experiment, the Forget set is the single inserted T-shirt image, while the Remain set is all other handwritten digits.

Evaluation Metrics To measure both the strength of unlearning and the quality of retained generations, the following metrics are employed:

- FID (Fréchet Inception Distance): evaluates the distributional similarity between generated and real images for the Remain set. Lower FID indicates better fidelity and diversity of generated samples.
- IS (Inception Score): measures the sharpness and diversity of generated images. Higher IS reflects higher quality and variety.
- CLIP-IQA: a perceptual quality score based on CLIP embeddings, used particularly in the Stable Diffusion experiments. Higher values indicate better alignment with natural image quality.

- SSCD (Self-Supervised Copy Detection): measures similarity between generated images and the original unlearned datapoints. Lower SSCD means the model has forgotten the target datapoints more effectively.
- NLL (Negative Log-Likelihood): the log-likelihood of the unlearned datapoints under the current model. Higher NLL indicates the model assigns lower probability to the forgotten examples, consistent with successful unlearning.
- Unlearning success rate (in Stable Diffusion): the percentage of curated prompts for which the model no longer reproduces memorized training examples.

Results For the experiments with the DDPM, SISS was tested across two different settings:

- **CelebA-HQ (celebrity faces):** Forgetting was performed one face at a time over 6 faces. A visual representation of the results for the experiments with the CelebA-HQ dataset is present in Figure 3.4
  - SSCD (similarity metric): reduced by more than 50% (from 0.87 to ~ 0.36).
  - FID: maintained at 25.1 (close to pretrained 30.3).
  - NLL: increased moderately (1.44 vs. 1.25), consistent with successful forgetting.
  - In an extended test, SISS remained stable when forgetting 50 faces sequentially, ending with FID 20.3.
- **MNIST + T-shirt:** Forgetting targeted a single inserted T-shirt example.
  - p (generation rate of T-shirts): reduced from 0.74% to 0%.
  - NLL: increased from ~ 1.0 to ~ 8.1, confirming that the T-shirt was fully unlearned.

Limitations SISS is tailored to forget specific entities or objects in the training distribution, as in CelebA-HQ experiments, where individual celebrity faces are unlearned. This setup assumes that only a small subset of identities must be removed, while all others remain available. In practice, this selectivity may be challenging: forgetting just a few samples can influence generalization, since the Remain set is composed of the other entities that should be preserved, and forgotten identities may share visual features with them. Furthermore, SISS does not directly address scenarios where larger groups of samples, or entire semantic categories, need to be removed, which could limit its usefulness in stricter privacy-sensitive applications.

3.6.4 Erase to Enhance (MRI Unlearning)

Erase to Enhance by Xue et al. [148] is, to our knowledge, the only work in the machine unlearning literature applied to medical imaging tasks of image reconstruction or generation. In contrast to most previous studies in the field of medical image, which mainly focus on classification or text-to-image generation, this work adapts the unlearning to an image-to-image setting. Specifically, the authors study MRI reconstruction from undersampled k-space data using a CNN-based architecture with cascaded

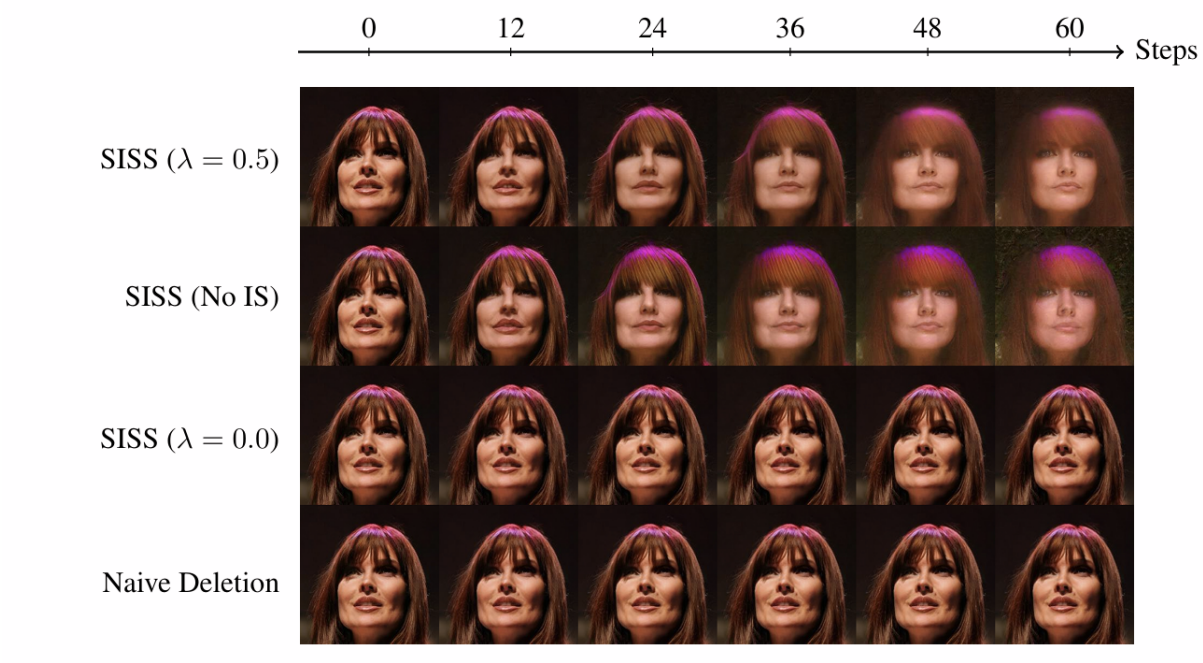


Figure 3.4: Visualization of celebrity unlearning with SISS method over fine-tuning steps [3].

UNets, and propose to remove the influence of undesired training distributions. This study also constitutes one of the first attempts to formally define and apply unlearning within the context of MRI reconstruction.

Unlearning scenario The authors design a proxy task where the retain set corresponds to brain MRI scans, while the Forget set consists of knee MRI scans. Training a reconstruction model jointly on both anatomies leads to spurious artifacts and hallucinations, which can compromise diagnostic utility. The unlearning objective is therefore to erase the knowledge of the knee distribution, ensuring the model behaves like a model trained only on brain data, without retraining from scratch.

Method Overview The authors of this paper adapt machine baseline unlearning strategies to MRI reconstruction: Fine-Tuning (FT), Gradient Ascent with ℓ_1 Regularization (GA-ℓ_1) to constrain weights for stability, and Noisy Labeling (NL). Furthermore, the authors propose hybrid methods that combine FT with GA-ℓ_1 or NL (denoted GA-ℓ_1-FT and NL-FT), leveraging a small retain subset to preserve reconstruction fidelity while performing targeted unlearning.

Forget Set and Retain Set: The retain set D_r is composed of 5,898 slices of brain MRI data from the FastMRI dataset, while the Forget set D_f consists of 536 knee MRI slices (retain-to-forget ratio 10:1). The model $\hat{G}$ is trained only on D_r, serving as the upper bound for performance after unlearning. In practice, the unlearning algorithms operate on the mixed model G (trained on $D_r \cup D_f$), with limited access to a subset of retain data D'_r (as low as 1–10% of D_r).

Evaluation Metrics: Performance is evaluated using standard image quality metrics:

- **Brain Test Accuracy (BTA):** Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) on brain test data, measuring retention of desired knowledge.

- **Knee Test Accuracy (KTA):** PSNR and SSIM on knee test data, lower values indicate stronger forgetting.
- **Unlearning Accuracy (UA):** PSNR/SSIM on the Forget set itself, confirming effective unlearning.
- **Run-Time Efficiency (RTE):** relative computational cost compared to retraining.

Results The experiments show that naive fine-tuning fails to erase knee knowledge, with the model still reconstructing knee images. Pure GA-ℓ_1 or NL degrade both forget and retain performance, indicating over-forgetting. By contrast, hybrid methods achieve the best balance: NL-FT reaches brain reconstruction performance, while significantly lowering KTA, indicating effective unlearning. Importantly, these results were achieved with only 10% of the retain data, underscoring the data efficiency of the approach. The authors also note that increasing retain data beyond this point does not necessarily improve unlearning, sometimes even degrading results, suggesting an optimal subset size for efficient MRI unlearning.

Overall, this study demonstrates that unlearning can be effectively applied to medical reconstruction models, mitigating artifacts from undesired distributions while maintaining images medical utility.

3.7 Summary

This research tried to define the current state of the art in the machine unlearning field to remove the data influence from generative models, and also to see the current trends of usage of machine unlearning in federative learning and in the medical imaging field. According to the papers retrieved with the PRISMA framework, and according to some authors [83], there are still few works that tackle the problem of specific data unlearning in generative models, especially in diffusion models. Most of the found works focus on concept unlearning, which is a different problem. The only found papers that tackle data unlearning in diffusion models are SISS [3] and KL–L_2 [83]. However, KL–L_2 leads to noise production in the images belonging to the Forget set, which is not the goal of this dissertation, and SISS tackles the removal of one specific entity when other entities are not required to be anonymous, which is not the case in the work in this dissertation, where a large number of patients must be forgotten. Therefore, there is still a gap in the literature for methods that can effectively remove the influence of specific data points from diffusion models, while preserving overall model quality and without resorting to full retraining. When it comes to machine unlearning in the medical imaging field, one work was found that tackles the problem of unlearning in image generation tasks [148], however, it tries to forget entire domains (knee MRI from the model that generates brain MRI), to stop hallucinations, which is not the case of this dissertation, where the goal is to forget specific patients, and therefore, the entanglement between the Forget set and the Remain set is higher.

Chapter 4

Literature Review: Anonymization Techniques in Medical Imaging

4.1 Overview

The protection of patient privacy in medical imaging has become a central concern in both clinical practice and research, as imaging data often contains biometric features that can be linked to individuals.

In Section 4.3, the main categories of anonymization techniques proposed in the literature are reviewed, ranging from classic methods such as masking or blurring, to more advanced approaches using generative deep learning models. Particular attention is given to the trade-off between preserving diagnostic utility and ensuring privacy, an ongoing challenge that drives most developments in this area.

In Section 4.4, deanonymization techniques are analyzed, which expose vulnerabilities in existing methods and highlight the limits of anonymization when confronted with increasingly sophisticated model attacks. Together, these perspectives provide a balanced overview of current progress and ongoing challenges in ensuring privacy in medical imaging.

4.2 Document Retrieval Methodology

Unlike the systematic literature review presented in Chapter 3, the approach used for this chapter was not based on a structured systematic review protocol. Instead, the collection of publications was initiated from a relevant subset of articles, which formed the basis of the literature review and was further extended through the references cited in those works. This method allowed for a more targeted exploration of key contributions in the field, ensuring that the most influential and technically relevant articles on medical image anonymization were included in the review. Although this approach does not claim to be exhaustive, it provides a foundation for understanding the main techniques and research directions in the area.

4.3 Image Anonymization Techniques

When anonymizing case-based explanations, many conventional privacy-preserving techniques are insufficient as they do not offer protection at the point of display, when images must be interpretable by

humans.

For example, Federated Learning keeps data on local devices and avoids central data storage, but does not anonymize the image itself. Metadata-Level Anonymization removes personal identifiers from image headers, such as patient name or scan date, but it leaves the visual content untouched, meaning identifiable features in the image are still present. Differential privacy adds noise during model training or to the output data itself, but excessive noise can degrade image quality, rendering the image clinically useless.

One other approach to generate private case-based explanations is to generate synthetic images with deep learning models and use them as case-based explanations instead of the real images. However, recent research has shown that these models tend to memorize some of the training images, leaking training data information.

To address this, researchers have proposed methods that modify the visual content of the image itself, ensuring that personally identifiable features are removed while retaining diagnostic utility. These can be grouped into two broad categories: classic anonymization techniques, such as blurring or masking sensitive regions, and deep learning-based approaches, delving into in-model and post-model approaches, which use generative models to selectively alter or replace identity features while preserving relevant clinical information.

4.3.1 Classic methods

Classic methods include deterministic pixel masking techniques, such as blurring and masking, to obscure identifiable features. Algorithms like k-Same [109] use blurred facial regions in head MRIs, making individuals indistinguishable from at least k others. Similarly, U-Net-based segmentation [121] can detect and paint distinctive body markings, such as scars or tattoos, to reduce the risk of reidentification.

Although simple to implement, these methods can degrade diagnostically relevant image content and are vulnerable to reconstruction attacks, such as the facial reconstruction methods described by Schwarz et al. [120].

4.3.2 Deep Learning Based Methods

Generative models synthesize new images that resemble the original data but with Personally Identifiable Information (PII) removed or altered. Compared to classic anonymization, these methods offer greater flexibility by disentangling identity from diagnostic content and enabling the creation of high-quality synthetic data.

Within this category, two broad strategies can be distinguished:

- **In-model methods:** Integrate privacy mechanisms directly into the generative process itself, for example, by adding differential privacy during training, disentangling latent features, or modifying the diffusion sampling procedure. These methods aim to ensure that the model never learns or reproduces identity-specific information in the first place.

- **Post-model methods:** Apply anonymization after generation, typically by filtering or modifying synthetic images that may still contain residual identity features. This strategy provides an additional safeguard but often at the cost of efficiency, since a portion of generated images may need to be discarded or altered.

The following subsections review methods from each category.

4.3.3 In-Model Methods

4.3.3.1 Differential Privacy

In medical imaging, differential privacy is most often implemented through differentially private stochastic gradient descent (DP-SGD) [1]. In this method, gradients are clipped to limit sensitivity, Gaussian noise is added before each parameter update, and a privacy accounting mechanism tracks the cumulative use of the privacy budget to ensure the entire training process remains within the specified guarantees.

Differential Privacy in GANs While DP-SGD is still considered a standard, its effectiveness is limited by its privacy–utility trade-off. In high-resolution imaging, the addition of noise can significantly degrade image quality, and since high-dimensional data require proportionally more noise, subtle pathologies end up being distorted or lost [85]. Several advances have been proposed to mitigate these drawbacks. For example, PATE-GAN by Jordon et al. [75] leverages ensembles of teacher models to generate private labels for training a student GAN, while adaptive clipping strategies [147] aim to reduce unnecessary distortion. However, PATE-GAN remains computationally intensive, scales poorly to large datasets, and tends to reduce the diversity of generated samples.

To address these challenges, Privacy-Preserving GANs (PP-GANs) [147, 152], integrate differential privacy directly into adversarial training by cutting discriminator gradients and adding Gaussian noise. This approach helps to prevent overfitting while maintaining sample diversity. However, achieving a balance between privacy and image fidelity, as well as avoiding mode collapse, continues to be a challenge. Recent improvements include conditional PP-GANs [6, 136], which use metadata to enhance realism, and multi-scale architectures.

Differential Privacy in Diffusion models Recently, attention has turned to diffusion models, which are known for producing high-quality synthetic images. Differentially Private Diffusion Models (DPDMs), proposed by Dockhorn et al. [33], incorporate differential privacy into diffusion training by applying DP-SGD with gradient clipping and noise injection, achieving state-of-the-art image generation performance on datasets such as MNIST and CIFAR-10, surpassing prior DP-GAN results. In the domain of medical imaging, Daum et al. [28] extended this idea to 3D cardiac MRI data. Their privately fine-tuned latent diffusion models, combined with pretraining on public data, significantly improved synthetic image realism at privacy budgets of $\varepsilon = 10$, and achieved a much lower FID than without pretraining.

4.3.3.2 Confounder-Based Image Encryption

ConfounderGAN by Tian et al. [131], is an encryption framework for image data privacy that injects imperceptible confounder noise, forming a spurious causal path $X \leftarrow N \rightarrow Y$, so that models trained on the altered images fail to learn the correct $X \rightarrow Y$ mapping, despite the images remaining visually natural. The generator crafts this noise, the discriminator ensures it is subtle, and a classifier ensures the noise correlates with labels. Once trained, the generator can quickly encrypt both in-distribution and out-of-distribution images, making it viable for real-time applications. However, although the images remain

visually natural, the injected confounder noise severely degrades the utility of the data for downstream learning, since models trained on encrypted images fail to recover the true causal relationships.

4.3.3.3 Disentanglement-Based Anonymization

Montenegro et al. [104] proposed a disentanglement framework tailored to the anonymization of medical case-based explanations. Their model addresses the key challenge that medical images, when used as explanations, risk revealing patient identity while still needing to preserve diagnostic content.

The method employs an autoencoder that decomposes each image into three independent latent vectors: (i) identity features, (ii) medical features, and (iii) residual features necessary for reconstruction. Independence is enforced via a disentanglement loss, ensuring that altering one vector does not affect the others. To maintain diagnostic fidelity, a disease classifier supervises the medical vector, while an identity classifier supervises the identity vector. A GAN-based discriminator further enforces the realism of reconstructed images, even when training data are scarce.

Anonymization is achieved by replacing the identity vector with a synthetic one, generated by a variational autoencoder trained to sample privacy-preserving identity features. This ensures that the new image retains its original medical characteristics, but no longer corresponds to the patient's identity. The framework also naturally enables the creation of counterfactual explanations by replacing the medical vector instead of the identity vector, showing how a case would appear under different pathological conditions.

Experiments on iris, chest X-ray, and face datasets demonstrate that this disentanglement approach preserves medical content while significantly reducing re-identification accuracy, outperforming previous anonymization methods for medical case-based explanations.

4.3.3.4 Classifier-Free Identity Guidance

Recent advances in diffusion models have motivated their adoption in privacy-preserving medical image generation because of their ability to produce high-resolution synthetic data.

At the sampling stage, classifier-free guidance (CFG), which blends conditional and unconditional diffusion estimates to steer generation without an external classifier, can be adapted to preserve privacy by steering the sampling away from identity-related features [60]. For example, Shaheryar et al. [122] adapts the CFG for identity suppression in diffusion models. Their method employs CFG during inference to steer generation away from unwanted identities by identifying and avoiding their latent "signatures" in the U-Net bottleneck, achieving anonymization without the need to retrain the network.

4.3.4 Post-Model Methods

Campos et al. [11] extended the work of Packhäuser et al. [112] and proposed a post-model anonymization strategy that incorporates identity-based similarity filtering into the synthetic image generation pipeline. This pipeline architecture is presented in Figure 4.1. The central idea is that, although latent diffusion models are capable of producing high-quality medical images, these samples can inadvertently replicate patient-identifiable characteristics. To mitigate this risk, a filtering stage is employed to detect and remove synthetic images that are too similar to real patient data.

This filtering stage requires training a Patient Retrieval Network, which is implemented as a Siamese Neural Network based on the ResNet-50 [57] architecture. The network takes two images as input and uses a contrastive loss function to generate embeddings, ensuring that images of the same patient are mapped close together in the embedding space. A separate verification model follows the same architecture, but instead of learning embeddings, it compares input images and classifies whether they belong to the same patient. For both models, the training data consist of positive pairs, formed from images of the same patient, and negative pairs, generated randomly from images of different patients [11].

The anonymization pipeline follows a three-step process:

- **Computation of identity embeddings:** Each training image is processed by a Siamese retrieval network, trained with a contrastive loss, to obtain an identity embedding. These embeddings are stored in a KD-tree, which allows efficient nearest-neighbor searches in the embedding space.
- **Nearest-neighbor retrieval:** For every synthetic image generated by the diffusion model, the system retrieves the three most similar training images. Then, these are ranked according to similarity to the Rank-1, Rank-2, and Rank-3 neighbors.
- **Verification and filtering:** A verification network, architecturally similar to the retrieval network but trained with a classification loss, is used to assess whether the Rank-1 synthetic–real image pair belongs to the same patient. If the similarity score exceeds a predefined threshold (set to 50% in the authors' implementation), the synthetic image is flagged as non-anonymous and removed from the catalogue.

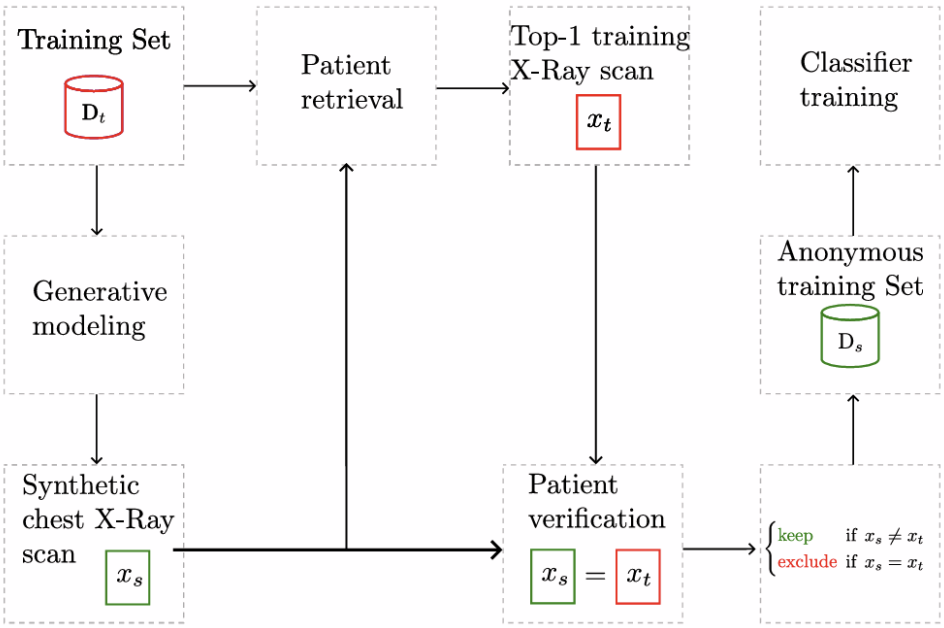


Figure 4.1: Anonymization Pipeline proposed by Campos et al. [11]: A generative model generates synthetic samples based on the real training data. From these samples, those closely resembling patients in the training set are removed based on a patient retrieval and a patient verification model.

Despite its effectiveness, this post-model filtering approach presents several drawbacks. First, it is inherently empirical. Identity-based filtering relies on similarity metrics and thresholds that may

fail in edge cases. Second, the process is computationally inefficient: a potentially large proportion of generated images can be discarded, wasting resources, and there is no upper bound on the number of images deleted. Finally, because filtering occurs after image generation, there is no direct control over the generative process itself, leaving open the possibility of privacy leakage through undetected identifiers and not guaranteeing the RTBF to patients.

4.4 Image De-Anonymization Techniques

While anonymization methods aim to protect patient privacy, a growing number of research studies show that medical images can often be re-identified or linked back to individuals. The feasibility of these attacks depends largely on the adversary's level of access to the system, which can be categorized as follows:

- **Explanation Access**: The adversary only sees anonymized explanations or synthetic outputs.
- **Black-Box Access**: The adversary can query the model through an API, observing inputs and outputs.
- **White-Box Access**: The adversary has full access to model parameters, as may occur in federated learning with untrusted participants.

Table 4.1 summarizes the main de-anonymization approaches proposed in the literature, organized by their required level of access. These threat models illustrate the variety of attack vectors available to adversaries depending on the information they can access. In the following subsections, we provide a detailed discussion of the most relevant categories of attacks.

Table 4.1: Overview of de-anonymization attacks proposed and used in the literature, grouped by attack type and required access level.

Attack Type	Method	Representative Papers	Access Type
Re-Identification	Multiclass Recognition	[104]	Explanation
	Siamese Recognition	[104, 112]	Explanation
	Linkage (MRI–Face)	[37]	Explanation
	Linkage (Iris)	[137]	Explanation
Model Exploitation	Model Inversion	[40]	Black-/White-box
	Membership Inference	[34, 63, 79, 97, 124]	Black-/White-box
	Data Extraction	[14]	Black-/White-box
	Data Reconstruction	[14]	Black-/White-box
Adversarial Queries	Generative Leakage (GANs)	[56]	Black-box
	Generative Leakage (DM)	[99]	Black-box
	Adversarial Querying	[111]	Black-box

4.4.1 Re-Identification Attacks (Explanation Access)

These attacks operate even when the adversary only has access to anonymized explanations or synthetic outputs, without interacting directly with the model. They exploit latent biometric traits present in

medical images, which remain identifiable despite the removal of explicit identifiers. For example, Packhäuser et al. [112] showed that chest radiographs encode anatomical signatures that enable patient-level matching using Siamese networks. Similarly, linkage attacks combine anonymized scans with external biometric data. For example, brain MRIs can be matched to facial photographs, as demonstrated by Molano et al. [37], and iris images can enable cross-modal identification, as shown by Trokielewicz et al. [137]. In addition, identity-recognition approaches such as multiclass classifiers and Siamese networks have been proposed for evaluating privacy-preserving methods. Montenegro et al. [104] demonstrated that both strategies can succeed with explanation-only access, making them useful benchmarks for re-identification risk assessment. These studies highlight that re-identification can succeed with explanation-level access alone.

4.4.2 Model Exploitation Attacks (Black-Box and White-Box Access)

Unlike re-identification, these attacks target the internal behavior of machine learning models. In black-box settings, the adversary can only query the model through its API, whereas in white-box scenarios, they may have access to the model parameters. Model inversion attacks attempt to reconstruct input samples from model outputs, first demonstrated by Fredrikson et al. [40]. Membership inference attacks [124] instead determine whether a given patient's data was part of the training set by exploiting the differences between the seen and unseen samples. More recent work has extended membership inference to generative models. Duan et al. [34] analyzed MIA techniques originally designed for GANs and studied their adaptability to diffusion models, while Kong et al. [79] proposed an attack on deterministic DDIM sampling. Matsumoto et al. [97] evaluated both black-box and white-box membership attacks, showing that their effectiveness decreases with larger training sets. Hu et al. [63] introduced loss-based and likelihood-based inference strategies, demonstrating the feasibility of MIAs even in diffusion models. Beyond membership inference, Carlini et al. [14] investigated data extraction and reconstruction attacks, showing that generative models can output samples closely resembling training data under both black- and white-box conditions. Together, these methods expose vulnerabilities at both the sample (inversion, reconstruction, extraction) and dataset (membership inference) levels.

4.4.3 Adversarial Queries (Black-Box Access)

This class of attacks assumes that the adversary can interact with a generative model as a black box, probing it to reveal hidden information about the training data. Generative models are particularly vulnerable because they may memorize and reproduce training samples. Hayes et al. [56] showed that GANs can leak identifiable images, while Meehan et al. [99] reported similar issues in diffusion models. Orekondy et al. [111] further demonstrated that adversarial queries can extract private samples even from models trained on anonymized or synthetic datasets. These findings underscore that black-box querying alone may be sufficient for de-anonymization.

4.5 Summary

This chapter reviewed the current state of anonymization techniques in medical imaging, highlighting both traditional and deep learning-based approaches. Classic methods offer simplicity and computational efficiency, but they often compromise diagnostic quality and remain vulnerable to reconstruction

attacks. Deep learning-based approaches, by contrast, enable more sophisticated privacy-preserving strategies. Generative adversarial networks (GANs), variational autoencoders, and more recently, diffusion models have been explored to synthesize realistic medical images or selectively remove personally identifiable features while preserving task-relevant content. Among these, diffusion models stand out for their ability to generate high-resolution and clinically realistic data, leading to the development of a wide range of privacy-aware adaptations, including classifier-free identity guidance and differentially private fine-tuning.

Despite these advances, the chapter also emphasized that anonymization cannot be viewed in isolation from de-anonymization processes. Re-identification attacks demonstrate that biometric traits can persist in anonymized or synthetic images, even when explicit identifiers are removed. Model exploitation attacks such as inversion, membership inference, and data extraction highlight the risks posed by both black-box and white-box access to generative models. Furthermore, adversarial queries reveal that generative models may memorize and leak training samples, raising concerns about their safe deployment in clinical settings.

Finally, several open challenges remain. There is a lack of standardized benchmarks and evaluation protocols for measuring both privacy protection and clinical image utility, which makes it difficult to compare methods fairly. Moreover, balancing privacy guarantees with image fidelity continues to be a central concern, as excessive anonymization can erode diagnostic relevance while insufficient anonymization leaves patients vulnerable to re-identification.

In summary, anonymization methods have developed, but persistent de-anonymization risks demand rigorous evaluation and a balanced focus on both privacy and utility.

Chapter 5

Machine Unlearning Experimental Setup

5.1 Overview

The primary objective of this dissertation is to ensure that deep learning models, particularly in the field of medical imaging, do not generate memorized images from the training set, or images containing memorized patients' Personally Identifiable Information (PII). Achieving this is essential for creating fully anonymous synthetic images that can serve as accurate and interpretable case-based explanations. In addition, it is crucial to uphold patients' "Right to be Forgotten" as outlined in the GDPR, by removing their influence from the model.

Practically, the experiments in this dissertation investigate whether applying machine unlearning to diffusion models enables the removal of patient-identifiable information (PII), thereby ensuring that the generated medical images are entirely anonymous.

Section 5.2 presents the unlearning scenario addressed, along with a method to estimate the difficulty of this particular setting. Section 5.3 presents the model selected to perform unlearning, while Section 5.4 describes the dataset, recognition tasks, and pre-processing steps performed. In Section 5.5, the unlearning methods implemented are presented and discussed. Finally, Section 5.6 introduces the metrics used to evaluate the images generated by these methods. A summary of this chapter is presented in Section 5.7.

5.2 Machine Unlearning Scenario

The unlearning scenario addressed in this dissertation falls into the scope of Data-Point Unlearning, since the goal is to forget specific images from the training set related to patients whose information was totally or partially memorized.

5.2.1 Difficulty Assessment of Unlearning Scenario

As noted by Fan et al. [38] and described in Subsection 3.3.1.1, not all unlearning requests pose the same level of difficulty. In their work, the most challenging cases correspond to sets containing (i) samples with a high entanglement score (ES), which means sets with the highest similarity to retained data, or (ii) samples that are strongly memorized by the model.

To operationalize this idea in our setting, we calculate the Entanglement Score (ES) using the cosine similarity of image embeddings.

Let $\mathbf{e}_f$ and $\mathbf{e}_r$ denote ℓ_2-normalized embedding vectors of a forget image I_f and a retain image I_r, respectively. The pairwise ES is given by:

$$\mathrm{ES}(I_f, I_r) = \frac{1 + \cos(\mathbf{e}_f, \mathbf{e}_r)}{2},$$

where

$$\cos(\mathbf{e}_f, \mathbf{e}_r) = \frac{\mathbf{e}_f \cdot \mathbf{e}_r}{\|\mathbf{e}_f\| \, \|\mathbf{e}_r\|}.$$

By construction, $\mathrm{ES} \in [0, 1]$: values close to 1 indicate strong entanglement (a forget image lies near a retain image in the embedding space), while values near 0 indicate strong separability.

In practice, the embeddings are extracted using the autoencoder (VAE) of the generative model chosen.

5.3 Model Choice

When viewed as a medical image anonymization strategy, the unlearning techniques explored in this dissertation fall into the category of post-model anonymization methods, since they are applied after the model has been trained by defining a Forget and a Remain set with the generated synthetic data.

For image generation, a latent diffusion model based on the Medfusion by Müller-Franzes et al. [106] architecture, described in Subsection 2.6.4, was chosen. This model has shown good performance in producing realistic synthetic medical images and has already been used in previous studies, such as in *Towards Case-based Interpretability for Medical Federated Learning* by Latorre et al. [80] and in *Latent Diffusion Models for Privacy-preserving Medical Case-based Explanations* by Campos et al. [11], to generate case-based explanations. Its suitability for this task makes it an appropriate backbone for the unlearning experiments presented here. Consequently, unlike the unlearning techniques analyzed in Chapter 3, which operate in pixel space, the techniques for unlearning in this model operate in latent space.

5.4 Dataset and Task Choice

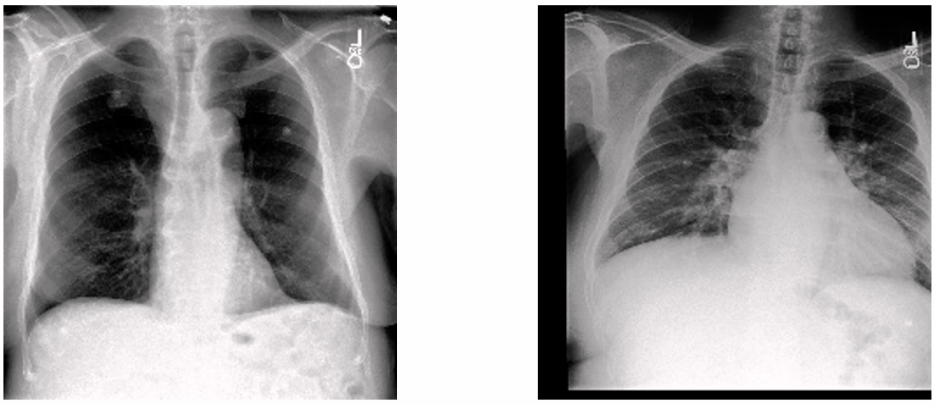
The problem addressed in this dissertation is task-agnostic, meaning that the proposed method can be applied to any two-dimensional medical imaging classification problem. For experimental evaluation, we focus on chest X-ray classification, given its widespread use in clinical practice and its importance as one of the most common imaging modalities.

Accordingly, this work primarily relies on the MIMIC-CXR-JPG dataset [73], a large-scale collection comprising 377,110 chest radiographs from 65,240 patients. Derived from the original MIMIC-CXR dataset [74], it has been used in prior research and is considered a well-established benchmark for medical imaging studies. The dataset is also considered highly trustworthy as the images were acquired during routine clinical practice at the Beth Israel Deaconess Medical Centre (BIDMC), representing authentic patient populations, and have undergone curation and HIPAA-compliant de-identification, such as removal of direct identifiers, to ensure both data quality and patient privacy.

The images were automatically annotated using the CheXpert tool [68], which extracts 14 common thoracic findings from free-text radiology reports. Each observation is labeled as positive (1), negative (0), uncertain (-1), or left blank if not mentioned. To ensure reliable supervision, a U-Ignore strategy was followed, using only clearly positive or negative labels while discarding uncertain and blank cases. The scale and clinical diversity of the dataset, which covers a wide range of demographics, conditions, and acquisition settings, make it a valuable resource to develop robust deep learning models.

For the classification and recognition task, Cardiomegaly was selected as the target disease due to its prevalence in the dataset and its clear visual presentation. Cardiomegaly is characterized by an enlarged heart, and Figure 5.1 illustrates the visual difference between a healthy case and one affected by this condition.

As this disease is most reliably diagnosed by posteroanterior chest X-rays (PA), since anteroposterior (AP) views tend to exaggerate heart size and hinder accurate detection, only PA images were considered to train the generative model. Consequently, the diffusion model was trained with *15223 images* from *10156* different patients.



(a) No Cardiomegaly. (b) Cardiomegaly.

Figure 5.1: Comparison between a healthy X-ray (a) and one that presents Cardiomegaly (b). Both images were obtained from the MIMIC-CXR-JPG [73] dataset.

5.4.1 Data Preprocessing: Eliminating Mislabeled Samples

During initial experiments with the MIMIC-CXR-JPG dataset, it was observed that the unlearning techniques led the models to generate synthetic images resembling lateral views rather than the intended postero-anterior (PA) views, as is possible to see in Figure 5.2. A closer inspection of the model's output before unlearning suggested the presence of mislabeled samples in the MIMIC-CXR-JPG dataset, where lateral chest scans had been incorrectly designated as PA views.

To address this issue, an initial attempt was made to identify mislabeled images using automated metrics, such as pixel distance between incorrect original model outputs and all training samples. However, this method did not reveal a sufficient number of images to explain the consistent generation of lateral-view images. As a result, a manual review of the training dataset was performed by arranging the images into grids for visual inspection. This process identified 222 mislabeled lateral view images,

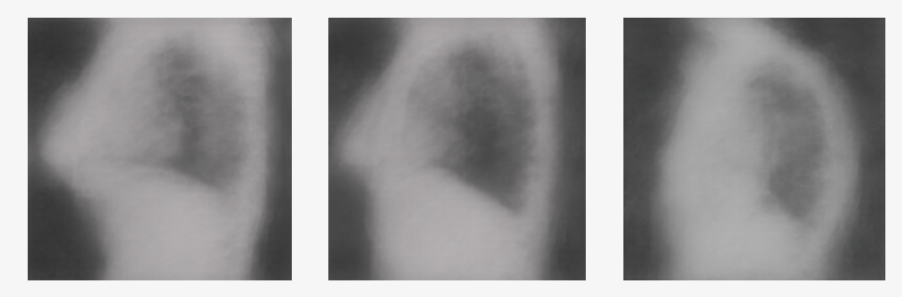


Figure 5.2: Examples of synthetic chest X-rays generated after applying unlearning techniques. Instead of producing postero-anterior (PA) views as intended, the model outputs resemble lateral views because of mislabeled samples in the MIMIC-CXR-JPG [73] dataset.

which were subsequently removed from the training set. The model was then retrained on the corrected dataset.

Figure 5.3 provides an example of the mislabeled images identified and excluded.

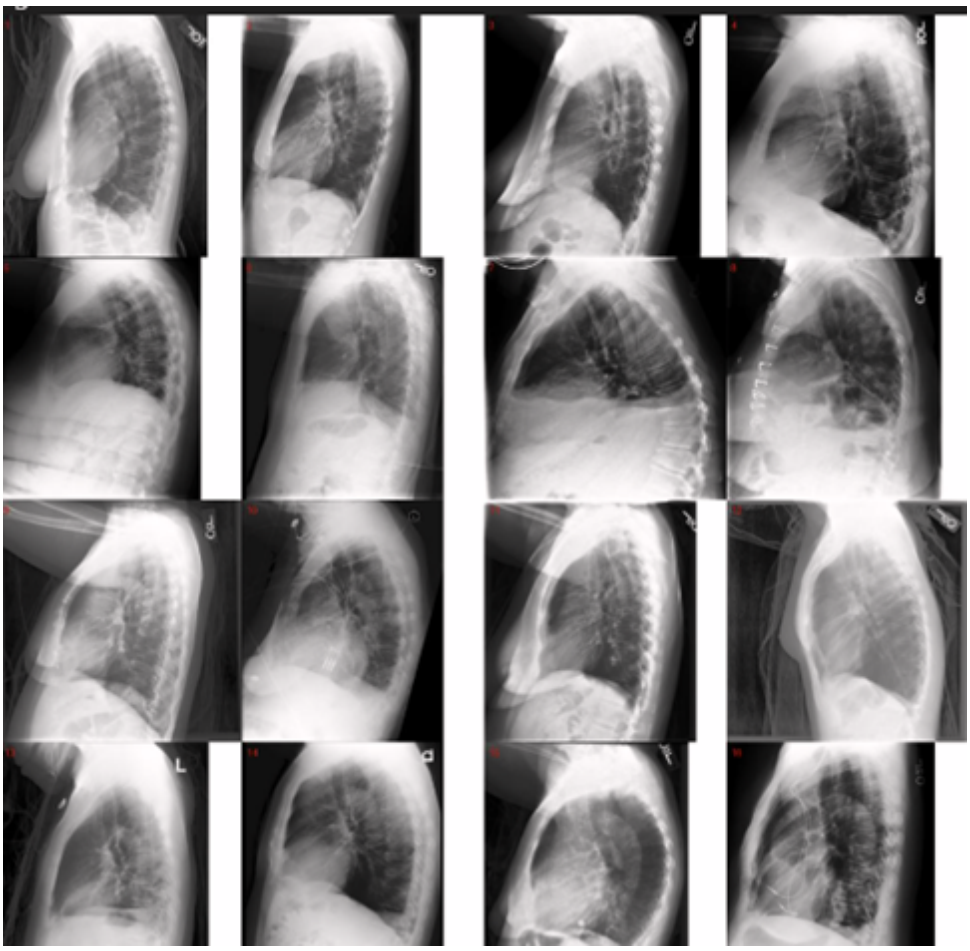


Figure 5.3: Examples of found mislabeled chest X-ray images in the MIMIC-CXR-JPG [74] dataset.

5.5 Machine Unlearning Methods

From the analyzed literature, it was chosen to explore different unlearning techniques that represent different categories of unlearning methods.

5.5.1 Gradient Ascent (GA)

Gradient ascent is a full parameter baseline unlearning method, described in Section 2.5.2, which involves performing a gradient ascent on the Forget set to increase the loss, effectively unlearning the data.

This technique formulates unlearning as maximizing the standard diffusion loss on the Forget set. Given a forget sample x_f at diffusion step t, the noise estimator predicts $\hat{\varepsilon}(x_f, t, c_f)$, where c_f is the original class label. The model is trained to maximize the prediction error by minimizing the negative of the standard diffusion loss.

The unlearning loss is defined as:

$$\mathscr{L}_{\text{GA}} = -\|\varepsilon - \hat{\varepsilon}(x_f, t, c_f)\|^2, \tag{5.1}$$

where ε is the true noise added during the forward diffusion process. By minimizing this negative loss, the model effectively performs gradient ascent on the original objective, increasing prediction errors on the set of images to be forgotten.

To prevent catastrophic forgetting and maintain generalizability, the Remain set is simultaneously used to compute the standard diffusion training loss $\mathscr{L}_{\text{remain}}$. The combined training objective is:

$$\mathscr{L} = \mathscr{L}_{\text{remain}} + \alpha \mathscr{L}_{\text{GA}}, \tag{5.2}$$

where α controls the trade-off between the unlearning objective and knowledge retention.

A pseudo-code of this approach's implementation is presented in Algorithm 1.

Algorithm 1 Gradient Ascent Unlearning for Diffusion Models

Require: Pretrained diffusion model θ, Forget set D_f, retain set D_r, hyperparameter α
Ensure: Unlearned model θ^*

for each training iteration **do**
  Sample batch from retain set: $(x_r, c_r) \sim D_r$
  Sample batch from Forget set: $(x_f, c_f) \sim D_f$
  // Standard diffusion loss on retain set
  $\mathscr{L}_{\text{remain}} \leftarrow \text{DiffusionLoss}(x_r, c_r, \theta)$
  // Gradient ascent loss on Forget set
  Sample noise: $\varepsilon \sim \mathscr{N}(0, I)$, timestep: $t \sim \text{Uniform}(0, T)$
  $\hat{\varepsilon} \leftarrow \text{NoiseEstimator}(x_f, t, c_f; \theta)$
  $\mathscr{L}_{\text{GA}} \leftarrow -\|\varepsilon - \hat{\varepsilon}\|^2$ **// negative loss**
  // Combined objective
  $\mathscr{L} \leftarrow \alpha \cdot \mathscr{L}_{\text{GA}} + \mathscr{L}_{\text{remain}}$
  Update θ using $\nabla_\theta \mathscr{L}$
end for

5.5.2 KL-Divergence based Unlearning (KL-away)

Inspired by the KL–L_2 method proposed by Li et al. [83], described in subsection 3.6.2, we formulate unlearning for diffusion models as a distributional divergence problem and propose KL-away.

Whereas KL–L_2 relaxes the KL-divergence to an L_2 loss by pushing forget samples toward Gaussian noise, KL-away encourages divergence from its own frozen predictions on the Forget set, thereby erasing previously learned information. This also differs from gradient-ascent unlearning, which enforces divergence from the ground-truth noise during training.

Concretely, given a forget sample x_f at diffusion step t, the current model predicts $\hat{\varepsilon}_\theta(x_f,t,c_f)$. A reference prediction $\hat{\varepsilon}_{\theta_0}(x_f,t,c_f)$ is obtained from a frozen copy of the original model θ_0. Under the Gaussian assumption, the KL-divergence between the current and reference distributions reduces to the squared distance between their predicted means (Appendix A.2 provides a detailed derivation).

The unlearning loss is therefore defined as:

$$\mathscr{L}_{\text{unlearn}} = -\tfrac{1}{2}\|\hat{\varepsilon}_\theta(x_f,t,c_f) - \hat{\varepsilon}_{\theta_0}(x_f,t,c_f)\|^2, \tag{5.3}$$

This encourages the current model to diverge from the frozen teacher's outputs on the forget data. Minimizing this loss, therefore, maximizes the distance from the teacher, effectively erasing the model's prior knowledge about D_f.

To maintain generalization, the Remain set is used simultaneously with the standard diffusion loss $\mathscr{L}_{\text{remain}}$. The combined objective becomes:

$$\mathscr{L} = \mathscr{L}_{\text{remain}} + \alpha\,\mathscr{L}_{\text{unlearn}}, \tag{5.4}$$

where α balance unlearning strength and knowledge retention.

A pseudo-code of this approach's implementation is available in Algorithm 2.

Algorithm 2 KL-Divergence Unlearning for Diffusion Models (KL-away)

Require: Pretrained diffusion model θ_0, Forget set D_f, retain set D_r, weights λ, α
Ensure: Unlearned model θ^*

Initialize current parameters $\theta \leftarrow \theta_0$
Freeze a teacher copy $\theta_{\text{teach}} \leftarrow \theta_0$ *// fixed reference, no gradients*
for each training iteration **do**
  Sample batch from Remain set: $(x_r, c_r) \sim D_r$
  Sample batch from Forget set: $(x_f, c_f) \sim D_f$
  // Retain loss (standard diffusion training)
  $\mathscr{L}_{\text{remain}} \leftarrow \text{DiffusionLoss}(x_r, c_r; \theta)$
  // Unlearning loss (diverge from frozen teacher, not ground-truth)
  $\hat{\varepsilon} \leftarrow \text{NoiseEstimator}(x_f, t, c_f; \theta)$
  $\hat{\varepsilon}_0 \leftarrow \text{NoiseEstimator}(x_f, t, c_f; \theta_{\text{teach}})$ *// teacher prediction, fixed*
  $\mathscr{L}_{\text{unlearn}} \leftarrow -\frac{1}{2}\|\hat{\varepsilon} - \hat{\varepsilon}_0\|^2$ *// encourages deviation from teacher outputs*
  // Combined objective
  $\mathscr{L} \leftarrow \mathscr{L}_{\text{remain}} + \alpha \cdot \mathscr{L}_{\text{unlearn}}$
  Update θ using $\nabla_\theta \mathscr{L}$
end for
return θ

5.5.3 Subtracted Importance Sampled Scores (SISS)

This technique is a full-parameter unlearning technique that aims to unlearn the Forget set while preserving the performance on the Remain set. It was chosen as it is the only unlearning technique found in literature that has the same unlearning scenario, data-point forgetting, in I2I Diffusion Models, and unlike KL–L_2, doesn't lead to full degradation of the content we aim to forget.

As it was described and formulated in Subsection 3.6.3, this technique combines naive deletion (fine-tuning on retained data) with gradient ascent on the Forget set through importance sampling, enabling computationally efficient unlearning with a single forward pass per training iteration. A pseudocode of this approach implementation is available in Algorithm 3.

Algorithm 3 SISS Unlearning for Diffusion Models

Require: Pretrained diffusion model parameters θ; Forget set D_f of size k; Retain set D_r with total dataset size n; mixture parameter $\lambda \in (0,1)$ (e.g., 0.5); superfactor $s \geq 0$;

Ensure: Unlearned model θ^*

1: **for** each training iteration **do**

2: Sample minibatch from retain set: $x_r \sim D_r$

3: Sample minibatch from forget set: $x_f \sim D_f$

4: Sample timestep: $t \sim \text{Uniform}\{0,\ldots,T\}$

5: Define the mixture density:

$$q_\lambda(m_t \mid x_r, x_f, t) \leftarrow (1-\lambda)\, q(m_t \mid x_r, t) + \lambda\, q(m_t \mid x_f, t)$$

6: Sample a noisy point from the mixture: $m_t \sim q_\lambda(\cdot \mid x_r, x_f, t)$

7: Compute importance weights:

$$w_r \leftarrow \frac{n}{n-k}\,\frac{q(m_t \mid x_r, t)}{q_\lambda(m_t \mid x_r, x_f, t)}, \qquad w_f \leftarrow \frac{k}{n-k}\,\frac{q(m_t \mid x_f, t)}{q_\lambda(m_t \mid x_r, x_f, t)}$$

8: Compute per-term targets (no extra forward pass needed):

$$\varepsilon_r \leftarrow \frac{m_t - \gamma_t x_r}{\sigma_t}, \qquad \varepsilon_f \leftarrow \frac{m_t - \gamma_t x_f}{\sigma_t}$$

9: **Single forward pass:** $\varepsilon_\theta \leftarrow \text{NoiseEstimator}(m_t, t; \theta)$

10: Loss:

$$\mathscr{L}_{\text{SISS}} \leftarrow w_r \left\|\varepsilon_r - \varepsilon_\theta\right\|_2^2 - (1+s)\, w_f \left\|\varepsilon_f - \varepsilon_\theta\right\|_2^2$$

11: Update parameters: $\theta \leftarrow \theta - \eta\, \nabla_\theta \mathscr{L}_{\text{SISS}}$

12: **end for**

5.5.4 Saliency Unlearning (SalUn)

Saliency-Guided Unlearning (SalUn), by Fan et al. [39], described in Subsection 3.6.1, is a localized unlearning method that focuses on unlearning specific parts of the model that are most influenced by the Forget set, based on saliency maps. It was chosen as it has been considered in the literature as a state-of-the-art method in classification and image generation tasks, for class and concept forgetting, and because it has been applied to medical data in classification tasks before by Nasirigerdeh et al.

[108]. For this technique, in addition to using the gradient ascent as proposed in the original paper, it was also experimented to combine the saliency mask with the presented KL-Away loss. Note that although random labeling is considered the best method to combine with the saliency mask according to the authors, label manipulation methods do not make sense in this case of data-point forgetting, as the images to forget are not related to a label. A pseudocode of saliency unlearning using gradient ascent and using KL-Away is available in Algorithm 4 and in Algorithm 5, respectively.

Algorithm 4 Saliency-Guided Unlearning (SalUn) with Gradient Ascent (GA)

Require: Pretrained diffusion model θ_0, Forget set D_f, retain set D_r, mask threshold γ, learning rate η, iterations T

Ensure: Unlearned model θ^*

// Phase 1: Saliency mask computation (fixed at initialization)
$g_S \leftarrow \nabla_\theta \ell_{\text{MSE}}(\theta; D_f)\big|_{\theta=\theta_0}$
$m_S \leftarrow \mathbb{1}(|g_S| \geq \gamma)$
$\theta \leftarrow \theta_0$
for $t = 1$ **to** T **do**
  Sample retain batch: $(x_r, c_r) \sim D_r$
  Sample forget batch: $(x_f, c_f) \sim D_f$
  // Retain loss
  $\mathscr{L}_{\text{retain}} \leftarrow \text{DiffusionLoss}(x_r, c_r; \theta)$
  // Forget loss (gradient ascent)
  $\mathscr{L}_{\text{forget}} \leftarrow -\text{DiffusionLoss}(x_f, c_f; \theta)$
  // Combined objective and masked update
  $\mathscr{L} \leftarrow \mathscr{L}_{\text{retain}} + \alpha \mathscr{L}_{\text{forget}}$
  $g \leftarrow \nabla_\theta \mathscr{L}$
  $\theta \leftarrow \theta - \eta \cdot (m_S \odot g)$
end for
return θ

Algorithm 5 Saliency-Guided Unlearning (SalUn) with KL-away Loss

Require: Pretrained diffusion model θ_0, Forget set D_f, retain set D_r, mask threshold γ, learning rate η, iterations T

Ensure: Unlearned model θ^*

// Phase 1: Saliency mask computation (fixed at initialization)
$g_S \leftarrow \nabla_\theta \ell_{\text{MSE}}(\theta; D_f)\,\big|_{\theta=\theta_0}$
$m_S \leftarrow \mathbb{1}(|g_S| \geq \gamma)$
$\theta \leftarrow \theta_0$
Freeze teacher $\theta_{\text{teach}} \leftarrow \theta_0$
for $t = 1$ **to** T **do**
  Sample retain batch: $(x_r, c_r) \sim D_r$
  Sample forget batch: $(x_f, c_f) \sim D_f$
  // Retain loss
  $\mathscr{L}_{\text{retain}} \leftarrow \text{DiffusionLoss}(x_r, c_r; \theta)$
  // Forget loss (KL-away, diverge from teacher)
  $\hat{\varepsilon} \leftarrow \text{NoiseEstimator}(x_f, t, c_f; \theta)$
  $\hat{\varepsilon}_0 \leftarrow \text{NoiseEstimator}(x_f, t, c_f; \theta_{\text{teach}})$
  $\mathscr{L}_{\text{forget}} \leftarrow -\frac{1}{2}\|\hat{\varepsilon} - \hat{\varepsilon}_0\|^2$
  // Combined objective and masked update
  $\mathscr{L} \leftarrow \mathscr{L}_{\text{retain}} + \alpha \mathscr{L}_{\text{forget}}$
  $g \leftarrow \nabla_\theta \mathscr{L}$
  $\theta \leftarrow \theta - \eta \cdot (m_S \odot g)$
end for
return θ

5.6 Evaluation Metrics

To evaluate if the proposed unlearning methods still ensure that the model produced anonymous, yet high-quality and clinically meaningful, it was chosen metrics that can be grouped into three main categories: **Image Quality Metrics**, which measure the realism and diversity of the synthetic images generated after unlearning; **Image Utility Metrics**, which evaluate if the generated images remain useful to serve as case-based explanations; **Image Anonymization and Unlearning Metrics**, which quantify the unlearning and anonymization of the generated images.

For Image Quality Metrics, the following metrics were chosen:

- **Fréchet Inception Distance (FID)**: Measures the distance between the distributions of real and generated images. Lower values indicate better quality.
- **Precision (P)**: Measures the fraction of generated images that are realistic (i.e., lie in the manifold of real images). Higher values indicate better quality.

To measure the image utility after unlearning, it was chosen to train a DenseNet121 classifier to detect Cardiomegaly in chest X-rays. The classifier was trained on the synthetic dataset generated by each unlearned model and evaluated on a test set with images from the MIMIC-CXR-JPG dataset. The following metrics were used to evaluate the classifier's performance:

- **Area Under the Receiver Operating Characteristic Curve (AUC)**: Measures the ability of the classifier to distinguish between classes. Higher values indicate better performance.

- **Accuracy (ACC) (%)**: Measures the overall accuracy of the classifier. Higher values indicate better performance.
- **F1 Score (F1) (%)**: Harmonic mean of precision and recall. Higher values indicate better performance.

For Unlearning Metrics and Image Anonymization metrics, the following metrics were chosen:

- **Number of Non-Anonymous Images Produced**: The total count of images in the dataset that remain non-anonymized after the anonymization process. Lower values indicate better anonymization.
- **Number of Patients Identified**: The number of unique patients that can still be identified within the dataset after anonymization.Lower values indicate better anonymization.
- **Number of Patients Re-Identified**: The number of patients that were originally anonymized but later re-identified by the unlearned model. Lower values indicate better anonymization.
- **Number of New Patients Identified**: The number of patients not originally present in the identification pool, but newly identified after model unlearning. Lower values indicate better anonymization.
- **Unlearning Accuracy (%)**: Corresponds to the proportion of patients anonymized when considering only those originally identified, computed as shown in Equation 5.5. Higher values indicate better anonymization.

$$\text{Unlearning Accuracy} = \frac{\text{Patients Originally Identified} - \text{Patients Re-Identified by the Unlearned Model}}{\text{Patients Originally Identified}} \tag{5.5}$$

- **Overall Patient Identifiability Ratio (%)**: The proportion of patients whose identity can be inferred from the output images, indicating residual information leakage. It is computed as shown in Equation 5.6. Lower values indicate better anonymization.

$$\text{Overall Patient Identifiability Ratio} = \frac{\text{Patients Identified}}{\text{Dataset Size}} \tag{5.6}$$

- **Overall Dataset Anonymization (%):** The proportion of anonymized images in the dataset, computed as demonstrated in Equation 5.7. Higher values indicate better anonymization.

$$\text{Overall Dataset Anonymization} = \frac{\text{Datase Size - Non-Anonymous Images Produced}}{\text{Dataset Size}} \tag{5.7}$$

- **Average Negative Log Likelihood of the Forget set (NLL):** The NLL is used to quantify the likelihood that a model considers the data points it was trained on. Higher average values in the Forget set serve as a direct signal of effective unlearning. This metric was described in Section 3.3.3.

Note that the metrics related to anonymization and unlearning (except for the NLL metric) are computed by performing a patient re-identification attack, inspired by the Anonymization Pipeline described in Section 4.3.4. This attack, which is also used to define the Forget and Retain sets for the unlearning methods, will be further detailed in Chapter 6.

5.7 Summary

This chapter presented the experimental setting adopted to perform and evaluate machine unlearning in the context of generating fully anonymous case-based explanations by removing the influence of specific patients from the diffusion model, guaranteeing the RTBF and compliance with data privacy laws.

As a model to perform unlearning, a latent diffusion model based on the Medfusion architecture was selected, as it has shown strong performance in producing realistic synthetic chest X-rays. The primary dataset used is MIMIC-CXR-JPG, restricted to postero-anterior (PA) views and further refined by removing mislabeled samples, and Cardiomegaly detection was chosen as the downstream utility task.

The unlearning scenario is framed as a Data-Point Forgetting Scenario, and its difficulty is assessed using an Entanglement Score (ES) that quantifies similarity between the Forget and Remain sets in the model's latent space. Four unlearning methods were adapted and implemented: Gradient Ascent (GA), KL-Divergence-based unlearning (KL-away), Subtracted Importance Sampled Scores (SISS), and Saliency-Guided Unlearning (SalUn).

Finally, the evaluation strategy uses image quality and utility metrics, as well as image anonymization and unlearning metrics, allowing evaluation of both effectiveness and side effects of unlearning in generative medical imaging.

Chapter 6

Generate anonymous case-based explanations with Machine Unlearning

6.1 Overview

This chapter investigates machine unlearning techniques to mitigate patient privacy leakage in synthetic medical image generation. We focus on identifying and removing memorized patient-identifiable information (PII) from diffusion-based generative models, while preserving diagnostic utility. To achieve this, we first detect memorized patients through a re-identification attack, and then apply different unlearning strategies to remove their influence from the model, aiming to achieve anonymization without compromising explanatory value.

6.2 Method

One of the main challenges when using machine unlearning to make models stop generating synthetic images leaking patients' identifiable information (PII) is to define the Forget set and Remain set properly. We considered it crucial to ensure that the Remain set is properly anonymized, so patients whose PII was not leaked initially would not be compromised during the generalization step.

To identify patients whose identity features have been memorized or leaked, we followed the post-model anonymization approach used by Campos et al. [11] and Packhäuser et al. [112], which employs a patient re-identification attack.

As described in Subsection 4.3.4, this approach trains two Siamese Networks: a Verification Network and a Patient Retrieval Network. For each synthetic image, the Patient Retrieval Network retrieves the three closest real training images (Rank-1, Rank-2, Rank-3) based on embedding similarity. The Verification Network then compares each synthetic image with its Rank-1 match and computes a similarity score. If this score exceeds a predefined threshold of 50%, the synthetic image is flagged as non-anonymous, since it may correspond to the same patient.

Therefore, to define the Forget and Remain sets, we adopt the following procedure: synthetic images flagged as anonymous are assigned to the Remain set, while the Forget set consists of all real patient training images for which at least one corresponding synthetic image exceeded the similarity threshold

of 50%. In simpler terms, if a synthetic image is found to be similar to a real image from a patient, the Forget set will include all images belonging to that patient that were used in training.

After proceeding with unlearning on the model, using the Forget set and Remain set created, we generate samples using the same seed used to generate the original synthetic dataset.

Then these generated images are submitted to the same patient re-identification attack for unlearning and anonymization evaluation.

In Figure 6.1 is possible to see the proposed pipeline to define the Forget Set and Remain set to allow the deletion of memorized patients' PII.

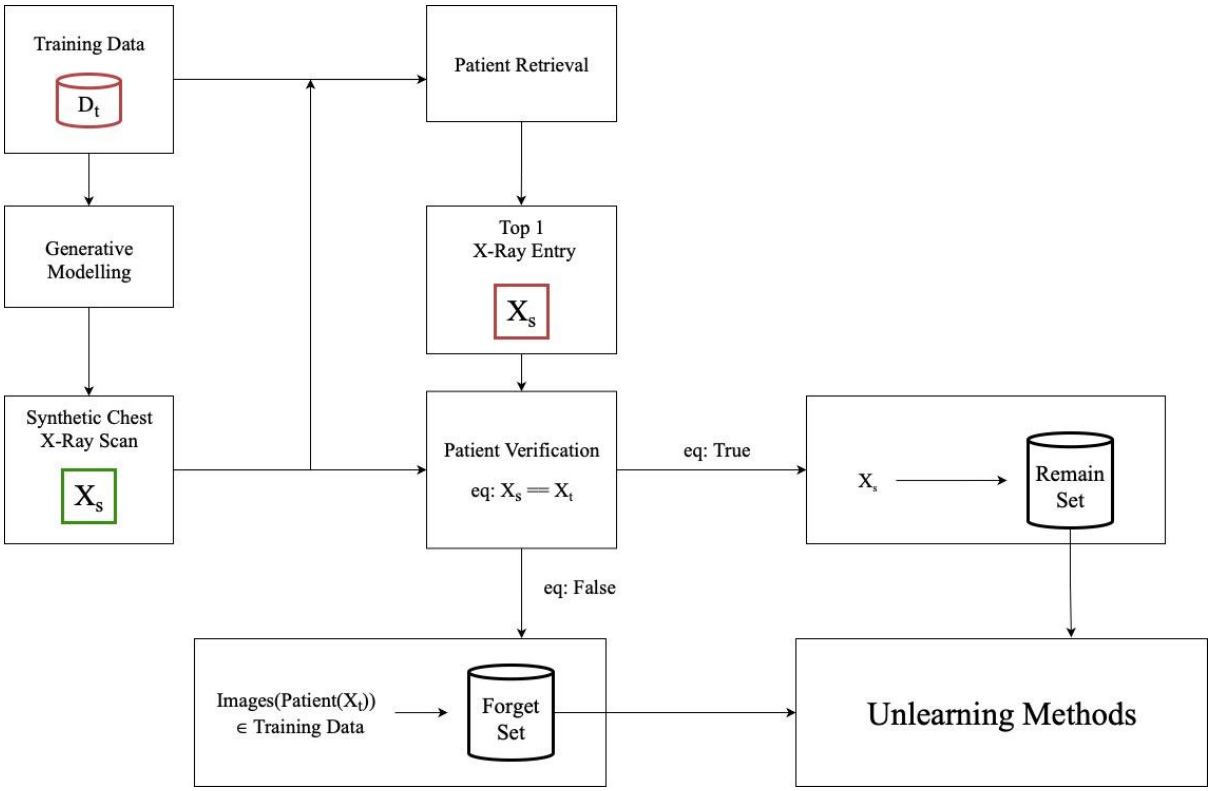


Figure 6.1: Pipeline to allow deletion of memorized patients' PII. Synthetic images from the original model are first evaluated using a patient re-identification attack with a Retrieval and a Verification Network. If the similarity between a synthetic image and its Rank-1 real match exceeds 50%, the corresponding patient is marked as memorized, and all its images used in training are added to the Forget set. Otherwise, the synthetic image is added to the Remain set. After unlearning, new synthetic samples are generated with the same seeds and re-evaluated to assess anonymization and unlearning effectiveness.

6.3 Experiments

Image Generation: Images were generated using the Medfusion model [106] architecture. Following the training settings of the authors, the latent embeddings are encoded using a VAE, which accepts $3 \times 512 \times 512$ pixel images and has 8 embedding channels. The diffusion model uses DDIM, has time and label embeddings of size 1024, and uses a scaled linear noise scheduler with $T = 1000$, which varies from $B_0 = 0.002, B_T = 0.02$. The model is trained for a maximum of 1001 epochs with an early stopping patience of 30 epochs. We generate 4000 synthetic images for the MIMIC-CXR-JPG [73], with a balanced split between positive and negative cases of Cardiomegaly.

Patient Retrieval and Verification Networks: Similar to what was employed by Packhäuser et al. [112] and Campos et al. [11], the verification model was trained for a maximum of 100 epochs with an early stopping criterion of 5 epochs, using the Adam optimizer with $3 \times 256 \times 256$ images and a learning rate of 0.0001. The retrieval network was trained for 30 epochs in the first phase, during which the model backbone is frozen, and 50 epochs during the second phase, where all parameters are unfrozen. In both phases, a learning rate of 0.158489 with weight decay of $1e^{-5}$ was used. The verification model was trained using the Adam optimizer with $3 \times 256 \times 256$ images and a learning rate of 0.0001, and for a maximum of 100 epochs with an early stopping patience criterion of 5 epochs.

Unlearning Methods: Each unlearning technique was limited to be performed for a maximum of 100 epochs, with an early stopping patience criterion of 10 epochs. This choice was motivated by the recommendations of Xue et al. [148], who argue that the unlearning process should be precise and efficient compared to a full retraining of the model and therefore suggest structuring the unlearning steps to approximately 10% of the total training epochs.

6.4 Results and Discussion

Image Generation: Tables 6.1 and 6.2 present the evaluation of image quality and utility, respectively, for images generated by the model before unlearning. The same metrics are used to assess images generated after applying the unlearning methods, allowing these pre-unlearning values to serve as a baseline for understanding the effects of the unlearning techniques.

Table 6.1: Quantitative evaluation of the images generated using the Medfusion model before unlearning.

Synthetic dataset size	FID (↓)	P (↑)
4000	53.50	80.57

Table 6.2: Utility evaluation of the images generated using the Medfusion model before unlearning.

Synthetic dataset size	ACC (↑)	F1-Score (↑)	AUC (↑)
4000	84.16%	72.21%	0.88

Retrieval and Verification Networks: The patient retrieval and verification models successfully allow patient re-identification, as shown in Tables 6.3 and 6.4, respectively. To evaluate the verification model, the Area under the Receiver Operating Characteristic (AUC), Precision (P), Recall (R), and F1 score were calculated. Equation 6.1 calculates R-precision as the number of relevant images retrieved (r) inside the first R images, where R represents the total number of relevant images in the dataset. Equation 6.2 shows the $mAP@R$, which is the average precision at R ($AP@R$) for each query. During this process, 885 images from the original set were identified as non-anonymous, containing personally identifiable information (PII) from 631 different patients. The total Remain set consisted of 3115 images.

$$\text{R-Precision} = \frac{r}{R} \tag{6.1}$$

$$mAP@R = \frac{1}{Q}\sum_{i=1}^{Q} AP_i@R \tag{6.2}$$

Table 6.3: Quantitative evaluation of patient retrieval model.

mAP@R (↑)	**P@1 (↑)**	**R-Precision (↑)**
89.47	97.13	89.95

Table 6.4: Quantitative evaluation of patient verification model.

AUC (↑)	**ACC (↑)**	**F1 (↑)**	**P (↑)**	**R (↑)**
99.06	96.06	92.29	93.12	91.47

Unlearning Methods The overall evaluation results for the unlearning methods employed are presented in Table 6.5. The metrics for the approaches using Gradient Ascent and Saliency Unlearning with Gradient Ascent are not included because, as it is possible to see in Figure 6.2, they led to over-forgetting. The images progressively degrade into unstructured noise or completely blacked-out areas, demonstrating that the unlearning procedure was too aggressive, causing the model to lose all meaningful representations rather than selectively forgetting the target information. These results indicate that Gradient Ascent may be inappropriate for machine unlearning in this context, as it lacks the necessary regularization to maintain model stability during the forgetting process.

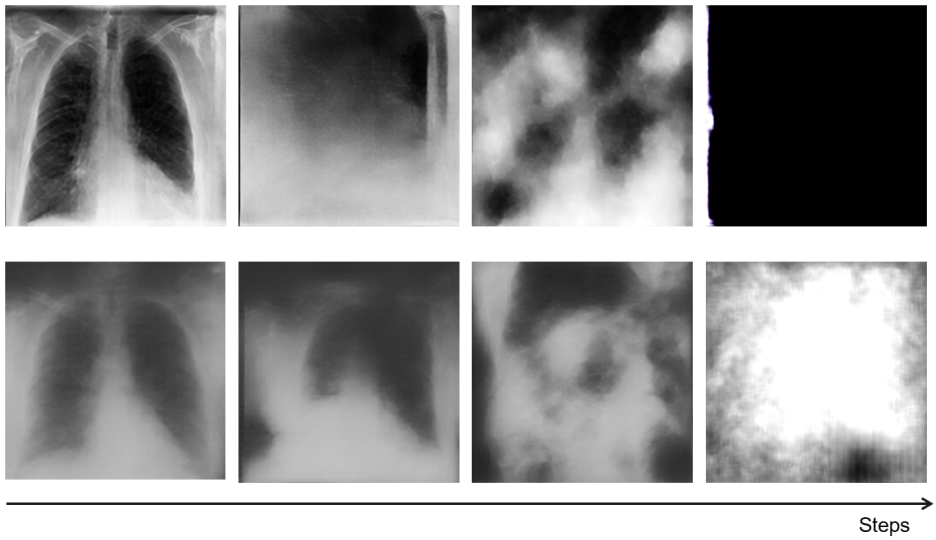


Figure 6.2: Images produced by the mixup of the SalUn localization strategy with Gradient Ascent (On the top row) and by Gradient Ascent Unlearning Algorithm (On bottom row).

Anonymization and Unlearning Analysis When it comes to anonymization and unlearning, it is possible to see that all unlearning strategies improved anonymization compared to the baseline model. Specifically, the number of non-anonymous images and re-identified patients decreased across all methods, with SISS achieving the highest values for unlearning and anonymization. Furthermore, the consistent increase in NLL across methods indicates that the models are effectively diverging from memorized patient-specific information and performing Unlearning. In Figure 6.5 is presented a visual representation of how different approaches anonymize the same synthetic image, associated with a patient.

Our results indicate that full-parameter approaches (KL-Away and SISS) achieved slightly better anonymization than localized unlearning methods such as those based on SalUn.

Even when looking at the results using SalUn with different sparsity masks, it is possible to see that lower sparsity led to better unlearning performance. This suggests that a less conservative selection of salient weights, where a larger fraction of parameters are updated during the unlearning steps, improves the model's ability to erase PII from the patients selected. In practice, higher sparsity corresponds to a threshold that restricts updates to a smaller subset of parameters. This suggests that restrictive, highly localized updates may be insufficient to allow the full removal of the PII from the patients selected.

Figure 6.3 shows examples where KL-Away achieves better anonymization than the KL-Away + SalUn variants.

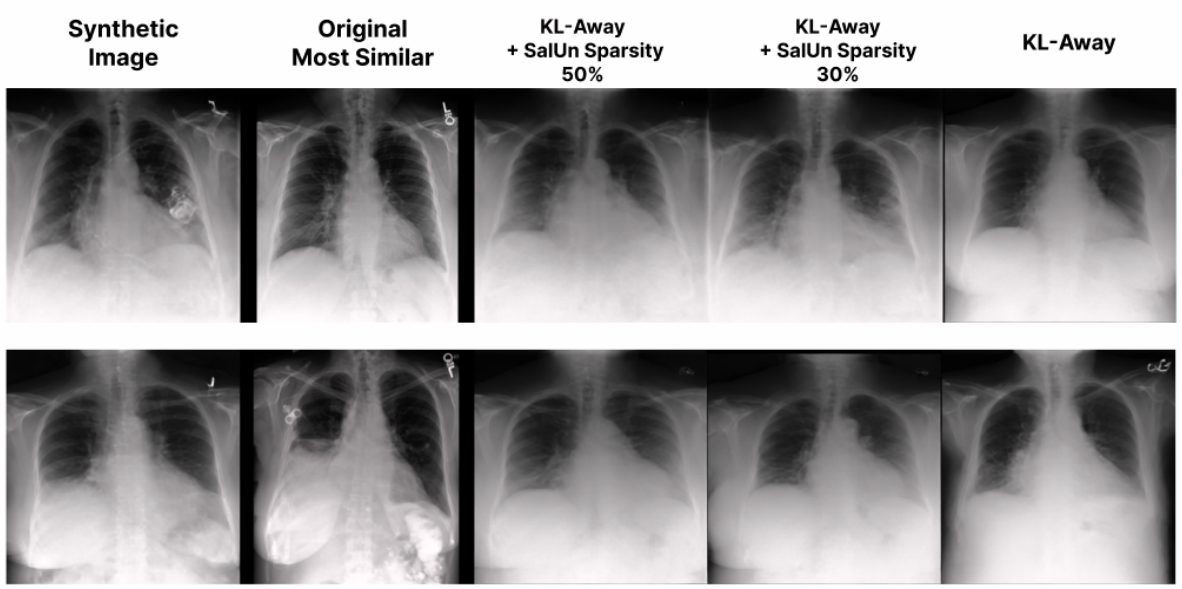


Figure 6.3: Comparison of anonymization quality. KL-Away effectively removes patient-specific features, while KL-Away + SalUn approaches fail to anonymize.

Image-Quality and Utility Analysis In general, we can conclude that all unlearning techniques led to a decrease in the image quality and utility of the generated datasets, when compared to the synthetic dataset produced by the model before unlearning. However, the results show that the SalUn method with 50% sparsity is the best at preserving diagnostic performance and image utility and fidelity among the different unlearning methods. In contrast, increasing sparsity to 30% of the mask or using KL-Away alone results in a gradual decline in both utility and quality metrics, with SISS showing the

strongest deterioration, with accuracy and F1 dropping below 50% and image quality metrics such as FID reflecting severe degradation.

Trade-Off Between Image Quality and Utility vs Anonymization Although SISS was initially expected to be the most suitable algorithm for this task, the results show that while it achieves the strongest anonymization performance, with the lowest identifiability rates and the highest unlearning accuracy (81.30%), this comes at a significant cost: as unlearning strength increases, particularly in SISS, there is a degradation in image utility metrics (ACC, F1, and AUC) and image quality, suggesting that stronger anonymization reduces diagnostic fidelity. The degradation of image quality after unlearning with the SISS method is showcased in Figure 6.4. In contrast, KL-Away and its variants (SalUn+KL-Away) offer a more favorable balance, attaining high anonymization levels while decreasing less the overall diagnostic performance. From all of these techniques, Salun+KL-Away with a 50% sparsity mask may be the most balanced method. These findings highlight a trade-off between anonymization and usability, indicating that although SISS maximizes privacy, approaches like Salun+KL-Away are preferable when both privacy and diagnostic fidelity must be maintained.

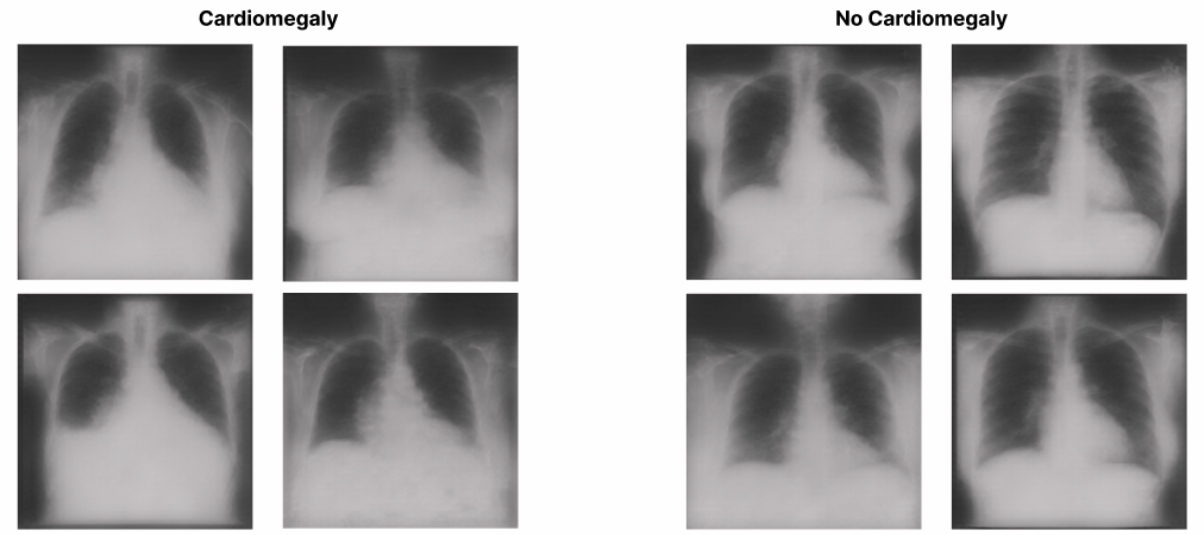


Figure 6.4: Examples of the images produced with the SISS unlearning technique, showcasing the image quality degradation.

Why couldn't we forget all patients originally identified? Although the overall anonymization of the dataset improved with unlearning, and the patient identifiability ratio decreased, it is noticeable that not all patients were successfully forgotten. This showcases the need for an unlearning technique that applies machine unlearning to a set of different entities at once in I2I diffusion models. The found literature that applied unlearning in Image-to-Image diffusion models [3] only performed the unlearning entity by entity, and not at once, which can be much less efficient when we want to forget a larger number of entities. Another explanation for the incomplete forgetting of patients is the inherent difficulty of this unlearning task. Forgetting specific, memorized PII is a much finer-grained task than forgetting an entire class like "car" or a concept like "nudity". In our setup, the Forget set consists of training images whose PII was memorized and leaked in the model's outputs. Moreover, the average entanglement score

between the Forget and Remain sets, computed as described in subsection 5.2.1, is 76%, which is a relatively high degree of entanglement. According to Fan et al. [38], sets with high entanglement scores or those with samples that are strongly memorized by the model are particularly challenging to unlearn, since their influence is more deeply embedded in the model. Our unlearning scenario satisfies both of these conditions, which may explain, at least in part, why some patients could not be fully forgotten.

Why are new patients being re-identified When analyzing the results, it is possible to see that the machine unlearning techniques compromise the PII of patients who were not identified initially. At first, this result seems inconsistent, given that the Remain set was explicitly designed to be anonymous to mitigate the risk of re-identification of new patients. A reason for this is that relying on a subset of synthetic images, treating only the patients revealed in that subset as the patients "memorized", logically does not guarantee that the other patients are not still memorized by the model. So, although these unlearning methods allow for achieving better anonymization on the subset of patients initially identified, it does not guarantee that complete anonymization for all the patients involved in training, and which PII may have been memorized, but not leaked in the initial subset generated. Moreover, our current approach to construct the Forget and Remain sets only considers the rank-1 (most similar) retrieved image to determine patient correspondence. However, synthetic images may leak information from multiple patients simultaneously. A synthetic image flagged as non-anonymous based on similarity to one patient may also contain features from other patients (rank-2, rank-3 matches) that are not captured in our Forget set. This means some patient information remains embedded in the model through these unaccounted associations, and is then revealed by the unlearning methods. Further experiments in Chapter 7 will investigate alternative strategies for Forget and Remain set definition to try to address this limitation.

6.5 Summary

In this chapter, we explored the use of machine unlearning to mitigate privacy risks in diffusion-based medical image generation. By defining Forget and Remain sets through a patient re-identification attack using retrieval and verification networks, we identified patients whose personally identifiable information (PII) was memorized by the model and selectively applied unlearning techniques to remove such information.

Our experiments showed that while unlearning consistently reduced patient identifiability and improved anonymization, it also introduced a trade-off with image quality and diagnostic utility. Full-parameter approaches such as KL-Away and SISS proved to be more effective at eliminating PII than localized strategies, such as SalUn, which may reflect the distributed nature of memorization in latent diffusion models. However, stronger anonymization methods, particularly SISS, also led to significant degradation in image fidelity and diagnostic performance.

Overall, KL-Away and its variants offered a more balanced compromise, achieving high levels of anonymization while retaining diagnostic relevance. However, their inability to completely forget all initially identified patients underscores both the need for new unlearning methods capable of forgetting multiple entities simultaneously and the inherent challenges of unlearning in this setting.

Table 6.5: Evaluation of the datasets' images by the different unlearning metrics across the various metrics.

Metric	**Synthetic**	**SalUn + KL-Away Sparsity 50%**	**SalUn + KL-Away Sparsity 30%**	**KL-Away**	**SISS**
Anonymization and Unlearning Metrics:					
Number of non-anonymous (↓) images produced	885	766	708	602	**525**
Number of Patients Identified (↓)	631	366	331	289	**271**
Number of Patients Re-identified (↓)	–	147	145	133	**118**
Number of New Patients Identified (↓)	–	192	186	156	**153**
Unlearning Accuracy (%) (↑)	–	76.70%	77.02%	78.92%	**81.30%**
Overall Dataset Anonymization (%) (↑)	77.88%	80.85%	82.23%	84.94%	**86.85%**
Overall Patient Identifiability Ratio (%) (↓)	15.77%	9.15%	8.28%	7.23%	**6.78%**
NLL (↑)	10.82	11.15	11.56	12.02	**21.01**
Image Utility Metrics:					
ACC (%) (↑)	84.16%	**80.09%**	77.16%	76.02%	41.22%
F1-SCORE (%) (↑)	72.21%	**74.83%**	71.22%	70.99%	40.90%
AUC (↑)	0.88	0.85	**0.87**	0.83	0.52
Image Quality Metrics:					
FID (↓)	53.50	**81.15**	85.13	84.91	179.2
P (↑)	80.57%	**78.02%**	75.72%	68.60%	21.02%

Another limitation of the pipeline presented was the fact that although the unlearning techniques employed reduced the number of non-anonymous images and re-identified patients, they also introduced new privacy risks. In particular, some patients who were initially considered anonymous became identifiable after unlearning. This outcome suggests that defining Forget and Remain sets solely based on the subset of synthetic images flagged as leaking PII is insufficient, as other patient identities may still be memorized but were not detected in the initially generated dataset. Therefore, while unlearning enhances privacy overall, it does not guarantee complete protection, highlighting the need for more comprehensive strategies for Forget–Remain set construction.

These findings motivate further exploration of machine unlearning techniques for generating anonymous case-based explanations, as other anonymization methods, such as Disentanglement or Differential

Privacy, currently offer better trade-offs.

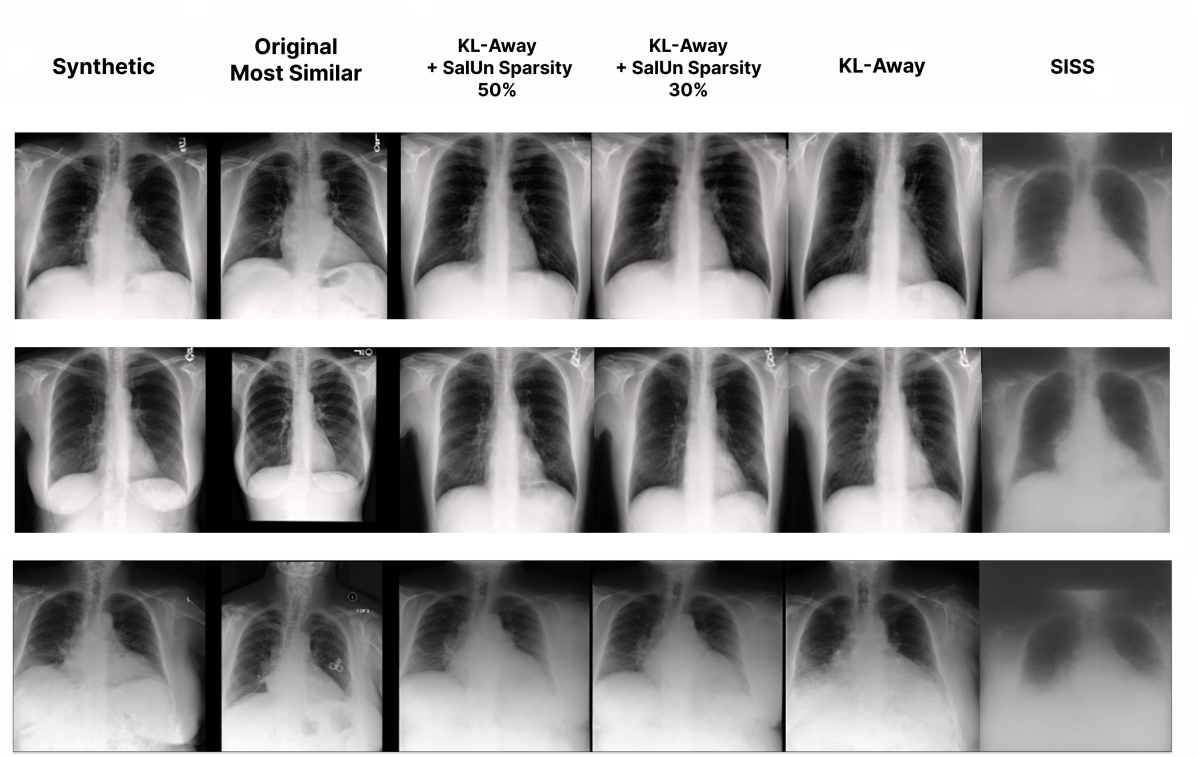


Figure 6.5: Examples where all identified patients could be anonymized by the different unlearning methods. In cases where all KL-Away variants succeed, the resulting images appear visually similar, although KL-Away without a saliency mask introduces more noticeable differences compared to the original synthetic image. The figure also highlights quality and utility degradation with SISS: in the middle row, SISS produces an image with cardiomegaly, reducing clinical utility, whereas the other methods preserve the absence of cardiomegaly.

Chapter 7

Improvement-Oriented Experiments

7.1 Overview

Based on the findings presented in Chapter 6, this chapter addresses the identified limitations and evaluates the hypotheses formulated in response, intending to improve the results obtained. Section 7.2 tests a different method to define the Forget and Remain set, to see if it is possible to mitigate the number of new patients being identified. In Section 7.3, an experiment is carried out to verify if it would be possible to forget some of the patients that could not be forgotten by any of the unlearning techniques tested when applying unlearning individually to them. The chapter ends with Section 7.4 presenting a summary and remarking notes.

7.2 Defining Different Forget and Remain Sets

In Chapter 6, one of the identified limitations was that new patients, who were not initially flagged as having their PII memorized by the model, were identified after unlearning. A possible reason for this outcome is that our method only considered the similarity score of the rank-1 image from the Retrieval Network when defining the forget and Remain sets, which could overlook cases where synthetic images leak information from multiple patients simultaneously. To investigate this hypothesis, we designed an experiment in which the Forget and Remain sets are redefined to account for higher-rank retrievals.

7.2.1 Method

We employed the same generative model, patient retrieval, and verification networks as in Chapter 6. The key difference was in the construction of the Forget and Remain sets: instead of relying only on the rank-1 retrieved image, we computed similarity scores between each synthetic image and its three closest retrieved images (rank-1, rank-2, and rank-3). If the similarity with any of these exceeded the 50% threshold, all corresponding patient images used in training were included in the Forget set. This adjustment aimed to capture cases where synthetic images may embed features from multiple patients simultaneously.

A visual overview of this method is present in Figure 7.1.

Using this method, 1412 images were identified as non-anonymous, resulting in a Remain set of 2588 images. Then, unlearning was performed with these sets, and new samples were generated using

the same seeds as the original synthetic dataset. The resulting datasets were evaluated with the same patient re-identification attack, along with image quality and utility metrics.

Given the collapse observed with Gradient Ascent and its saliency-based variant (SalUn + GA) in Chapter 6, these techniques were excluded from the present experiments.

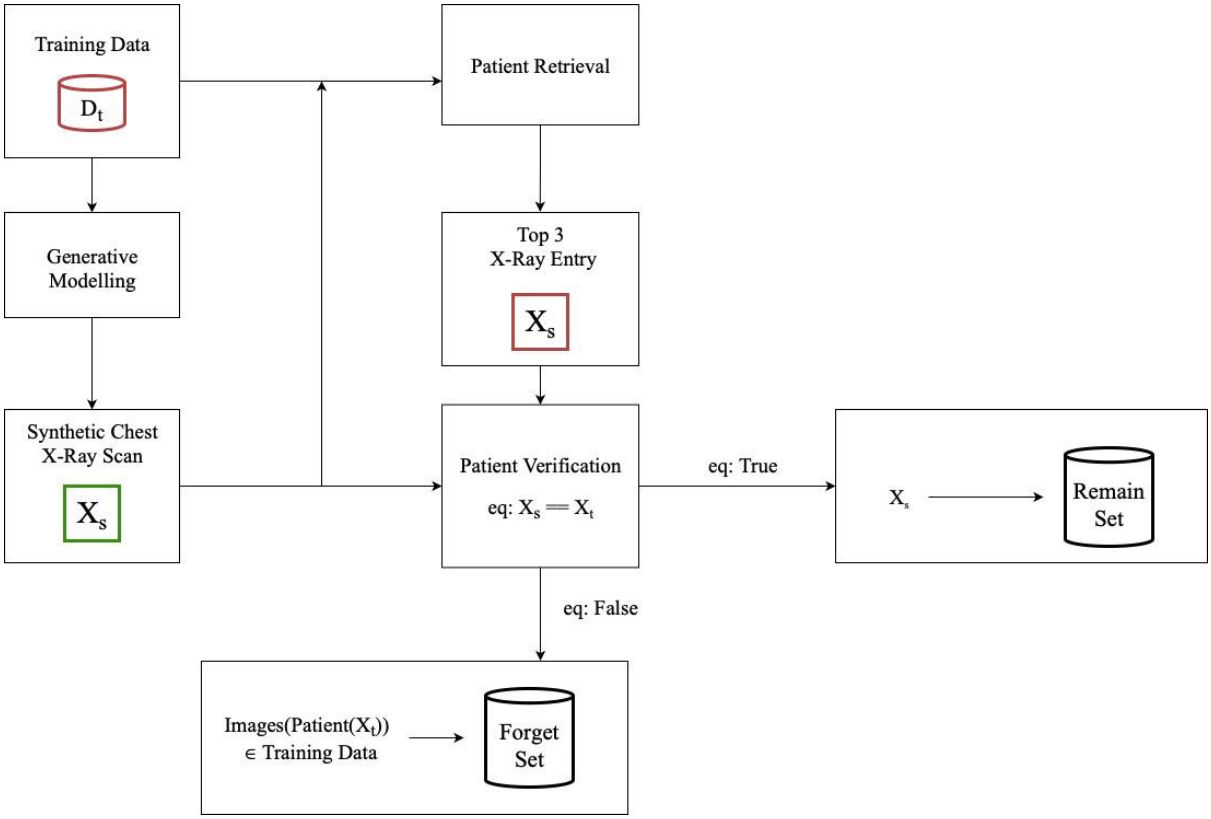


Figure 7.1: Redefined pipeline for detecting memorized patients' PII. Unlike the initial approach, which only considered the rank-1 retrieved image, this method evaluates similarity scores with the top-3 retrieved images. Patients associated with any retrieval above the similarity threshold are added to the Forget set, capturing cases where synthetic images embed features from multiple patients simultaneously.

7.2.2 Results and Discussion

The overall results obtained using this method are presented in Table 7.1.

This experiment helped clarify that some synthetic images were indeed leaking information from multiple patients simultaneously, as reflected in the larger number of non-anonymous images and patients identified compared to the original experiment, and that some patients were not being captured by the previous method.

However, as this method was also not able to fully prevent the identification of new patients, it reinforced the fact that memorized patients cannot be captured only through a subset of generated images, highlighting the need for methods that more comprehensively detect all memorized patients to construct a Forget set that contains all memorized patients. This also highlighted that the unlearning methods could not forget all memorized entities by only identifying a subset of them. Future work could explore

the use of attacks such as membership inference attacks (MIA) to more effectively identify and retrieve memorized patients.

Moreover, the increase in the number of patients identified to be unlearned, made the trade-off between localized and full-parameter approaches more evident: localized methods preserved good image quality and utility, while full-parameter techniques degraded even more their image quality compared to the original experiment, while achieving higher unlearning accuracy.

This behavior can be explained by the scope of parameter updates during unlearning. Localized methods restrict modifications to a subset of parameters most responsible for memorization, which allows them to perform unlearning while preserving the broader representational capacity of the model, while full-parameter approaches update the entire network, which can amplify forgetting but also disrupt unrelated regions of the network. With a larger Forget Set, this broad disruption became more evident, resulting in stronger degradation of image quality and utility compared to localized techniques.

Lastly, these results foster the development of new unlearning techniques that guarantee better deletion of patients identified in this specific scenario of medical image anonymization, as we were unable to forget all identified patient information. Unfortunately, with the tested unlearning methods, we could conclude that other anonymization methods, such as Disentanglement or Differential Privacy, currently offer better trade-offs.

7.3 Forgetting a Single Patient

Alberti et al. [3] performed data-point unlearning on diffusion models by targeting one instance at a time. Building on this, we hypothesized that the limited effectiveness in removing the influence of all originally identified patients may be due to the techniques tested struggling to achieve full unlearning accuracy when the Forget set contains multiple patients, rather than a single one.

7.3.1 Method

To assess whether certain patients could not be forgotten due to this limitation, we selected three patients whose images resisted forgetting across all previous methods. We then performed unlearning individually on each of these patients. For each case, the Forget set consisted of all images belonging to the target patient, while the Remain set was the one obtained in Section 7.2. This experiment was performed with the unlearning technique of KL-Away, as it seemed the most suitable, considering the previously obtained results.

7.3.2 Results and Discussion

Using this setup, we were able to successfully unlearn one of the three selected patients. This outcome highlights two important insights: (i) unlearning at the level of a single patient can in some cases be more effective than attempting to forget multiple patients at once, supporting our hypothesis about scalability limitations of current methods, and (ii) the inability to forget two of the patients underscores the high entanglement of the data, and the need for new unlearning strategies specifically tailored for data-point unlearning in diffusion models.

Table 7.1: Evaluation of the datasets' images by the different unlearning metrics across the various metrics.

Metric	**Synthetic**	**SalUn + KL-Away Sparsity 50%**	**SalUn + KL-Away Sparsity 30%**	**KL-Away**	**SISS**
Anonymization and Unlearning Metrics:					
Number of non-anonymous (↓) images produced	1412	919	846	608	**567**
Number of Patients Identified (↓)	1197	795	718	529	**497**
Number of Patients Re-identified (↓)	–	403	383	317	**314**
Number of New Patients Identified (↓)	-	392	335	212	**183**
Unlearning Accuracy (%) (↑)	–	66.33%	68.00%	73.52%	**73.77%**
Overall Dataset Anonymization (%) (↑)	64.70%	77.03%	78.85%	84.80%	**85.08%**
Overall Patient Identifiability Ratio (%) (↓)	29.93%	19.86%	17.95%	13.12%	**12.43%**
NLL (↑)	10.84	12.19	12.59	13.18	**20.12**
Image Utility Metrics:					
ACC (%) (↑)	84.16%	**80.09%**	75.11%	60.18%	41.78%
F1-SCORE (%) (↑)	72.20%	**74.29%**	70.91%	65.12%	41.02%
AUC (↑)	0.88	**0.87**	0.86	0.83	0.53
Image Quality Metrics:					
FID (↓)	53.50	**85.57**	91.54	101.86	189.23
P (↑)	80.57%	**77.20**	74.85	61.17	20.45

Future research should focus on developing techniques that can guarantee patient-level forgetting even in challenging and highly entangled scenarios, and also techniques that can do this with all memorized patients at once, since applying unlearning individually, instance by instance, would not be practical or efficient in this scenario.

Figure 7.2 illustrates the anonymization process for the patient we were able to forget in this experiment, showing the real image, its most similar synthetic image before unlearning, and the anonymized images generated after unlearning.

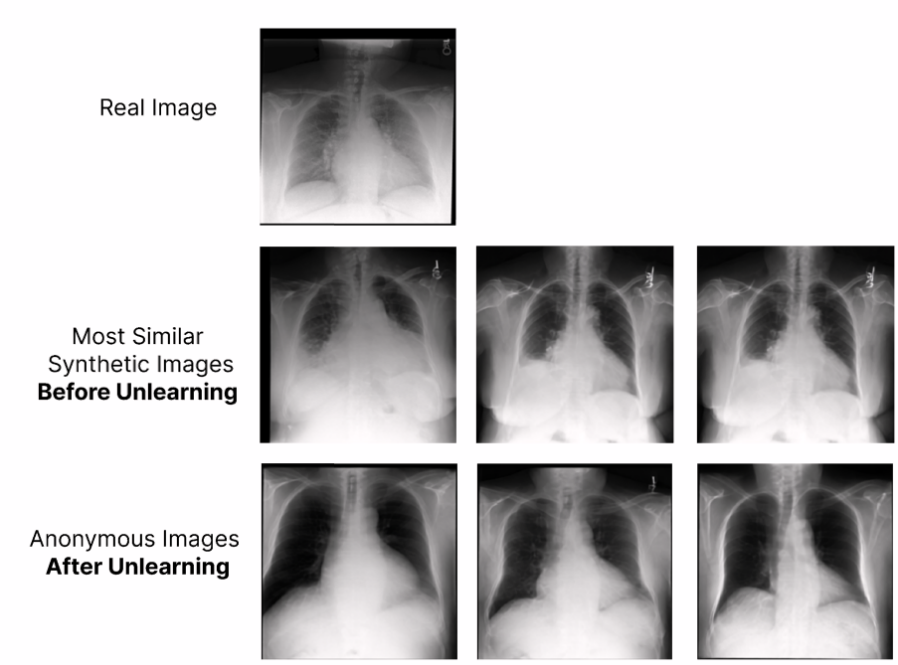


Figure 7.2: Visual example of the successful example in the single-patient unlearning experiment. The top row shows the real image of the target patient; the middle row presents the most similar synthetic images (non-anonymous) before unlearning, and the bottom row shows the anonymized images obtained after unlearning with KL-Away.

7.4 Summary

This chapter explored improvement-oriented experiments designed to address the limitations identified in Chapter 6. First, we redefined the Forget and Remain sets by considering higher-rank retrievals, which revealed that synthetic images can indeed leak information from multiple patients simultaneously. Although this adjustment increased the number of identified patients and reduced the Remain set, it did not eliminate the appearance of new patients after unlearning. These findings highlighted the need for more comprehensive methods to detect memorization and construct effective Forget sets, with membership-inference attacks (MIA) proposed as a promising direction.

The second experiment investigated unlearning at the patient level by applying unlearning individually to patients who had previously resisted unlearning under the different techniques. The results demonstrated that single-patient unlearning can, in some cases, improve effectiveness. However, the persistence of patients resistant to forgetting underscores the limitations of current approaches and highlights the need for better unlearning methods capable of achieving patient-level forgetting.

In general, these experiments clarified trade-offs between localized and full-parameter approaches, emphasized the challenges of completely removing memorized patient information, and motivated future research toward developing unlearning strategies that can achieve reliable patient-level deletion in diffusion-based medical image models.

Chapter 8

Conclusion and Future Work

This dissertation explored how machine unlearning (MU) can be applied to diffusion-based generative models to enable the generation of privacy-preserving, case-based explanations in federated learning (FL) environments. The motivation arose from two pressing challenges: (i) the risk of data leakage through memorization in generative models, and (ii) regulatory requirements for anonymization and compliance with the "Right to Be Forgotten" (RTBF) under privacy regulations such as the GDPR.

The systematic review confirmed the novelty of this research by revealing a lack of prior work on MU for medical imaging synthesis with generative models, and a very limited number of studies addressing data-point unlearning in diffusion models. In particular, no previous research has focused on forgetting at the patient level in generative models for medical imaging, despite the privacy risks associated with these.

In this work, we used a re-identification attack to identify patients whose features were memorized by a diffusion model and applied unlearning to those patients. Two approaches based on this re-identification attack were tested to create the Forget Set. However, neither was able to capture all the patients memorized by the model. As a result, although we successfully unlearned most of the patients initially identified, the model continued generating synthetic images containing identifiable information from other patients. This limitation highlights the need for improved and more efficient methods to retrieve all memorized patients for inclusion in the Forget set.

We also proposed a new method tailored for machine unlearning in I2I diffusion models: Kl-Away. KL-Away leverages the Kullback–Leibler divergence between the current model and a frozen copy of the original model on the Forget set. By maximizing this divergence, the method explicitly pushes the model's predictions away from its previously memorized outputs, thereby erasing prior knowledge of forgotten samples. Simultaneously, training on the Remain set with the standard diffusion loss helps preserve performance and generalization. The experimental results demonstrated that this new technique achieved competitive performance: it successfully reduced patient identifiability and achieved good Unlearning Accuracy.

However, neither of the machine unlearning techniques tested allowed the complete removal of all initially identified patients, and all techniques led to a decrease in the image quality and utility of the samples generated. This limitation can be partly explained by the high entanglement between Forget and Remain sets in the latent space, which makes patient-level forgetting in diffusion models particularly challenging.

Nevertheless, these results also highlight the need for new unlearning techniques, especially tailored for diffusion models, that can perform robust and scalable data-point unlearning in these models. In particular, there is a literature gap on methods capable of removing multiple identities at once in image-to-image diffusion models.

Unfortunately, attending to the results obtained with the tested machine unlearning techniques, other anonymization methods, such as Disentanglement or Differential Privacy, currently offer better trade-offs.

Future research could build on this dissertation in several directions:

- **Use different methods to evaluate memorization and create the Forget Set.** To allow the model to forget all memorised patients' images, these must be, in the first place, included in the Forget set. Future research could investigate the use of membership inference attacks (MIA), adversarial auditing techniques, or hybrid strategies that combine signal-based and latent-space analyses. This would allow for a more complete Forget set, thereby improving the effectiveness of unlearning.

- **Evaluate with other metrics:** Future work could evaluate unlearning methods with (i) attack-based audits such as MIA, (ii) system-level measures such as GPU memory consumption and runtime, and (iii) privacy-utility trade-offs, comparing MU directly with alternatives like differential privacy or disentanglement.

- **Focus on Unlearning a Single Entity:** Future research could extend the proposed methods to the unlearning of a single entity (e.g., a specific patient). This scenario has not yet been explored in the context of medical image generation and could provide valuable insights into the feasibility and limitations of unlearning in the medical context.

- **Development of unlearning techniques that forget multiple entities at once.** Current methods in the literature, such as the original SISS by Alberti et al. [3], perform unlearning individually on each entity they want to remove from the model, making it inefficient to remove multiple entities at once. Future work should focus on designing unlearning strategies capable of forgetting groups of entities memorized in bulk, which would be more efficient and more relevant to clinical privacy requirements.

- **Extension beyond chest X-rays or medical imaging.** Although this work focused on cardiomegaly detection in chest X-rays, the framework should be validated across other modalities such as CT or MRI and across different non-medical datasets. This would help assess the generalizability and robustness of the unlearning in diffusion models.

References

[1] Martin Abadi et al. “Deep Learning with Differential Privacy”. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2016, pp. 308–318.

[2] Amro Abdalla et al. *GIFT: Gradient-aware Immunization of diffusion models against malicious Fine-Tuning with safe concepts retention*. _eprint: arXiv:2507.13598. 2025.

[3] Silas Alberti et al. *Data Unlearning in Diffusion Models*. _eprint: arXiv:2503.01034. 2025.

[4] Laman Aliyeva et al. “Deep Unlearning of Breast Cancer Histopathological Images for Enhanced Responsibility in Classification”. In: *2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT)*. 2024, pp. 1–6. DOI: 10.1109/AICT61888.2024.10740413.

[5] Alejandro Barredo Arrieta et al. “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information Fusion* 58 (2020), pp. 82–115.

[6] Brett K Beaulieu-Jones et al. “Privacy-preserving generative deep neural networks support clinical data sharing”. In: *Nature Medicine* 25.7 (2019), pp. 950–954.

[7] Stefanie Becker et al. “Evaluating and Testing Machine Unlearning Algorithms”. In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2022.

[8] Peter J. Bevan and Amir Atapour-Abarghouei. *Detecting Melanoma Fairly: Skin Tone Detection and Debiasing for Skin Lesion Classification*. 2022. arXiv: 2202.02832 [eess.IV]. URL: https://arxiv.org/abs/2202.02832.

[9] Shristi Das Biswas, Arani Roy, and Kaushik Roy. *CURE: Concept Unlearning via Orthogonal Representation Editing in Diffusion Models*. _eprint: arXiv:2505.12677. 2025.

[10] Hugo Bourtoule et al. “Machine Unlearning”. In: *Proceedings of the 42nd IEEE Symposium on Security and Privacy (S&P)*. 2021. URL: https://arxiv.org/abs/1912.03817.

[11] Filipe Campos et al. “Latent Diffusion Models for Privacy-preserving Medical Case-based Explanations”. English. In: *CEUR Workshop Proceedings* 3831 (Nov. 2024). Publisher Copyright: © 2024 Copyright for this paper by its authors.; 1st Workshop on Explainable Artificial Intelligence for the Medical Domain, EXPLIMED 2024 ; Conference date: 20-10-2024. ISSN: 1613-0073.

[12] Chen Cao, Wei Li, and Yisen Wang. “FUCRT: Class-wise Federated Unlearning with Class-aware Representation Transformation”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[13] Yinzhi Cao and Junfeng Yang. “Towards making systems forget with machine unlearning”. In: *IEEE Symposium on Security and Privacy*. IEEE. 2015, pp. 463–480.

[14] Nicholas Carlini et al. “Extracting training data from diffusion models”. In: *IEEE Symposium on Security and Privacy (S&P)*. 2023.

[15] Finn Carter. *ACE: Attentional Concept Erasure in Diffusion Models*. _eprint: arXiv:2504.11850. 2025.

[16] Finn Carter. *TRACE: Trajectory-Constrained Concept Erasure in Diffusion Models*. _eprint: arXiv:2505.23312. 2025.

[17] Ruchika Chavhan, Da Li, and Timothy Hospedales. *ConceptPrune: Concept Editing in Diffusion Models via Skilled Neuron Pruning*. _eprint: arXiv:2405.19237. 2024.

[18] Die Chen et al. *Growth Inhibitors for Suppressing Inappropriate Image Concepts in Diffusion Models*. _eprint: arXiv:2408.01014. 2024.

[19] Kongyang Chen et al. *Federated Unlearning for Human Activity Recognition*. _eprint: arXiv:2404.03659. 2024.

[20] Ling-Hao Chen et al. *ERASE: Error-Resilient Representation Learning on Graphs for Label Noise Tolerance*. _eprint: arXiv:2312.08852. 2023.

[21] Ruidong Chen et al. *TRCE: Towards Reliable Malicious Concept Erasure in Text-to-Image Diffusion Models*. _eprint: arXiv:2503.07389. 2025.

[22] Siyi Chen et al. *The Dual Power of Interpretable Token Embeddings: Jailbreaking Attacks and Defenses for Diffusion Model Unlearning*. _eprint: arXiv:2504.21307. 2025.

[23] Ikhyun Cho, Changyeon Park, and Julia Hockenmaier. *ViT-MUL: A Baseline Study on Recent Machine Unlearning Methods Applied to Vision Transformers*. _eprint: arXiv:2403.09681. 2024.

[24] Cornell University. *arXiv: Open-access repository*. https://arxiv.org/. 2025.

[25] Council of European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504. 2016.

[26] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2nd. Wiley, 2012. ISBN: 9781118585771.

[27] Florinel-Alin Croitoru et al. "Diffusion Models in Vision: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.9 (2023), pp. 10850–10869. DOI: 10.1109/TPAMI.2023.3261988.

[28] Deniz Daum et al. "On Differentially Private 3D Medical Image Synthesis with Controllable Latent Diffusion Models". In: *Proceedings of the MICCAI Workshop on Deep Generative Models*. Vol. 14224. Lecture Notes in Computer Science. Springer, 2024, pp. 78–89. DOI: 10.1007/978-3-031-72302-8_8. URL: https://arxiv.org/abs/2407.16405.

[29] Peng Deng et al. "Cooperation Among Multiple Medical Institutions on Retinal Disease Identification Based on Federated Learning and Unlearning". In: *2024 4th Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*. 2024, pp. 533–537. DOI: 10.1109/ACCTCS61748.2024.00100.

[30] Zhipeng Deng, Luyang Luo, and Hao Chen. *Enable the Right to be Forgotten with Federated Client Unlearning in Medical Imaging*. 2024. arXiv: 2407.02356 [eess.IV]. URL: https://arxiv.org/abs/2407.02356.

[31] Nicola K. Dinsdale, Mark Jenkinson, and Ana I. L. Namburete. "Unlearning Scanner Bias for MRI Harmonisation in Medical Image Segmentation". In: *Medical Image Understanding and Analysis*. Ed. by Bartłomiej W. Papie\.z et al. Cham: Springer International Publishing, 2020, pp. 15–25. ISBN: 978-3-030-52791-4.

[32] Nicola K. Dinsdale, Mark Jenkinson, and Ana IL Namburete. *FedHarmony: Unlearning Scanner Bias with Distributed Data*. arXiv:2205.15970 [cs]. May 2022. DOI: 10.48550/arXiv.2205.15970. URL: http://arxiv.org/abs/2205.15970 (visited on 08/12/2025).

[33] Tim Dockhorn et al. “Differentially Private Diffusion Models”. In: *arXiv preprint arXiv:2210.09929* (2022). URL: https://arxiv.org/abs/2210.09929.

[34] Yucheng Duan, Jinghui Chen, and Bo Li. “Membership Inference Attacks on Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 4036–4045.

[35] Khaoula ElBedoui. “ECG Classifiction Based on Federated Unlearning”. In: *2023 International Symposium on Networks, Computers and Communications (ISNCC)*. 2023, pp. 1–5. DOI: 10.1109/ISNCC58260.2023.10323758.

[36] Elsevier. *Scopus Digital Library*. https://www.scopus.com/. 2025.

[37] Luis Carlos Molano Esmeral and Andreas Uhl. “Low-effort re-identification techniques based on medical imagery threaten patient privacy”. In: *Medical Image Understanding and Analysis*. Springer. 2022, pp. 719–733.

[38] Chongyu Fan et al. *Challenging Forgets: Unveiling the Worst-Case Forget Sets in Machine Unlearning*. _eprint: arXiv:2403.07362. 2024.

[39] Chongyu Fan et al. *SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation*. arXiv:2310.12508 [cs]. 2024. DOI: 10.48550/arXiv.2310.12508. URL: http://arxiv.org/abs/2310.12508 (visited on 12/25/2024).

[40] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “Model inversion attacks that exploit confidence information and basic countermeasures”. In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 2015, pp. 1322–1333.

[41] Masane Fuchi and Tomohiro Takagi. *Erasing Concepts from Text-to-Image Diffusion Models with Few-shot Unlearning*. _eprint: arXiv:2405.07288. 2024.

[42] Shihua Gai et al. “AproMFL: Adaptive Prototype-based Multimodal Federated Learning”. In: *arXiv preprint arXiv:2502.04400* (2025). URL: https://arxiv.org/abs/2502.04400.

[43] Daiheng Gao et al. *EraseAnything: Enabling Concept Erasure in Rectified Flow Transformers*. _eprint: arXiv:2412.20413. 2024.

[44] Kuofeng Gao et al. *Towards Dataset Copyright Evasion Attack against Personalized Text-to-Image Diffusion Models*. _eprint: arXiv:2505.02824. 2025.

[45] Xiangshan Gao et al. “VeriFi: Towards Verifiable Federated Unlearning”. In: *IEEE Transactions on Dependable and Secure Computing* 21.6 (2024), pp. 5720–5736. DOI: 10.1109/TDSC.2024.3382321.

[46] Lingyue Ge. “Erasing memories: Implementing client unlearning in medical image analysis”. In: *Proceedings of SPIE - The International Society for Optical Engineering*. Vol. 13213. Type: Conference paper. 2024. DOI: 10.1117/12.3035404. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85200458882&doi=10.1117%2f12.3035404&partnerID=40&md5=8eb97e3bc561d338190eeed161d9c75c.

[47] Antonio Ginart et al. “Making AI forget you: Data deletion in machine learning”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. 2019.

[48] Chao Gong et al. *Reliable and Efficient Concept Erasure of Text-to-Image Diffusion Models*. _eprint: arXiv:2407.12383. 2024.

[49] Keltin Grimes et al. *Gone but Not Forgotten: Improved Benchmarks for Machine Unlearning*. _eprint: arXiv:2405.19211. 2024.

[50] Hanlin Gu et al. *A few-shot Label Unlearning in Vertical Federated Learning*. _eprint: arXiv:2410.10922. 2024.

[51] Qi Guo et al. *Forgetting Through Transforming: Enabling Federated Unlearning via Class-Aware Representation Transformation*. _eprint: arXiv:2410.06848. 2024.

[52] Misgina Tsighe Hagos, Kathleen M. Curran, and Brian Mac Namee. *Unlearning Spurious Correlations in Chest X-ray Classification*. arXiv:2308.01119 [eess]. Aug. 2023. DOI: 10.48550/arXiv.2308.01119. URL: http://arxiv.org/abs/2308.01119 (visited on 08/12/2025).

[53] Mengde Han et al. *Vertical Federated Unlearning via Backdoor Certification*. _eprint: arXiv:2412.11476. 2024.

[54] Seungyub Han et al. *Learning to Learn Unlearned Feature for Brain Tumor Segmentation*. arXiv:2305.08878 [eess]. May 2023. DOI: 10.48550/arXiv.2305.08878. URL: http://arxiv.org/abs/2305.08878 (visited on 08/12/2025).

[55] Xiaoxuan Han et al. *Probing Unlearned Diffusion Models: A Transferable Adversarial Attack Perspective*. _eprint: arXiv:2404.19382. 2024.

[56] Jamie Hayes et al. "LOGAN: Membership inference attacks against generative models". In: *Proceedings on Privacy Enhancing Technologies* 2019.1 (2019), pp. 133–152.

[57] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[58] Luis Hernandez-Cruz et al. "Federated learning in medical imaging: privacy-preserving collaborative learning for AI in healthcare". In: *arXiv preprint arXiv:2503.00581* (2024). URL: https://arxiv.org/abs/2503.00581.

[59] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 6840–6851.

[60] Jonathan Ho and Tim Salimans. "Classifier-Free Diffusion Guidance". In: *arXiv preprint arXiv:2207.12598* (2022).

[61] Andreas Holzinger et al. "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" In: *Brain Informatics*. Springer. 2017, pp. 1–13. DOI: 10.1007/978-3-319-70772-3_1.

[62] Mohammad Hosseini et al. "Privacy-preserving federated learning using secure multi-party computation and homomorphic encryption". In: *arXiv preprint arXiv:2503.00581* (2025). URL: https://arxiv.org/abs/2503.00581.

[63] Hexin Hu et al. "Membership Inference Attacks against Generative Diffusion Models". In: *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2023, pp. 1659–1673.

[64] Chi-Pin Huang et al. *Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers*. _eprint: arXiv:2311.17717. 2023.

[65] Mark He Huang, Lin Geng Foo, and Jun Liu. *Learning to Unlearn for Robust Machine Unlearning*. _eprint: arXiv:2407.10494. 2024.

[66] Matt Hudson and Mark I. Johnson. "Split-Second Unlearning: Developing a Theory of Psychophysiological Dis-ease". In: *Frontiers in Psychology* 12 (2021). Type: Article. DOI: 10.3389/fpsyg.2021.716535. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85121370284&doi=10.3389%2ffpsyg.2021.716535&partnerID=40&md5=fe13ff2d3e8856b90b2b0621e467d056.

[67] IEEE. *IEEE Xplore Digital Library*. https://ieeexplore.ieee.org/. 2025.

[68] Jeremy Irvin et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 590–597. DOI: 10.1609/aaai.v33i01.3301590.

[69] Anubhav Jain et al. *TraSCE: Trajectory Steering for Concept Erasure*. _eprint: arXiv:2412.07658. 2024.

[70] Yu Jiang, Chee Wei Tan, and Kwok-Yan Lam. *FedUHB: Accelerating Federated Unlearning via Polyak Heavy Ball Method*. _eprint: arXiv:2411.11039. 2024.

[71] Yu Jiang et al. *Efficient Federated Unlearning with Adaptive Differential Privacy Preservation*. _eprint: arXiv:2411.11044. 2024.

[72] Xiaomeng Jin et al. "Unlearning as Multi-Task Optimization: A Normalized Gradient Difference Approach with an Adaptive Learning Rate". In: *NAACL 2025 - Long Papers*. 2025.

[73] Alistair E. W. Johnson et al. *MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs*. arXiv preprint arXiv:1901.07042. 2019.

[74] Alistair E. W. Johnson et al. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific Data* 6 (2019), p. 317. DOI: 10.1038/s41597-019-0322-0.

[75] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. "PATE-GAN: Generating synthetic data with differential privacy guarantees". In: *International Conference on Learning Representations (ICLR)*. 2018.

[76] Georgios A Kaissis et al. "Secure, privacy-preserving and federated machine learning in medical imaging". In: *Nature Machine Intelligence* 2.6 (2020), pp. 305–311.

[77] M. R. Kapadia. "Content Based Medical Image Retrieval for Accurate Diagnosis Support". In: *The Open Biomedical Engineering Journal* 15 (2021). Euclidean distance is used to determine similarity between query and database features, pp. 235–245.

[78] Changhoon Kim, Kyle Min, and Yezhou Yang. *R.A.C.E.: Robust Adversarial Concept Erasure for Secure Text-to-Image Diffusion Model*. _eprint: arXiv:2405.16341. 2024.

[79] Zhikai Kong et al. "Diffusion Model Inversion Attacks and Defenses". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

[80] Laura Latorre et al. "Towards Case-based Interpretability for Medical Federated Learning". In: *arXiv preprint arXiv:2408.13626* (2024). URL: https://arxiv.org/abs/2408.13626.

[81] LearnOpenCV. *Denoising Diffusion Probabilistic Models*. https://learnopencv.com/denoising-diffusion-probabilistic-models/. Accessed: 2025-08-09.

[82] Byung Hyun Lee, Sungjin Lim, and Se Young Chun. *Localized Concept Erasure for Text-to-Image Diffusion Models Using Training-Free Gated Low-Rank Adaptation*. _eprint: arXiv:2503.12356. 2025.

[83] Guihong Li et al. *Machine Unlearning for Image-to-Image Generative Models*. 2024. arXiv: 2402.00351 [cs.LG]. URL: https://arxiv.org/abs/2402.00351.

[84] Ouxiang Li et al. *SPEED: Scalable, Precise, and Efficient Concept Erasure for Diffusion Models*. _eprint: arXiv:2503.07392. 2025.

[85] Xiang Li, Lei Ping, and Fei Wang. "Differentially private deep learning for medical imaging". In: *Medical Image Analysis* 82 (2022), p. 102617.

[86] Yuyuan Li et al. *Class-wise Federated Unlearning: Harnessing Active Forgetting with Teacher-Student Memory Generation*. _eprint: arXiv:2307.03363. 2023.

[87] Feng Liu et al. "SIFU: Sequential Informed Federated Unlearning". In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. 2023.

[88] Hengzhu Liu et al. *Game-Theoretic Machine Unlearning: Mitigating Extra Privacy Leakage*. _eprint: arXiv:2411.03914. 2024.

[89] Rui Liu et al. “In-Context Learning for Zero-shot Medical Report Generation”. In: *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*. Type: Conference paper. 2024, pp. 8721–8730. DOI: 10.1145/3664647.3680760. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85204904023&doi=10.1145%2f3664647.3680760&partnerID=40&md5=208f41a509540cfdba07a4d694fbec83.

[90] Yang Liu et al. “Secure multi-party computation for federated learning: A comprehensive survey”. In: *Frontiers of Computer Science* (2024). DOI: 10.1007/s11704-023-3282-7.

[91] Yihan Liu et al. “Fast Federated Machine Unlearning with Certified Guarantees”. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023.

[92] Zhili Liu et al. *Implicit Concept Removal of Diffusion Models*. _eprint: arXiv:2310.05873. 2023.

[93] Kevin Lu et al. *When Are Concepts Erased From Diffusion Models?* _eprint: arXiv:2505.17013. 2025.

[94] Shilin Lu et al. *MACE: Mass Concept Erasure in Diffusion Models*. _eprint: arXiv:2403.06135. 2024.

[95] Yifan Lu et al. “Neural Dehydration: Effective Erasure of Black-box Watermarks from DNNs with Limited Data”. In: (2023). _eprint: arXiv:2309.03466. DOI: 10.1145/3658644.3690334.

[96] Kasu Mamatha and Debajyoty Banik. “Machine Unlearning in MRI Reconstruction to Balance Privacy Protection and Model Performance Trade-Offs”. In: *Lecture Notes in Networks and Systems* 1424 LNNS (2025). Type: Conference paper, pp. 436–453. DOI: 10.1007/978-3-031-92605-1_27. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105009401669&doi=10.1007%2f978-3-031-92605-1_27&partnerID=40&md5=07db0e682ba1784e2591565f45e99d2d.

[97] Soichiro Matsumoto, Ryota Takahashi, and Kenji Kono. “Membership Inference Attacks Against Diffusion Models”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, pp. 1–5.

[98] H Brendan McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2017, pp. 1273–1282.

[99] Casey Meehan et al. “A non-parametric approach to detecting memorization in diffusion models”. In: *arXiv preprint arXiv:2205.14135* (2022).

[100] Michel Meintz et al. *Radioactive Watermarks in Diffusion and Autoregressive Image Generative Models*. _eprint: arXiv:2506.23731. 2025.

[101] Zheling Meng et al. *Concept Corrector: Erase concepts on the fly for text-to-image diffusion models*. _eprint: arXiv:2502.16368. 2025.

[102] Zheling Meng et al. *Dark Miner: Defend against undesired generation for text-to-image diffusion models*. _eprint: arXiv:2409.17682. 2024.

[103] Grégoire Montavon et al. “Explaining nonlinear classification decisions with deep Taylor decomposition”. In: *Pattern Recognition* 65 (2017), pp. 211–222. DOI: 10.1016/j.patcog.2016.11.008.

[104] Helena Montenegro and Jaime S. Cardoso. “Anonymizing Medical Case-Based Explanations Through Disentanglement”. In: *Medical Image Analysis* 95 (2024), p. 103209. DOI: 10.1016/j.media.2024.103209.

[105] Alessio Mora et al. *Federated Unlearning Made Practical: Seamless Integration via Negated Pseudo-Gradients*. arXiv:2504.05822 [cs]. Apr. 2025. DOI: 10.48550/arXiv.2504.05822. URL: http://arxiv.org/abs/2504.05822 (visited on 08/12/2025).

[106] Gustav Müller-Franzes et al. “A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis”. In: *Scientific Reports* 13.1 (July 2023). ISSN: 2045-2322. DOI: 10.1038/s41598-023-39278-0. URL: http://dx.doi.org/10.1038/s41598-023-39278-0.

[107] George R. Nahass et al. *Targeted Unlearning Using Perturbed Sign Gradient Methods With Applications On Medical Images*. arXiv:2505.21872 [eess]. May 2025. DOI: 10.48550/arXiv.2505.21872. URL: http://arxiv.org/abs/2505.21872 (visited on 08/12/2025).

[108] Reza Nasirigerdeh et al. *Machine Unlearning for Medical Imaging*. 2024. arXiv: 2407.07539 [cs.LG]. URL: https://arxiv.org/abs/2407.07539.

[109] Erika Newton, Latanya Sweeney, and Bradley Malin. “Preserving privacy by de-identifying face images”. In: *IEEE Transactions on Knowledge and Data Engineering*. Vol. 17. 2. IEEE, 2005, pp. 232–243.

[110] Quang H. Nguyen, Hoang Phan, and Khoa D. Doan. *Unveiling Concept Attribution in Diffusion Models*. _eprint: arXiv:2412.02542. 2024.

[111] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. “Gradient leakage in machine learning models”. In: *arXiv preprint arXiv:1805.12213*. 2018.

[112] Konstantin Packhäuser et al. “Deep learning-based patient re-identification can exploit the biometric nature of medical chest x-ray data”. In: *Scientific Reports* 12.1 (2022), p. 14851.

[113] Matthew J Page et al. “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews”. In: *BMJ* 372 (2021). DOI: 10.1136/bmj.n71. eprint: https://www.bmj.com/content/372/bmj.n71.full.pdf. URL: https://www.bmj.com/content/372/bmj.n71.

[114] Martin Pawelczyk et al. “From prototypes to counterfactuals: A survey of case-based explanation methods”. In: *arXiv preprint arXiv:2408.06679* (2024).

[115] Jie Ren et al. *SoK: Machine Unlearning for Large Language Models*. _eprint: arXiv:2506.09227. 2025.

[116] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV]. URL: https://arxiv.org/abs/1505.04597.

[117] Matan Rusanovsky et al. *Memories of Forgotten Concepts*. _eprint: arXiv:2412.00782. 2024.

[118] Parisa Saat et al. “A domain adaptation benchmark for T1-weighted brain magnetic resonance image segmentation”. In: *Frontiers in Neuroinformatics* 16 (2022). Type: Article. DOI: 10.3389/fninf.2022.919779. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140098939&doi=10.3389%2ffninf.2022.919779&partnerID=40&md5=a132d7d0285645a3c580fb48f0cd69cb.

[119] Roy Schwartz et al. “Green AI”. In: *Communications of the ACM* 63.12 (2020), pp. 54–63. URL: https://arxiv.org/abs/1907.10597.

[120] Daniel Schwarz, Bernhard Egger, and William A.P. Smith. “Reconstructing faces from anonymized medical images: A privacy threat”. In: *Medical Image Analysis* 76 (2022), p. 102310.

[121] Artem Sevastopolsky, Alexey Ivanov, and Nikolay Petrov. “Removing Identifiable Body Markings Using Deep Learning-Based Inpainting”. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2022.

[122] Adeel Shaheryar et al. “Black Hole–Driven Identity Absorbing in Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025.

[123] Aakash Sen Sharma et al. *Unlearning or Concealment? A Critical Analysis and Evaluation Metrics for Unlearning in Diffusion Models*. _eprint: arXiv:2409.05668. 2024.

[124] Reza Shokri et al. "Membership inference attacks against machine learning models". In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 3–18.

[125] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising Diffusion Implicit Models". In: *arXiv preprint arXiv:2010.02502* (2020). URL: https://arxiv.org/abs/2010.02502.

[126] Koushik Srivatsan et al. *STEREO: A Two-Stage Framework for Adversarially Robust Concept Erasing from Text-to-Image Diffusion Models*. _eprint: arXiv:2408.16807. 2024.

[127] Minghui Sun, Benjamin A. Goldstein, and Matthew M. Engelhard. *CLEAR: Unlearning Spurious Style-Content Associations with Contrastive LEarning with Anti-contrastive Regularization*. arXiv:2507.18794 [cs]. July 2025. DOI: 10.48550/arXiv.2507.18794. URL: http://arxiv.org/abs/2507.18794 (visited on 08/12/2025).

[128] Kahou Tam et al. *Towards Federated Domain Unlearning: Verification Methodologies and Challenges*. _eprint: arXiv:2406.03078. 2024.

[129] Yan Tan et al. "FedProto: Federated prototype learning across heterogeneous clients". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 8. 2022, pp. 8432–8440. DOI: 10.1609/aaai.v36i8.20819.

[130] Anshuman Tarun et al. "Fast yet effective machine unlearning". In: *International Conference on Learning Representations (ICLR)*. 2021.

[131] Qi Tian et al. "ConfounderGAN: Protecting Image Data Privacy with Causal Confounder". In: *Advances in Neural Information Processing Systems (NeurIPS 2022)*. 2022.

[132] Qi Tian et al. "ConfounderGAN: Protecting Image Data Privacy with Causal Confounder". In: *Advances in Neural Information Processing Systems*. Vol. 35. Type: Conference paper. 2022. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85150746068&partnerID=40&md5=de75a26f2b1e59eadd4a59baefff4673.

[133] Zhihua Tian et al. *Sparse Autoencoder as a Zero-Shot Classifier for Concept Erasing in Text-to-Image Diffusion Models*. _eprint: arXiv:2503.09446. 2025.

[134] Yu Tong et al. *Training-Free Watermarking for Autoregressive Image Generation*. _eprint: arXiv:2505.14673. 2025.

[135] Reihaneh Torkzadehmahani et al. "Improved Localized Machine Unlearning Through the Lens of Memorization". In: *arXiv preprint arXiv:2412.02432* (2024). URL: https://doi.org/10.48550/arXiv.2412.02432.

[136] Reza Torkzadehmahani, Peter Kairouz, and Benedict Paten. "DP-CGAN: Differentially private conditional generative adversarial networks for releasing real world health data". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 2757–2761.

[137] Mateusz Trokielewicz, Adam Czajka, and Piotr Maciejewicz. "Implications of ocular pathologies for iris recognition reliability". In: *Image and Vision Computing* 58 (2017), pp. 158–167.

[138] Jiahang Tu et al. *CE-SDWV: Effective and Efficient Concept Erasure for Text-to-Image Diffusion Models via a Semantic-Driven Word Vocabulary*. _eprint: arXiv:2501.15562. 2025.

[139] U.S. Department of Health and Human Services. *The Health Insurance Portability and Accountability Act of 1996 (HIPAA)*. https://www.hhs.gov/hipaa/index.html. https://www.hhs.gov/hipaa/index.html. 1996.

[140] Ruipeng Wang et al. *ACE: Concept Editing in Diffusion Models without Performance Degradation*. _eprint: arXiv:2503.08116. 2025.

[141] Yuan Wang et al. *Precise, Fast, and Low-cost Concept Erasure in Value Space: Orthogonal Complement Matters*. _eprint: arXiv:2412.06143. 2024.

[142] Zihao Wang et al. *ACE: Anti-Editing Concept Erasure in Text-to-Image Models*. _eprint: arXiv:2501.01633. 2025.

[143] Jing Wu and Mehrtash Harandi. *MUNBa: Machine Unlearning via Nash Bargaining*. _eprint: arXiv:2411.15537. 2024.

[144] Jing Wu and Mehrtash Harandi. *Scissorhands: Scrub Data Influence via Connection Sensitivity in Networks*. _eprint: arXiv:2401.06187. 2024.

[145] Jing Wu et al. *Erasing Undesirable Influence in Diffusion Models*. _eprint: arXiv:2401.05779. 2024.

[146] Yongliang Wu et al. *Unlearning Concepts in Diffusion Model via Concept Domain Correction and Concept Preserving Gradient*. _eprint: arXiv:2405.15304. 2024.

[147] Liyang Xie et al. "Differentially private generative adversarial network". In: *arXiv preprint arXiv:1802.06739*. 2018.

[148] Yuyang Xue et al. *Erase to Enhance: Data-Efficient Machine Unlearning in MRI Reconstruction*. 2024. arXiv: 2405.15517 [eess.IV]. URL: https://arxiv.org/abs/2405.15517.

[149] Tianyun Yang, Juan Cao, and Chang Xu. *Pruning for Robust Concept Erasing in Diffusion Models*. _eprint: arXiv:2405.16534. 2024.

[150] Takato Yasuno et al. "Bridge Slab Anomaly Detector Using U-Net Generator with Patch Discriminator for Robust Prognosis". In: *Structural Health Monitoring 2021: Enabling Next-Generation SHM for Cyber-Physical Systems - Proceedings of the 13th International Workshop on Structural Health Monitoring, IWSHM 2021*. Type: Conference paper. 2021, pp. 341–348. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85139235529&partnerID=40&md5=d47e2f6825c991b49204c1d09f2dc38e.

[151] Fuyao Zhang et al. *Verifiably Forgotten? Gradient Differences Still Enable Data Reconstruction in Federated Unlearning*. _eprint: arXiv:2505.11097. 2025.

[152] Yang Zhang et al. "GANs for generating private medical data". In: *arXiv preprint arXiv:1801.09308*. 2018.

[153] Yimeng Zhang et al. "To Generate or Not? Safety-Driven Unlearned Diffusion Models Are Still Easy To Generate Unsafe Images ... For Now". In: *European Conference on Computer Vision (ECCV)*. 2024.

[154] Mengnan Zhao et al. *AdvAnchor: Enhancing Diffusion Model Unlearning with Adversarial Anchors*. _eprint: arXiv:2501.00054. 2024.

[155] Mengnan Zhao et al. *Separable Multi-Concept Erasure from Diffusion Models*. _eprint: arXiv:2402.05947. 2024.

[156] Yuyao Zhong. "Federated unlearning for medical image analysis". In: *Fourth Symposium on Pattern Recognition and Applications (SPRA 2023)*. Ed. by Shien-Kuei Liaw. Vol. 13162. International Society for Optics and Photonics. SPIE, 2024, p. 1316206. DOI: 10.1117/12.3030004. URL: https://doi.org/10.1117/12.3030004.

[157] Jialiang Zhou et al. "FedAU: Federated Unlearning During Learning". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023.

Appendix A

Appendix A

A.1 Declarations

The author acknowledges the use of AI tools in the preparation of this dissertation. Specifically, Writefull, Grammarly, DeepL, and ChatGPT were used to improve readability, grammar, spelling, and the presentation of mathematical equations and formulas. The intellectual content and conclusions presented remain entirely the author's own work.

A.2 Derivation of KL Simplification under Gaussian Assumption

In this appendix, we provide the derivation showing how the KL divergence between two diffusion posteriors reduces to an L_2 distance between predicted noise vectors under the Gaussian assumption.

KL divergence between Gaussians with equal variance

At diffusion step t, the reverse-process posterior in a DDPM-style model is Gaussian:

$$p_\theta(x_{t-1} \mid x_t, c) \;=\; \mathcal{N}\big(\mu_\theta(x_t,t,c),\, \sigma_t^2 I\big),$$

where the mean μ_θ depends on the noise predictor $\hat{\varepsilon}_\theta$:

$$\mu_\theta(x_t,t,c) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\hat{\varepsilon}_\theta(x_t,t,c)\right).$$

Consider the current model θ and a frozen teacher θ_0. Their conditional distributions are

$$p_\theta(\cdot \mid x_t, c) = \mathcal{N}(\mu_\theta, \sigma_t^2 I), \quad p_{\theta_0}(\cdot \mid x_t, c) = \mathcal{N}(\mu_{\theta_0}, \sigma_t^2 I).$$

For two Gaussians with the same covariance $\sigma_t^2 I$, the KL divergence simplifies to

$$D_{\mathrm{KL}}\big(\mathcal{N}(\mu_\theta, \sigma_t^2 I) \,\|\, \mathcal{N}(\mu_{\theta_0}, \sigma_t^2 I)\big) = \frac{1}{2\sigma_t^2}\|\mu_\theta - \mu_{\theta_0}\|^2.$$

Relation to noise predictions

From the mean parameterization, the difference in means is linear in the noise predictions:

$$\mu_\theta - \mu_{\theta_0} = -\kappa_t\big(\hat{\varepsilon}_\theta - \hat{\varepsilon}_{\theta_0}\big), \quad \kappa_t = \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}.$$

Substituting gives

$$D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta_0}) = \frac{\kappa_t^2}{2\sigma_t^2}\,\|\hat{\varepsilon}_\theta - \hat{\varepsilon}_{\theta_0}\|^2.$$

Final form

Therefore, under Gaussian diffusion assumptions with fixed variance, the KL divergence between teacher and student distributions is proportional to the squared distance between their noise predictions:

$$D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta_0}) \;\propto\; \|\hat{\varepsilon}_\theta - \hat{\varepsilon}_{\theta_0}\|^2.$$

In practice, the proportionality constant depends only on the noise schedule and can be absorbed into the weighting factor λ. This motivates the use of the surrogate unlearning loss

$$\mathscr{L}_{\text{unlearn}} = -\tfrac{1}{2}\|\hat{\varepsilon}_\theta - \hat{\varepsilon}_{\theta_0}\|^2,$$

which maximizes divergence from the frozen teacher's predictions on the forget set.