



HAL
open science

Health data : Exploring emerging privacy enhancing mechanisms

Thomas Lebrun

► **To cite this version:**

Thomas Lebrun. Health data : Exploring emerging privacy enhancing mechanisms. Artificial Intelligence [cs.AI]. INSA de Lyon, 2024. English. NNT : 2024ISAL0114 . tel-04943229v2

HAL Id: tel-04943229

<https://theses.hal.science/tel-04943229v2>

Submitted on 12 Feb 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSA

N° d'ordre NNT : 2024ISAL0114

THESE de DOCTORAT DE L'INSA LYON, membre de l'Université de Lyon

École Doctorale N° 512
INFORMATIQUE ET MATHÉMATIQUES

Spécialité/ discipline de doctorat : INFORMATIQUE

Soutenue publiquement le 05/12/2024, par :

Thomas Lebrun

Health Data: Exploring Emerging Privacy Enhancing Mechanisms

Devant le jury composé de :

Présidente: Sonia BEN MOKHTAR

BEN MOKHTAR	Sonia	Directrice de Recherche	CNRS/INSA-Lyon	Examinatrice
LESTYAN	Szilvia	Docteure-Ingénieure de R.	INRIA	Examinatrice
DECOUCHANT	Jérémie	Professeur des universités	Université de Delft	Examineur
NGUYEN	Benjamin	Professeur des universités	INSA-CVL	Rapporteur
VINCENT	Emmanuel	Directeur de Recherche	INRIA	Rapporteur
CUNCHE	Mathieu	Professeur des universités	INSA-Lyon	Directeur de thèse
BOUTET	Antoine	Maître de conférence	INSA-Lyon	Encadrant de thèse
MAOUCHE	Mohamed	Chargé de recherche	INRIA	Co-Encadrant

Département FEDORA – INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
ED 206 CHIMIE	CHIMIE DE LYON https://www.edchimie-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr	M. Stéphane DANIELE C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne directeur@edchimie-lyon.fr
ED 341 E2M2	ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION http://e2m2.universite-lyon.fr Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.e2m2@univ-lyon1.fr	Mme Sandrine CHARLES Université Claude Bernard Lyon 1 UFR Biosciences Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69622 Villeurbanne CEDEX e2m2.codir@listes.univ-lyon1.fr
ED 205 EDISS	INTERDISCIPLINAIRE SCIENCES-SANTÉ http://ediss.universite-lyon.fr Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.ediss@univ-lyon1.fr	Mme Sylvie RICARD-BLUM Laboratoire ICBMS - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 sylvie.ricard-blum@univ-lyon1.fr
ED 34 EDML	MATÉRIAUX DE LYON http://ed34.universite-lyon.fr Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 yann.de-ordenana@ec-lyon.fr	M. Stéphane BENAYOUN Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 stephane.benayoun@ec-lyon.fr
ED 160 EEA	ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE https://edeea.universite-lyon.fr Sec. : Philomène TRECOURT Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 secretariat.edeea@insa-lyon.fr	M. Philippe DELACHARTRE INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 philippe.delachartre@insa-lyon.fr
ED 512 INFOMATHS	INFORMATIQUE ET MATHÉMATIQUES http://edinfomaths.universite-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	M. Dragos IFTIMIE Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 direction.infomaths@listes.univ-lyon1.fr
ED 162 MEGA	MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE http://edmega.universite-lyon.fr Sec. : Philomène TRECOURT Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon mega@insa-lyon.fr	M. Etienne PARIZET INSA Lyon Laboratoire LVA Bâtiment St. Exupéry 25 bis av. Jean Capelle 69621 Villeurbanne CEDEX etienne.parizet@insa-lyon.fr
ED 483 ScSo	ScSo¹ https://edsciencessociales.universite-lyon.fr Sec. : Mélina FAVETON Tél : 04.78.69.77.79 melina.faveton@univ-lyon2.fr	M. Bruno MILLY (INSA : J.Y. TOUSSAINT) Univ. Lyon 2 Campus Berges du Rhône 18, quai Claude Bernard 69365 LYON CEDEX 07 Bureau BEL 319 bruno.milly@univ-lyon2.fr

¹ ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Remerciements

Tout d'abord je souhaite remercier Antoine Boutet, mon encadrant, pour son accompagnement soutenu, pour les projets et opportunités professionnels, pour sa patience même lorsque j'ai dû me montrer entêté et enfin pour son exigence en terme de direction et communication de mes travaux de recherche.

Je tiens à remercier Mathieu Cunche, mon directeur de thèse, pour son écoute, pour avoir trouvé des solutions et pour nos discussions passionnées sur le modélisme.

J'aimerais également remercier Mohamed Maouche pour son appui en encadrement et en méthodologie ainsi que plus particulièrement pour avoir su m'encourager dans les moments difficiles.

À mes collègues de laboratoire, à l'équipe PRIVATICS et compagnons de thèse, en particulier Jan Aalmoes et Samuel Le Pelissier, je suis heureux que nous ayons réussi, tous les trois, à mener à bien nos trois années de thèse. Merci de m'avoir tenu compagnie pendant les pauses café, d'avoir relayé des informations vitales concernant les réinscriptions, et surtout pour votre positivité et votre soutien.

Finalement merci à mes proches pour m'avoir soutenu et encouragé, à mes amis et à mes parents.

LEBRUN Thomas

Résumé

Les données de santé représentent une grande quantité d'informations, générées quotidiennement et sensibles par nature. Cependant, leur partage est essentiel pour l'avancement de la recherche et, en fin de compte, l'amélioration des soins aux patients. L'utilisation des données médicales est confrontée à des limitations dues à leur sensibilité et à la nécessité de garantir la confidentialité, encadrée par les réglementations en vigueur. Cela nécessite une protection renforcée. L'intérêt pour des alternatives au partage de données brutes, telles que la pseudonymisation ou l'anonymisation, augmente avec les besoins d'accès à des données d'apprentissage pour l'utilisation de l'intelligence artificielle, qui requiert de grandes quantités de données pour fonctionner efficacement en tant qu'assistant médical.

Dans cette thèse, nous examinons de nouveaux mécanismes respectant la vie privée, rendus possibles par les avancées rapides de l'intelligence artificielle. Plus spécifiquement, mon analyse porte sur l'amélioration d'alternatives à la centralisation de données sensibles : l'apprentissage fédéré, une méthode décentralisée d'entraînement des modèles d'Intelligence Artificielles qui ne nécessitent pas le partage de données, ainsi que de la génération de données synthétiques, qui crée des données artificielles avec des propriétés statistiques similaires aux données réelles. Considérant l'absence de consensus pour l'évaluation de la confidentialité de ces nouvelles approches, nous avons axé notre travail sur la mesure méthodique de la fuite de confidentialité ainsi que la balance avec l'utilité des données synthétiques ou du modèle d'apprentissage fédéré. Mes travaux incluent un mécanisme pour améliorer les propriétés de confidentialité de l'apprentissage fédéré ainsi qu'une nouvelle méthode de génération conditionnelle de données synthétiques. Cette thèse vise à contribuer au développement de cadres plus robustes pour le partage sécurisé des données de santé, en conformité avec les exigences réglementaires, facilitant ainsi des innovations en matière de santé.

Mots-Clés:

Données Synthétiques, Apprentissage Automatique, Apprentissage Fédéré, Confidentialité, Ré-identification, Attaques d'Attributs Sensibles, Attaques d'Appartenance, Enclaves Sécurisées, Données Personnelles, Données de Santé

Abstract

Health data represents a large volume of information, generated daily and sensitive by nature. However, sharing this data is essential for advancing research and, ultimately, improving patient care. The use of medical data faces limitations due to its sensitivity and the need to ensure confidentiality, which is governed by current regulations. This necessitates enhanced protection. Interest in alternatives to sharing raw data, such as pseudonymization or anonymization, is increasing alongside the growing need for access to training data for the use of artificial intelligence, which requires large amounts of data to function effectively as a medical assistant.

In this thesis, we explore new privacy-preserving mechanism made possible by the rapid advancements in artificial intelligence. More specifically, my analysis focuses on improving alternatives to the centralization of sensitive data: federated learning, a decentralized method of training artificial intelligence models that do not need sensitive data sharing, as well as synthetic data generation, which creates artificial data similar statistical properties to real data. Given the lack of consensus on evaluating the privacy of these new approaches, our work focuses on the systematic measurement of privacy leakage and the balance with the utility of synthetic data or the federated learning model. My contributions include a mechanism to enhance the privacy properties of federated learning, as well as a new method for conditional synthetic data generation. This thesis aims to contribute to the development of more robust frameworks for the secure sharing of health data, in compliance with regulatory requirements, thereby facilitating innovations in healthcare.

Keywords:

Synthetic data, Avatar-based generation, Machine Learning, Federated Learning, Privacy, Re-identification, Sensitive Attribute Attacks, Membership Attacks, Secured Enclave, Personal Data, Health Data

List of Publications and Outputs

- [98] Thomas Lebrun, Antoine Boutet, Jan Aalmoes, and Adrien Baud. MixNN: protection of federated learning against inference attacks by mixing neural network layers. In Proceedings of the 23rd ACM/IFIP International Middleware Conference, Middleware '22. ACM, November 2022. doi: 10.1145/3528535.3565240. URL <http://dx.doi.org/10.1145/3528535.3565240>.
Core A
- [99] Thomas Lebrun, Louis Béziaud, Tristan Allard, Antoine Boutet, Sébastien Gambis, and Mohamed Maouche. Synthetic data: Generate avatar data on demand. URL <https://hal.science/hal-04715055>.
Core B [Currently in publication at Wise2024Qatar December 2024 - Special Track 2: Privacy, Security and Trust in the Digital Space @ WISE-2024]
- Presentation Poster on MixNN - Summer School 2023 on Federated Learning. Villeurbanne

List of Acronyms

Here is an exhaustive list of acronyms used in this manuscript:

- **AE:** Auto-Encoders
- **AI:** Artificial Intelligence
- **AIA:** Attribute Inference Attack
- **CHU:** *Centre Hospitalier Universitaire* (University Hospital Center)
- **CNIL:** *Commission nationale de l'informatique et des libertés* (National Commission for Information Technology and Civil Liberties)
- **DP:** Differential Privacy
- **DP-SGD:** Differentially Private Stochastic Gradient Descent
- **EHR:** Electronic Health Record
- **FL:** Federated Learning
- **GAN:** Generative Adversarial Networks
- **GDPR:** General Data Protection Regulation
- **HIPAA:** Health Insurance Portability and Accountability Act
- **IoT:** Internet of Things
- **LLE:** Locally Linear Embedding
- **MIA:** Membership Inference Attack
- **ML:** Machine Learning
- **MLaaS:** Machine Learning as a Service
- **MRI:** Magnetic Resonance Images
- **MST:** Maximum Spanning Tree
- **PCA:** Principal Component Analysis
- **PETs:** Privacy Enhancing Technologies
- **PIR:** Private Information Retrieval
- **PHI:** Protected Health Information

- **SDV**: Synthetic Data Vault
- **SGD**: Stochastic Gradient Descent
- **SGX**: Intel Software Guard Extensions
- **SNDS**: *Système National des Données de Santé* (National Health Data System)
- **TEE**: Trusted Execution Environment
- **VAE**: Variational Auto-Encoders
- **WSN**: Wireless Sensor Networks

Contents

1	Introduction	1
1.1	Context: Health Data and AI	2
1.1.1	Health Data at the Heart of Economic and Political Challenges	3
1.1.2	Health Data, Privacy Issues and Regulation	4
1.1.3	Emerging Methods to Protect Sensitive Data	5
1.2	Research Problematic	6
1.3	Contributions: Improving Privacy-Enhancing Machine Learning for Healthcare . .	6
1.3.1	Protecting Federated Learning Against Inference Attacks by Mixing Neural Network Layers	7
1.3.2	Synthetic Data Generation by Conditional Local Modelling	7
1.4	Outline	8
2	Background and Related Work	10
2.1	Introduction	11
2.2	Machine Learning Fundamentals	11
2.2.1	The Data Paradigm in Research and the Impact of Massive Data Collection	11
2.2.2	Neural Networks and loss functions	12
2.3	Privacy Preserving Mechanisms	14
2.3.1	Limits of Classical Anonymization	14
2.3.2	Differential Privacy	15
2.4	Privacy Enhancing Methods for Machine Learning	16
2.4.1	Federated Learning	17
2.4.2	Synthetic Data Generation	19
2.5	Health Data: Types and Applications	21
2.5.1	Applications on Textual Health Data	22
2.5.2	Applications on Genomic Data	23
2.5.3	Applications on Image Data and Magnetic Resonance Imaging	24
2.5.4	Applications on Time Series and Wearable and Sensor Data	24
2.5.5	Applications on Tabular Health Data	25
2.5.6	On the difficulty to centralize Health Data	26
2.6	Evaluating the Effectiveness and Applicability of Privacy-Enhancing Machine Learning	27
2.6.1	Impact of Federated Learning on Utility	27
2.6.2	Applicability and Quality of Anonymized and Synthetic Health Data . . .	28
2.7	Privacy Risks Associated with Health Data	30
2.7.1	Privacy Risk for Machine Learning	31
2.7.2	Federated Learning Privacy Risks	34
2.7.3	Anonymization and Synthetic Data Privacy Risk	35
2.8	Synthesis of Key Concepts	38

3	MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers	39
3.1	Introduction	40
3.2	Background for our Contribution	43
3.2.1	Privacy Risks in Federated Learning	43
3.2.2	Mixnets	44
3.2.3	Intel SGX	45
3.3	System and adversary model	45
3.4	Contribution: The MixNN Framework	46
3.4.1	Overview	46
3.4.2	Utility Equivalence	47
3.4.3	Implementation	48
3.5	Evaluation Setup	49
3.5.1	Dataset	49
3.5.2	Evaluation Metrics	50
3.5.3	Baselines	51
3.5.4	Methodology	52
3.6	Evaluation	52
3.6.1	No compromise with utility	53
3.6.2	Prevent information leakage	54
3.6.3	Robustness of the protection	55
3.6.4	System performance	56
3.7	Related work	56
3.8	Conclusion	57
4	M-Avatar: On-Demand Privacy-Enhancing Synthetic Data Generation Using Local Modeling	59
4.1	Introduction	60
4.2	Avatar Data: Limitations and Improvement	61
4.2.1	The Avatar Approach	62
4.2.2	Our Contribution: The M-Avatar Model	64
4.3	Evaluation Setup	65
4.3.1	Datasets	65
4.3.2	Evaluation metrics	66
4.3.3	Comparative baselines	67
4.3.4	Methodology	68
4.4	Evaluation	69
4.4.1	Understanding the data topology	69
4.4.2	Quantifying the utility loss	70
4.4.3	Measuring the privacy gain	72
4.5	Conclusion	75
5	Perspectives and Conclusion	77
5.1	Thesis Contributions	78
5.2	Privacy Challenges and Future Directions in Healthcare AI	79
5.2.1	Short-Term Perspectives	79
5.2.2	Long-Term Perspectives	80
5.3	Conclusion	82
A	Annexes	S

List of Figures

1.1	Thesis Overview: Reasoning Framework, Research Questions and Contributions	3
2.1	Example of Neural Network.	13
2.2	Federated Learning between Hospitals.	17
2.3	The Generator and Discriminator are learning in an adversarial way.	19
2.4	The auto-encoder compress the information on a reduced space and then rebuilds it.	20
3.1	Operating flow of Federated Learning.	43
3.2	MixNN introduces a proxy which receives the parameter updates from each participant, shuffle them to remove attribute footprint before to route them to the aggregation server.	45
3.3	Implementation and data-flow of the MixNN proxy.	47
3.4	MixNN provides the same utility than a standard FL scheme, noisy gradient however decreases significantly the utility and slows down the convergence.	50
3.5	Most of the participants have an accuracy with noisy gradient smaller than MixNN for all datasets.	51
3.6	MixNN better prevents the membership attack compared to a classical FL and a pruning strategy.	52
3.7	With MixNN, most participants benefit from a protection comparable to that provided by the use of a noisy gradient against a membership inference attack.	53
3.8	MixNN better prevents sensitive attribute leakage compared to using noisy gradient.	54
3.9	Reconstruction of the model updates is costly and gives a poor accuracy.	55
3.10	Time spent by MixNN to process the updates is linear according to the number of updates.	56
4.1	Avatars are generated for each record through stochastic averaging of its neighborhood.	62
4.2	To generate synthetic data of arbitrary size, M-Avatar conditionally samples in each of the first d dimensions of the latent space.	64
4.3	The long tail in the distribution of the distance to barycenter highlights the presence of few data at the edge which is more marked on AIDS.	69
4.4	Avatars tend to be near their original data.	69
4.5	Survival curves: the faithfulness of results is correlated to the size of the latent space of SAIPH, the larger, the better.	70
4.6	The survival curve provided by SynthPop and M-Avatar are very close to the one obtained with the original data.	70
4.7	The statistical properties of the original data (captured by the SDV score) are preserved by all synthetic data generation schemes, except for K-anonymity which degrades significantly the data.	71
4.8	The balanced accuracy of a prediction task trained from synthetic data varies according to the method used: MST and the avatar-based approach provide the best performance.	72

4.9	The risk of attribute inference is slightly higher for avatar data on AIDS than for other synthetic data.	72
4.10	The risk of singling out is very low for all synthetic data generation schemes.	73
4.11	The risk of linkability is generally greater for avatar data than for synthetic data from other methods.	73
4.12	The risk of re-identification is much more important for avatar data at the edge.	74
4.13	No record is re-identified on each generation on AIDS.	74
4.14	The risk of membership inference for Avatar is homogeneous for all AIDS data.	75
4.15	M-Avatar reduces the risk of membership inference compared to Avatar . (Here 0.5 represent the null risk).	75

Chapter 1

Introduction

Contents

1.1	Context: Health Data and AI	2
1.1.1	Health Data at the Heart of Economic and Political Challenges	3
1.1.2	Health Data, Privacy Issues and Regulation	4
1.1.3	Emerging Methods to Protect Sensitive Data	5
1.2	Research Problematic	6
1.3	Contributions: Improving Privacy-Enhancing Machine Learning for Healthcare	6
1.3.1	Protecting Federated Learning Against Inference Attacks by Mixing Neural Network Layers	7
1.3.2	Synthetic Data Generation by Conditional Local Modelling	7
1.4	Outline	8

1.1 Context: Health Data and AI

Artificial Intelligence (AI) is at the heart of today's technological innovation, even modifying deeply the research paradigm [72] focusing on data driven approaches. Its growth has been fueled by the increasing collections of information, which serve as the backbone for Machine Learning (ML). The healthcare field benefits as well from AI rising, that helps for decision assist on complex tasks such as diagnosis, treatment plans, and overall patient care [141, 177, 52].

In France, public healthcare institutions such as the *Centres Hospitaliers Universitaires* (CHU) play a central role in this transformation by collecting medical information. The sensitive nature of this data, which includes personal health information, necessitates careful management to protect patient privacy. This is overseen by the *CNIL*¹ (*Commission nationale de l'informatique et des libertés*), which ensures compliance with the GDPR² (General Data Protection Regulation). However, the collection and utilization of such sensitive information pose significant ethical and practical challenges.

Recent events underscore the need for enhanced protection of health data. In 2023 alone, at least eleven major health data breaches occurred³, including incidents involving large organizations like PharMerica and Welltok in the USA, where millions of patient records were exposed due to cyberattacks. These breaches highlight the severe privacy risks associated with health data and the critical importance of robust data protection measures. Nonetheless, this data leakage does not only happen overseas, in France, here are several major cyberattacks in the last years:

1. 17th March 2022 - French "*Assurance Maladie*"⁴ and over 510 000 personal data of individuals were stolen.
2. 31st July 2023 - *CHU Rennes*⁵ where over 300 Go of medical data were found back over the following days on the black market.
3. 15th January 2024 - *CHU NANTES*⁶ closing the hospital for several days.
4. 11th February 2024 - *CH Armentières*⁷ where hospital data were ransomed.

Those leakage hold two major issues, they not only compromise patient privacy but also harm public trust in the systems designed to protect and enhance the value of this sensitive information. Moreover, the controversy surrounding Google's collaboration with Ascension, one of the largest health-care networks in the United States, rises questions over how health data is shared and used by large tech companies [100]. Those concerns about how transparency and consent are handled reinforce the need for clear regulations and data practices that guarantee patient rights while enabling innovation.

¹<https://www.cnil.fr/fr>

²<https://gdpr-info.eu/>

³<https://www.chiefhealthcareexecutive.com/view/these-are-the-11-biggest-health-data-breaches-of-2023>

⁴<https://incyber.org/article/assurance-maladie-fuite-de-donnees-de-510-000-assures/>

⁵<https://www.usine-digitale.fr/article/chu-rennes.N2157752>

⁶[https://www.ouest-france.fr/societe/cyberattaque/le-chu-de-nantes-victime-dune-cyberat\[... \]](https://www.ouest-france.fr/societe/cyberattaque/le-chu-de-nantes-victime-dune-cyberat[...])

⁷[https://www.francetvinfo.fr/internet/securite-sur-internet/cyberattaques/nord-1-hop\[... \]](https://www.francetvinfo.fr/internet/securite-sur-internet/cyberattaques/nord-1-hop[...])

While AI is transforming healthcare, it also introduces challenges. How can we balance the need for innovation while protecting individual privacy? What measures can be implemented to ensure that health data is used securely?

These questions are central to AI integration into healthcare and form the basis of this thesis. In the following sections, we will underline the importance of health data and explore privacy-enhancing machine learning technologies, such as Federated Learning—which allows multiple healthcare providers to collaboratively train AI models without sharing raw patient data—and synthetic data generation techniques that create realistic medical datasets while enhancing patient privacy. We will describe our contributions to enhancing their privacy protections, including the development of secure aggregation protocols and Differential Privacy mechanisms, and demonstrate how these methods maintain their effectiveness in healthcare applications through extensive evaluations and case studies.

Our **thesis overview** (approach and research questions) and the two contributions of this thesis are illustrated in Figure 1.1, providing a comprehensive overview of our research process.

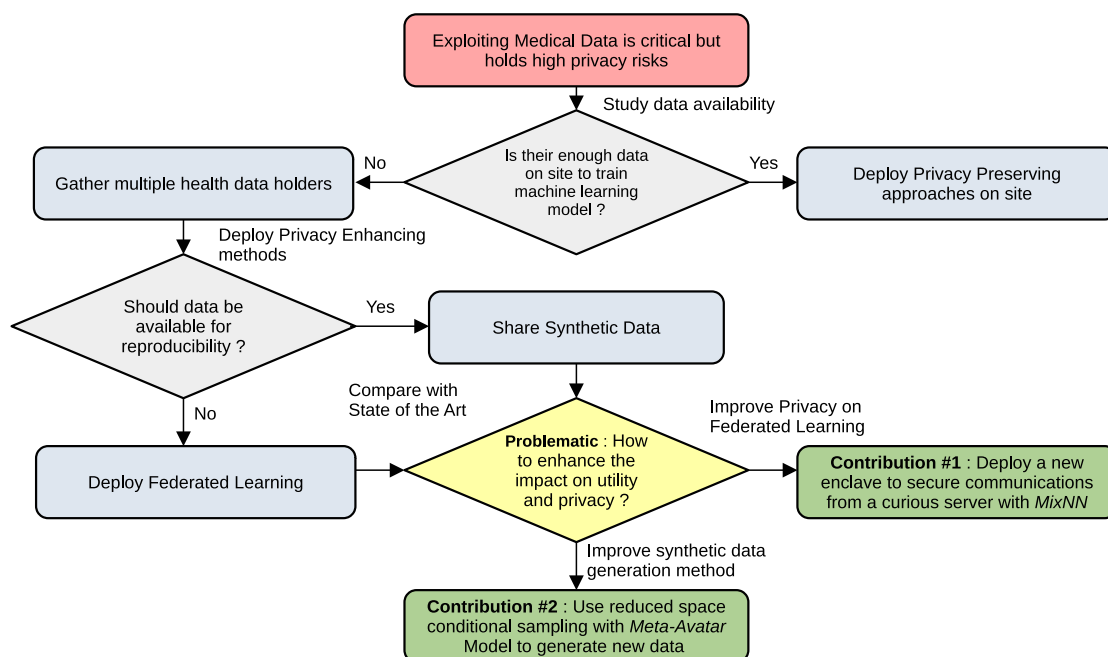


Figure 1.1: Thesis Overview: Reasoning Framework, Research Questions and Contributions

1.1.1 Health Data at the Heart of Economic and Political Challenges

The importance of health data sharing has been demonstrated in recent times, particularly during the COVID-19 pandemic [56, 174]. It is critical for public health by enabling researchers and policymakers to make informed decisions to protect and improve public health: tracking the spread of diseases, evaluating the effectiveness of treatments, assessing the impact of healthcare campaigns and others.

During the pandemic, data driven approaches such as France's *TousAntiCovid* app⁸, launched during the COVID-19 pandemic, helped to assist in contact tracing and managing the virus's spread in a privacy preserving way. Data collection during the pandemic allowed real-time analysis, supporting resource allocation and vaccines development. Nevertheless, health data is also utilized beyond emergency situations to as well support medical research and healthcare decisions and services. In France, the *Système National des Données de Santé* (SNDS) is a pioneering initiative dedicated to centralizing health data from hospitals, health insurance, and mortality records. The SNDS provides a centralized, secure repository for health data, enabling researchers to access large datasets while ensuring strict data protection protocols. This initiative is further developed in section 2.5.6 and demonstrates that centralized health data management is garnering significant interest from governments. Additionally, the establishment of the Health Data Hub (HDH)⁹ in France and the European Health Data Space (EHDS)¹⁰ at the European level demonstrate the political commitment to advancing centralized health data infrastructures.

As well as governments, global tech giants such as Amazon (Amazon One Medical¹¹) and Google (Google Health¹²) have also shown their interest by deploying specialized branches about health data. These companies are investing to reach the healthcare sector and provide personalized medicine services and predictive analytics which both heavily rely on data collection. The centralization of such sensitive data by private entities, the lack of transparency of their processes for data management (including security) and usages is a concern for people privacy which could increase with the development of digital wearable in the healthcare sector (watches, scales, blood pressure monitor...).

1.1.2 Health Data, Privacy Issues and Regulation

The GDPR [1] is a regulation of the European Parliament and of the Council of 27 April 2016. It guarantees the protection of personal data with regard to its processing and forbid its communication without explicit consent. Personal Data is any kind of information that can be linked to an individual. Thanks to the GDPR, sharing health data is strictly regulated while the CNIL¹³ [34] classifies those data as the most sensitive. Therefore, the GDPR (Article 9¹⁴) prohibits the processing and sharing of health data without the explicit consent of the individual, except in specific cases such as those justified by medical necessity, public health concerns, and even then, only with appropriate safeguards to protect data privacy.

Similarly, to the GDPR in Europe, the Health Insurance Portability and Accountability Act (HIPAA) [139] in 1996, defines rules for protecting the privacy and security of individuals' medical information in the U.S. As well as the GDPR, HIPAA establishes national standards to guarantee that sensitive health data is not disclosed without the individual's explicit consent or awareness. It outlines strict guidelines for how healthcare providers, insurance companies, and other entities must handle,

⁸<https://www.campusfrance.org/en/tousanticovid-a-new-application-to-fight-the-epidemic>

⁹<https://www.health-data-hub.fr/>

¹⁰<https://www.european-health-data-space.com/>

¹¹<https://health.amazon.com/onemedical>

¹²<https://health.google/>

¹³<https://www.cnil.fr/en/cloud-risks-european-certification-allowing-foreign-authorities-access-...>

¹⁴<https://gdpr-info.eu/art-9-gdpr/>

transmit, and store Protected Health Information (PHI). The Privacy Rule under HIPAA restricts the use and sharing of PHI, while the Security Rule requires robust measures to protect electronic health data from breaches and unauthorized access.

In addition to the GDPR, European authorities introduced the AI Act¹⁵ in early 2024: a landmark regulatory framework designed to guarantee safe and ethical use of artificial intelligence, such as in sensitive sectors like healthcare. The AI Act categorizes AI systems based on their risk levels, with unacceptable risk systems, such as social scoring, being completely prohibited; high-risk AI systems, such as those used in medical applications, are allowed but are subject to regulations to protect patient privacy and ensure data security. The Act imposes compliance responsibilities on AI providers (developers), particularly when dealing with medical data, underlying the necessity for robust safety measures, transparency, and accountability. Additionally, General Purpose AI models, which could also be applied in healthcare, must meet specific transparency and cybersecurity standards, especially when presenting systemic risks, to ensure that personal health data remains protected.

Health data, being highly personal, can reveal identifying and sensitive information about an individual, which could lead to discrimination or other violations of their rights (assurances adapting their prices, loan being refused). Therefore, the GDPR and HIPAA impose strict conditions on the handling of such data, including the need: 1. to minimize data collection, 2. to ensure anonymization when possible or pseudonymization at least, and 3. to implement robust security measures to prevent unauthorized access or data breaches.

1.1.3 Emerging Methods to Protect Sensitive Data

Given that health data is on one hand, strictly protected by regulations such as HIPAA, GDPR and with extension AI Act, and on the other hand also crucial for medical research and health services development, the key question becomes: How to protect health data to facilitate their sharing without reducing their value? Personal data refers to any sensitive information that can be linked to an individual, and anonymizing such data involves removing this link to protect privacy.

To share health data in a privacy-preserving manner, it must be anonymized according to the criteria outlined by the Group of National Data Protection Authorities (G29)¹⁶. These criteria include techniques such as data masking, pseudonymization, and aggregation, which are designed to prevent the re-identification of individuals. Anonymized data no longer falls under the regulation of GDPR nor HIPAA because it loses its identifying personal features. However, enhancing privacy often comes at the cost of reducing the data's utility [181], as crucial information may be removed in the process. This highlights that high privacy and data utility generally do not coexist well.

There is a trade-off, and the objective is to obtain the minimum confidentiality necessary to provide the highest utility. This trade-off is at the root of our research problem.

¹⁵<https://artificialintelligenceact.eu/high-level-summary/>

¹⁶[https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216\[... \]](https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216[...])

1.2 Research Problematic

The aim of this research is to enhance privacy protection in the use of Machine Learning on medical health data without compromising its re-usability for various purposes, including medical research. Achieving this balance is a challenge due to the trade-off between privacy and data utility. Specifically, two obstacles are limiting effective data sharing and learning: protecting data privacy during the sharing process and preserving data utility during the machine learning phase. This thesis investigates emerging approaches, including Federated Learning, a decentralized machine learning algorithm, and synthetic data generation, to address these challenges. As illustrated in our **thesis overview** on Figure 1.1, our objective is to evaluate the extent to which these methods protect privacy and balance the trade-off with data utility. During my work of research, I used an empirical approach and evaluated my results on real-world data to ensure their validity and applicability in practical scenarios.

1.3 Contributions: Improving Privacy-Enhancing Machine Learning for Healthcare

As previously explained, the collection and sharing of health data, which is highly sensitive by nature, are not legally feasible without patient consent. Since obtaining such consent is unlikely and difficult, requiring thrust in technologies used to protect privacy, this highlights the need to evaluate the privacy properties of emerging solutions: a central theme in my research. Two alternative approaches are drawing major interests.

The first approach, Federated Learning (FL), involves distributing the learning task to each data source (participants/clients) and exchanging model updates between a central server and the participants. This method limits the leakage of sensitive information as no data is collected in the process. However, despite not sharing raw data, sensitive information can still be inferred from the exchanged models. Therefore, it is essential to evaluate and quantify these potential leaks to ensure that FL maintains robust privacy protections. This is where my research work intervenes by assessing the privacy properties of FL and proposing enhancements to mitigate these risks.

The second approach, synthetic data generation and sharing, allows for the dissemination of data that retains the statistical properties and information of the original dataset while providing better control over the amount of information shared. Sharing synthetic data facilitates the republishing of information for new health studies, as this data falls outside the scope of GDPR, thereby promoting open science and ensuring the reproducibility of results. My research critically evaluates the privacy guarantees of synthetic data generation methods to ensure that they offer sufficient protection without compromising data utility.

Such approaches are reducing the privacy leakage as the original data with sensitive information is not directly shared. However, the privacy risk is not fully mitigated, as there is no theoretical guarantee of protection, and this risk is often underestimated in the literature. There is still progress available for a better protection and privacy evaluation for privacy enhancing machine learning.

The list of my main contributions is further detailed below:

- [98] Thomas Lebrun, Antoine Boutet, Jan Aalmoes, and Adrien Baud.

MixNN: protection of federated learning against inference attacks by mixing neural network layers. In Proceedings of the 23rd ACM/IFIP International Middleware Conference, Middleware '22. ACM, November 2022. doi: 10.1145/3528535.3565240. URL <http://dx.doi.org/10.1145/3528535.3565240>.

- [99] Thomas Lebrun, Louis Béziaud, Tristan Allard, Antoine Boutet, Sébastien Gambs, and Mohamed Maouche. Synthetic data: Generate avatar data on demand. URL <https://hal.science/hal-04715055>. [Currently in publication at Wise2024Qatar December 2024 - Special Track 2: Privacy, Security and Trust in the Digital Space @ WISE-2024]

1.3.1 Protecting Federated Learning Against Inference Attacks by Mixing Neural Network Layers

Federated Learning (FL) was introduced by McMahan et al. in 2016 [114]. This decentralized approach to machine learning is gaining traction in healthcare as it enhances privacy, which is critical in both health and research applications, as demonstrated by initiatives like FedBioMed¹⁷ and startups such as Tune Insight¹⁸. Unlike traditional methods that require centralizing data in a single location, FL allows models to be trained locally on decentralized data sources: sensitive patient information never leaves its original location (sources also known as participants or clients). This approach is seen as a significant improvement over centralized data collection without anonymization, as it Differential Privacy the risk of exposing sensitive information.

Nonetheless, FL is not without limitations. Although data is not directly shared, the FL framework can still be vulnerable to attacks such as model inversion and gradient leakage, where adversaries attempt to infer sensitive information from the model updates. Those risks have been widely proved in the literature [50, 169, 45].

To resolve such challenges, solutions on server side exist to secure the communications from inference attacks [136]. However, this approach requires trusting the aggregation server, which is not always feasible if the server may be curious and attempt to infer sensitive information from the aggregated updates.

To further enhance privacy without relying on trust in the aggregation server, we propose MixNN: an intermediary enclave between clients and the aggregating server that allows clients to improve their privacy without trusting the server. MixNN encrypts updates between the enclave and the clients, decrypts the updates within a secure computation enclave, and mixes the updates before returning them to the server. This process protects the updates from privacy leakage by a curious server while maintaining short execution times and preserving model performance.

1.3.2 Synthetic Data Generation by Conditional Local Modelling

Synthetic data sharing as an alternative to give access to sensitive data is also emerging as a promising solution for ML in healthcare. Synthetic data is any form of data that was not issued from the measure of a real event: it is adjusted to mimic real data to create not real but realistic data samples [79]. Additionally, synthetic data enhances the reproducibility of research by enabling open science, where datasets can

¹⁷<https://fedbiomed.org/>

¹⁸<https://tuneinsight.com/>

be freely shared and allow the scientific community to reproduce results of pairs. As healthcare increasingly relies on data and machine learning, synthetic data provides a promising solution. It serves as a compromise between traditional anonymization approaches: pseudonymization, which offers poor privacy protection, and Differential Privacy, a solution that does not offer a compromise compatible with healthcare needs. While synthetic data is not a perfect solution, it offers a better balance by striving to maintain both utility and privacy. This makes it possible to share and utilize data across different institutions and jurisdictions, addressing stringent privacy regulations like HIPAA, GDPR and AI Act.

Nevertheless, there is a strong misconception that not directly sharing real data completely eliminates privacy leakage for the original data. In reality, privacy risks still exist and are challenging to evaluate due to the evolving nature of privacy attacks. The field of synthetic data generation lacks standardized evaluation methods, hindering the generalization of privacy assessments. Furthermore, privacy risks and anonymization are not binary properties; instead, they exist on a spectrum. Additionally, the level of privacy risk is often not uniform across all data records, as some records may contain more sensitive information than others. This heterogeneity in risk necessitates nuanced approaches to privacy protection to ensure that more sensitive data are adequately safeguarded without excessively compromising overall data utility. To tackle this challenge, we provided a large framework of evaluation of privacy and utility to evaluate the compromise proposed by various generative approaches in the state-of-the-art as there is still no general approach in the literature. We then propose a novel generation approach to create realistic synthetic tabular data with a high trade-off of privacy and utility. Preserving the relationships between variables is often challenging when generating data. To address this, we generate new data in a reduced Principal Component Analysis (PCA) space. In this lower-dimensional space, the data is encoded using fewer variables. We then conditionally generate each variable and subsequently project the data back into the original high-dimensional space. We then compared our solution with existing ones and found a competitive privacy-utility trade-off in regard of the state-of-the-art while keeping a reasonable computation time.

1.4 Outline

This thesis will be divided as follows.

- At first, in Chapter 2, we will provide background knowledge to better contextualize and understand the privacy enhancing machine learning approaches on health data. The paradigm of data-driven research as well as basics concepts and behaviors of machine learning will be presented. Then we will have an overview of privacy enhancing methods, how Synthetic Data and Federated Learning fit among them, as well as their applications to health data. We will then provide insights from the state-of-the-art to answer the problematic in our **thesis overview** (Figure 1.1) about the evaluation of the privacy and utility of both types of approaches.
- Then, in Chapter 3, we will analyze our first contribution in Federated Learning: we will explain our method of mixing neural networks layers after providing context, before evaluating its impact.

- Next, in Chapter 4, we will position our innovation in the context of synthetic data, describe our approach of conditional local modelling in detail and how to evaluate it.
- Finally, in Chapter 5, we will bring together our contributions in a final section where we will discuss their strengths as well as their limitations, consider future work and conclude.

Chapter 2

Background and Related Work

Contents

2.1	Introduction	11
2.2	Machine Learning Fundamentals	11
2.2.1	The Data Paradigm in Research and the Impact of Massive Data Collection	11
2.2.2	Neural Networks and loss functions	12
2.3	Privacy Preserving Mechanisms	14
2.3.1	Limits of Classical Anonymization	14
2.3.2	Differential Privacy	15
2.4	Privacy Enhancing Methods for Machine Learning	16
2.4.1	Federated Learning	17
2.4.2	Synthetic Data Generation	19
2.5	Health Data: Types and Applications	21
2.5.1	Applications on Textual Health Data	22
2.5.2	Applications on Genomic Data	23
2.5.3	Applications on Image Data and Magnetic Resonance Imaging	24
2.5.4	Applications on Time Series and Wearable and Sensor Data	24
2.5.5	Applications on Tabular Health Data	25
2.5.6	On the difficulty to centralize Health Data	26
2.6	Evaluating the Effectiveness and Applicability of Privacy-Enhancing Machine Learning	27
2.6.1	Impact of Federated Learning on Utility	27
2.6.2	Applicability and Quality of Anonymized and Synthetic Health Data	28
2.7	Privacy Risks Associated with Health Data	30
2.7.1	Privacy Risk for Machine Learning	31
2.7.2	Federated Learning Privacy Risks	34
2.7.3	Anonymization and Synthetic Data Privacy Risk	35
2.8	Synthesis of Key Concepts	38

This chapter introduces essential background on machine learning, the evaluation of its performance and associated privacy risks. It also introduces Federated Learning and synthetic data as alternatives to sharing sensitive information and knowledge. This overview sets the context for the research conducted in this thesis.

2.1 Introduction

Data is central in the healthcare sector, providing insights for clinical decision-making, medical research, and public health policy. Yet, its utilization comes with significant privacy concerns because of the sensitivity of personal data.

On one hand, data holders may wish to share this information to better address public health issues, but on the other hand, the impact on patient privacy could be unacceptable. Currently, as presented by Qayyum et al. [136] as well as Khalid et al. [87] in their surveys, two methods stand out as alternatives to sharing sensitive data: Federated Learning (as no data is shared) and anonymized synthetic data sharing as briefly described in the Introduction chapter. Federated Learning as well as most of popular Synthetic Data generation belong to Machine Learning and therefore inherits of its risks. They do not propose a perfect share of general information and a perfect protection of sensitive information so such trade-off has to be considered through studying risks inherent to machine learning.

This chapter provides background knowledge on general machine learning, their performance evaluation and their risk for user privacy. Given this overview, we further explain Federated Learning and synthetic data as alternatives to directly share sensitive information. Then we present the types of health data, their crucial role in medical research and patient diagnosis. The privacy risks associated with their publication will be described. Finally, the application on health data of these two alternative sharing methods and their often under evaluated impact on privacy will be explained. This will help to better situate the research work carried out in this thesis.

2.2 Machine Learning Fundamentals

This section introduces the basic and general elements used in this chapter. First, we describe how the research paradigm change with the massive collection of data from which the machine learning gained in interest. Next, we present the basis of most machine learning approaches used in this research work: the neural network and their convergence.

2.2.1 The Data Paradigm in Research and the Impact of Massive Data Collection

The current research paradigm is focused on the exploitation of massive datasets, from which statistical or AI models are derived. This approach has had a significant impact on all disciplines and has led to an emphasis on large-scale data collection, as data has become the lifeblood of research. These advancements have been facilitated by the increasing digitization of healthcare records and the widespread adoption of EHRs, which store vast amounts of patient data. The accumulation of such data enables the

development of advanced AI models but also raises significant privacy concerns, as highlighted by Vijayan et al. [164].

The collection of this vast amount of data is subject to user consent, with the likelihood of consent being studied by Shandhi et al. [146]. While the abundance of data has permit advancements in AI and personalized medicine, it also poses significant threats to patient privacy. The aggregation and analysis of sensitive health information increase the risk of data breaches, unauthorized access, and misuse of personal data. For instance, even anonymized datasets can be vulnerable to re-identification attacks, compromising individual privacy. This shift from a hypothesis-driven approach to a data-driven approach in research has led to the evolution of privacy protection laws. These developments underscore the need for robust privacy-enhancing technologies to mitigate threats and protect individual rights in the era of big data.

2.2.2 Neural Networks and loss functions

An ML model is a function $f_\theta : x \mapsto y$ parameterized by a set of parameters θ , where x denotes the input (or feature) space ($x = x_1, x_2, \dots, x_k$), and y the output space ($y = y_1, y_2$). Training an ML model consists of finding the optimal set of parameters θ that fits the training data. This is done by optimizing an objective function (loss, see below) which penalizes the model when it is wrong. For instance, if we consider a classification task trained through a supervised learning, parameters θ are updated if the model misclassifies training data.

The loss function, often referred to as the cost function or objective function, is the function that the model optimizes during training. One common loss function used in classification tasks is the cross-entropy loss, which is defined as follows:

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c})$$

Where:

- N is the number of training examples.
- C is the number of classes.
- $y_{i,c}$ is the ground truth label for the i th example and class c , where $y_{i,c} = 1$ if the example belongs to class c , otherwise $y_{i,c} = 0$.
- $\hat{y}_{i,c}$ is the predicted probability that the i th example belongs to class c .

This loss function is minimized during the training process to improve the accuracy of the model's predictions. For regression tasks, other losses are used: it is quite current to see the average of the L1 or L2 distance, respectively MAE for mean absolute error and MSE for mean squared error. The L^i distance between two vectors \mathbf{x} and \mathbf{y} in R^n being defined as:

$$\|\mathbf{x} - \mathbf{y}\|_i = \left(\sum_{k=1}^n |x_k - y_k|^i \right)^{\frac{1}{i}}$$

Where:

- $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$ are vectors in R^n .
- i is the order of the norm.
- n is the dimension of the vectors.

Neural networks are a family of ML models which have become popular for a variety of ML tasks. They were introduced during the second half of the twentieth, firstly by Rosenblatt et al. [140] on its most elemental state: the perceptron and then generalized on larger networks by Rumelhart et al. [142]. A neural network is composed of multiple layers of non-linear mappings from input to intermediate hidden states (or hidden layers) and then to output where each layer transforms the output of the preceding layer to produce input for the next layer. The topology of the connections between layers and the type of considered transformation function are task-dependent and impact the accuracy of the model.

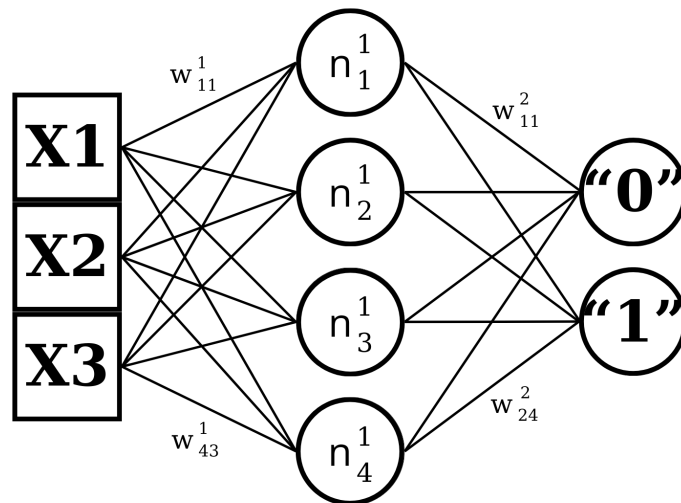


Figure 2.1: Example of Neural Network.

The above Neural Network structure has two layers, one first of size 4 and one second of size two used for a binary classification: it learns to classify three-dimensional data as 0 or 1.

A neural network f is composed of a collection of n hidden layers ($f = (l_1, l_2, \dots, l_n)$). Each layer l_i is composed of a set of m neurons ($l_i = (n_1^i, n_2^i, \dots, n_m^i)$). For input x , the output of the neural network, can be formally written as:

$$f_{\theta}(x) = F_n(F_{n-1}(\dots F_2(F_1(x))))$$

Where x is the input, n the number of layers, F_i represents a transformation function of the layer l_i , and θ is the set of floating-point weights associated with each connection between two neurons of different layers. Considering a fully connected neural network (as depicted in Figure 2.1), θ_{ab}^t represents the weight connecting the node n_a^t to the node n_b^{t-1} .

These weights are updated during training according to Stochastic Gradient Descent (SGD) to optimize the objective function. SGD is an iterative approach where the optimizer receives a batch of training data and updates the model parameters θ at each iteration according to both the direction of the gradient of the objective function

and a learning rate η which scales the update. Once the gradient is close to zero, the model has converged to a local minimum and the training is finished. The model is evaluated through its accuracy over testing data points not used to train the model. The hyperparameters refer to the set of tunable parameters not related to the neural network (e.g., weights associated with connections), such as the number of training iterations, the size of the training batch, the learning rate, and the number of hidden layers.

2.3 Privacy Preserving Mechanisms

In Privacy and Freedom [166], Alan Westin defines privacy as the right of individuals to control the access from others to their personal information. From this short definition, it appears clearly that increasing privacy implies the loss of information shared. Machine Learning models are using this information to solve complex tasks so it also follows that there is a trade-off between utility of models and the privacy. Therefore, there is a difficulty to correctly build privacy without destroying the utility as pointed by Maaten et al. [107]. Increasing privacy, by using anonymization of privacy by design algorithms (as Differential Privacy), is a complex task and will be presented in this section as well as their limits.

2.3.1 Limits of Classical Anonymization

Data anonymization implies to remove any personal and sensitive information that can lead to re-identification. A first approach would be to make indistinguishable group of records to ensure anonymity, this is how the three following approaches anonymize the data: k-anonymity, l-diversity and t-closeness. Each of them is an improvement of the precedent and their mechanisms as well as their limitations will be presented in the following subsection.

K-Anonymity, as Sweeney pointed out in [154], removing obvious identifiers in medical records—such as name, address, and telephone number—does not guarantee patient anonymity. There remains a risk that combining other pieces of information can lead to the re-identification of patients. Latanya Sweeney directly proved it by re-identifying Massachusetts Governor William Weld’s health records by purchasing Cambridge voter rolls and cross-referencing them with anonymized Group Insurance Commission of Massachusetts data. To overcome the limitations of those insufficient anonymization techniques, Sweeney et al. [155] presented k-anonymity in 2002. Their approach is about ensuring that each data record is indistinguishable from k-1 other records. Nevertheless, this technique being simple to compute and reducing the re-identification risk, it still holds privacy concern. First this approach provides a uniform privacy protection independently of the privacy risk of the records that is not uniform, which lead the second limitation of this approach. The k-anonymity does not protect against attribute disclosure as it does not take into account the diversity of sensitive attributes within each anonymized group. Despite each record being hidden behind an anonymized group, the group can still leak sensitive information due to the potential lack of diversity. This is to manage this issue that l-diversity was developed.

L-Diversity, to reduce the privacy risks of k-anonymity, Machanavajjhala et al. [108] introduced l-diversity in 2007. It enhances the process by ensuring that each group of records that share the same quasi-identifiers contains at least l different values for the

sensitive attribute. Despite the attribute disclosure risk being reduced, it still holds some drawbacks. It strongly distorts the data reducing its utility for practical use. Moreover, it still holds vulnerabilities to inference attacks as highlighted by Li et al. [102] as they introduced t-closeness to overcome those shortfalls.

T-Closeness, Li et al. introduced t-closeness [102] after l-diversity to limit the inference attack risk. It requires that the distribution of a sensitive attribute within each group of records is close to the distribution of the attribute in the entire dataset. Despite that its deployment is compromised when there is multiple sensitive attributes or large attribute domains and has a computational overhead. The tuning of the t parameter, which serves as the threshold for the maximum allowable difference between the distribution of sensitive attributes within an equivalence class and the overall dataset, is difficult and limits its practical utilization. Furthermore, it also degrades the data utility as bad as l-diversity does. Bindschaedler et al. [19] introduces the concept of plausible deniability as a formal privacy guarantee for releasing sensitive datasets, ensuring that an output record can only be released if it is indistinguishable from a certain number of input records, independent of an adversary's background knowledge. It therefore provides a legal interpretation for GDPR and anonymous data.

Anonymization is a complex task, and privacy risks are often evaluated based on specific attacks, making the actual risk hard to assess and often underestimated. Not all techniques have theoretical bounds on the protection mechanisms they provide. To address such limitations, Differential Privacy has been developed. It is explained in the following section.

2.3.2 Differential Privacy

As data-driven research continues to expand, there is an emerging need in machine learning for better privacy-preserving learning techniques. This is in this context that Differential Privacy was developed by Dwork et al. in 2006 [46]. Differential Privacy is a mathematical framework that comes with strong theoretical guarantees on the privacy impact of their method to statistical inference on sensitive information. The main concept of this method is that the learning process stays the same indifferently if a point is added or removed from the dataset. Doing so, it drastically reduces the risk that an adversary infers whether an information was present or not in such dataset.

A strong interest in Differential Privacy comes into the theoretical guarantee as it follows: given two neighboring datasets D_1 and D_2 that differ by at most one element, and a randomized algorithm (mechanism) \mathcal{M} , the algorithm \mathcal{M} is said to satisfy ϵ -Differential Privacy if for all possible outputs S of the algorithm, the following bound holds:

$$\Pr[\mathcal{M}(D_1) \in S] \leq e^\epsilon \cdot \Pr[\mathcal{M}(D_2) \in S] \quad (2.1)$$

Where:

- \mathcal{M} represents a randomized mechanism or algorithm applied to the datasets.
- D_1 and D_2 are neighboring datasets, differing by at most one element (e.g., one individual's data).
- S is any possible subset of the output space of \mathcal{M} .

- ϵ (epsilon) is the privacy parameter that controls the trade-off between privacy and accuracy. Smaller ϵ values correspond to stronger privacy guarantees.

This bound ensures that the presence or absence of a single individual's data in the dataset does not significantly affect the output distribution of the algorithm, thereby protecting individual privacy.

Dwork et al. [47] extensively discuss the core techniques for achieving Differential Privacy, the applications of these techniques, and the inherent limitations of any method that seeks to prevent complete privacy breakdowns. Their work stress that achieving strong privacy guarantees often requires to rethink rather than to adapt existing algorithms to be privacy-preserving.

For Neural Networks 2.2.2, Differential Privacy is applied on the stochastic gradient descent (DP-SGD) [2] which add calibrated noise during gradient update in the training phase. Such process strongly deteriorates the prediction performance of the model as Heo et al. [70] point out. To overcome this issue, they underline the fact that not all data points have the same privacy risk and the learning process can be improved on points with lower privacy requirements.

By design, Differential Privacy reduces significantly the risk of membership inference attacks, as explained in Section 2.7.1.1. This is because Differential Privacy ensures that the inclusion or exclusion of any single data point does not substantially affect the outcome of a computation, as formalized in both the definitions of membership inference attacks 2.3 and Differential Privacy 2.1. Kairouz et al.[83] provide a bound on the success of MIA as a function of the epsilon parameter. Consequently, it provides theoretical guarantees that protect against such privacy breaches.

While Differential Privacy offers strong privacy assurances, it is also known to compromise the usability of the methods to which it is applied, especially in complex domains like medical data analysis where low accuracy can be unacceptable. This trade-off between privacy and utility making real-world deployment challenging. Current research trends in Differential Privacy focus on mitigating this trade-off by developing advanced techniques that better balance privacy and utility. However, for medical applications, achieving a satisfactory balance remains an unresolved issue.

As a result, alternative approaches have emerged—some leveraging Differential Privacy and others not—such as sharing synthetic data and employing decentralized learning methods. These alternatives aim to provide practical solutions for privacy-preserving data analysis and will be explored in the following section.

2.4 Privacy Enhancing Methods for Machine Learning

Although the previous section introduced the fundamentals of privacy, these methods have limitations when applied to practical scenarios where both utility and privacy are critical. We will now explain Federated Learning as a way to deploy machine learning without communication of data records, to provide a better compromise between privacy and utility. Finally, we detail known approaches to generate synthetic data as a novel method of data anonymization and also improve the utility/privacy trade-off.

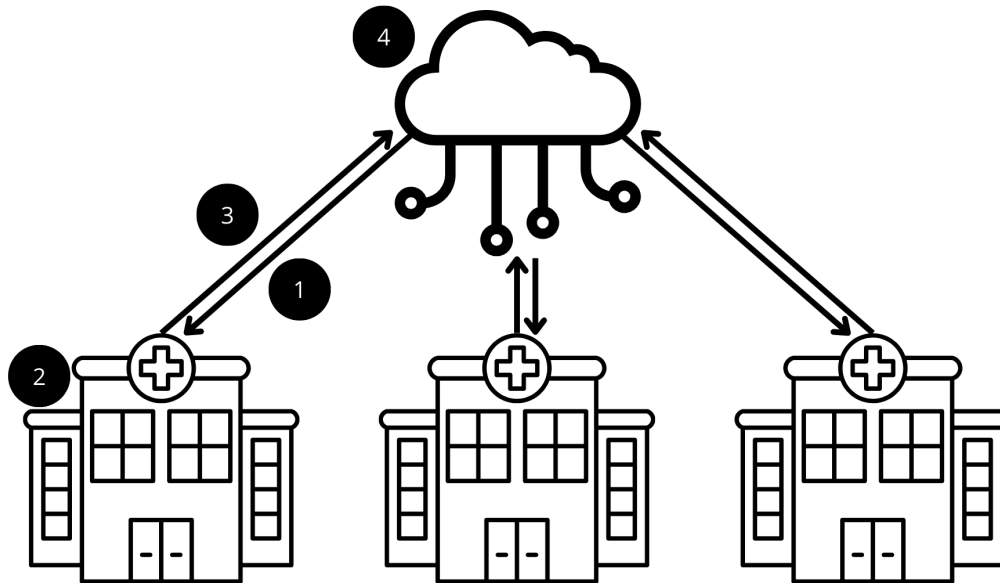


Figure 2.2: Federated Learning between Hospitals.

2.4.1 Federated Learning

Federated Learning is a very promising field of machine learning on the side of decentralized approaches. Here decentralized means that instead of having one device that will store the data and compute the models (known as the classical or centralized approach), the data will stay split among multiple storage and the learning task will be decentralized among them. This method was first proposed by McMahan et al. [114] from Google and further improved for scalability by Bonawitz et al. [22] also from Google and aimed to lower communication costs in machine learning tasks: the communication of large datasets being too costly. The solution here is to split the learning process between clients and to aggregate updates sent by clients to a server. More precisely, here is how this process works, the server initializes the learning process then, while the model is not converged, the following is repeated:

1. The server sends the current model to the clients.
2. The clients update their model on their local data.
3. The clients send back their update to the server.
4. The server aggregates the updates.

More formally here is the mathematical definition of the aggregation on the server. Let:

- K be the total number of clients.
- n_k be the number of data points held by client k .
- $N = \sum_{k=1}^K n_k$ be the total number of data points across all clients.
- \mathbf{w}_t be the global model parameters at the server at round t .
- \mathbf{w}_t^k be the local model parameters of client k after the local update at round t .

- $\Delta \mathbf{w}_t^k = \mathbf{w}_t^k - \mathbf{w}_t$ be the update (change) in model parameters at client k after its local update.

The server updates the global model parameters \mathbf{w}_{t+1} for the next round as follows:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \frac{1}{N} \sum_{k=1}^K n_k \Delta \mathbf{w}_t^k \quad (2.2)$$

Where:

- \mathbf{w}_{t+1} represents the updated global model parameters after aggregation at round $t + 1$.
- $\Delta \mathbf{w}_t^k = \mathbf{w}_t^k - \mathbf{w}_t$ is the local update from client k .
- $\frac{n_k}{N}$ is the weight assigned to the local update from client k , proportional to the size of its local dataset relative to the total dataset across all clients.

This weighted averaging ensures that updates from clients with more data have a greater influence on the global model update, promoting a balanced and fair aggregation across all participating clients. Now that the process has been formalized, here is the reason why this method reduces communication costs: sharing the learning model, even multiple time, is very likely to be less bandwidth consuming than sending large data. A classical application for Federated Learning is mobile keyboard prediction presented by Hard et al. [68] where isolated datasets fail to have enough data to correctly predict the next word, collecting all the datasets is impossible for privacy and bandwidth concerns so Federated Learning is an efficient solution.

However, many challenges remain for Federated Learning practical implementation, such as ensuring privacy, reducing communication costs, improving robustness, and handling heterogeneous data across clients. To address these challenges, various aspects of FL can be modified or enhanced. For example, the aggregation server can perform a median aggregation instead of an average to improve robustness against outliers or malicious updates from compromised clients. FL is inherently a multidisciplinary field involving statistics, machine learning, communication protocols, encryption techniques, and more. Its objectives can vary widely, including optimizing prediction accuracy, minimizing communication costs (as explored by Konečný et al. [91]), and reducing energy consumption (as discussed by Damaskinos et al. [39]).

Despite its advantages, FL is susceptible to attacks from both the server side and through communication channels, similar to vulnerabilities in classical machine learning systems. To mitigate such security risks, Mitchell et al. [120] introduced secure aggregation. The key idea is that the server can decrypt only the aggregated model updates, not the individual updates from each client, thereby preserving privacy. However, this approach introduces computational and communication overhead, which must be effectively managed to maintain the efficiency of the system.

As another alternative to share sensitive data to perform a large centralized approach of machine learning, one other as promising approach than Federated Learning is synthetic data generation which is about sharing data similar yet different to the sensitive original data and will be presented in the following subsection.

2.4.2 Synthetic Data Generation

This subsection provides a broad overview of the families of synthetic data generators that are the focus of this thesis. First, generative adversarial networks and auto-encoders will be introduced, both of which utilize two sets of neural networks. Additionally, attention will be given to other approaches, such as SynthPop, spanning trees, and Bayesian networks.

2.4.2.1 Generative Adversarial Networks (GAN)

Generative Adversarial Networks are one of the most popular approach to generate fake images from zero and one thing that made them popular for the public may be to convert images to a different style like a photo to the style of a painting from Monet [159]. Originally introduced by Goodfellow et al. [58], they consist in a set of two neural networks that compete against the other. The first network, called the generator, learns to produce fake samples that resemble the original data to deceive the second network, the discriminator, which learns to distinguish generated data from authentic data within a dataset. This process is known as adversarial training because both networks are continuously improving in response to the other's advancements. The generator learns to create realistic samples starting from random noise. Once the model is trained, new data is generated by generating noise and passing it through the generator to create a realistic synthetic sample.

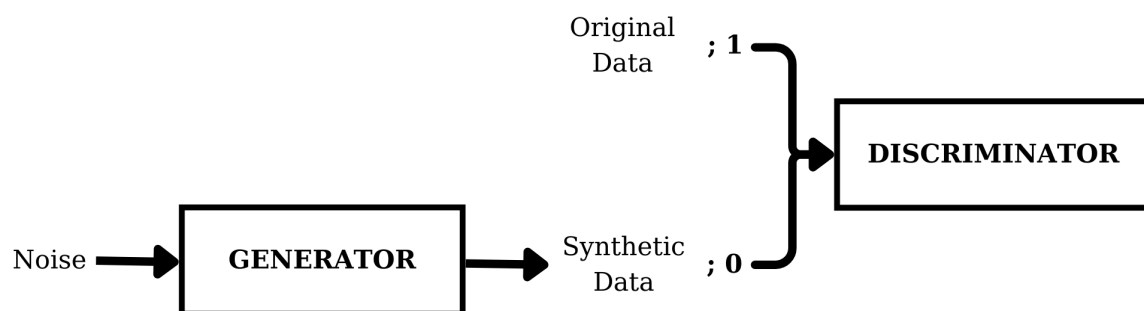


Figure 2.3: The Generator and Discriminator are learning in an adversarial way.

In extension to GANs, Mirza et al. propose CGAN [119], which provide additional information to both generator and discriminator to the generated data. One example they provide is for the MNIST dataset about classifying handwritten numbers for a computer vision and image classification task, the type of number generated is the conditional information provided to both models. CTGAN is an improvement of CGAN, introduced by Xu et al. [170], it incorporates a mode-specific normalization technique, which helps the model to handle rare occurrences within the data and ensures better coverage and representation of all modes in the synthetic data. CTtab GAN+ introduced by Zhao et al. [180] improves further the method by handling mixed-type data and are using an augmented conditional vector to help the generator building synthetic data. They also include Differential Privacy on the stochastic gradient descent (DP-SGD) during the training process. Similarly Fang et al. [51] are as well featuring Differential Privacy in CTGAN to ensure strong privacy guarantees and also studying its deployment with Federated Learning.

2.4.2.2 Auto Encoders

Auto-encoders are a family of machine learning models where two models are learning in cooperation to generate data. They were first introduced by Hinton et al. [75] in 1993 and are consisting in two neural networks, an encoder and a decoder (Figure 2.4), that respectively encode the input data into a reduced space (the latent space) and decode the reduced information from the latent space back to its original shape: the reconstructed data. The loss function that is optimized here is the distance between the reconstructed data and the original one. Once trained the data can be generated by sampling randomly a new data record in the latent space and decode it back to have a synthetic data. Although this method was first designed for dimension reduction and not for relevant synthetic data generation, this is to tackle this shortfall that the following article were written.

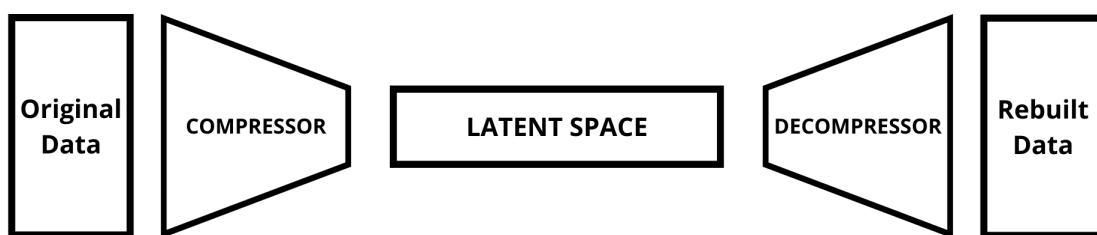


Figure 2.4: The auto-encoder compress the information on a reduced space and then rebuilds it.

The first innovation to improve the generation process of encoders was Variational Auto Encoders (VAE) introduced by Kingma et al. [89] in 2014. The original auto-encoders provided synthetic data with low utility because the latent space was unstructured, the sampling method might generate points in low probability so unrealistic areas. The main contribution of VAE is to structure the latent space by using a Kullback-Leibler divergence in the loss function so the latent space looks more like a multi-dimensional Gaussian distribution. Doing so the sampling process in the latent space is improved and the coherence of the generated data as well.

Improving further VAEs, Xu et al. [170] are proposing tabular VAEs or TVAE which are an hybridization of GANs presented above and VAEs. They have a deeper focus on tabular data by improving the management of mixed data-types and scalability for large datasets. The generator network in the GAN model is here replaced by a VAE, so the generation process is the same as for auto-encoders.

2.4.2.3 Other Approaches for Synthetic Tabular Datasets

In this part of the synthetic data section, tree approaches to generate synthetic tabular data with high compromise between utility and privacy will be presented. For this kind of approaches, that does not belong to the large families of AE and GAN, one of the most promising by its performance is Synthpop developed by Nowok et al. in 2016 [129]. This method relies on a conditional and sequential generation of the synthetic data: given a dataset of k variables (or columns), the model will train $k-1$ prediction trees, the i -th model being trained to predict the column $i+1$ from the i first columns. Once trained the generation is as follows: the first column is easily generated from the original distribution and then the next columns are generated iteratively. From our

experience in the later of this thesis, this model holds a good trade-off between utility and privacy and for a short computational time.

As presented earlier, differentially private synthetic data generation has become an essential approach in privacy-preserving data analysis. The general methodology of Maximum Spanning Tree (MST) presented by Mckenna et al. [113], relying on their earlier work on graphical model [112], involves the three following key steps: selecting low-dimensional marginals, measuring these marginals with noise addition mechanisms, and generating synthetic data that closely preserves the measured marginals. This approach has been the winning mechanism of the 2018 NIST Differential Privacy synthetic data competition [128]. It also demonstrates its broad applicability by preserving the utility of synthetic data and shows a high compromise with privacy.

From Bayesian approaches, PrivBayes was developed by Zhang et al. [176]. as a differentially private method tailored for high-dimensional data release such as tabular data. By constructing a Bayesian network that models the correlations among attributes and injecting noise only into low-dimensional marginals, PrivBayes effectively mitigates the curse of dimensionality, enabling the generation of accurate synthetic datasets. Which as well as for SynthPop demonstrate the need to handle the correlation between columns when generating synthetic data to ensure the quality of the synthetic data generated.

Finally, a last novel approach is emerging this last year [62]: the Avatar approach developed by a french startup named Octopize¹. This method is deeply analyzed in section 4.2.1. The avatar approach aims to generate realistic and similar looking data but on the record level. Neighbors of each data points are found in a reduced data space (PCA) and are randomly averaged to generate a close yet different new record from the original one. Hyperparameters of the method can be adjusted to get different trade-offs between utility and privacy.

2.5 Health Data: Types and Applications

Health data is any information related to health conditions, reproductive outcomes, causes of death, and quality of life for individuals or populations. It includes clinical metrics along with environmental, socioeconomic, and behavioral information pertinent to health and wellness [90]. A substantial amount of health data is collected when individuals interact with healthcare systems. This data, gathered by healthcare providers, typically includes records of services received, the conditions under which those services were provided, and the resulting clinical outcomes. Following this definition, health data can be unstructured and of multiple types. To provide a short example: John had a bike accident and described to his doctor how he lost control on a slippery road and fell, landing heavily on his left leg (text data). At the hospital, a radiography revealed a fracture in his femur (image data). Over the next few months, he underwent physiotherapy and relearned to walk, with sensors tracking his progress in real-time (time series data). The hospital recorded various metrics such as age, injury severity, and therapy frequency, which were analyzed to predict his recovery time (tabular data). The electronic health record (EHR) of John will have multiple type of data we will further describe during this section.

¹<https://www.octopize.io/>

2.5.1 Applications on Textual Health Data

Textual Data are the most present type of data in the EHR. Generally transcribing a discussion between a health professional and a patient and written by the professional itself, it holds a lot of sensitive information such as the name of the patient. Textual data is too the most unstructured data type because it is often in free writing. For both privacy preserving and structuring issues, it is difficult to be shared and used for machine learning.

2.5.1.1 Textual Health Data and Predictive Machine Learning

Machine learning methods can be used on textual data to extract information from the EHR or even to detect health issues. Kraljevic et al. [92] developed an approach of text extraction to enable large scale clinical analysis and show the transferability of their method between hospitals. In the same way Nuthakki et al. [130] demonstrates the effectiveness of a deep learning model, ULMFiT, in predicting medical codes from unstructured clinical notes, achieving high accuracy and showing potential to improve efficiency and reduce errors in the healthcare industry. On a more pathology focused objective, Hong et al. [76] highlights the use of an attention-based deep learning model to identify cognitive concerns from electronic medical records, more particularly on Dementia, hardly diagnosed, with crucial information frequently hidden in unstructured clinician notes. Their method outperforms baseline models that rely solely on structured diagnosis codes and medication data.

2.5.1.2 Textual Health Data and Federated Learning

To overcome the issue of too small or too biased dataset due to the impossibility to share sensitive information inside the medical system, an alternative may be the use of Federated Learning which mitigates the privacy risks for the patients. Peng et al. [134] study the performance of deploying a large language model with Federated Learning in comparison to learning on an individual database. Their approach outperforms their comparative baseline, is faster and more resilient. Nonetheless, they do not study the impact on privacy and consider that the Federated Learning is private by itself. Similarly, Shohman et al. [147] explore the deployment of BEHRT, a large language model from BERT and specialized on EHR and are comparing the impact on text prediction performance of an already trained model. Once again, the positive privacy impact is announced but not evaluated.

2.5.1.3 Synthetic Textual Health Data

To publish privacy-preserving data, generating synthetic data and sharing it instead of the sensitive, real data appears to be a promising alternative. Guan et al. [61] investigate whether generated EHR text can be as informative as real EHRs. They highlight that synthetic data can retain similar information to the original text while maintaining a consistent structure, unlike real EHRs, and effectively protecting privacy. On a similar yet different direction, text-generative IA such as ChatGPT² to perform text mining has been used on EHR by Tang et al. [157]. They highlight the privacy risks associated with sharing sensitive information with third parties, such as OpenAI,

²<https://www.openai.com/chatgpt>

as well as the lack of performance in traditional text mining approaches using this method. To address these issues, they propose using ChatGPT to generate new data that can be leveraged by a local model to enhance learning and improve performance in classical text mining tasks, such as named entity recognition and relation extraction. This approach significantly boosts the text mining performance of the model on EHRs

2.5.2 Applications on Genomic Data

With recent advancements in accelerating and reducing the cost of sequencing methods, genomic data has become an invaluable component of health data. It encompasses DNA sequences and related information, which can be utilized to understand genetic predispositions to diseases such as cancer.

2.5.2.1 Genomic Data and Classical Machine Learning

Libbrecht et al. [103] explore the applications of genomic data in conjunction with machine learning, underlining its potential for medical diagnostics and personalized treatment plans. Kim et al. [88] investigate how deep learning can predict the behavior of AsCpf1 guide RNAs, which are crucial in genetic modification and gene editing technologies. Gurovich et al. [64] demonstrate the use of deep learning to identify syndromes of genetic disorders through facial recognition, illustrating another innovative application of machine learning in genomics.

Nevertheless, the highly informative nature of genomic data prevents it from being shared without consent because it would violate privacy policies, a consent surely hard to obtain as genomic data may reveal tendencies to grievous disease. This limitation poses significant challenges for collaborative research and data-driven medical advancements. To overcome such limitations alternative to directly publishing sensitive genomic data have been studied

2.5.2.2 Federated Learning on Genomic Data

Both Alvarellós et al. [7] and Raimondi et al. [137] propose the deployment of Federated Learning over decentralized genomic data to better respect privacy policies. The second one study how a federated approach can match the performance of a centralized approach to detect the Crohn's Disease and with respect of privacy. Yet, none evaluate the privacy risk of using Federated Learning and are considering it privacy preserving because the sensitive data do not leave the participants storage. Literature on attacks on unprotected federated is pointing out the opposite as it will be shown further in this thesis. In a different direction, Chen et al. [28] proposed a more privacy oriented Federated Learning with trusted computing to detect rare disease from genomic datasets such as Kawasaki disease and compared their solution to other privacy preserving approaches.

2.5.2.3 Synthetic Genomic Data

Oprisanu et al. [131] study how useful genomic synthetic data can be for genomic research and what risks it still holds. It evaluates six state-of-the-art models for generating synthetic genomic data, finding that while some models produce high-

utility synthetic data, privacy issues persist, drawing attention to the need for careful assessment of synthetic data before deployment.

2.5.3 Applications on Image Data and Magnetic Resonance Imaging

Health data also cover images, such as X-rays, CT scans, and Magnetic Resonance Imaging (MRI). MRI, in particular, is used for detecting diseases in a non-invasive way, such as Alzheimer's disease, cancers, or sclerosis, to name a few. Despite that, its slow acquisition process and high cost limit its utilization. In recent decades, interest has focused on accelerating image capture to reduce the cost of using such equipment. Machine learning techniques have been employed to reconstruct incomplete images. Nonetheless, their clinical adoption has been limited due to slow computation times and the creation of unnatural-looking images [123]. While promising, these methods need improvements in evaluation practices, training datasets, and model reliability [85].

Overcoming the problem of a small dataset of MRIs is a challenging task due to the cost of constructing such images and the impossibility of sharing sensitive information between clinics and hospitals. One approach could be transfer learning [96], where sharing a model between actors instead of sharing data reduces the privacy risk.

Similar to approaches explained in detail later in this thesis, alternatives to personal MRI data sharing include Federated Learning on MRI data [43] and Synthetic MRI generation. With Federated Learning, a model is shared between multiple actors to learn a common pattern or task such as image reconstruction [49] or cardiovascular disease detection [104], but it is not limited to these types of diseases.

Clinics could also benefit from using synthetic data to share their data for open science objectives, as seen in [132, 115], and with less privacy risk in [53] because the patients were rats.

2.5.4 Applications on Time Series and Wearable and Sensor Data

Data collected from wearable devices and sensors include vital signs, physical activity levels, and other health metrics. This real-time data supports chronic disease management and proactive health interventions as described by Vijayan et al. [164]. A large quantity of information can be gathered in real time to potentially rise alerts about body dysfunctions: An unusual electrocardiogram may be the sign of a heart failure, a high temperature on the thermometer can be a sign of a fever, sudden variations on an accelerometer may indicate that the patient fell, etc... Such information is extremely sensitive and rich and holds further challenges: as the sensors often have a limited memory and battery, the machine learning on the device is an unreasonable choice so the data is very likely to be communicated on another device that can perform both data exploitation and storage.

Banaee et al. [14] reviews the methods and algorithms for analyzing data from wearable sensors, focusing on common data mining tasks like anomaly detection, prediction, and decision making, and highlights the challenges for data mining methods in health monitoring systems which is critical for disease and health issues detection. On a more privacy focused track, Ahmed et al. [4] explores Wireless Sensor Networks (WSNs) for monitoring and data transmission, studying their applications in healthcare, national security and disaster management. It also examines data privacy

challenges in WSNs, discussing encryption, authentication, and the adoption of privacy-enhancing technologies.

MotionSense: Malekzadeh et al. [110, 109] and MobiAct: Vavoulas et al. [163] are both about Human Activity Recognition. Data has been gathered on multiple patients by using both accelerometer and gyroscopic time series data by patient smartphones. Both datasets are a viable setup to deploy Federated Learning. On [109], Malekzadeh et al. also propose a sanitation approach to remove the impact of a meta information such as gender passively encoded on patient walking signals for a better protection of patients' privacy.

Lange et al. [95] introduces a method using Generative Adversarial Networks (GANs) and Differential Privacy (DP) safeguards to create privacy-aware synthetic health data, to improve data availability and model performance in stress detection tasks, showing significant boosts in F1-scores while maintaining data integrity. Dahmen et al. [38] introduce SynSys, a machine learning-based synthetic data generation method that uses hidden Markov models and regression models trained on real datasets to produce more realistic synthetic time series data, validated against a real smart home dataset and demonstrating improved activity recognition accuracy using semi-supervised learning.

As real-time health monitoring data is closely linked to communication, there appears to be a heightened focus on securing communications and protecting user privacy. This increased emphasis contrast with other fields of health care data where privacy guarantees are less present in the literature.

2.5.5 Applications on Tabular Health Data

The last type of data is one of the most common in data science: tabular datasets. Here the information is encoded on continuous or categorical variables in form of data records (or row) that contain for every variable (or column) a value. Large clinical trials reuniting cohorts of patients can be encoded in a tabular form describing health information about participants. This help to drive large statistical approaches to determine how different factors are impacting the evolution of a disease or what is the impact of a treatment on the survival or remission of patients.

Several studies have been conducted to structure large datasets and utilize them in classical machine learning experiments. The Medical Expenditure Panel Survey (MEPS) [3, 35] collects data on healthcare use, payments, and insurance coverage of Americans, supporting policy-relevant research. This dataset is used by research and governance actors to analyze healthcare delivery and financing in the U.S. for informing healthcare policies and practices, with ongoing efforts to enhance its accuracy and utility for research and policy analysis. The Wisconsin Breast Cancer dataset [168] is about breast cancer prediction: it is used to distinguish between benign and malignant breast cytology samples based on 11 cytological characteristics, with nine showing significant differences and was acquired from the University of Wisconsin Hospitals, Madison. On a similar direction, the ACTG 175 trial (AIDS [65]) was a randomized, double-blind, placebo-controlled study that compared monotherapy and combination therapy involving zidovudine, didanosine, and zalcitabine and how they impact the development of the HIV. Those data records include participants from multiple clinical sites across the U.S. and Puerto Rico. This dataset is rich in sensitive attributes such as the gender, the race, the sexual orientation and the use of injection drugs.

Similarly tabular data from electronic health records (EHR) [106] can be effectively used in machine learning models to predict heart disease stages. These models analyze structured data to identify key features and improve the accuracy of heart disease diagnosis and monitoring, showcasing the relevancy of machine learning application in tabular health data.

To provide a privacy preserving approach on tabular health data, Kerkouche et al. [86] propose to deploy Federated Learning with Differential Privacy applied on the gradients and also focus on reducing the bandwidth cost in communication by only communicating the sign of each weight instead of its complete value: the server then only aggregates the sum of small steps values multiplied by the sign of each clients.

As presented in previous sections, synthetic data generation also extend to tabular data and therefore has an application in health tabular data. Mendelevitch et al. [117] introduces a framework to assess the statistical fidelity and privacy preservation of synthetic datasets, which is prevalent for sharing useful and privacy preserving health data. To evaluate the quality of the synthetic health data, they use various approaches such as data visualization to validate that both original and synthetic data looks the same, they also use summary and comparative statistics between datasets and then confirm the clinical consistency on the synthetic data to ensure that similar results can be found on synthetic data instead of using the original one. They also evaluate the impact on privacy of sharing synthetic data instead of the real one such as membership and attribute inference attacks presented in 2.7. Similarly Yale et al. [171] also focus on generating synthetic health data with GAN (HealthGAN) while maintaining a high utility/privacy trade-off.

In a similar direction, both Hernandez et al. [71] and Murtaza et al. [124] are publishing overviews highlighting that despite the rising number of contributions using synthetic tabular data, there is no generalization or common ground between papers to evaluate correctly the quality of synthetic data. Murtaza et al. provide further incentives by classifying the usefulness of metrics for synthetic data in healthcare.

Azizi et al. [12] explores the challenges and solutions associated with sharing health data for research across international jurisdictions, focusing on privacy concerns. Their study compares Federated Learning deployment and synthetic data generation, evaluating their relative strengths and weaknesses. The research aims to assess differences between several countries (Canada and Austria) in the role of sex on cardiovascular health using a combined dataset. They indicate that deploying synthetic data was more efficient than federated making it a better alternative to open sensitive health data and provided significant statistics about population while guarantying privacy following their evaluation.

2.5.6 On the difficulty to centralize Health Data

As explained earlier in this thesis, health data are crucial for public good but are compromised to be gathered across health data holders such as hospitals without deploying great means to protect the individual's privacy. Nonetheless, there are some examples where large actors such as states have undertaken this project. The SNDS³ (Système National des Données de Santé), which is France's National Health Data System is a large data repository that contains health records covering over 99% of the French population, making it one of the most extensive health data systems globally.

³<https://www.cnil.fr/fr/snds-systeme-national-des-donnees-de-sante>

Established by the modernization law of January 26, 2016, and expanded by subsequent legislation, the SNDS integrates multiple datasets⁴ including the SNIIRAM (*Système National d'Information Inter Régimes de l'Assurance Maladie*), PMSI (*Programme de Médicalisation des Systèmes d'Information*), and BCMD (*Base de Causes Médicales de Décès*). These databases collectively contain information on healthcare usage, hospital stays, prescription drugs, medical procedures, and causes of death, all while ensuring the pseudonymization of personal identifiers to protect individual privacy.

The main objective of the SNDS is to sustain medical research⁵ and public health in a secure and centralized way. Researchers and public health officials can request access to the SNDS to use this data to improve patient care, track health trends, and evaluate medical treatments. This access is strictly regulated by the CNIL, ensuring that sensitive health information is protected while still being available for use research purposes restricting its use to a strict and defined framework on secure repositories. To centralize and secure sensitive health data is still a difficult task, however governments and institutions are recognizing the value of large-scale health data in advancing medical research and improving public health outcomes. In this direction the SNDS is a model for how health data can be managed to balance accessibility with stringent data protection standards.

2.6 Evaluating the Effectiveness and Applicability of Privacy-Enhancing Machine Learning

During this section, we study how to control the information shared and used to build machine learning model on health data to limit the privacy leakage, yet this generally comes with a reduction of prediction performance of such model. To measure the trade-off between privacy and utility it is therefore central to assess how to evaluate the impact on utility.

2.6.1 Impact of Federated Learning on Utility

Evaluating the impact on utility of Federated Learning is a trivial task if the evaluation of the original model is already known. Federated Learning slightly reduce the models performance in comparison to centralized approaches, the main reason is that: the model being unable to process all the data in one centralized place, the model has more difficulty to converge. Hannemann et al. [67] are providing a comparison centralized/decentralized on health data showing that Federated Learning comes with a cost of model utility. Another issue that comes usually with real decentralized datasets is that the data distribution among clients may be unbalanced in size and highly non-identically distributed. This means that models that provide the best prediction results are not the same among the clients. Milasheuski et al. [118] compare different methods of FL to tackle heterogeneous data on health data split between clients. They also evaluate that in the case of decentralized data, FL provides better results than no federation at all.

⁴[https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/le-systeme-national-des-don\[... \]](https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/le-systeme-national-des-don[...])

⁵<https://ecosysteme-snds.health-data-hub.fr/>

To handle unbalanced distributions among the clients personalization methods in FL aims to provide each clients a model that can learn generalities from the federation of clients and specificity from each individual clients to provide a performing model when a consensus model from the original FL approach might provide poor results. Arivazhagan et al. [10] are describing and evaluating personalization layers in FL where first layers of the model are aggregated with the FL process so feature extraction is common among the clients, and last layers are not aggregated so the client keep the specificities of its data. This approach found an application with the MELLODY project [73] where pharmaceuticals properties of molecules were found by federating the feature extraction.

Then providing supplementary privacy enhancing techniques to FL participate to further deteriorate model's utility, Choudhury et al. [32] measured that Differential Privacy on FL reduces utility on health data.

On the first hand, evaluating impact of Federated Learning is simple as the evaluation task is already define as every classical machine learning application. On the other hand, with synthetic data generation, as the task is not as clearly defined, such evaluation becomes difficult and therefore lacks of generalization in the literature.

2.6.2 Applicability and Quality of Anonymized and Synthetic Health Data

When it comes to anonymous as well as synthetic data, the evaluation of the impact on quality is a non-trivial task, in a perfect world, we would like to have the same results for every learning task performed on anonymized data or the original one but without leaking any sensitive information about individuals. Comparing how much two datasets (the original and the anonymized one) have in common might be hard. A first approach is to compare empirically performance differences on a specific task such as disease prediction on a dataset about strokes: we split the original data into a learning data set and a control data set, the first set is used to generate synthetic data and then two models are trained (a first one on the learning set, the second one on the synthetic data) then we compare both prediction performances of models on the control set. Nevertheless, this evaluation is task dependent so results may vary and generally no specific task is given so this evaluation might be impossible or meaningless to deploy. On a more general and statistical approach, there is distances that allows to compute gaps between uni-variate data distributions such as the Wasserstein Distance⁶ (or Earth Mover Distance - EMD) which are for example used in WGAN [63], GANs with Wasserstein Distance, to evaluate similarity between real and synthetic samples and therefore through convergence generate better samples. Another general approach is the one proposed by Patki et al. [133] with their contribution Synthetic Data Vault (SDV) that aims to provide tools⁷. To evaluate quality between two datasets⁸, they compare first the similarity between original and synthetic columns: a correlation similarity is used for quantitative columns, a contingency similarity is used for qualitative columns. Second, they compare cross correlation between original and synthetic columns, when columns are of the same type, they use the same approach as described before and when the pair of compared columns are of different types (one

⁶https://en.wikipedia.org/wiki/Wasserstein_metric

⁷<https://sdv.dev/>

⁸<https://docs.sdv.dev/sdmetrics/reports/quality-report/whats-included>

is qualitative, the other is quantitative) they discretize the quantitative variable and apply a contingency similarity. By both scores they compute first the "Column Shapes" (direct comparison) and then the "Column Pair Trends" (paired comparison). The overall quality score of the synthetic data, in comparison to its original real dataset, is the mean of those two scores.

To provide a better evaluation of health synthetic data, El Enam et al. [48] publish: Seven ways to evaluate the utility of synthetic data, and brings a framework for assessing its utility.

1. **Replication of Studies:** The utility of synthetic data may be validated by replicating analyses performed on real data and ensuring that similar conclusions are drawn from both datasets, as presented earlier in this section.
2. **Subjective Assessment by Domain Experts:** Experts evaluate synthetic data by assessing it as real or synthetic based on its plausibility.
3. **General Utility Metrics:** Automated metrics are used to compare distributions, correlations, and statistical properties between real and synthetic data. SDV Metrics are an example.
4. **Bias and Stability Assessment:** Repeatedly generating synthetic datasets helps assess the bias and variability across different generations.
5. **Structural Similarity:** This ensures synthetic data has the same format, variable types, and metadata as the real data, allowing analysts to run the same code on both datasets.
6. **Comparison With Public Aggregate Data:** In cases like COVID-19 data, comparing synthetic data with publicly available aggregate statistics provides insight into utility.
7. **Comparison With Other Privacy-Enhancing Technologies (PETs):** This allows the evaluation of synthetic data against methods like federated analysis and homomorphic encryption to determine the best approach for data privacy.

They underline that utility assessments should be performed every time synthetic data is generated as well as a focus on balancing privacy and data utility otherwise further analyses might be meaningless.

In a similar direction of statistical evaluations, Alaa et al. [5] propose 3 metrics inspired from recall and precision for classification in their paper:

- The α -**Precision** to measure the fidelity of synthetic data to the original one: how much the generated data resembles to real data samples, depending on their presence in the original data.
- The β -**Recall** to evaluate the diversity of synthetic data: how much the synthetic data cover the real data variability.
- The **Authenticity** to assess the overfitting on synthetic data on real one: how much synthetic data has memorized the original data instead of generating new samples.

In regard of those metrics, the authors audit models to improve post generation synthetic data by removing low-quality of over-fitted generated samples and therefore improving both utility and privacy.

Dankar et al. [41] presents an in-depth comparison of synthetic data generation methods. It categorizes utility metrics into four key dimensions:

- **Attribute Fidelity:** Measures how well attributes in the synthetic data resemble the original data. This measure is a Hellinger⁹ distance which is a metric of the utility conservation between two uni-variate distributions. This provides insight that data structure and uni-variate distribution are similar in both datasets. This metric is a direct comparison and provide an information coherent to the first part of the SDV quality score.
- **Bi-variate Fidelity:** Evaluates the correlations between attribute pairs to maintain the information encoded between variables by computing the pairwise correlation distance. This metric is a paired comparison and provide similar information to the second part of the quality score of SDV.
- **Population Fidelity:** Assesses the overall distribution of the synthetic data compared to real data, ensuring large-scale statistical properties are similar. Metrics like propensity score are used here to gauge how distinguishable real data is from synthetic data by computing the propensity of synthetic data in the real data, if the propensity is not balanced then the Population Fidelity lowers.
- **Application Fidelity:** Focuses on the performance of synthetic data when used in machine learning tasks, such as classification. It assesses whether models trained on synthetic data can achieve similar accuracy when tested on real data and rejoin the task-oriented metric described in the introduction of this section.

Through a large experimental field, they conclude that SynthPop provide the highest utility score, a results similar to our evaluation in our contribution about synthetic data. Evaluating utility is crucial to determine whether a privacy enhancing machine learning model is relevant. Nonetheless, this is only a half of the evaluation to assess plenty its performance, its privacy impact needs to be studied in parallel.

2.7 Privacy Risks Associated with Health Data

The GDPR defines private data as any information that can be linked to an individual. As such, if data is classified as private, it cannot be shared without the explicit consent of the individual concerned. Federated Learning and Synthetic Data Sharing are often considered outside the scope of the GDPR, as no personal information is directly shared. In Federated Learning, only the model is shared, and in synthetic data sharing, the synthetic information can be considered a form of anonymization since no direct personal information is shared. However, the literature shows that neither solution is without risk: in Federated Learning, the communicated models can leak information, and servers can be curious, potentially inferring sensitive information about clients participating in the learning process [44]; with synthetic data, information about the original data it mimics can also be inferred [153]. In this section we provide further

⁹https://en.wikipedia.org/wiki/Hellinger_distance

information on how to evaluate empirically the privacy impact of such protection methods and what is the remaining risk.

2.7.1 Privacy Risk for Machine Learning

The rising utilization of machine learning to build models and solve complex tasks is also rising privacy concerns about the data records such model are training on. Can a model leak sensitive information about the data it used to learn? To measure such vulnerability, privacy metrics have a critical role to assess the impact on information leakage of a mechanism (privacy preserving or not). Yet, the diversity and complexity of privacy metrics in the literature make it challenging to select appropriate measures, often leading to the development of new metrics and complicating the comparability of privacy studies. There exists no strict definition of what a privacy risk is as there exist a mathematical definition for a circle: the reason is because the subject is too complex to be simply defined. Wagner and Eckhoff [165] provide a comprehensive survey of over 80 privacy metrics, categorizing them based on the aspects of privacy they measure and their deployment requirements. In this section we will focus on two main privacy risks evaluated on machine learning: Membership and Attribute Inference Attacks.

2.7.1.1 Membership Inference Attack

One of the most popular privacy risks currently in the machine learning and privacy domain is the membership inference risk. As explained during the introduction chapter, the current research paradigm is to collect as much data as possible and to train a model with the best predictive performance on a given task. It is current that in such approach the model is left to learn every possible information to provide the best predictive performance (to reduce its training loss the lowest possible). This method leads to overfitting where the model learns specificity about the data it was trained on and therefore have significant predictive performance differences on data seen during the training and other unseen data. This is membership risk for machine learning. First introduced by Shokri et al. [148], a membership inference attack seeks to determine, given a data record and a machine learning model, whether this record was part of the model's training data. More formalized:

Given a machine learning model \mathcal{M} trained on a dataset D_{train} , a Membership Inference Attack (MIA) aims to determine whether a specific data point x belongs to D_{train} . Formally, the attacker seeks to infer the binary membership status m_x of a data point x where:

$$m_x = \begin{cases} 1 & \text{if } x \in D_{\text{train}} \\ 0 & \text{if } x \notin D_{\text{train}} \end{cases} \quad (2.3)$$

The attacker typically has access to the model \mathcal{M} and can query it with the data point x to obtain the model's output $\mathcal{M}(x)$, which includes the predicted label and possibly the confidence score or probability distribution over all classes. The attacker's goal is to build an inference function $A(\mathcal{M}(x))$ such that:

$$A(\mathcal{M}(x)) \approx m_x$$

Where $A(\mathcal{M}(x))$ is the decision rule or classifier that the attacker uses to predict the membership status of x . The effectiveness of the attack is measured by the accuracy

of A in correctly predicting m_x .

Depending on authors, there is a taxonomy on context where MIA is deployed. Here is a general description that fit their global interpretation in the literature:

- **White Box:** The attacker has access to the model and their parameters. This set provide the highest level of information and is for example a usual set in the basic setting of Federated Learning when the adversary is the server.
- **Grey Box:** This setting is uncommon and designate a limited access to a restricted part of the model and its parameters.
- **Black Box:** The Adversary has no access to the model but can still request it. This happens when requesting an online service with data.
- **No Box:** There is no access to the model. This setting is usual when attacking a synthetic dataset as the generating model might not be shared (for a better privacy control).

As explained upper, Yeom et al. [173] are pointing a correlation between over fitting to the training data and membership inference as well as for attribute inference presented in the following subsection. This contribution provides as well a link between both risk which help to lift some criticisms made about the membership inference risk. Generally, the link for this risk to a real privacy concern is often difficult to make and this risk is usually evaluated in unrealistic scenarios to point out an upper bound of privacy for a protection mechanism. However, establishing connections with attribute inference, which is directly identified as a privacy risk by both GDPR and HIPAA and will be explained in the following, underline that evaluating membership risk is relevant to assess a privacy concern in machine learning.

Yet, Song and Mittal [152] are pointing out that the membership risk is often under evaluated in the literature and are proposing a more accurate approach that is not using Neural Networks but rather the information entropy¹⁰ of the predicting model. They also insist on: privacy risk is not uniform on the dataset, some records hold higher risks. Additionally, they demonstrate that early stopping, that ends the learning process when the model shows signs of overfitting, is more effective to reduce privacy risks than most complex defense methods.

2.7.1.2 Attribute Inference Attack

The other most popular privacy risk is the sensitive attribute inference. According to Article 9¹¹ of the GDPR, these sensitive attributes include any "personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation". Removing the sensitive information in a data record can be a difficult task as it can be also coded even partially in other features: for example, removing the gender from a record might not be enough as it can still be inferred from the age and the weight/size of an individual.

¹⁰[https://en.wikipedia.org/wiki/Entropy_\(information_theory\)](https://en.wikipedia.org/wiki/Entropy_(information_theory))

¹¹<https://gdpr-info.eu/art-9-gdpr/>

As well as for membership inference, there is a link between over fitting or overfitting and sensitive attribute inference as studied by Song et al. [151] which is still for the same reason as stated before: the more the model will learn specificity from its training data instead of learning generalities, the more it will leak sensitive information about those training records. Finally, Yeom et al. [173] also establish this link and also similar behavior between attribute inference risk and membership inference risk. Nevertheless, Zhao et al. [179] underline an opposite direction by demonstrating that even with models sensitive to membership inference, attribute inference fails to provide convincing results and are therefore infeasible.

2.7.1.3 Outliers and population with higher risks

The privacy risk is not uniform among the records and as stated by several privacy articles from the literature, outliers tend to be more at risk than the general population. Nonetheless, there is no clear and universal definition of what an outlier is. Smiti proposes a survey [150] analyzing several publications about outliers' definition and their field of application. Here are the main families of approaches presented in this thesis but this list is not exhaustive.

- **Statistical-based Methods:** These methods assume an underlying distribution (parametric or non-parametric) and detect outliers as points that deviate significantly from this distribution. Gaussian-based methods, such as boxplots and regression-based approaches are a classical approach of it. Their implementation does not hold any difficulty, their evaluation is fast and they handle qualitative data. Yet, their deployment is compromised with high-dimensional datasets and unknown distributions which are usual with real data applications.
- **Distance-based Methods:** These approaches detect outliers by evaluating the distance between data points and their neighbors. Examples include the Solving Set approach, ABOD (Angle-Based Outlier Detection), and LDOF (Local Distance-based Outlier Factor). Distance-based methods work well with multivariate datasets and are easy to understand but are less effective in high-dimensional spaces due to the curse of dimensionality¹²: Datasets with a large number of dimensions tend to have longer computation time for calculus such as the distance and tend to have vast empty areas. They also meet difficulties to handle data stream as every distance to the other records should be recomputed each time a point is added.
- **Density Based Methods:** These techniques calculate the density around a data point and compare it to that of its neighbors to identify outliers. The LOF (Local Outlier Factor) method is a classic example. These methods are effective in identifying local outliers than the two previous type of approaches and are agnostic of data distribution but are computationally expensive and struggle with large datasets. Nonetheless, they tend to be the approach providing the records with higher risk with machine learning. As an insight, as the model often adjust their parameters depending on the data located in an area, the less point there is, the less the area contains general information about the distribution locally and will provoke model overfit.

¹²https://en.wikipedia.org/wiki/Curse_of_dimensionality

- **Clustering-based Methods:** These methods detect outliers by clustering data and identifying points that do not belong to any cluster or are far from other points in a cluster. Algorithms like DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and ODC (Outlier Detection and Clustering) fall under this category. They are unsupervised approaches that are robust to data types. Nevertheless, clustering methods are sensitive to parameter choices but are well-suited for noise identification but they tend to have large computation time. Outlier-ness here is also considered as a binary property which may lack of subtlety in practice.

Following those definitions, exploratory research should be deployed on the data to verify what characteristics and properties might lead to identify areas with higher privacy risks.

As pointed out during this section, not handling the overfitting problem of machine learning or not controlling what information is extracted during the learning process leads to privacy concerns. To handle such problem several solutions have emerged to both reduce the sensitive information directly present in the training data (known as anonymization) or to directly limit the learning of the model, such as Differential Privacy. Even if sharing synthetic data or using decentralized approaches such as Federated Learning reduce the privacy risks because the original and sensitive data is not directly shared, there is still some residual risk for privacy for those methods and they will be presented below.

2.7.2 Federated Learning Privacy Risks

As pointed out in previous sections of this chapter, Federated Learning without protections share similar risk to Neural Network model's attacks: even if the data is not shared, Federated Learning still holds privacy risks for the clients. The clients sending their model update to the server, this last one could use this protocol or information to provoke privacy leakage from their clients. The server can be malicious and send fake updates to the clients to maximize information leakage in the clients' response. It can be honest but curious and still perform some attribute or membership inference attacks on clients' updates. Secure aggregation [21] presented earlier cancel the risk that a server could infer sensitive information by being honest but curious. If the communication is not protected, a server can still deploy classical machine learning attacks on updates communicated by the clients such as membership or attribute inference attack as demonstrate following articles from the literature.

Nasr et al. [126] analyzes the privacy risks of deep learning models, focusing on white-box membership inference attacks in both centralized and Federated Learning. They show that through exploiting models' parameters they can effectively perform membership inference, even in well-generalized models. The study also introduces active attacks in Federated Learning (this is the malicious server case), where adversaries manipulate training to increase inference accuracy of their attack. In a similar direction Melis et al. [116] show similar results and are also indicating that attribute inference risks rise similarly with active attacks from the server.

Zhang et al. [178] highlights privacy risks in collaborative learning, showing that sensitive attributes not used in model training are still at risk to be inferred by adversaries through model outputs. The authors propose a novel attack using shadow models to predict the distribution of sensitive attributes from other participants'

datasets, even with black-box access. Their results show vulnerabilities for learning models to those attacks. They also underline that a sensitive information that is absent or removed from training data is not guaranteed to be protected as it can still be coded in other combination of variable.

All those three articles share a same conclusion: Federated Learning is not private by design; it needs a systematic privacy impact assessment and additional privacy preserving mechanism.

2.7.3 Anonymization and Synthetic Data Privacy Risk

As stated in previous sections, generating synthetic data may be considered as an anonymization protocol. Bellovin et al. [18] explores the potential of synthetic data to protect privacy while preserving the utility of datasets, particularly in the legal context (HIPAA). The article evaluates whether synthetic data can effectively substitute real data in research while guaranteeing privacy standards.

They conclude that synthetic data generation has an interesting privacy-utility trade-off but is not a perfect solution, especially in a sensitive field like healthcare; it requires further development and refinement, both legally and technologically, to be fully effective.

Privacy risks exist on a spectrum and must be evaluated even with anonymized or synthetic data, as true anonymity is rarely achieved. According to the G29 Working Party¹³, the CNIL identifies three primary privacy threats associated with anonymization: Singling Out, Linkability, and Attribute Inference [33]. Giomi et al.[57] implemented an evaluation framework to assess the extent of privacy protection achieved during the anonymization process. In addition to these three risks, we will also discuss and describe Re-identification and Membership Inference. By addressing these criteria, we aim to balance the necessary confidentiality with the utility of the data, ensuring that anonymization processes do not excessively compromise data usefulness.

2.7.3.1 Risk of singling out an individual

Despite efforts to anonymize data, there remains a potential risk that an individual can be singled out. This occurs when specific characteristics or patterns within the anonymized data have been directly copied from the original data, allowing the isolation of a marginal individual with a unique combination of features. Such an attack might be performed by iteratively searching for identifying combinations of variables. However, the implementation proposed in [57] is not time-efficient, as every combination has to be verified sequentially.

2.7.3.2 Risk of linking records related to an individual

There remains the possibility of linking different records that belongs to the same individual. Even if data is anonymized, subtle correlations or unique identifiers might allow an attacker to connect separate datasets, effectively re-identifying the person behind the separated data. The implementation proposed in [57] relies on nearest neighbors' identification with Gower's coefficients [59].

¹³https://www.cnil.fr/sites/cnil/files/atoms/files/wp260_guidelines-transparence-fr.pdf

2.7.3.3 Risk of Inferring Attribute about an Individual

There is also a risk that sensitive or personal information can be inferred about an individual based on the available anonymized data (containing information about this sensitive attribute). Even if the data is anonymized or synthetic, correlation or distance-based methods techniques might predict removed (or hidden) sensitive information about individuals. The implementation in [57] is similar to the one for linkability: nearest neighbor attack for original data in the synthetic one allows to infer an obfuscated sensitive attribute.

To protect against attribute inference in synthetic data, Ping et al. [135] propose DataSynthesizer. This tool generates privacy-preserving synthetic datasets by adding noise to attribute distributions, ensuring that sensitive attributes are protected while limiting utility degradation. As some attributes are by definition of the CNIL more at risk than other, voluntarily removing the risk from the data generation is relevant. DataSynthesizer operates in three modes: correlated, independent, and random to control the privacy-utility trade-off depending on the needs. Additionally, it uses Differential Privacy techniques, such as Laplace noise, to further increase privacy and Bayesian Networks to ensure that correlation between variables are kept.

2.7.3.4 Risk of Re-identification

The re-identification risk exists when the anonymization mechanism is a one-to-one mechanism: this happens when for one original record there is one linked anonymized record like in K-anonymity. In contrast, machine learning approaches like Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs) generate synthetic data that do not have a one-to-one mapping to the original data, removing the re-identification risk.

Re-identification refers to the process of matching anonymized or pseudonymized data back to the individuals to whom they originally pertained. Even when explicit personal identifiers such as names, social security numbers, or addresses are removed from a dataset, the remaining information—such as demographic details, behavioral patterns, or aggregated statistics—can sometimes be combined with other data sources or analyzed in ways that enable the identification of specific individuals.

It is important to note that re-identification risk might be related to other privacy risks presented above, such as linkability and singling out. There is ongoing debate as terminologies are not always the same in the literature. Re-identification, linkability, and singling out are interconnected, as they all involve the potential to associate data with specific individuals: Sweeney’s example (explained in section 2.3.1) where combined linkability and singling out techniques were used to re-identification attack.

2.7.3.5 Risk of Membership Inference

As presented earlier, membership inference risk refers to the potential for an attacker to determine whether a specific individual’s data was included in a dataset, even if the data has been anonymized or synthesized. In the context of anonymized or synthetic data, this risk arises when patterns or statistical properties of the data inadvertently reveal information that allows for such inferences. For example, even if direct identifiers are removed, subtle differences between the synthetic data and the general population data might allow an attacker to guess with high confidence whether a particular individual’s

information was used to generate the dataset. For example, if an anonymization process leaves too much information, an unusual but similar anonymized record to another original record might indicate that the second one belonged to the original dataset.

An extension to this attack that is typically made for machine learning approaches can be made only by comparing both synthetic/anonymized data to a non-anonymized one. However, if there is access to a model that generates synthetic data, it is also possible to perform membership inference attack. Chen et al. [27] highlight vulnerabilities on GAN models whether an attacker has access to the generator or discriminator models, Hayes et al. [69] shows also similar results.

To evaluate whether a given real data belonged to the original data that led to the generation of a given synthetic dataset (MIA on synthetic data), Hilprecht et al. [74] propose a Monte Carlo Black-Box MIA based on density for tabular data. Their attack relies on the hypothesis that the attacked dataset is split in half: member/non-member which is an unrealistic assumption, and on the insight that synthetic data records are closer (denser) in the member data than in the non-member data. This is a threshold attack based on the distance between a real record and its neighbors in the synthetic data. They also propose a reconstruction attack MIA based on how well a VAE will reconstruct a data record used during training based on threshold decision and demonstrate that those models are at risk of privacy leakage.

Hyeong et al. [78] explore the vulnerability of tabular data synthesis models to MIA. It evaluates four models, including CTGAN and TableGAN, under both black-box and white-box attack scenarios. Their results indicate that models are particularly susceptible to white-box attacks (as this is the most permissive scenario), with the Earth Mover's Distance serving as a strong predictor of attack success. Differential Privacy techniques like DP-GAN provide some defense against MIAs, but they come with a trade-off in data utility. On a similar direction Stadler et al. [153] evaluates the privacy and utility trade-offs of synthetic data, assessing its protection against privacy attacks. They show that synthetic data is not inherently safer than traditional anonymization, particularly against linkage (formalized as membership inference in their contribution) and attribute inference attacks. Outliers remain especially vulnerable in synthetic datasets, as they are more easily re-identified because of their uncommon characteristics. Their evaluation of differentially private synthetic data generation (PrivBayes and PateGan) shows that it mitigates these risks but significantly reduces data utility, similarly to the rest of the literature. Both articles share a similar conclusion: Synthetic data alone is not a reliable solution for privacy-preserving data sharing and it requires additional techniques to balance privacy and utility. Moreover, Stadler et al. [153] also provide an important metric in terms of methodological evaluation, the privacy gain: the impact on privacy of sharing synthetic data should always be compared to the privacy risk of sharing the unprotected original information.

Kuppa et al. [93] proposes a new metric: the Privacy Score to evaluate the risk of privacy leakage in synthetic data. The Privacy Score calculates a memorization coefficient for each synthetic records, indicating the probability that a model has memorized that record, indicating a vulnerability to membership inference. Their approach is agnostic of model utilized: it can be applied to any synthetic data model (No-Box scenario) as the attack is directly performed on the synthetic data. They finally also underline the difficulty to balance privacy with data utility and suggest that further research is needed.

Even if sharing synthetic data instead of sharing the original one is reducing privacy risks, the privacy of the original data is not guaranteed by default. By constructing realistically efficient privacy attacks and by defining correctly what is a risk and a sensitive information, one can evaluate methodically the privacy risks that last when using anonymization mechanism such as synthetic data. This evaluation is difficult because the risks evaluated are always depending on an attack and as the attackers become more efficient, new privacy risks are appearing and what is protected today is not guaranteed to be protected tomorrow. The same approach has to be applied to every privacy enhancing mechanisms and Federated Learning and Synthetic data as well should follow this direction.

2.8 Synthesis of Key Concepts

During this section, we provided background knowledge on privacy-enhancing machine learning applied to health data. Key points include:

- **Understanding Classical Machine Learning Convergence:** It's crucial to comprehend how classical machine learning models converge to provide good prediction performance. This understanding helps identify potential privacy leaks.
- **Privacy Leaks from Model Behavior:** From the convergence behavior of machine learning models, privacy leaks can emerge and need to be handled when models are applied to sensitive data like health records.

Improving the compromise between privacy and utility is at the heart of the problematic of the **thesis overview** (Figure 1.1). Having a deep understanding of how to evaluate both aspects of machine learning on health data is central to the research work we provided with our following contributions to improve Federated Learning and synthetic data generation.

Chapter 3

MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers

Contents

3.1	Introduction	40
3.2	Background for our Contribution	43
3.2.1	Privacy Risks in Federated Learning	43
3.2.2	Mixnets	44
3.2.3	Intel SGX	45
3.3	System and adversary model	45
3.4	Contribution: The MixNN Framework	46
3.4.1	Overview	46
3.4.2	Utility Equivalence	47
3.4.3	Implementation	48
3.5	Evaluation Setup	49
3.5.1	Dataset	49
3.5.2	Evaluation Metrics	50
3.5.3	Baselines	51
3.5.4	Methodology	52
3.6	Evaluation	52
3.6.1	No compromise with utility	53
3.6.2	Prevent information leakage	54
3.6.3	Robustness of the protection	55
3.6.4	System performance	56
3.7	Related work	56
3.8	Conclusion	57

Machine Learning (ML) has emerged as a core technology to provide learning models to perform complex tasks such as their application on medical data which is critically sensitive by nature. Boosted by Machine Learning as a Service (MLaaS), the number of applications relying on ML capabilities is ever increasing. However, ML models are the source of different privacy violations through passive or active attacks from different entities, even with Federated Learning (FL) where the sensitive data is not gathered. In this section, we present MixNN a proxy-based privacy-preserving system for Federated Learning to protect the privacy of participants against a curious or malicious aggregation server trying to infer sensitive information (membership and attribute inferences). MixNN receives the model updates from participants and mixes layers between participants before sending the mixed updates to the aggregation server. This mixing strategy drastically reduces privacy leaks without any trade-off with utility. Indeed, mixing the updates of the model has no impact on the result of the aggregation of the updates computed by the server, facilitating therefore their deployment on decentralized health data. We report on an extensive evaluation of MixNN using several datasets and neural networks architectures to quantify privacy leakage through membership and attribute inference attacks as well the robustness of the protection. This chapter is an extension of our paper published at Middleware 2022 [99].

We show that MixNN significantly limits both the membership and attribute inferences compared to a baseline using model compression and noisy gradient (well known to damage the utility) while keeping the same level of utility as classic Federated Learning, answering the problematic illustrated in Figure 1.1, we aim to improve the utility/privacy trade-off in a Federated Learning scheme in enhancing privacy.

3.1 Introduction

As presented in the beginning of this work of research, the collection of personal data is a subject firmly grounded in public debates and even more when it comes to health data, classified as critical by the CNIL. The growing awareness of the population on privacy issues led to stronger regulations on data protection (e.g., GDPR, HIPAA) and contributed to the appearance of new services making privacy an incentive vector such as privacy-based search engine (e.g., Duckduckgo, Qwant), web browsing (e.g., Web Proxy, Tor, Brave), or mailing (e.g., Protonmail). These services rely on infrastructures setup and maintained by companies, nonprofit organizations promoting privacy, or are fully peer-to-peer involving devices of end-users.

However, personal and private data is still the fuel of all desires given the large domain of applications and promising results data-driven approaches are allowing. In this context, Machine Learning (ML) has emerged as a core technology to analyze and provide learning models from large volumes of data and to perform complex tasks such as classifications, predictions or clustering providing powerful tools to help healthcare professionals to provide personalized medicine. The success of ML has driven different providers to launch Machine Learning as a Service (MLaaS) engines to make ML operation easier for anyone, without the cost and time to build in-house infrastructures. As already demonstrated in Section 2.5, these new services has led to an ever increasing number of new applications or services relying on ML capabilities in different domains such as computer vision, health analytic and speech recognition to name a few , while leaking information about the data used for training them [54, 152, 179]. The fact

that many applications using this technology involve the collection and processing of personal and sensitive data has raised privacy concerns [37].

As presented in Section 2.7, the memorization of training data by a ML model is the source of different privacy violations such as membership, property and attribute inference through passive or active attacks [143], which retain their deployment on health data. Membership inference [126, 81] refers to the capacity of an adversary to identify if a data point (or the data of an individual) has been used to train the target model. This attack has a serious privacy implication if the model is training with sensitive information (e.g., data from people with certain health status). Property inference [55, 178], in turn, corresponds to the inference by an adversary of the properties of training data such as the features that characterize each class. This property inference can also concern a subset of the training inputs. This ability to learn from training data is desired if the inference is directly related to the main task of the model. By contrast, attribute inference [116] corresponds to the fact that an adversary is able to infer an unintended and undesired attribute not correlated to class's characteristic feature. Root causes related to these attack surfaces as well as the link between utility (e.g., through model overfitting [151, 173]) and privacy are not well understood.

With the recent development of Federated Learning, hospitals can securely collaborating in the learning process, allowing all participants to benefit from shared medical insights while maintaining patient privacy. It can also include networks of smaller devices or sensors that monitor patient health metrics, sharing their learning to enhance personalized treatments and support patient recovery. This new ML scheme has attracted many attentions these last years, not only from the research community but also from major Internet companies, suggesting future deployments. For instance, Google already exploits FL for next-word prediction in a virtual keyboard for smartphones [68]. While the FL scheme is a clear step forward towards enforcing users' privacy, it still suffers from a large ML-based attack surface including membership, property and attribute inference from participants or from the server. Different protection mechanisms to limit inference capabilities of an adversary have been proposed [125] and still have a cost in utility, all these solutions are reducing the accuracy of the model and its capacity to converge [107, 175].

For instance, some solutions [162, 175] are based on perturbation in order to reveal only a noisy information to the server, such as Differential Privacy. However, these solutions significantly damage the accuracy of the model and its capacity to converge [158]. Secure aggregation relying on a cryptographic scheme has been also proposed [17, 20, 21]. Similar to MixNN, this solution ensures that the server is only aware of the aggregate of all models, keeping the model of each participant (and the associated inference) private. Although the overhead of this solution remains low, the underlying cryptographic scheme requires the participation of the server in the protection. We argue that such solutions are not deployed in practice. Indeed, few companies accept to afford the additional cost of the protection. For instance, Private Information Retrieval (PIR) protocols which follow similar cryptographic scheme to protect the profile of users are not widely adopted in practice. Moreover, a curious or malicious server trying to infer information from participants will certainly not adopt such a protection. While hospitals may have the influence and resources to demand privacy protocols such as secure aggregation, networks of individual patients using smartphones to track their physical activities might not prioritize requesting additional

privacy enhancements from the server.

In this chapter, we present *MixNN*, a new privacy-preserving service for FL against inference attacks from a curious aggregation server. To achieve that, *MixNN* relies on a proxy mixing the layers of the model updates (also named parameter updates) among participants before sending them to the aggregate server. Like *Mixnets* to ensure anonymity in information routing [25], mixing the layers of the participants' updates of neural network prevents inference attacks (both membership and attribute inference attacks) without decreasing the accuracy of the aggregated model. This solution, albeit simple, leads to drastically improving the privacy without any trade-off with utility, which is critical for an application on health data. In addition, *MixNN* is transparent to the FL service, participants only need to configure a web proxy for the associated traffic. To make the deployment of *MixNN* easier by anyone (e.g., operate by an individual or non-profit organizations willing to protect privacy) and possibly on an untrusted infrastructure, the proxy mixing the neural network layers is running inside an SGX enclave ensuring confidentiality and attestation on its behavior. Interestingly enough, the behavior of the proxy can be adapted according to the expected security and privacy guarantees. For instance, the proxy can aggregate itself the model updates or can adopt another aggregation scheme to improve the robustness against model poisoning or backdoor attacks [13] by replacing averaging with robust estimators such as geometric median. In this case, the utility-privacy-performance trade-off can change.

To illustrate the capability of *MixNN* to protect privacy while maintaining the same level of utility, we implemented *MixNN* and experimentally evaluated it with several datasets (two with real medical application and two others) and neural network architectures. We also implemented both membership and attribute inference attacks to quantify and compare privacy leakage of *MixNN* against classical Federated Learning scheme, a model compression scheme, and a baseline using perturbation (noise) to protect the model updates (widely used in Differential Privacy). We show that *MixNN* drastically reduces the membership inference compared to other baselines (on average up to 73.9%, 73.8%, and -0.2% less inference against a classical FL, model compression and LDP, respectively), and limits the attribute inference (on average up to 13.8%, 14.6%, and 12.9% less inference against a classical FL, model compression and LDP, respectively) without decreasing the accuracy of the global aggregated model. Moreover, we show that reconstructing the update of a participant (by identifying the layers of an individual among the mixed updates) is costly and a difficult task which gives poor accuracy. Finally, we show that the *MixNN* proxy is scalable and introduces only a small latency on the model updates. The source code of *MixNN* as well as the experiments and datasets are publicly available ¹.

The remainder of this chapter is organized as follows.

- Section 3.2 presents background, Section 3.3 defines the problem and the threat model,
- Section 3.4 explains the design and the implementation of *MixNN*,
- Section 3.5 presents our evaluation setup,
- Section 3.6 reports the evaluation of *MixNN*,
- Section 3.7 reviews related work,

¹<https://gitlab.inria.fr/abaud/mixnn>

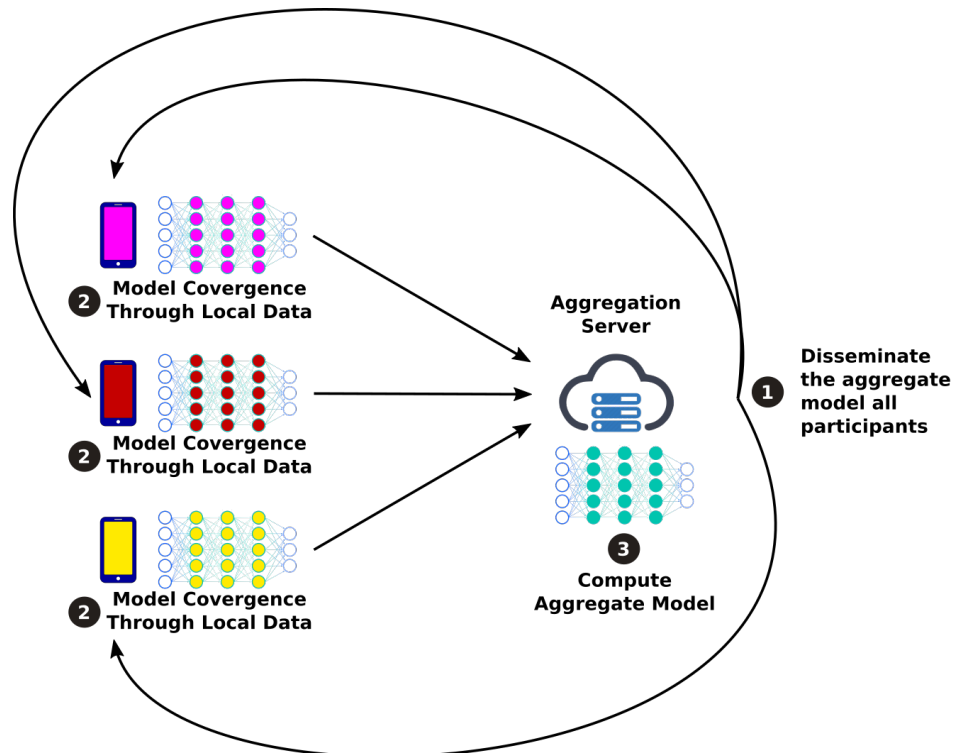


Figure 3.1: Operating flow of Federated Learning.

- Section 3.8 concludes this chapter.

3.2 Background for our Contribution

In this section, in completion to sections of the background Section 2.2.2 about Neural Networks, we provide a short reminder about Federated Learning and its privacy risks 3.2.1 and we review background related to Mixnet 3.2.2 and Intel SGX 3.2.3.

3.2.1 Privacy Risks in Federated Learning

This Section provide a summary of Federated Learning functioning 2.4.1 and evaluation 2.6.1, as well as privacy risks evaluations 2.7.1 applied on FL 2.7.2.

Federated Learning is a collaborative learning scheme to train an ML model [114, 22]. In such a scheme, personal data never leaves the device of participants. Instead, devices train a ML model locally and interact with a central server to build a global learning model.

The iterative-based operating flow of classical FL is depicted Figure 3.1. Each iteration contains three steps. First, the aggregation server disseminates a global model to participants (step ❶ in the figure). Each participant then trains and refines this model with its own data stored locally (step ❷). After this local training, each participant holds its own variation of the model sent by the server. Participants then send their updated model parameters to the aggregation server. Finally, the server aggregates all these updates to generate a new global model (step ❸) which will be disseminated to participants in the next iteration. Iteratively, the global model maintained on the server converges without requiring access to the personal data of participants.

By keeping locally data of the users on their device, FL improves privacy by design. However, FL can disclose sensitive information via model updates that are based on the training data. Indeed, any useful ML model reveals something about the population from which the training data was drawn. Indeed, a classifier model for instance may reveal the features that characterize a given class or help construct data points that belong to this class. The first privacy violation is property inference: identification of the features that characterize each class, making it possible to construct representatives of these classes through model inversion attacks [54]. Another privacy violation is attribute inference [151]: the leak of personal and unintended information (properties that hold for certain subsets of the training data, but not generically for all class members). The last privacy violation in our setting is membership inference [148]: given an exact data point, determine if it was used to train the model.

Memorization of training data by deep neural networks enables an adversary to conduct all these privacy violations. Firstly, this memorization usually combined with overfitting of the model are exploited by an adversary to conduct a membership inference attack in order to discriminate if a user has been part of the training or not [126]. This attack has a serious privacy implication especially if the learning model is related to sensitive information (e.g., presence of a certain pathology). Secondly, as deep-learning models come up with separate internal representations of all kinds of features, some of which are unpredictable and independent of the task being learned, the memorization of the training data can be leveraged by an adversary to infer a sensitive attribute [116]. In addition, due to the distributed nature of FL, passive and active inference attacks can be conducted by any participant or by the server.

Introducing Differential Privacy in the Stochastic Gradient Descent (DP-SGD) [2, 125] has been proposed to reduce the inference capability of an adversary, however this solution significantly damages the accuracy of the model and its capacity to converge [107, 175]. In addition, the noise calibration and the management of the privacy budget is not trivial. Other defenses propose to reduce the overfitting [143] but inherently decrease the utility. Secure aggregation relying on a cryptographic scheme has been also proposed [17, 20, 21]. Although the overhead of the underlying cryptographic schemes tends to be reduced, the management of this overhead and the participation of the server raises questions.

3.2.2 Mixnets

The concept of mixing information to make them indistinguishable or unlinkable is not new. Mix networks (Mixnets) [25] uses this concept to provide a proxy-based anonymity system. This system aims to provide unlinkability between the message sent by a user, and the message received by the destination. More precisely, to prevent traffic analysis attacks, Mixnets route each message of the user through a set of anonymity servers called mixes. Mixes collect and shuffle (or mix) many messages before to route them to the destination. A variety of mixnets have been proposed including Aqua [97], Riffle [94], and Mixminion [40] addressing differently the trade-off between anonymity, latency and bandwidth.

The limitation of these systems is similar to Tor, it is difficult for a user to determine which edge is uncompromised and powerful adversaries controlling both ends of the circuit can still deanonymize clients.

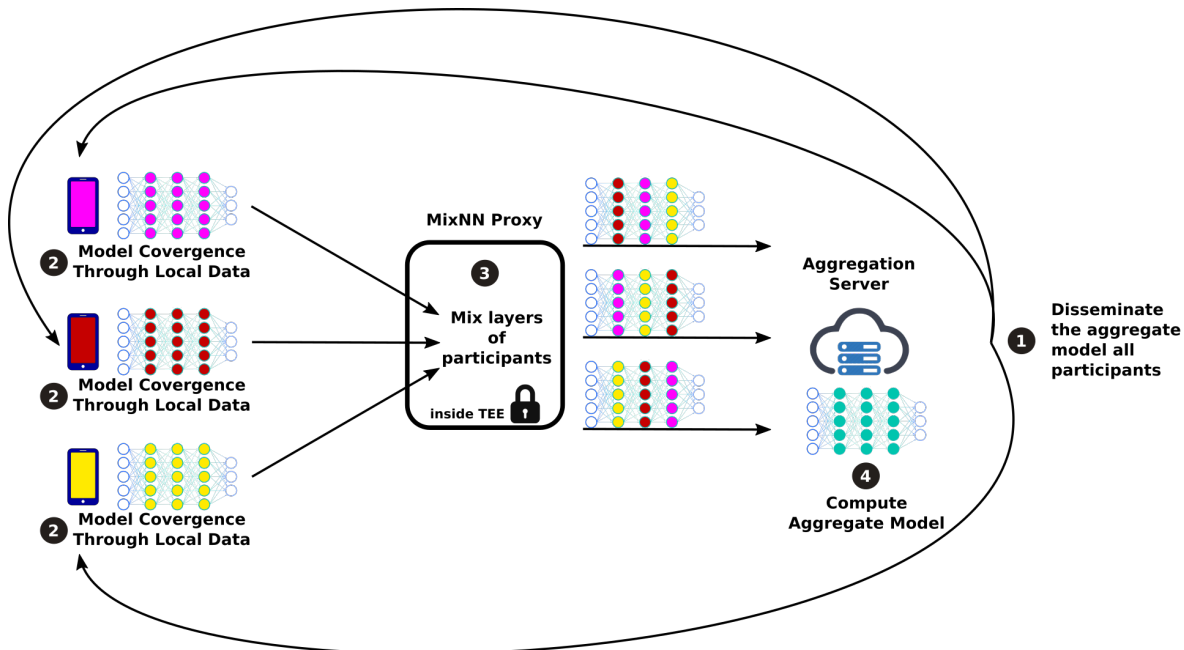


Figure 3.2: MixNN introduces a proxy which receives the parameter updates from each participant, shuffle them to remove attribute footprint before to route them to the aggregation server.

3.2.3 Intel SGX

The MixNN proxy relies on a Trusted Execution Environment (TEE), which leverages custom microprocessor zones, to enforce isolation, confidentiality and integrity of code and data. Specifically, we use Intel Software Guard Extensions (SGX) [36, 8] which defines the concept of enclave. The memory of an enclave is encrypted and cannot be directly accessed by other system software even by privileged code (e.g., the operating system or hypervisor). Enclaves can be attested to prove that the code running in the enclave is the one intended, and that it is running on a genuine Intel SGX platform. Once attested, enclaves can be provisioned with secret data by using authenticated secure channels. Moreover, enclaves can persist secret data outside the trusted zone by using a sealing mechanism. However, such protection comes with resource constraints. More precisely, only 96 MB out of the 128 reserved for the enclave can be used by applications. Although virtual and dynamic memory support is available [29, 28, 111], it incurs significant overheads in paging (the sealing and unsealing operations used an encryption key derived from the CPU hardware). However, [121] evaluates the usage of TEEs on mobile devices and reports a small system overhead at the client-side.

3.3 System and adversary model

Before presenting MixNN, we describe our assumptions and the considered threat model. The operating flow of MixNN involves three premises with different level of trust, namely: (i) the client machine; (ii) the MixNN proxy; and (iii) the aggregation server.

First, we assume that the client machine is trusted. This includes the training data and all the computations performed locally. We do not consider malicious users trying to poison the model or to introduce backdoors.

Second, we assume that the MixNN proxy is running inside an Intel SGX enclave on an untrusted node. An adversary is thus not able to compromise the behavior or the data of the proxy. In addition, the remote attestation provided by Intel allows any participants to control that the MixNN proxy conforms to the expected behavior and has not been tampered with, and that it is running securely within an enclave on an Intel SGX enabled platform. However, an adversary can monitor the node where the MixNN proxy is deployed, possibly physically (e.g., monitoring network traffic, power consumption or memory access patterns). Consequently, an adversary can leverage side channel attacks [24] to infer information. We assume that the SGX enclave has generated a public and private key pair (k_{pub} and k_{priv}). As this keys generation is part of the proxy behavior and can be verifiable by the remote attestation, we do not assume collusion between the enclave and any party (e.g., the aggregation server).

Lastly, we consider an honest but curious aggregation server. This server builds a model for a main classification task through a Federated Learning scheme but also aims to infer membership and sensitive attributes from participants. This aggregation server conducts passive attacks. Specifically, it passively follows the FL operational flow and exploits auxiliary knowledge to infer sensitive information from participants. We consider an adversary with two types of auxiliary knowledge. First, we consider an adversary able to collect or to use a public dataset with similar raw data and statistical properties (including the sensitive attribute). This auxiliary knowledge allows the adversary to train a model to infer the sensitive attribute from the model updates sent to the aggregation server by each participant. Second, we also consider an adversary able to collect raw data from each participant. This second auxiliary knowledge, in turn, is used by the adversary to evaluate each received model update through the data of each participant in order to link a model update to a specific participant. More details about the inference attacks are described in Section 3.5.2. Finally, we consider protected exchanges between participants and the MixNN proxy (parameter updates are encrypted by using the public k_{pub} of the SGX enclave).

3.4 Contribution: The MixNN Framework

In this section, we first present an overview of the MixNN (Section 3.4.1), the equivalence in terms of utility with a classical FL (Section 3.4.2), and then give implementation details (Section 3.4.3).

3.4.1 Overview

To avoid inference attacks during the learning process of a service using a Federated Learning scheme, MixNN operates as depicted in Figure 3.2. To use MixNN, users have only to configure its system to use a proxy for the associated traffic (e.g., through the configuration of its browser). As such, users seamlessly get protected without changing their habits. More precisely, compared to the classical FL pipeline (Figure 3.1), all parameter updates will be sent to the MixNN proxy instead of the aggregation server. To secure these updates, they are encrypted with the public key of the enclave (k_{pub}) to ensure that only the MixNN proxy is able to read and process them. Once loaded in the enclave, the proxy decrypts and stores the parameter updates of each layer in different lists. The proxy then picks at random one update for each layer in the associated list to generate the message containing the parameter updates to send to the aggregation

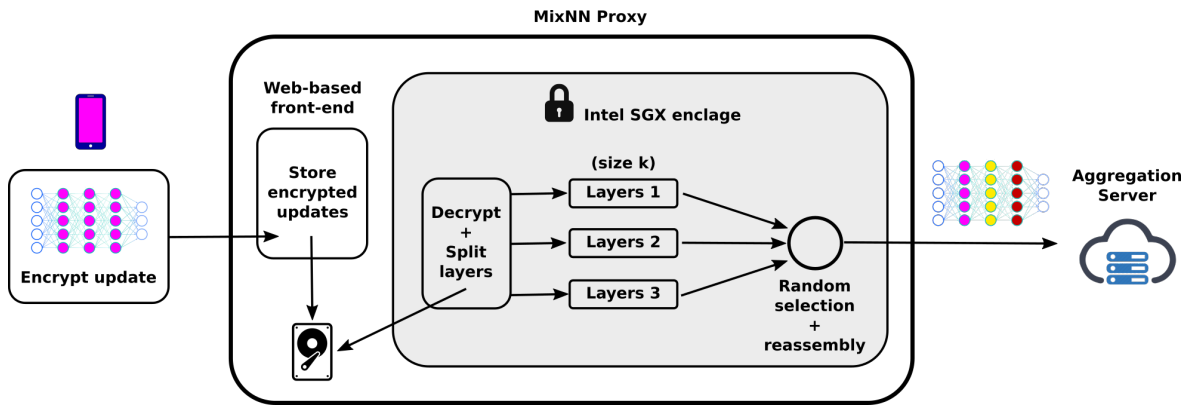


Figure 3.3: Implementation and data-flow of the MixNN proxy.

server. Note that the proxy needs to initialize first each list with k updates before sending updates to the aggregation server.

The rest of the workflow remains unchanged compared to the classical one. The server aggregates the parameter updates to generate a global model which will be disseminated to all participants. Participants will then refine this model locally with their personal data before sending the parameter updates to the MixNN proxy.

The accuracy of the global model remains unchanged with or without using MixNN. Indeed, whether mixed or not, the aggregation of parameter updates of each layer is identical. In contrast, the privacy leakage through the footprint of parameter updates returned by participants is drastically reduced. Specifically, by receiving an update mixing information from different users (breaking potential footprints), the aggregation server is not able to infer any sensitive information. Consequently, MixNN is able to drastically improve privacy without compromising the accuracy of the system (no trade-off between utility and privacy). Indeed, by design, MixNN provides the same utility than a classical FL scheme. It is worth mentioning that the behavior of the proxy can adopt another strategy. For instance, the proxy can aggregate itself batch of model updates through averaging scheme or more robust estimators against model poisoning or backdoor attacks [13] such as geometric median. In this case, the utility-privacy-performance trade-off can change.

3.4.2 Utility Equivalence

By design, MixNN provides the same utility than a classical FL scheme. In this section, we prove this equivalence.

Let C be the number of participants sending their updates to the proxy. We show in this section that whether the participants use MixNN or not, the resulting aggregated model is the same. We assume that the considered MixNN proxy has enough information to send L updates to the server. Then the proxy creates a sequence (M_{ij}) such that $\forall (j_1, j_2) \in \{1, \dots, n\}^2$ with $j_1 \neq j_2$ and $\forall i \in \{1, \dots, L\}$ $M_{ij_1} \neq M_{ij_2}$. And also such that $\forall (i_1, i_2) \in \{1, \dots, L\}^2$ with $i_1 \neq i_2$ and $\forall j \in \{1, \dots, n\}$ $M_{i_1j} \neq M_{i_2j}$.

According to previously defined notation for the nodes of a neural network, we define the t -th layer of the c -th participant of the proxy by $(\theta_{.t}^c)$. In the following matrix, each line is a model sent by the proxy. We remark that each combination of participant/layer appears once and only once in the matrix. This is a fundamental assumption regarding the equality of accuracy level between traditional FL and MixNN.

$$A = \begin{pmatrix} (\theta_{\cdot,1})^{M_{11}} & (\theta_{\cdot,2})^{M_{12}} & \dots & (\theta_{\cdot,n})^{M_{1n}} \\ (\theta_{\cdot,1})^{M_{21}} & (\theta_{\cdot,2})^{M_{22}} & \dots & (\theta_{\cdot,n})^{M_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ (\theta_{\cdot,1})^{M_{L1}} & (\theta_{\cdot,2})^{M_{L2}} & \dots & (\theta_{\cdot,n})^{M_{Ln}} \end{pmatrix}$$

Now, with the regular FL procedure, the information sent by the participant is:

$$B = \begin{pmatrix} (\theta_{\cdot,1})^1 & (\theta_{\cdot,2})^1 & \dots & (\theta_{\cdot,n})^1 \\ (\theta_{\cdot,1})^2 & (\theta_{\cdot,2})^2 & \dots & (\theta_{\cdot,n})^2 \\ \vdots & \vdots & \ddots & \vdots \\ (\theta_{\cdot,1})^C & (\theta_{\cdot,2})^C & \dots & (\theta_{\cdot,n})^C \end{pmatrix}$$

We note $Agr : \mathcal{M}(C \times n) \longrightarrow \mathcal{M}(1 \times n)$ the aggregation function which makes the mean of the columns. We show that $Agr(A) = Agr(B)$.

$$Agr(A) = \left(\frac{1}{L} \sum_{i=1}^L (\theta_{\cdot,1})^{M_{i1}}, \frac{1}{L} \sum_{i=1}^L (\theta_{\cdot,2})^{M_{i2}}, \dots, \frac{1}{L} \sum_{i=1}^L (\theta_{\cdot,n})^{M_{in}} \right)$$

$$Agr(B) = \left(\frac{1}{C} \sum_{c=1}^C (\theta_{\cdot,1})^c, \frac{1}{C} \sum_{c=1}^C (\theta_{\cdot,2})^c, \dots, \frac{1}{C} \sum_{c=1}^C (\theta_{\cdot,n})^c \right)$$

We make the additional assumption that $L = C$ which means that the MixNN proxy waits for the C participants to send their updates before mixing. Which gives us that:

$$Agr(A) = Agr(B) \iff \left[\forall l \in \{1, \dots, L\} \quad \sum_{c=1}^C (\theta_{\cdot,l})^{M_{cl}} = \sum_{c=1}^C (\theta_{\cdot,l})^c \right]$$

Which is true since our assumption on (M_{ij}) gives us that $\varphi : \{1, \dots, C\} \longrightarrow \{1, \dots, C\} \quad c \mapsto M_{cl}$ is a bijective mapping.

3.4.3 Implementation

MixNN is implemented inside an Intel SGX enclave to protect its behaviors and confidentiality even if it is deployed on an untrusted node. In our implementation, the SGX enclave first generates an RSA key pair including a private and a public key of 4,096 bits, and parameters update is encrypted by participants using the public key of the enclave with AES-256-GCM algorithm, ensuring only the enclave could access its content. The data-flow of the MixNN proxy is depicted Figure 3.3. A web-server is used as a front-end of our proxy, receiving parameter updates from participants and saving them on disk. To improve performance, these updates are sent in a binary format. The SGX enclave then scans the file system for new updates. Each new update is decrypted and split by cutting the binary block into pieces corresponding to the size of each layer. The parameters associated to each layer are then stored in different lists. The size of these lists (noted k) and the memory allocation according to the considered neural network models are initialized at the creation of the enclave. Once k parameter updates are received, those lists are full and each subsequent update will generate a mixed update, picking one random layer from each list. This generated update can then

be forwarded to the aggregating server. Also, our implementation takes advantage of the multi-threaded capabilities of Intel SGX with each thread processing one incoming parameter update and generating a mixed update if necessary.

To avoid side-channel attacks against SGX [24], the cost (the execution time) to process an update is constantly the same. Depending on the considered model, the size of a model can be important and not fit into the memory limit of the enclave (96MB), requiring encrypted storage outside the enclave. To avoid side-channel attack based on memory access, ORAM mechanisms (e.g., ZeroTrace [144]) can be adopted to carry out secure and oblivious access of data. The associated overhead is negligible in our context where updates are sent only periodically.

3.5 Evaluation Setup

In this section we present the experimental setup used to evaluate MixNN, which includes datasets (Section 3.5.1), metrics (Section 3.5.2), baselines we compared against (Section 3.5.3), and the considered methodology (Section 3.5.4).

3.5.1 Dataset

We used two image recognition benchmark datasets (CIFAR10 and LFW) and two motion datasets for activity recognition (MotionSense and MobiAct), that are our real case application on health data, to assess MixNN.

CIFAR10 is a major image classification benchmarking dataset where the data records are composed of 60,000 32×32 RGB images where each record is mapped to one of 10 classes of common objects such as airplane, bird, cat, dog. There are 50,000 training images and 10,000 test images. The main task is the classification of the images. We artificially define 20 participants split into three groups with different preferences. We define 3 types of preference which corresponds to specific and non-overlapping categories of images. The dataset is slightly balanced, two groups gather 6 participants and the last one gathers 8 participants. The profile of the participant is composed of 80% of images corresponding to its preferred classes, and the remaining 20% is composed of random images from other classes. The sensitive attribute is the preferences of the user.

MotionSense [138] contains data captured from an accelerometer (acceleration and gravity) and gyroscope at a constant frequency of 50Hz collected with an iPhone 6s kept in the front pocket. Overall, a total of 24 participants have performed six activities (going downstairs, going upstairs, walking, jogging, sitting and standing) during 15 trials in the same environment and conditions. The main classification task is the activity detection and the sensitive attribute is the gender of the users.

MobiAct [163] records the motion data from 58 subjects during more than 2,500 trials, all captured with a Samsung Galaxy S3 in the front pocket. This dataset includes signals recorded from the accelerometer and gyroscope at 20Hz. We only used the trials corresponding to the same activities as MotionSense in order to do the evaluation with the same settings. Similar to MotionSense dataset, the main classification task is the activity detection and the sensitive attribute is the gender of the users.

Labeled Faces in the Wild (LFW) [77] contains face images for face recognition with 13,233 total samples with images for 5,749 people. The dataset additionally has attributes such as age, race, gender, smile, facial hair, glasses etc. The main

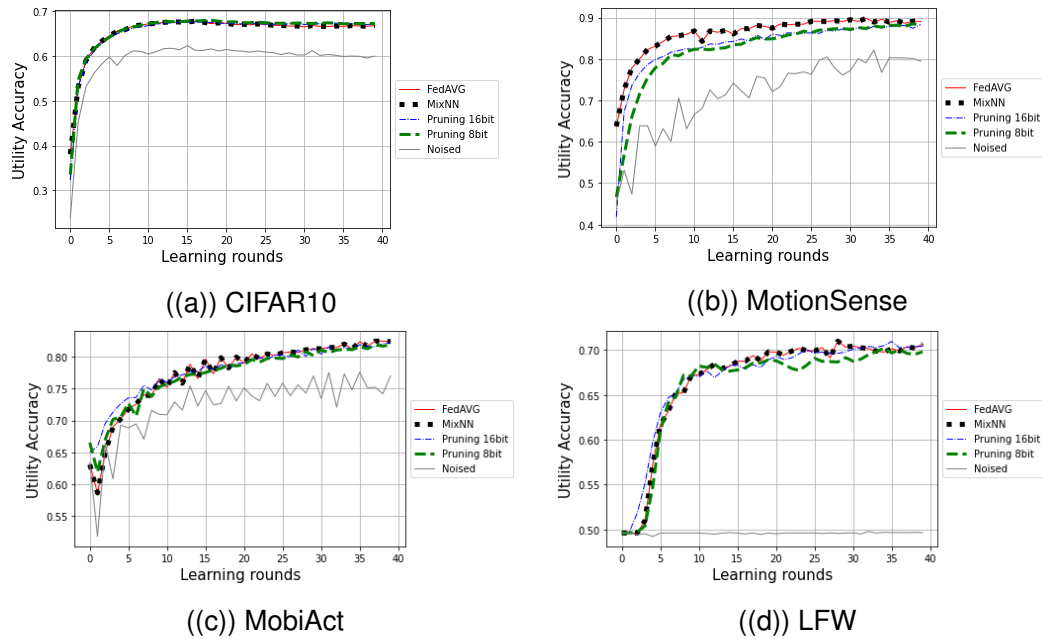


Figure 3.4: MixNN provides the same utility than a standard FL scheme, noisy gradient however decreases significantly the utility and slows down the convergence.

classification task is smile’s detection and the sensitive attribute is the gender of the users.

For CIFAR10, MotionSense and MobiAct datasets, we use a neural network composed of two convolutional layers and three fully connected layers for the classification task. For LFW, in turn, we use a more complex architecture provided by Facebook, named Deep Face [156]. This neural network is composed of multiple convolutional, locally connected, maxpooling, and fully connected layers.

3.5.2 Evaluation Metrics

We evaluate MixNN through three complementary dimensions: utility, privacy and system performance.

To evaluate the utility of the target model, we consider the classification accuracy for the main task (e.g., the activity detection), noted *Model Accuracy*, measuring the ratio of number of correct predictions over the total number of predictions made.

To assess the privacy, we implemented both a membership and an attribute inference attack. First, we revisit the membership inference attack to compute an upper-bound of the risk of linkability between a model update and a participant. We consider here an adversary with an auxiliary knowledge about each participant (some raw data). The adversary (the aggregation server) is thus able to evaluate a received model update with the auxiliary data of each participant in order to link this (anonymous) model update to a specific participant. Specifically, the adversary predicts the link between a participant and the participant for which its auxiliary data produced the lowest loss. We report the *Linkability Accuracy* measuring the precision of this linkability. As an update produced by MixNN is composed of layers from different participants, we report an upper-bound by counting a good prediction if the most sensitive participant (the layer with the lowest loss) included in the update is linked with the correct participant.

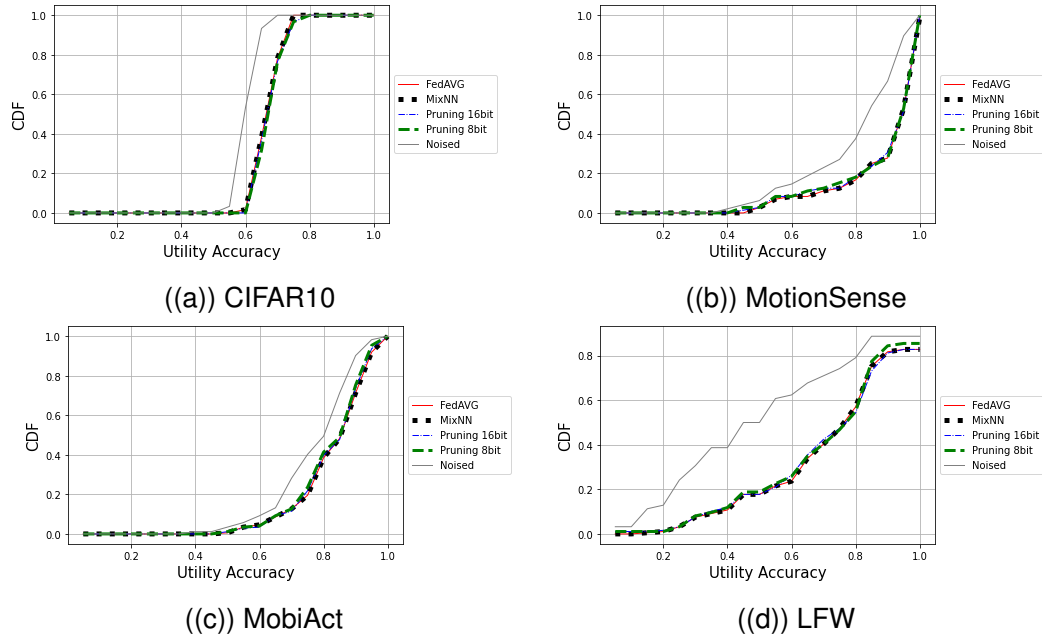


Figure 3.5: Most of the participants have an accuracy with noisy gradient smaller than MixNN for all datasets.

Second, for the attribute inference attack, we train a random forest classifier to infer the sensitive attribute from the parameter updates. This classifier leverages auxiliary knowledge, specifically the raw data and the sensitive attribute of few participants (3 males and 3 females in our case). To increase the number of model updates used for the training of this random forest classifier, each received model update is refined with each auxiliary data. More precisely, a model update received at one learning round by the adversary (the aggregation server in our case) is refined with the raw data of each participant part of the auxiliary knowledge and labeled with the same sensitive attribute. We report the success of the attribute inference, noted *Inference Accuracy*. This value indicates a data leakage according to the number of classes. For instance, with a balanced dataset over the gender, an accuracy different than 50% indicates that the adversary is able to identify the gender of a participant with an accuracy higher than random guess.

To evaluate the behavior of MixNN from a systems perspective, we consider the end-to-end latency which is the time spent by the proxy to route the parameter updates to the aggregation server.

3.5.3 Baselines

We compare the utility and privacy provided by MixNN against different comparative approaches. Firstly, we consider an approach using noisy gradient widely used in Differential Privacy studies [125, 182].

We use an implementation based on an introduction of Gaussian noise to the updates computed through a classical local training such as using in local Differential Privacy [125].

Secondly, we consider an model compression scheme using Pruning. Pruning the network results in reducing the format of the parameters of the neural network (e.g., from 32 bits to 16 bits). A compressed format reduces the overall memory

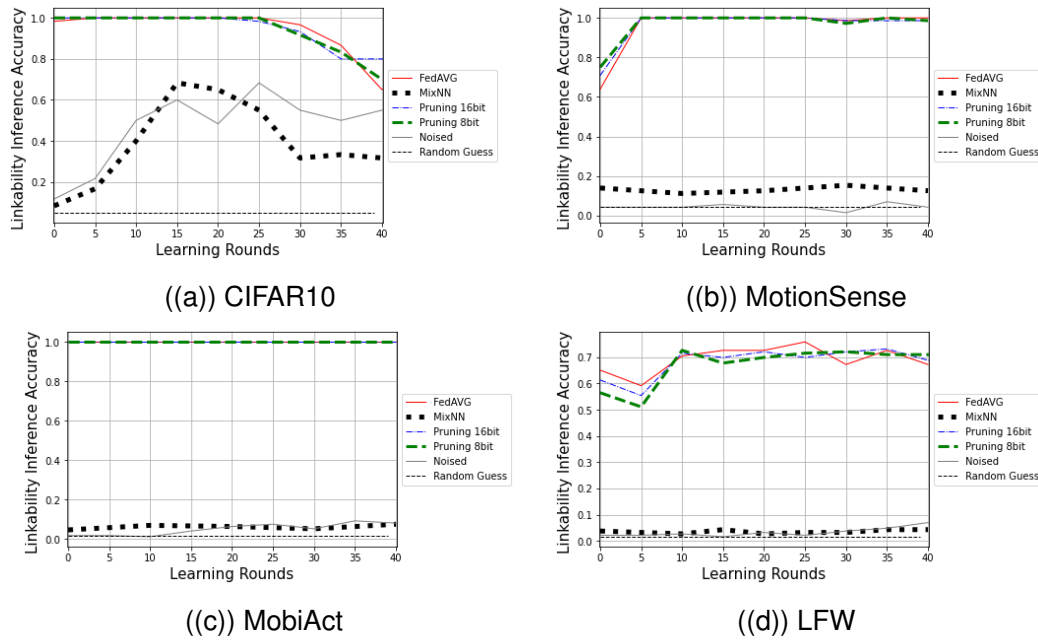


Figure 3.6: MixNN better prevents the membership attack compared to a classical FL and a pruning strategy.

bandwidth [66]. Lastly, we also consider a classical Federated Learning scheme.

3.5.4 Methodology

The dataset is split between training and testing, with 5/6 of trials used for training and validation and 1/6 for testing. For CIFAR10, the Federated Learning model is trained on 3 local epochs for a size of data batch of 32 samples on each learning rounds, the server aggregates 20 users on each of the 40 learning rounds. For MotionSense (and MobiAct), the training is (respectively) done on 2 (and 3) local epochs for batches of 256 samples for each of the 40 learning rounds, and the server aggregates 24 users for MotionSense and 58 users for MobiAct. For LFW, the training is done on 2 local epochs for batches of 8 samples for each of the 40 learning rounds, and the server aggregates 62 users. For every dataset, we use the "Adam" optimizer proposed by Tensorflow. We use 3-fold cross-validation in which the testing set is randomly generated from 1/3 of the users. Reported results correspond to average over 3 repetitions of each experiment. The experiments have been computed on Grid5000². The noise introduced consists on adding a Gaussian noise $\mathcal{N}(0,0.1)$ on each scalar of the neural network weights. Finally, MixNN has been implemented within the SGX enclave of an Intel i5-8500 CPU.

3.6 Evaluation

We now report the results in terms of utility (Section 3.6.1) and privacy (Section 3.6.2) provided by MixNN under the considered experimental setup (Section 3.5). We also analyze the robustness of MixNN against an aggregation server trying to defeat the

²<https://www.grid5000.fr/w/Grid5000:Home>

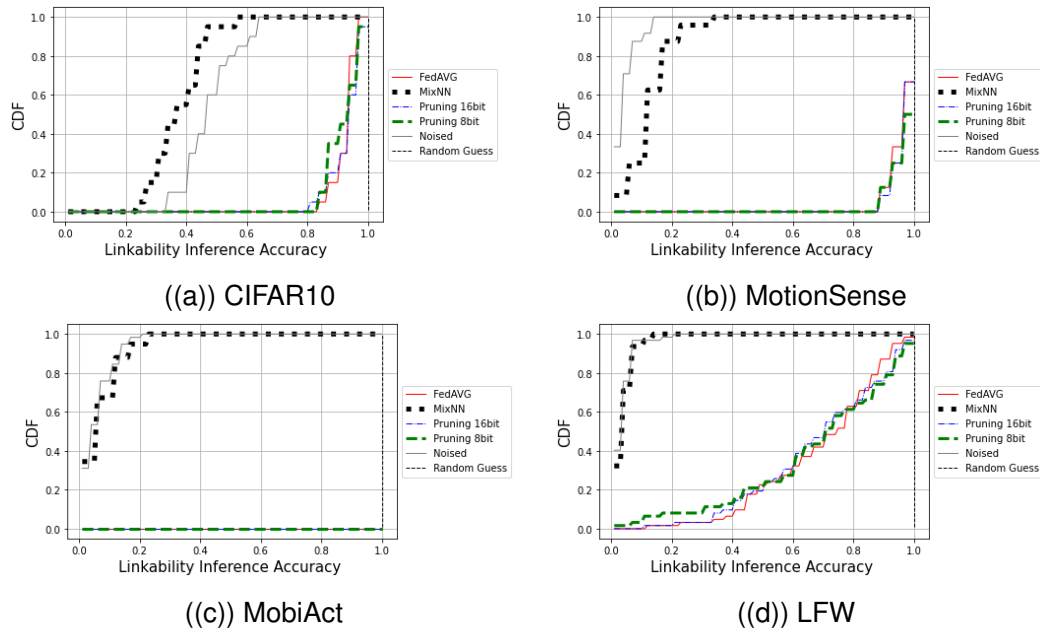


Figure 3.7: With MixNN, most participants benefit from a protection comparable to that provided by the use of a noisy gradient against a membership inference attack.

protection. (Section 3.6.3) as well as its performance from a systems perspective (Section 3.6.4).

Our results show that MixNN efficiently reduces the information leakage through both a membership and an attribute inference attack without compromise on the accuracy of the model. We also show that MixNN introduces a negligible end-to-end latency.

3.6.1 No compromise with utility

In this section, we evaluate the capacity of MixNN to protect privacy without compromising the utility. We compare the accuracy performance for the main classification task provided by MixNN against a classic FL scheme (without MixNN proxy), a baseline using noisy gradient such as using in local Differential Privacy and a pruning strategy. Figure 3.4 reports the accuracy according to the learning round for all datasets. First, the results show that the same level of accuracy is provided by a standard FL scheme and MixNN with an accuracy growing according to the learning rounds. This result is expected due to the aggregation equivalence of both approaches. Second, the results show that noisy gradient provides 10% lower accuracy on average and slows down the convergence. Noisy gradient even breaks the learning capability in the case of the LFW dataset. Other approaches based on pruning slightly deteriorate the model's accuracy.

Figure 3.5 reports the cumulative distribution of the accuracy over the population of participants at the learning round 40. Results show that most of the participants have an accuracy with noisy gradient smaller than MixNN for all datasets (on average 0.67 for noisy gradient against 0.78 for MixNN).

In practice, node churn (some participants do not participate to all learning rounds) results by the reception of fewer number of updates than expected. As the MixNN proxy waits the reception of k updates before to prepare the mixed updates

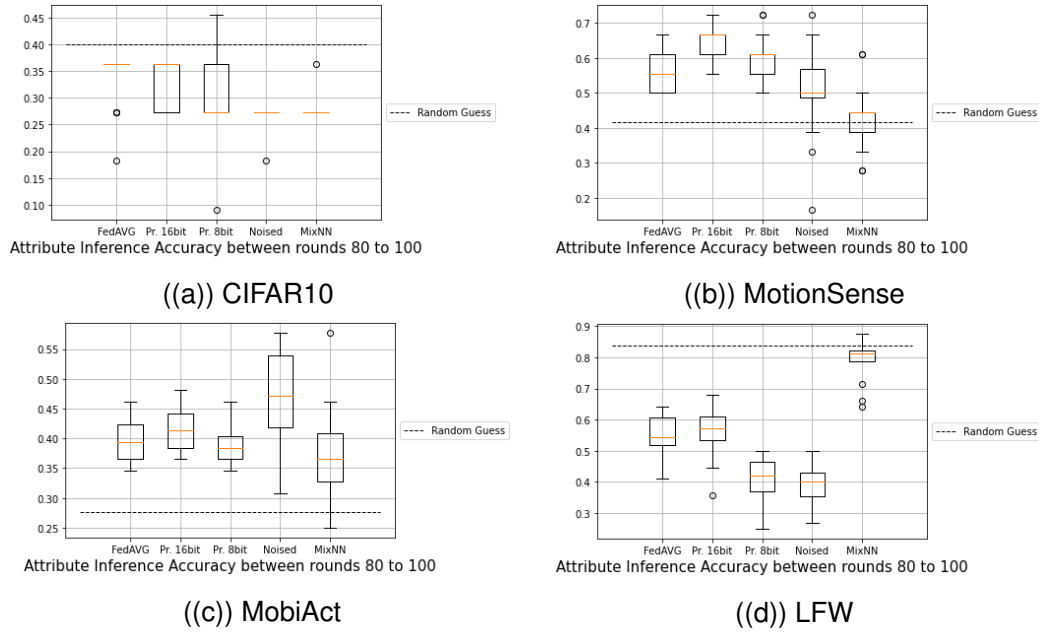


Figure 3.8: MixNN better prevents sensitive attribute leakage compared to using noisy gradient.

(Section 3.4.3), lists of updates (split by layer) maintained in the MixNN proxy can gather information from different learning rounds, and might slow the convergence of the learning. However, this effect has only a limited impact (e.g., with 20% of updates missing at each round, a slowdown in convergence was observed only on the LFW dataset).

3.6.2 Prevent information leakage

In this section, we evaluate the privacy leakage through both a membership and a sensitive attribute inference attack.

3.6.2.1 Membership inference attack

For the membership inference attack, Figure 3.6 reports for all datasets the accuracy of the linkability between an update and a participant for MixNN, a classical FL, a pruning strategy using 16 and 8 bits and noisy updates according to a growing number of learning rounds. In all datasets, the pruning approaches and the classical Federated Learning provide roughly the same poor results: the updates are successfully linked to the correct participant (around 100% of accuracy), except for LFW with an accuracy around 70%. This means that the gradient vector returned by participants can be exploited to provide an efficient footprint to re-identify the associated participant. Our approach MixNN and the noisy gradient provide a near perfect protection (close to a random guess) for all datasets except for CIFAR10 which slightly reduce the linkability accuracy. Figure 3.7 reports the cumulative distribution of the accuracy of the linkability over the population of participants. Results show that with MixNN, most participants benefit from a protection comparable to that provided by the use of a noisy gradient against a membership inference attack. However, this protection does not come at the cost of the utility as shown in the previous section.

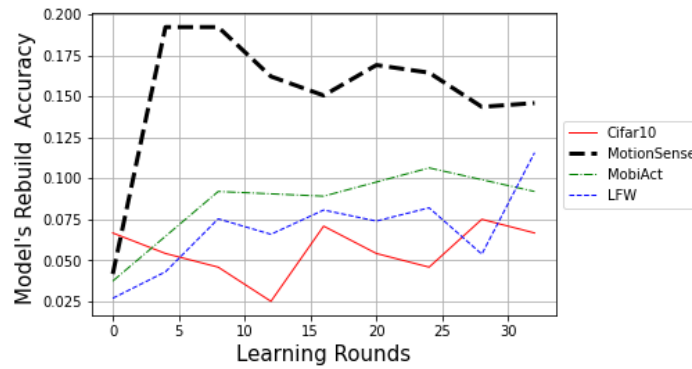


Figure 3.9: Reconstruction of the model updates is costly and gives a poor accuracy.

3.6.2.2 Attribute Inference Attack

For the attribute inference attack, the Figure 3.8 reports for all datasets the accuracy of the inference for MixNN, a classical FL, pruned updates at 16 and 8 bits and noisy updates according to a growing number of learning rounds. This attribute inference attack trends to provide better accuracy on fully converged models, so we evaluate the attack between the rounds 80 and 100 of the Federated Learning process. For all datasets, the accuracy of the attack for MixNN is closer to the random guess (the attacker is unable to learn the sensitive attribute) compared to other baselines approaches. The other comparative baselines, in contrast, fail to protect the sensitive attribute inference (e.g., the classical Federated Learning update leaks up to 52% of the client's gender in comparison to the random guess). For CIFAR10, MixNN provides the best protection and fits to the random guess. The attack provides poor results, the reason is that the clients have artificial data distribution: there is model weights that can be optimal on every client's dataset. The retrain on the auxiliary data is then useless and the random forest classifier fails to learn such pattern on the sensitive attribute.

3.6.3 Robustness of the protection

MixNN shuffles model updates sent by participants. A malicious aggregation server could then attempt to break the protection by enumerating all possible combinations of the shuffled update items (a layer-by-layer brute force approach) in order to reconstruct the update of origin. We evaluate a worst-case scenario where the aggregation server holds participants data and is able to learn a learning model for each of them. The malicious server then exploits these models to evaluate layer by layer the received pieces of update. More precisely, for each participant and for each layer of the model learned with auxiliary data, the server replaces the layer with each element received at this layer and selects the one which gives the lowest loss on it to build back the model. We evaluate then how accurate the server was in this rebuilding process. In the best case, this process fails in more than 80% of the participants across all datasets (Figure 3.9). Breaking the mixing strategy of MixNN seems to be difficult to achieve and is very costly in computation time.

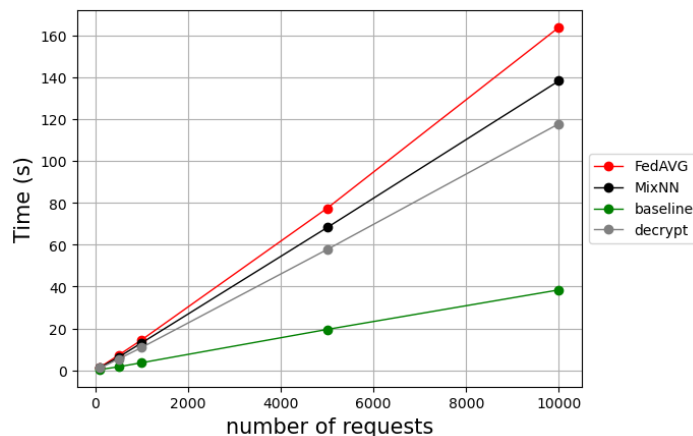


Figure 3.10: Time spent by MixNN to process the updates is linear according to the number of updates.

3.6.4 System performance

We implemented MixNN proxy inside an SGX Enclave to evaluate its system performance. As described Section 3.4.3, this MixNN proxy receives the update parameters as an encrypted binary file, decrypts inside the enclave each layer and stores them in the trusted memory of the enclave before to prepare new updates with mixed layers and sends them to the untrusted aggregation server. Figure 3.10 depicts the time spent by the MixNN proxy to process a growing number of updates. We compare the scalability of MixNN against three comparison baselines. Specifically, these comparison baselines include a federated average aggregation (the enclave receives and decrypts the updates and then does an average of all received updates and send all updates with the averaged values to the untrusted server), a decryption only proxy (the enclave decrypts the received updates and forwards them to the aggregation and untrusted server), and a simple proxy (the proxy sends every received update to the aggregation and untrusted server without passing them to the enclave). Results show that time spent by MixNN to process the updates is linear according to the number of updates. As the learning round of classical FL scheme is usually not conducted at high rate (e.g., only when the device is plugged in and has a WiFi connection), this short delay introduced by MixNN is negligible. In addition, the constant processing time over all updates for a given model (the MixNN proxy waits to receive a constant number of model updates before mixing them) reduces the surface for side channel attacks.

This experiment uses a model designed for activity recognition (MotionSense dataset), with two convolutional layers and four fully connected layers. Each update of this model uses 3.3MB inside the enclave, 4.4MB while encrypted.

3.7 Related work

The incentive behind using FL is to collectively build a learning model with better accuracy than if each user trained a model with their own data. The goal is to improve the accuracy as much as possible but several dimensions have an influence. The standard FL scheme [114] learns one global model and replicates it locally on every client. However, heterogeneity of data across user devices can severely degrade

performance of standard federated averaging for ML learning applications, especially for atypical users. Indeed, one unique model cannot cope with the heterogeneity of data and provide the best utility for all users [23]. To address this data heterogeneity, several approaches have been proposed such as local adaptation [175, 10] and clustering [145]. Specifically, the clustering mechanism proposed in [145] also leverages a similarity metrics between the model updates sent back by participants (similar to MixNN) to cluster the population.

Nasr et al. [126] designs passive and active inference algorithms for Federated Learning. However, this work only targets membership inference. In addition, while this attack also exploits the privacy vulnerabilities of the SGD algorithm, authors used a neural network to classify if a participant is a member of the training data or not a member. Zhu et al. [182] also exploits the gradient exchange to infer private training data of participants. To do that, authors iteratively optimize "dummy" inputs and labels to minimize the distance between dummy gradients and real gradients. Once the optimization finished, the dummy data is close to the private training data. Jagielski et al. [80], in turn, investigates the guarantees of differentially private SGD but via data poisoning attacks.

Running MixNN in an Intel SGX enclave improves trust and confidentiality through an isolated execution environment. However, this TEE is still vulnerable to side channel attacks [24, 161]. The most common countermeasure is to use data oblivious algorithms. The objective of this technique is to eliminate the link between the nature of data inputs and the execution of the program (e.g., through the execution time or memory footprints). To achieve that, the obfuscation technique consists to hide potential patterns by making them all uniform regardless of the considered data. To reduce its inherent cost, the considered data oblivious algorithm needs to be chosen carefully according to the application [6].

3.8 Conclusion

We presented MixNN, a proxy-based privacy-preserving framework to prevent both membership and sensitive attribute inference attacks conducted from a curious aggregation server exploiting the model updates. MixNN breaks the footprint leaked in the model updates by mixing layers between multiple participants. As this mixing strategy does not impact the result of the model aggregation performed by the server, the privacy improvement of MixNN does not compromise the utility of the model learned collaboratively. We experimentally evaluated MixNN with different benchmark datasets, including real-world healthcare applications with connected sensors, and compared it against a state-of-the-art baselines using local Differential Privacy, a pruning strategy and a classical FL. Results show MixNN provides the same model accuracy than a classical FL scheme (the same utility) while providing a better protection (a better privacy) compared to other baseline approaches, allowing the deployment of a protected Federated Learning approach on health data where both good utility and high privacy are critical.

This conclude our contribution to better protect Federated Learning against a

curious aggregation server. The next chapter will present our contribution about Synthetic Data generation: M-Avatar.

Acknowledgment

This work has been partly supported by ANR grant ANR-20-CE23-0013 and by the Face Foundation under the Trusty-AI project.

Chapter 4

M-Avatar: On-Demand Privacy-Enhancing Synthetic Data Generation Using Local Modeling

Contents

4.1	Introduction	60
4.2	Avatar Data: Limitations and Improvement	61
4.2.1	The Avatar Approach	62
4.2.2	Our Contribution: The M-Avatar Model	64
4.3	Evaluation Setup	65
4.3.1	Datasets	65
4.3.2	Evaluation metrics	66
4.3.3	Comparative baselines	67
4.3.4	Methodology	68
4.4	Evaluation	69
4.4.1	Understanding the data topology	69
4.4.2	Quantifying the utility loss	70
4.4.3	Measuring the privacy gain	72
4.5	Conclusion	75

Anonymization is crucial for the sharing of personal data in a privacy-aware manner yet it is a complex task that requires to set up a trade-off between the robustness of anonymization (the privacy level provided) and the quality of the analysis that can be expected from anonymized data (the resulting utility). Synthetic data has emerged as a promising solution to overcome the limits of classical anonymization methods while achieving similar statistical properties to the original data.

In this domain, projection-based synthetic data methods [11, 105, 84, 26] are a specific type of synthetic data generation that rely on local stochastic simulation modeling for data generation. For example, *Avatar* [62] (Section 4.2.1) protocol specifically generates an avatar for each original record depending on its local neighborhood. While these approaches have been used in healthcare, their attack surface is not well documented and understood. In this chapter, which is an extension of our paper currently in publication [98], we address this issue by providing an extensive assessment of such approaches and comparing them against other data synthesis methods and standard anonymization schemes such as k -anonymity. Our empirical analysis using various datasets show that avatar-generated data are subject to the same utility and privacy trade-off as other data synthesis approaches with a privacy risk more important on the edge data, which correspond to records that have the fewest alter egos in the original data. Finally, we propose an improvement, *M-Avatar* (section 4.2.2) based on local modeling, which leads to significant decrease in the attacks' success rates while maintaining high quality of generated synthetic data; responding to our problematic on Figure 1.1: Improving the utility/privacy trade-off of Synthetic Data generation as a privacy enhancing method.

4.1 Introduction

As presented earlier in this thesis, the collection of personal data has grown to a tremendous proportion and is done through diverse sources such as credit cards, medical records, digital photographs, emails, websites, social media, Internet of Things (IoT), smartphones, wearable technologies, to name a few [149, 122, 164]. All of this data has enormous value for improving the understanding of human behavior and creating useful societal applications, but it also raises serious privacy concerns. For instance, healthcare generates massive amounts of data whose sharing and re-using is essential for accelerating research and to develop robust machine learning algorithms methods that can be deployed in clinical settings. Specifically, this health data can be used to improve the quality of care and knowledge of the health system, identify disease risk factors, assist in diagnosis, monitor of the effectiveness of treatments, deliver personalized healthcare value, etc [101]. However, this data is very sensitive and must be anonymized before it can be used beyond the purpose of its initial collection.

Anonymization is a complex task that requires calibrating a trade-off between the privacy guarantees (e.g., robustness to privacy attacks) and the remaining usefulness of anonymized data, which is difficult to control and depends on the data and the analysis considered. Thus, in practice, a high privacy protection often results in a limited utility. To overcome this limitation, the use of synthetic data that resemble the real data (which preserves global statistical properties and task-specific performance) is increasingly recognized as a promising way to enable such reuse while addressing personal data privacy concerns [31]. For example, some projections predict that synthetic data will

completely eclipse real data in AI models by 2030¹. However, there is still no consensus on a standard approach to systematically and quantitatively assess the privacy gain and residual utility of synthetic data, which slow their adoption.

Nonetheless, to shed some light on the real guarantees of synthetic data and help hospitals position themselves on this new technology, some papers have started to assess the privacy [9] and utility [160] of synthetic data for medical data analyses.

Recently, new approaches based on local projection have attracted attention for generate synthetic patient-data [62]. For each individual observation, this approach identifies the k nearest neighbors in a latent space and leverages this neighborhood to generate an avatar through a local stochastic modeling. While appealing these projection-based approaches lack a proper privacy assessment [82]. To overcome this limitation, in this chapter we present an extensive utility and privacy assessment of avatar data based on a wide variety of metrics using multiple real-world datasets. More precisely, we quantify the privacy of synthetic data through criteria used to evaluate anonymization schemes according to the GDPR, namely singling out, linkability and inference. In addition, we have also implemented a re-identification attack (mapping a synthetic data record to a close raw data record) and a membership inference attack (inferring data records leveraged to generate a synthetic dataset). These two metrics might be related to the criteria established by the Article 29 Working Party (G29) on data protection. Specifically, re-identification can be aligned with the concept of linkability as it allows the association of synthetic and raw data, while membership inference may resemble a form of inference as it potentially reveals sensitive information about belonging to the original dataset. Moreover, we evaluate the utility through an extensive set of different metrics, and compared the avatar approach to different synthetic data generation methods as well as anonymization schemes. Our main objective is to provide a comprehensive assessment of the utility and privacy of avatar data to subsequently facilitate their use in the medical field under the best conditions. We also discuss the main limits of this approach and propose an improvement to overcome them. Specially, this improvement is based on local modeling in the latent space, depicts utility and privacy trade-off better aligned with the state-of-the-art.

The outline of this chapter is as follows.

- Section 4.2.1 we describe the Avatar approach as well as our solution M-Avatar,
- Section 4.3 present our experimental setup,
- Section 4.4 present our evaluation and the associated results,
- Section 4.5 conclude our contribution.

4.2 Avatar Data: Limitations and Improvement

In this section we will first present the Avatar model and then our contribution M-Avatar. The notions of PCA, FAMM and KDE are needed to understand the following models and will not be deeply explained in this thesis as they are basic yet complex. However here is a short description:

¹<https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai>

- PCA² (Principal Component Analysis) is a dimensionality reduction technique that transforms data into a set of uncorrelated variables called principal components and ordered following the quantity of information they hold.
- FAMD³ (Factor Analysis for Mixed Data) extends PCA to datasets containing both categorical and continuous variables.
- KDE⁴ (Kernel Density Estimation) is a non-parametric method used to estimate the probability density function of a random variable. For M-Avatar we use it to transform a data distribution into a continuous function to sample it with a Monte Carlo approach: by generating uniformly a floating number between 0 and 1, we can generating its corresponding value in the original distribution by inverting its KDE function.

4.2.1 The Avatar Approach

The Avatar method [62] has been designed for biomedical analysis from tabular data.

The original dataset is composed of n entries of p variables. Each entry represents an individual and each variable can be continuous, categorical, Boolean or represent a date. The Avatar method aims to create a new dataset of n synthetic observations and p variables with consistent yet different values compared with those of the original dataset. The Figure 4.1 (similar to the Figure 5 in the article [62]) and its following pseudo-code deeply describe its functioning.

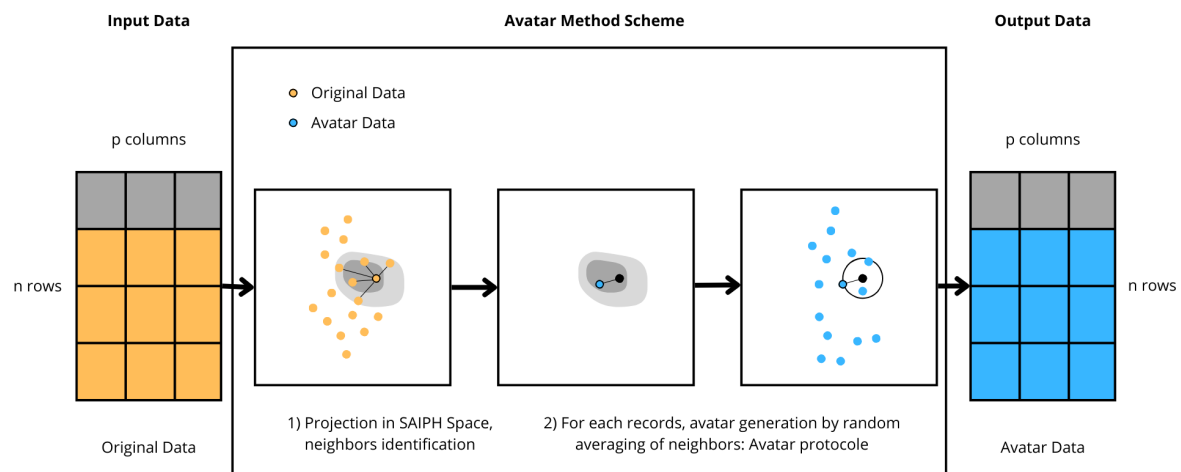


Figure 4.1: Avatars are generated for each record through stochastic averaging of its neighborhood.

To achieve this, Avatar relies on three main steps:

1. Each individual's data record is projected from the original space into a latent space using a factor analysis technique (e.g., PCA, FAMD))

²https://en.wikipedia.org/wiki/Principal_component_analysis

³https://en.wikipedia.org/wiki/Factor_analysis_of_mixed_data

⁴https://en.wikipedia.org/wiki/Kernel_density_estimation

2. Using the first d dimensions of this space, pairwise Euclidean distances are calculated between all projections associated with the individuals' data to find the k nearest neighbors.
3. for each individual, a single avatar is created by pseudo-stochastically weighting the attributes of its k nearest neighbors and is then de-projected from the latent space to the original one.
4. Synthetic data (all avatar data) are then shuffled to change the order between the original individuals and the avatars.

More precisely, here are the mathematical expression of the **Avatar** process, following the equations wrote in [62]:

$$\text{For } i \in \{1, \dots, k\}, P_i = D_i \times R_i \times C_i \quad (1)$$

Where:

- D_i is the inverse of the distance between O and its i -th neighbor V_i in the latent space.
- $R_i \sim \xi(1)$ is a random weight following an exponential distribution with $\lambda = 1$.
- $C_i = \frac{1}{2^j}$ is a contribution, where j is the value at the i -th index of the randomly shuffled vector $[1, 2, \dots, k]$.

Each weight is then normalized using the following equation:

$$W_i = \frac{P_i}{\sum_{j=1}^k P_j} \quad (2)$$

Where W_i is the normalized weight for the i -th nearest neighbor.

The avatar A of the original record O of neighborhood $V_{i \in [1, k]}$ is then:

$$A = \sum_{j=1}^k V_j * W_j$$

It is important to note that the neighborhood is chosen on the first d axis, but all the calculus are done on all the axis afterward, especially for de-projection part in step 3.

Avatar is not the only method exploiting the neighborhood as for instance the Local Linear Embedding (LLE) [30] first computes the nearest neighbors before doing the projection in an embedding. Also, similarly to **Avatar** Chawla et al. [26]: introduce the SMOTE (Synthetic Minority Over-sampling Technique) algorithm that generates synthetic samples to address class imbalance in datasets, particularly for the minority class. The steps for generating a synthetic instance are as follows:

1. Randomly select a minority class observation, called the "initial" observation.
2. Identify its k nearest neighbors among the minority class observations (where k is a user-defined parameter).
3. Randomly select one of the k nearest neighbors.

4. Generate a random coefficient α between 0 and 1 (excluding 1).
5. Create a new synthetic instance between the initial observation and the selected nearest neighbor using the value of α . For example, if $\alpha = 0.5$, the new synthetic instance will be located halfway between the initial observation and the selected nearest neighbor.

This technique helps create new, diverse synthetic instances by interpolating between minority class data points, effectively addressing class imbalance issues.

Although the *Avatar* method depicts an interesting trade-off between utility and privacy, several issues remain. More specifically, the evaluation of privacy is carried out globally and with ad-hoc metrics, which does not make it possible to properly capture the real risk for certain atypical and vulnerable individuals. To improve the utility and privacy trade-off, the value of k could also be dynamically defined according to the context of each data point to adapt the utility and privacy trade-off for each of them and thus limit the degradation for profiles which are already well protected because they are located in a dense area. The most limiting aspect of *Avatar* is the fact that the input data has the same size as the output data and that each avatar comes from a single raw data and its neighborhood.

This bijective nature opens up the risk of re-identification (mapping an avatar to a raw data), which is not the case when a model is build and then exploited to generate synthetic data of arbitrary size.

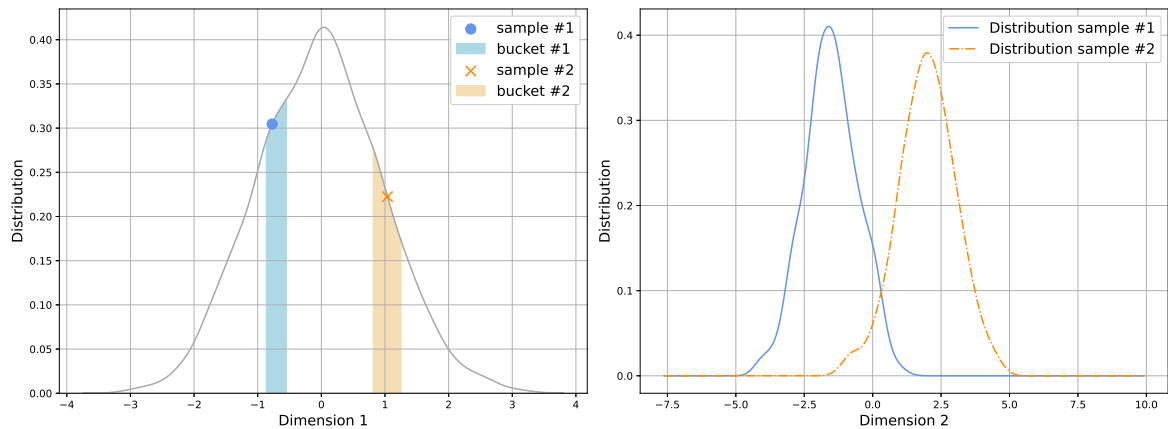


Figure 4.2: To generate synthetic data of arbitrary size, M-Avatar conditionally samples in each of the first d dimensions of the latent space.

Figure 4.2 illustrates a toy example created using a dataset based on two conditionally generated Gaussian noises: $(X_i^1, X_i^2)_{i \in \{1, \dots, 1000\}}$. Here, $X_i^1 \sim \mathcal{N}(0, 1)$ and $X_i^2 \sim X_i^1 + \mathcal{N}(0, 1)$. This toy example helps to better illustrate the functioning of our new approach M-Avatar, which is presented below.

4.2.2 Our Contribution: The M-Avatar Model

To overcome the limitations of *Avatar*, we propose an alternative method, called M-Avatar, which builds a global model that makes it possible to generate synthetic data on-demand, while removing the constraints of producing one avatar data for each original profile. To achieve this goal, we:

1. First, we construct the data distribution of the projections in the first d dimensions of the latent space and then we use Kernel Density Estimation (KDE).
2. Afterwards to generate synthetic data, we first sample a value in the distribution of the first dimension of the latent space (Figure 4.2) before building the conditional distribution to this sample in the second dimension (we consider a bucket gathering 10% of data around the sample) and sample again a value in this distribution. This operation is then repeated for the first d dimensions. This conditional construction of the distribution makes it possible to preserve the neighborhood information (local modeling) in the first dimensions that contain the most information by sampling in dimension d_i a value consistent with the sample chosen in dimension $d_i - 1$.
3. For dimensions greater than d , the quantity of data respecting the constraints of previous sampling being considerably reduced, sampling from a distribution that is too sparse would reduce the utility too much. To avoid this, we randomly choose a value among the projection values of the original data at the considered dimension. This random choice makes it also possible to mix the influence of different data while maintaining a good level of utility.

The closest state-of-the-art method is Local Resampler [84], which samples locally from the original data distribution (compared to M-Avatar which conditionally samples from each dimension of the latent space) to create an avatar data.

4.3 Evaluation Setup

4.3.1 Datasets

We consider seven real-world datasets covering broad application domains. To improve the readability, results are illustrated only through the AIDS and WBCD dataset. However, our evaluation results are presented in tabular format in the annex (see A). These also include the min-max performance range for each method, along with their median values (Tabular 4.1), providing an overview of the overall performance of each method.

The datasets considered in our study are covering several fields: healthcare (MEPS, AIDS, and WBCD), criminal justice (COMPAS), income prediction (ADULT), school admission (LAW), and credit card (CREDIT).

- Medical Expenditure Panel Survey (MEPS) contains 15,830 samples with 138 features of different patients using medical services by capturing the trips made to clinics and hospitals [35].
- Acquired Immunodeficiency Syndrome (AIDS) gathers 2,139 patients and 26 variables for HIV-infected patients who participated in a clinical trial published in 1996 in the New England Journal of Medicine [65]. This dataset contains highly sensitive information such as the race, the gender, the homosexuality and the use of injection-drugs for its patients.
- Wisconsin Breast Cancer Diagnosis (WBCD) comprises 683 observations and 10 variables computed from a digitized image of a breast cancer sample [168].

- Recidivism dataset (**COMPAS**) containing the criminal history, jail and prison time, demographics and COMPAS risk scores for defendants in Florida from 2013 and 2014. This dataset contains 6,172 criminal entries with 7 attributes [15].
- US adult income dataset (**ADULT**) comprises 30,940 data records with 95 attributes about individuals from 1994 US Census data. These attributes include marital status, education, occupation, job hours per week among others [16].
- Law school dataset (**LAW**) contains four attributes on 21,790 law students such as their entrance exam scores, their grade-point average collected prior to law school as well as their first-year average grade. The data was collected based on a survey conducted by Law School Admission Council across 163 law schools in the United States [167].
- UCI Credit Card dataset (**CREDIT**) is from the UCI Machine Learning dataset repository and contains information about different credit card applicants. The dataset contains 30,000 records with 24 attributes for each profile [172].

4.3.2 Evaluation metrics

There are numerous ways to evaluate synthetic data [48, 42, 5] such as utility metrics that measure the quality of synthetic data and its ability to faithfully reproduce the original data as well as privacy measures, which quantifies the leakage of personal information. More precisely, to evaluate the utility, we considered the SDV quality score [133], which captures the overall assessment of synthetic data's quality, combining various aspects like statistical similarity, data characteristics, and correlations between pairs of attributes. We also considered the predictive balanced accuracy of the synthetic data by examining the performance of a learning model trained with original data or trained with the synthetic data, this is the Task Accuracy (see specification below). In addition, for the AIDS dataset we considered the survival curve, which is a healthcare metric adapted for this dataset.

The classification tasks considered for the Task Accuracy depend on the dataset: AIDS is about predicting if patients have immune deficiencies, for WBCD is about determining if there are malignant cells for breast cancer, MEPS is about predicting the need for medical expenditures, COMPAS is about predicting recidivism, ADULT is about salary prediction, LAW is about predicting if students will obtain their law diploma, and CREDIT is about predicting if clients will have credit defaults.

4.3.2.1 Anonymeter Metrics

To assess the privacy guarantees, we leverage on **Anonymeter** [57], a statistical framework to jointly quantify different types of privacy risks in synthetic tabular datasets (Section 2.7.3 provides further attacks description). This framework includes attack-based evaluations for the singling out, linkability, and inference risks, which are the three key indicators of factual anonymization according to data protection regulations (e.g., GDPR).

- Singling out attack determines whether attributes (or combinations thereof) that are rare or unique in the synthetic data might also be rare or unique in the original data.

- The inference attack determines whether the synthetic dataset can be exploited to make inferences about attributes of target original records. In practice, this attack infers the value of the attribute to the value associated with the closest avatar data.
- Finally, the linkability attack captures whether the synthetic dataset can be leveraged to determine whether or not two disjoint sets of original attributes belong to the same individual. In practice, this attack can be used to test if the avatar data is the closest to the two disjoint set of original data associated with the same individual.

For all these three attacks, the risk assessment quantifies whether an adversary has an advantage in attacking a person that participated in the construction of the synthetic data (leads to a leak of personal information) compared to attacking a person from the general population (control dataset).

4.3.2.2 Membership and Re-Identification Metrics

Finally, to complete the privacy evaluation, we have also implemented a re-identification and a membership inference attack. For each avatar data, a re-identification is inferred with the original data whose projection in the latent space is closest to the avatar's projection.

For *Avatar*, we designed an attack to better highlight the membership inference of the method. We perform SAIPH (we perform a PCA or a FAMD depending on data type) on the synthetic data, reducing it to five principal components (same as *Avatar*), and project both the real data (members and non-members) and the synthetic data onto this latent space. For each synthetic data point, we identify its c nearest real data points (c , also known as 'Filter Size', varying from 1 to 20 based on distance quantiles) and increment their MIA risk scores by 1. Finally, we predict the top 50% of real data points with the highest risk scores as members and the rest as non-members. We simulate other avatar generations to determine the best c value depending on the quantile. In other words, the c original data whose projections are closest to an avatar's projection are inferred as members. The value of c varies depending on the data density from 1 (as for re-identification) for dense data, to 20 for edge data.

4.3.3 Comparative baselines

To place it in relation with other state-of-the-art approaches, we evaluate the *M-Avatar* model against *Avatar* (thanks to the API provided by Octopize) as well as the following alternatives:

- SAIPH⁵: First, in order to evaluate only the impact of the latent space used by *Avatar* (without exploiting nearest neighbors to generate avatar data), we consider a solution that projects the original point into a low-dimensional latent space (SAIPH type, like the one used by *Avatar*, with a dimension limited up to 20) and reconstruct the data point in the original space from this projection. Indeed, passing through this latent space and keeping only the n first axis compresses the information and induces a loss of utility. However this comparison

⁵<https://github.com/octopize/saiph>

is limited as Avatar keeps all the dimension when de-projecting unlike our use of SAIPH here.

- The MST [113]⁶ algorithm came first in the 2018 NIST Differential Privacy Synthetic Data Competition⁷. It consists of three steps: (1) select a collection of low-dimensional marginals, (2) measure these marginals with an additional noise (we considered for our experiments $\epsilon = 3$) and (3) generate synthetic data that preserve well the noisy marginals.
- SynthPop [129]⁸ generates data from the conditional distributions. Variables are synthesized one-by-one using sequential regression modeling and are conditioned on the original variables that are earlier in the synthesis sequence.
- CT-GAN [170]⁹ uses a conditional generative adversarial network to generate synthetic tabular data that contains a mix of discrete and continuous columns.
- K-anonymity [155]¹⁰ is not a data generation scheme but rather a data anonymization technique that is used to protect individuals' privacy in a dataset. More precisely, a dataset is considered k -anonymous when, for every combination of identifying attributes in a dataset, there are at least $k - 1$ other people with the same attributes (we considered $k = 20$ as the number of considered neighbors in Avatar paper [62]).

4.3.4 Methodology

To conduct the experiments, we followed the protocol outlined below.

1. First, we have split the data into two equal-sized sets (50-50). We perform this split 25 times to obtain a statistical representation. The first set, referred to as "original data", is used to generate a synthetic dataset of the same size while the second set, the "control data" is kept aside to compute the baseline metrics. Thus, the creation of synthetic data is not influenced by the control data.
2. Then we generate the synthetic data with each pair of (generative method, original data).
3. For both utility and privacy, the control data is used to ensure that we are exclusively evaluating the impact of the synthetic data generation method.
4. Each metric result represents the average of 25 evaluations on different original/control splits for a given generation method of synthetic data.

We used the API of Octopize to generate avatar data¹¹ with $k = 20$ and $d = 5$ which are the parameters used in the original article of the method [62].

Each other generative method (CT-GAN, SynthPop, MST) are used from their respective libraries, the SAIPH approach is available on the git repository of Octopize. As for Octopize, all method parameters have been left to their default values for fair evaluation.

⁶<https://docs.smartnoise.org/synth/synthesizers/mst.html>

⁷[https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-dif\[... \]](https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-dif[...])

⁸<https://cran.r-project.org/web/packages/synthpop>

⁹<https://github.com/sdv-dev/CTGAN>

¹⁰<https://github.com/Nuclearstar/K-Anonymity>

¹¹<https://www.octopize.io/>

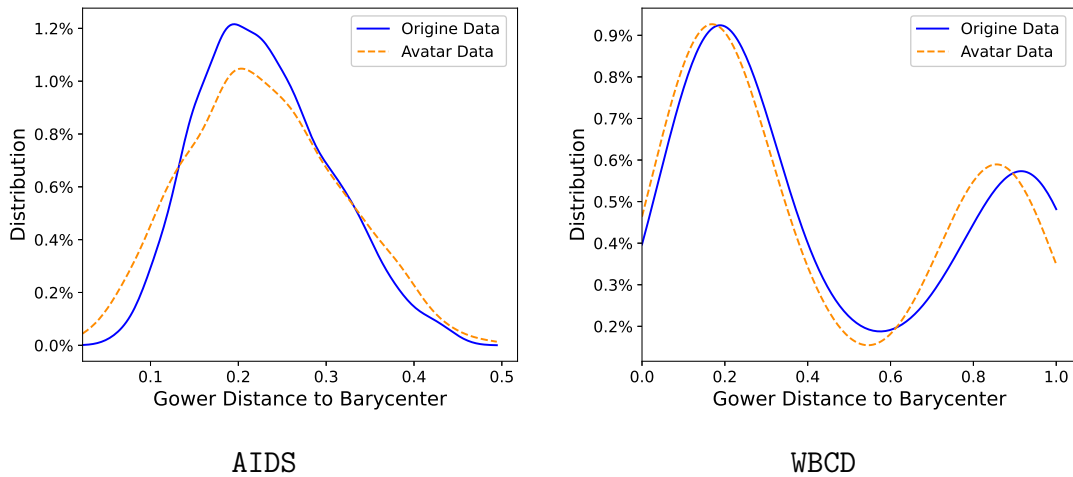


Figure 4.3: The long tail in the distribution of the distance to barycenter highlights the presence of few data at the edge which is more marked on AIDS.

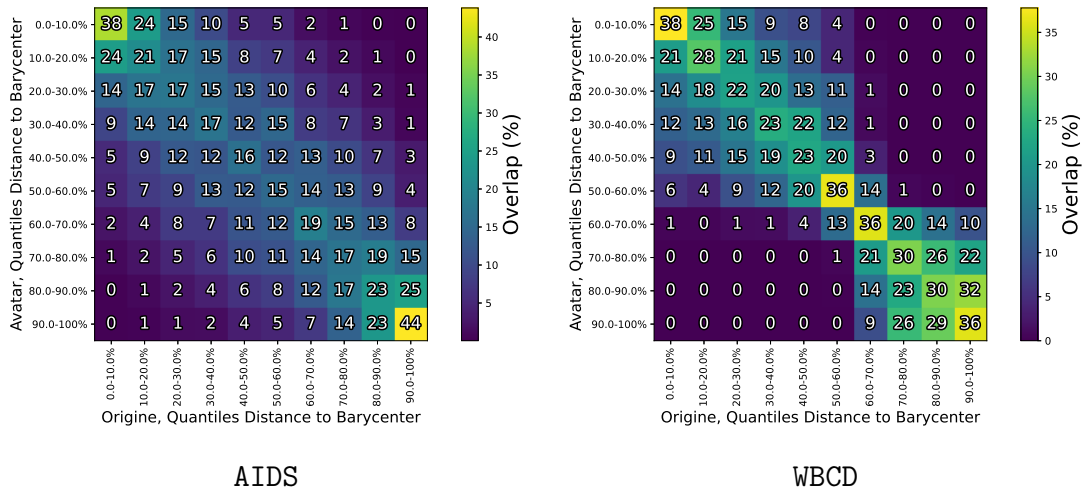


Figure 4.4: Avatars tend to be near their original data.

4.4 Evaluation

4.4.1 Understanding the data topology

In this section, we first aim at analyzing the topology and the relationship between both the original and the avatar data. To achieve this, we measure the distance of each original and avatar data to the barycenter of the data, focusing in particular our attention on the edge data. We consider the Gower and the Euclidean distance for the original and the latent space, respectively. Figure 4.3 depicts for the AIDS dataset the distribution of the data centroid (the barycenter) of the original data as well as the avatar data. We can observe that both distributions are similar and contain a long tail showing that only a few data that are far from the barycenter.

Additionally, Figure 4.4 also highlights for both AIDS and WBCD datasets that the data at the edge in the original data tends to remain at the edge in the avatar data, and vice versa (result similar on all datasets). For instance, 44% of the original data for AIDS (and 36% for WBCD) that are the farthest from the barycenter are also the farthest from the barycenter in the avatar data. Moreover, the matrix clearly shows that the distance to the center is generally preserved for all data. However, focusing on edge

data, it can be observed that they are easily distinguishable and in small numbers, which makes them more vulnerable to re-identification (Section 4.4.3).

	Orig.Data	Avatar	SAIPH	M-Avatar	CT-GAN	SynthPop	MST	K-anon.
Task Acc.	.672	.662	.603	.685	.571	.651	.608	.518
SDV Score	1.	.803	.627	.777	.700	.804	.748	.560
Linkability	.974	.191	.006	.006	.004	.007	.005	.003
Singling Out	.992	.042	.007	.013	.015	.014	.013	.010
MIA	.751	.558	.502	.501	.501	.498	.504	.501
AIA Risks								
Gender	.970	.205	.019	.022	.021	.026	.022	.035
Race	.975	.191	.029	.028	.030	.037	.028	.020

Table 4.1: Median Values of Utility and Privacy Metrics Across Datasets

4.4.2 Quantifying the utility loss

Afterwards, we have evaluated the quality of the synthetic data. First, we assess the impact of the size of the latent space for SAIPH on the utility. As described in Section 4.3.3, our application of SAIPH only involves projecting the original data point into a latent space, keeping n axis and removing the last ones and then projecting back to its original space. This baseline to better understand Avatar: they choose only the 5 first axis in SAIPH to build the neighborhood of each record, which will be largely different (wider) as a neighborhood evaluated on all axis, see the information reduction. Figure 4.5 depicts for the AIDS dataset, the survival curve according to a growing size of the latent space, from 2 to 20 over the 26 dimensions of the original data. The results show that the larger the size of the latent space, the closer the survival curve is to the one from the original data.

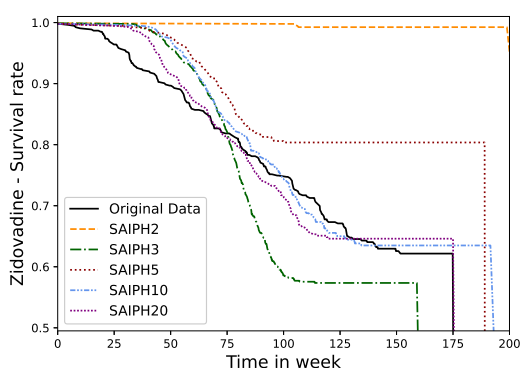


Figure 4.5: Survival curves: the faithfulness of results is correlated to the size of the latent space of SAIPH, the larger, the better.

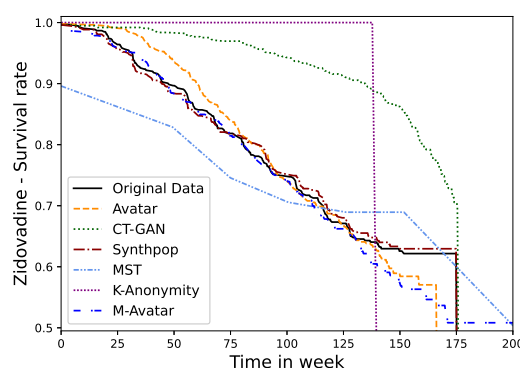


Figure 4.6: The survival curve provided by SynthPop and M-Avatar are very close to the one obtained with the original data.

We have also compared the survival curve from the avatar data against the data from other comparative baselines (Figure 4.6). The results obtained show that both SynthPop and M-Avatar produce a survival curve that closely matches the one from the

original data. Conversely, the results show that the data from CT-GAN and K-anonymity provide survival indicators that are not usable. Similarly, MST also strongly deteriorates the survival rate. The survival curve from the avatar data and from the SAIPH latent space is slightly impacted with the difference between these two curves coming from the exploitation of local neighbors for the generation of avatar data. Taking advantage of this neighborhood improves the fidelity of the survival curve compared to the original data.

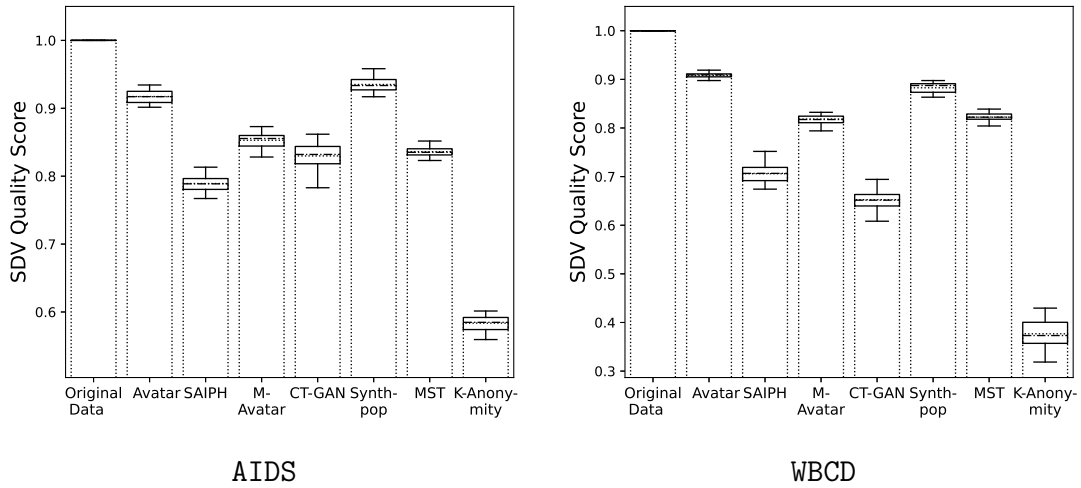


Figure 4.7: The statistical properties of the original data (captured by the SDV score) are preserved by all synthetic data generation schemes, except for K-anonymity which degrades significantly the data.

To evaluate the impact on statistical properties (e.g., statistical similarity, data characteristics and correlations between attributes), we then compute the SDV average quality score. Figure 4.7 reports for the AIDS dataset this quality score for the avatar data and for the other comparative baselines. The results show that apart from K-anonymity, which clearly deteriorates the statistical properties of the data, all other approaches maintain an SDV quality score close to 0.7, in which a score of 0.95 is achieved with data close to the original data. We find similar results on WBCD and on other datasets in the annex.

Finally, to evaluate the impact of synthetic data on predictive tasks, we compare the accuracy of the classification on a test data of a Random Forests model trained from original data compared to one trained on synthetic data, which we call Task Accuracy (tasks of predicting if patients have immune deficiencies and determining if there are malignant cells for breast cancer for AIDS and WBCD as described in Section 4.3.2). In both evaluations, the test was the same, distinctly separated from the training data that served to generate the synthetic data.

Figure 4.8 displays the balanced accuracy of the classifier trained from all the considered synthetic data generation schemes in the case of the AIDS dataset. Results show that the balanced accuracy provided by MST is close to the accuracy from the original data, however those results are exceptional on the AIDS dataset and are not present on other dataset like WBCD and the other datasets. The Avatar approach is just behind with a balanced accuracy slightly higher than 0.8, followed by the other methods. The general order of baselines (Tabular 4.1) in term of utility are generally Avatar, M-Avatar and SynthPop with comparative performances on SDV and Task Accuracy, then the order is MST, CT-GAN- SAIPH and finally K-anonymity.

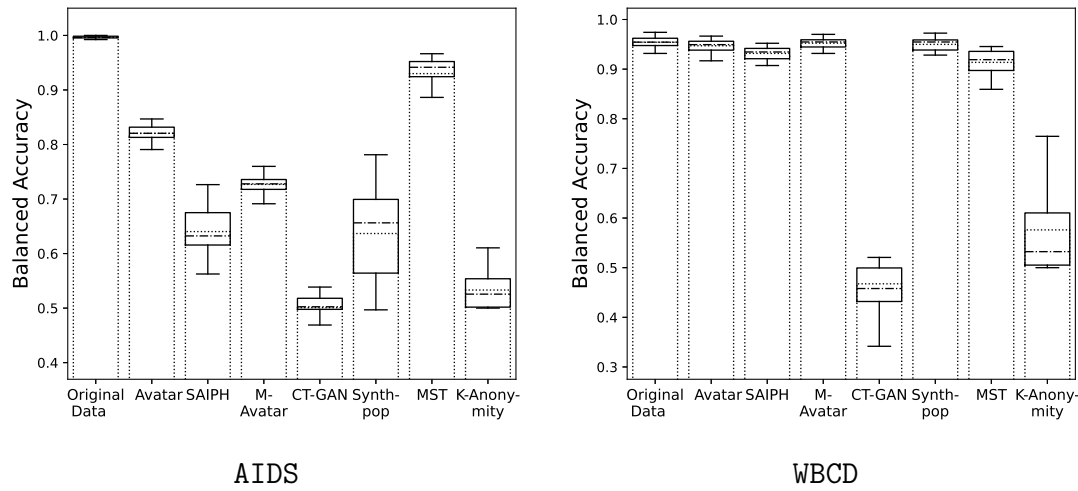


Figure 4.8: The balanced accuracy of a prediction task trained from synthetic data varies according to the method used: MST and the avatar-based approach provide the best performance.

4.4.3 Measuring the privacy gain

In this section, we evaluate the privacy gain brought by synthetic data methods. Specifically, we quantify the privacy risk associated with the disclosure of synthetic data against a singling-out, linkage, attribute inference, re-identification and membership inference attack. Figure 4.9 depicts the risk of inference for AIDS. The solution that displays the highest risk is the *Avatar* approach (privacy risk around 0.3 for the Sexual Orientation and Gender Inference). We believe that the high risk for the avatar data comes from the fact that both *Avatar* and the implementation of the attack exploit neighborhood information. The other baselines display a similar inference risk level below 0.1 for the Sexual Orientation or around 0.15 for Gender. We did not display WBCD here as it does not contains sensitive attributes.

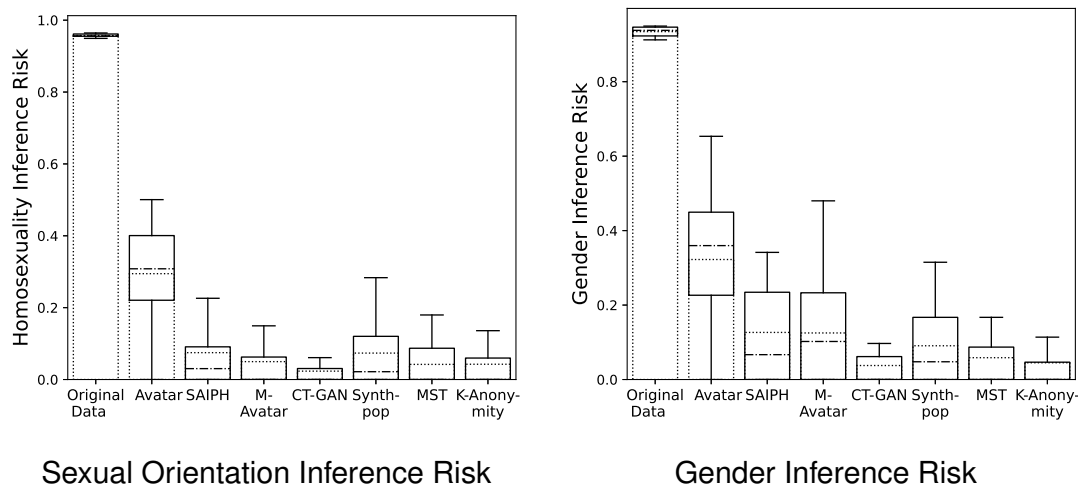


Figure 4.9: The risk of attribute inference is slightly higher for avatar data on AIDS than for other synthetic data.

Figures 4.10 and 4.11, in turn, displays respectively for the AIDS dataset the risk of singling out and the risk of linkability. The results show that the risk of singling out remains very low for all baselines, which means that all these baselines significantly reduce the uniqueness of synthetic data compared to the original data which are highly

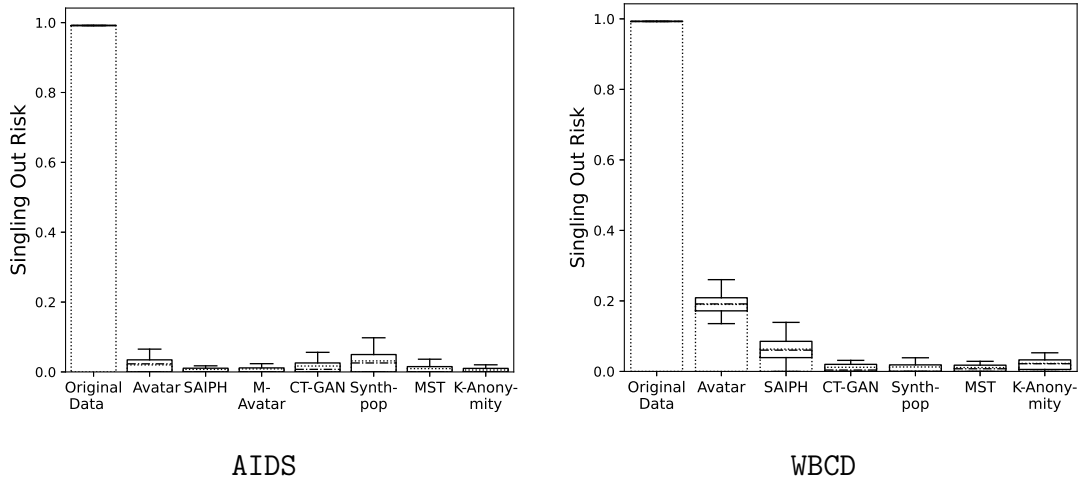


Figure 4.10: The risk of singling out is very low for all synthetic data generation schemes.

unique, except for *Avatar* on WBCD, which depicts a privacy risk around 0.2, on all datasets (Tabular 4.1) the risk for avatar is usually 4 times higher than any other baselines. The results also show that the risk of linkability remains very limited for all baselines except for *Avatar*, which depicts a privacy risk around 0.3 and usually near 20 times higher than other baselines. The high risk for this approach comes from the fact that this attack (as *Avatar*) leverages the closest neighbors to infer the linkability. The risk for *M-Avatar* is higher on WBCD however this is an exception as it is not the case on other datasets.

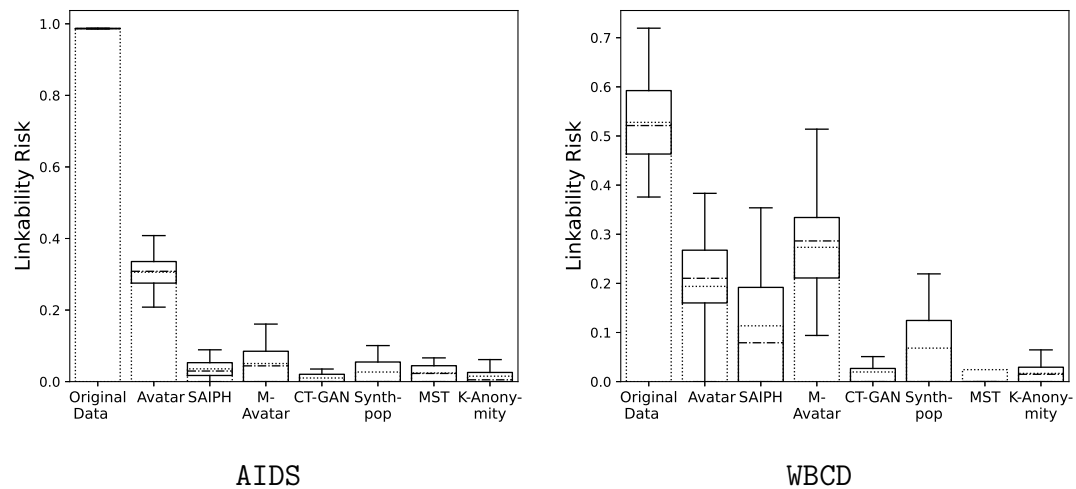


Figure 4.11: The risk of linkability is generally greater for avatar data than for synthetic data from other methods.

Then, we evaluate the risk of re-identification according to the distance of the avatar data to the barycenter (Figure 4.12). As explained in Section 4.4.1, the original data which is at the edge tends to remain at the edge also in the avatar data. The results on AIDS show that the avatar’s edge data is more likely to be re-identified than the data in the center of the point cloud. More precisely, the edge data in the last quantile (more distinguishable) exhibits a risk close to 30% while data belonging to the densest part (less distinguishable) displays a re-identification risk of 8%. As it is easy to identify edge data, an average risk of re-identification (here the dotted line close to 10%) does not sufficiently reflect the real risk of re-identification. Figure 4.12 also depicts the risk

of re-identification for SAIPH, which is much more important than Avatar (up to 60% for edge data). Similar tendencies can be found on WBCD but less marked.

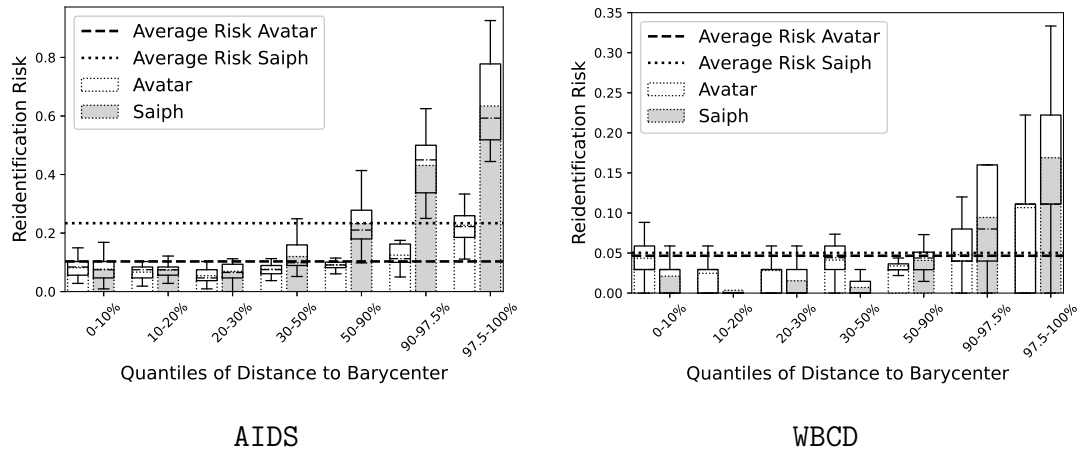


Figure 4.12: The risk of re-identification is much more important for avatar data at the edge.

A complementary result on the re-identification is assessing whether a given record is more likely to be re-identified between multiple iterations of the Avatar process. Through 25 runs we compute the probability to be re-identified for each record. Figure 4.13 highlights that even the outlier population with the highest chance of re-identification, this risk is rarely over 0.5 (a record is re-identified half of the time), underlying that this risk is stochastic.

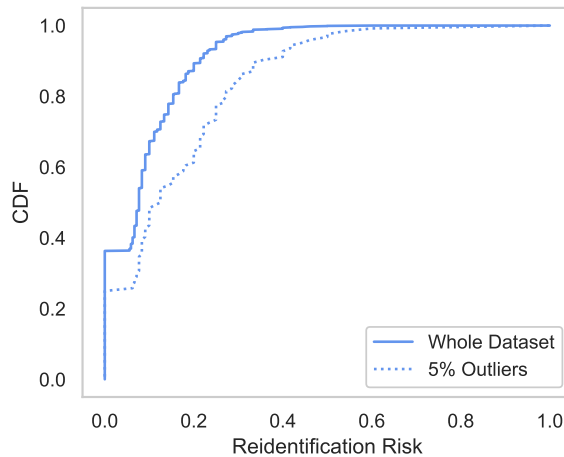


Figure 4.13: No record is re-identified on each generation on AIDS.

On the figure 4.14, we illustrate the impact of k , referred to here as 'Filter Size,' in relation to the distance from the data center. For better attack performance on the AIDS dataset, k must be increased as the data becomes less dense. By learning the optimal k based on the data topology from other Avatar datasets, we manage to achieve a uniform risk across the entire dataset. However, the results differ for the WBCD dataset, as this attack depends on the topology, and has only categorical variables, the data distribution is inherently different.

Finally, we evaluated the risk of membership inference. Figure 4.14 depicts for the AIDS dataset the balanced accuracy depending on the position of the avatar data relative to the barycenter, with the results showing that the membership inference is

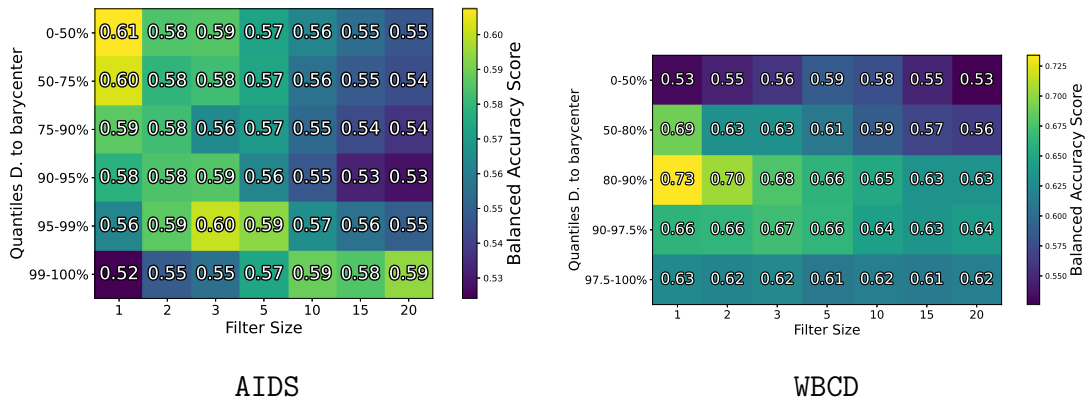


Figure 4.14: The risk of membership inference for Avatar is homogeneous for all AIDS data.

homogeneous for all data, around 0.6. Figure 4.15 compares the risk of membership inference for all synthetic data generation methods for the AIDS and WBCD datasets. The results demonstrate that only Avatar and SAIPH introduce a risk, while the others including M-Avatar significantly reduce this risk. The other takeaway from this figure is that while our attack (tailored to target projection-based synthetic data approaches such as Avatar) is ineffective against the other baselines, this does not imply that the other baselines are fully protected against membership inference attacks.

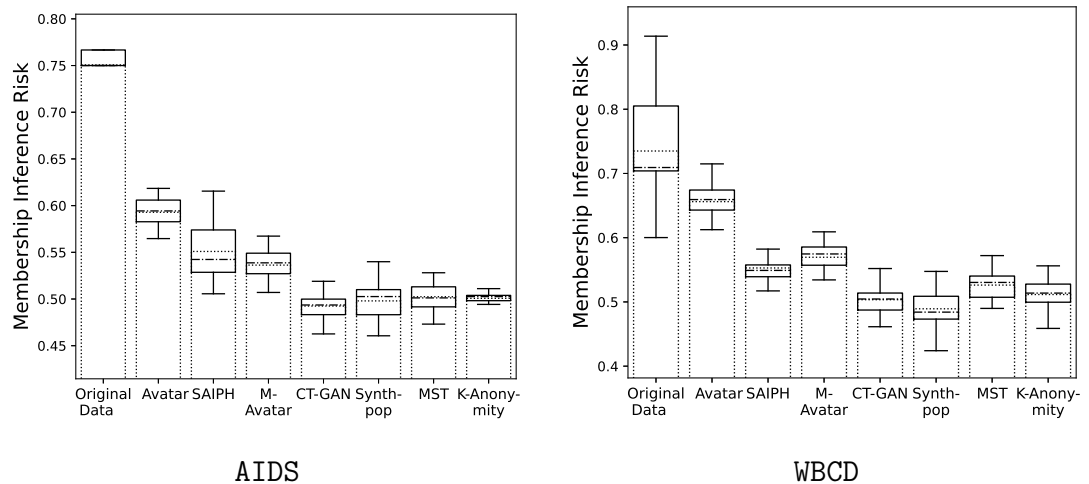


Figure 4.15: M-Avatar reduces the risk of membership inference compared to Avatar. (Here 0.5 represent the null risk).

4.5 Conclusion

In this chapter, we have conducted an in-depth utility and privacy assessment of the projection-based synthetic data approaches. We have found that edge data in the original data tends to remain at the edge in the avatar data, which favors the probability of being re-identified compared to data that is less distinguishable. We also propose an alternative method (called M-Avatar) based on conditional sampling in the latent space, which allows synthetic data to be generated on-demand. Specifically, by removing the bijective nature of avatar data (a raw data produces an avatar, a constraint that only concerns certain use cases), M-Avatar can generate synthetic data

of arbitrary size. Finally, in terms of utility and privacy compromise, MST, SynthPop, and the proposed M-Avatar solution comes out on top in our comparison.

Octopize Acknowledgment

Finally, we would like to thank Octopize that made the application available for this study and to have taken the time for explaining us their solution. We also want to emphasize that our feedback has been taken into account for future versions of the method, particularly regarding the handling of outliers, the reduction of the bijective aspect of the method, as well as the improvement of their privacy metrics.

Chapter 5

Perspectives and Conclusion

Contents

5.1	Thesis Contributions	78
5.2	Privacy Challenges and Future Directions in Healthcare AI	79
5.2.1	Short-Term Perspectives	79
5.2.2	Long-Term Perspectives	80
5.3	Conclusion	82

5.1 Thesis Contributions

In this research, my goal was to assess and enhance the protection offered by privacy-enhancing mechanisms for health data, such as Federated Learning and Synthetic Data generation, and without compromising their performance to answer our research problematic (Figure 1.1): how to improve privacy on privacy enhancing technologies without harming their usefulness.

To reach this goal, we introduced two important contributions: **MixNN** and **M-Avatar**.

Firstly, we developed **MixNN**, a proxy integrating a secured enclave positioned between the clients and the aggregating server for Federated Learning. **MixNN** enables clients to enhance their privacy independently of the server by encrypting the updates exchanged between the enclave and the clients. The enclave then decrypts these updates using secure computation techniques and without having access to the original communication. By mixing the updates at a layer level, **MixNN** returns them to the server while maintaining the model's performance and with low execution times. **MixNN** improves privacy by reducing the risk of sensitive information leakage of model updates from a curious aggregation server.

To quantitatively demonstrate **MixNN**'s ability to protect privacy while preserving utility, we implemented the method and experimentally evaluated it using datasets relevant to healthcare applications and neural network architectures. We performed membership and attribute inference attacks to quantify and compare the privacy leakage of **MixNN** against traditional Federated Learning schema, a model compression approach, and a baseline using perturbation (noise) as in Differential Privacy. Our results show that **MixNN** reduces membership inference attacks compared to other baselines—on average, up to 73.9%, 73.8%, and -0.2% less inference against classical Federated Learning, model compression, and Local Differential Privacy, respectively—and limits attribute inference—on average, up to 13.8%, 14.6%, and 12.9% less inference against the same baselines—without decreasing the accuracy of the globally aggregated model.

Secondly, we proposed **M-Avatar**, a novel synthetic data generation method focused on local modelling to better capture relationships between variables in healthcare datasets. To handle high dimensionality and complex data distributions, we first reduce the dimensionality using dimension reduction mechanism (PCA, FAMD) to create a latent space. In this latent space, we generate the first coordinate randomly and then conditionally generate the subsequent dimensions based on the previous ones. Once generated, the synthetic samples are mapped back to the original space. As well as the representation provided by the reduction mechanism retains meaningful structure, this method allows us to generate complex distributions through local distribution extrapolation. **M-Avatar** achieves a balance between statistical representation and reduce sensitive information leakage in the synthetic data.

We evaluated **M-Avatar** on several healthcare datasets, as well as on other more general. We measured both the utility and privacy of the generated synthetic data, comparing **M-Avatar** against other baselines including **Avatar**, **SynthPop**, **MST**, and others. Our results show that **M-Avatar** enhances data utility without compromising privacy. Specifically, **M-Avatar** achieved a median Task Accuracy of 0.685, surpassing other baselines such as **Avatar** (0.662), **SynthPop** (.651) and **MST** (.608), which shows that models trained on **M-Avatar**'s synthetic data retain predictive performance close

to models trained on the original data. In term of statistical properties conservation, **M-Avatar** has among the highest median SDV score of all baselines (0.777) and is competitive with the highest median values such as **Avatar** (0.803) and **SynthPop** (0.804). In terms of privacy, **M-Avatar** remove the re-identification risk in comparison to **Avatar** as it is an on-demand mechanism. The median membership inference risk is nearly removed between **Avatar** (0.558) and **M-Avatar** (.501, 0.5 being the best result). For Inference Attacks (AIA) on sensitive attributes like Gender and Race, **M-Avatar** significantly lowers the inference risks. The median AIA risk for Gender on **M-Avatar** (.022) is ten times lower than **Avatar** (0.205), other baselines hold similar results. These results indicate that **M-Avatar** effectively enhances data privacy by significantly lowering the risks associated with re-identification and sensitive attribute inference, without decreasing the accuracy of machine learning models trained on the synthetic data. Compared to ongoing methods such as **Avatar**, the trade-off between privacy and utility is greatly improved and reach the state of the art such as **SynthPop**. Our contributions are applicable to healthcare data systems, where privacy concerns are critical due to the sensitive nature of personal health information. We provide our codes in open access to guarantee the reproducibility of our results and facilitate the evaluation of privacy risks as their evaluation should be systematic. By integrating **MixNN** into Federated Learning processes, healthcare institutions can collaboratively train models without exposing individual patient data. Similarly, **M-Avatar** enables the sharing of synthetic healthcare datasets that maintain utility for research and analysis while protecting patient privacy.

5.2 Privacy Challenges and Future Directions in Healthcare AI

While our work made progresses on enhancing privacy in machine learning for healthcare, we also considered different directions for future research to keep improving privacy protections and address emerging challenges. In this section, we will discuss both short-term and long-term perspectives.

5.2.1 Short-Term Perspectives

5.2.1.1 Enhancing MixNN with Additional Protections

To further secure **MixNN**, we plan to integrate Differential Privacy (DP) mechanisms. By adding calibrated noise to the updates within the enclave, we can provide formal privacy guarantees while controlling the impact on model performance. The DP could be deployed with different modes: on client side or on enclave side depending on the level of protection desired (a general one or a targeted one). Furthermore, it would allow us to provide a more refined baseline compared to our LDP, which merely adds noise to the updates and drastically reduces model performance. Additionally, combining **MixNN** with secure aggregation techniques can offer stronger protection against potential server-side adversaries.

We also aim to extend **MixNN** to handle client poisoning and backdoor attacks. Although **MixNN** is compatible with aggregation strategies that mitigate such attacks (e.g., median-based aggregation), we will explore the possibility to deploy adaptive mechanisms within the enclave to detect and counteract malicious updates.

Moreover, we are considering implementing mixing at the bit level (0-1) on updates, greatly improving privacy without compromising utility but might have a performance issue. Depending on the level of mixing, protecting against malicious behavior might become more difficult and should be evaluated. Handling client collusion with the server scenarios is another aspect we plan to investigate, ensuring that *MixNN* keeps its robustness even when multiple clients and the server attempt to compromise the system collaboratively.

5.2.1.2 Further Testing and Validation

Evaluating *MixNN*'s performance in real-world healthcare applications is also to study, particularly in scenarios with asynchronous updates, client selection strategies, and highly imbalanced data distributions to still verify its applicability. This will help us understand practical applications and their difficulties and optimize *MixNN* for deployment in an already existing system and in diverse healthcare settings.

Additionally, we will investigate whether certain data distributions may still result in data leakage despite the use of *MixNN*, aiming to strengthen its robustness against various types of privacy attacks. Further exploration of the impact of different parameters on *MixNN*'s performance and privacy guarantees will enhance its generality and applicability.

5.2.1.3 Extending M-Avatar's Capabilities

For *M-Avatar*, we intend to generalize the method by exploring automatic selection of parameters to optimize the quality and privacy of the synthetic data.

Integrating Differential Privacy into *M-Avatar* is another short-term goal. By incorporating DP mechanisms during the data generation process by adding calibrated noise before mapping back the data, we can provide formal privacy guarantees for the synthetic data. This will enhance *M-Avatar*'s suitability for sensitive healthcare applications where strict privacy assurances are required.

We also hope to collaborate further with industry partners like Octopize to refine *Avatar* and facilitate its adoption in real-world healthcare data sharing scenarios by assuring its risks are rigorously estimated.

5.2.2 Long-Term Perspectives

5.2.2.1 Federated Synthetic Data Generation

Combining Federated Learning with synthetic data generation presents a relevant long-term research direction. We envision developing methods for generating synthetic data in a federated manner, where institutions collaboratively create synthetic datasets without sharing raw data. This approach could enhance privacy and facilitate large-scale healthcare data analyses. The key to deploy *M-Avatar* with Federated Learning would be to study the deployment of dimension reduction such as PCA and FAMD as explained in section 4.2.2. PCA with Federated Learning has been proposed by Grammenos et al. [60] and Federated FAMD is still to be explored but should be reachable as FAMD is a generalization of PCA.

Deploying *M-Avatar* in a secure Federated Learning environment using *MixNN* can further protect sensitive information during synthetic data generation. This integration

would involve adapting MixNN to handle the generation process and ensuring that the synthetic data is securely shared among participating institutions. However as Federated Learning and Synthetic Data are both reducing the utility of potential which could represent a serious limitation.

5.2.2.2 Addressing Adversarial Behaviors

Detecting and mitigating adversarial behaviors, such as client poisoning or backdoor attacks, might be difficult. In the context of privacy-preserving models, where individual updates are obscured, traditional detection mechanisms may be less effective. Future research will focus on developing robust strategies to identify and counteract malicious activities without compromising privacy.

For MixNN, detecting malicious updates should become more difficult depending on the granularity of the mixing. However instead of detecting, protection mechanism such as Federated Median Aggregation are still deployable.

For and as well for Synthetic Data, detecting the consequence of poisoned records in a model and in generated data are two different tasks. To correct a poisoned synthetic data model, the field of Machine Unlearning [127] could have some interesting direction to explore. Letting the model forget the impact of a poisonous data record should remove its impact. For an already generated synthetic data the solution might be less evident. Studying whether a synthetic data record represent a true outlier data or a poisonous data should be similar as in regular data: this domain is still open and usually outlier detection methods are used to detect poisonous data and except from expert review it seems compromised to different.

5.2.2.3 Bridging Theoretical Gaps in Privacy-Preserving Machine Learning

Machine Learning, particularly in privacy-preserving contexts, often lacks strong theoretical foundations for predicting performance and understanding privacy risks. Developing formal models and theoretical frameworks that define the limits and capabilities of methods like Differential Privacy in specific use cases (e.g., Federated Learning or synthetic data generation) is essential. This will provide clear guarantees and help generalize risk evaluations beyond specific attack models.

On the MixNN side, studying the impact of applying Differential Privacy either on the client or on the enclave is the first step to assess de empirical epsilon. This could help to further protect the global model against privacy leakage which MixNN do not handle. This direction is in development and scripts to deploy it are ready.

On the side, adding Differential Privacy have been less studied. A first direction could be to add noise on generated data in the latent space before de-projecting them. The noise variance should be a new parameter to increase the privacy and reduce utility. The empirical epsilon of such approach can be assessed afterward.

5.2.2.4 Expanding Privacy Metrics

To comprehensively evaluate privacy risks, we aim to expand our set of privacy metrics and propose a general framework for privacy evaluation, building upon tools like Anonymeter [57]. This framework will help standardize privacy assessments across different methods and datasets, providing clearer insights into the effectiveness of various privacy-preserving techniques. Federated Learning and Synthetic Data are

generally improving privacy but one of the key messages of this thesis is to always assess privacy risks. This is where a generalized framework to evaluate utility and privacy of models and dataset is relevant. Providing further metrics should be done through regular study of the state of the art. To achieve this, we would pay a particular attention to the simplicity of its deployment: for data analysis, the input should only be a csv file for example (case of tabular data). This framework should continually evolve as new attacks and vulnerabilities are discovered: protections that are effective today may not be sufficient against tomorrow's threats.

5.2.2.5 Regulatory Compliance and Ethical Considerations

With evolving regulations like the GDPR and the EU AI Act, aligning our methods with legal requirements is crucial. Future work will focus on integrating compliance mechanisms into the design of privacy-preserving algorithms. Additionally, addressing ethical considerations, such as fairness and transparency, will be essential to ensure that our methods are not only legally compliant but also socially responsible.

There is often a gap between the legal and scientific aspects of privacy. While the legal side focuses on identifying the victims and determining responsibility, data science concentrates on evaluating metrics to quantify information leakage. Bridging these two perspectives can be difficult, as translating legal requirements into technical measures and demonstrating compliance is often complex.

5.3 Conclusion

Our research focused on preserving privacy while maintaining utility in machine learning models applied to healthcare data. Federated Learning and synthetic data generation offer relevant alternatives to centralized learning models but come with their own privacy challenges. Through our contributions, *MixNN* and *M-Avatar*, we aimed to provide practical solutions to enhance privacy protections without compromising performance.

However, privacy issues in healthcare AI are not solely technical problems; they encompass regulatory, ethical, and societal dimensions. Ensuring robust privacy requires interdisciplinary collaboration among researchers, healthcare professionals, policymakers, and legal experts. By acknowledging this broader context, we emphasize that advancing privacy-preserving AI in healthcare is a collective effort.

Looking ahead, there is still progress to be done in developing comprehensive privacy protections, understanding the theoretical foundations of privacy risks, and creating frameworks that balance utility and privacy. Our future work is to keep improving our contributions to develop privacy preserving AI for modern healthcare systems and personalized medicine.

Long-term collaboration among stakeholders is essential to sustain innovation in healthcare and AI while safeguarding sensitive data. By continuing to develop and refine privacy-preserving technologies, we can enable the benefits of AI in healthcare without compromising individual privacy rights.

Bibliography

- [1] General Data Protection Regulation. Official Journal of the European Union, L119:1–88, 2016.
- [2] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS '16, pages 308–318, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-4139-4. doi: 10.1145/2976749.2978318. URL <https://doi.org/10.1145/2976749.2978318>. event-place: Vienna, Austria.
- [3] Agency for Healthcare Research and Quality. Medical Expenditure Panel Survey. URL <https://meps.ahrq.gov/mepsweb/>.
- [4] Akinsola Ahmed, Ejiofor Oluomachi, Akinde Abdullah, and Njoku Tochukwu. Enhancing data privacy in wireless sensor networks: Investigating techniques and protocols to protect privacy of data transmitted over wireless sensor networks in critical applications of healthcare and national security. 16(2):47–63. ISSN 09752307. doi: 10.5121/ijnsa.2024.16204. URL <http://arxiv.org/abs/2404.11388>.
- [5] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In International Conference on Machine Learning, pages 290–306. PMLR, 2022.
- [6] A. K. M. Mubashwir Alam, Sagar Sharma, and Keke Chen. SGX-MR: Regulating Dataflows for Protecting Access Patterns of Data-Intensive SGX Applications. arXiv preprint arXiv:2009.03518, 2020.
- [7] Maria Alvarellos, Hadley E. Sheppard, Ingrid Knarston, Craig Davison, Nathaniel Raine, Thorben Seeger, Pablo Prieto Barja, and Maria Chatzou Dunford. Democratizing clinical-genomic data: How federated platforms can promote benefits sharing in genomics. 13: 1045450. ISSN 1664-8021. doi: 10.3389/fgene.2022.1045450. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9871385/>.
- [8] Ittai Anati, Shay Gueron, Simon Johnson, and Vincent Scarlata. Innovative technology for CPU based attestation and sealing. Technical report, Intel, 2013.
- [9] Arno Appenzeller, Moritz Leitner, Patrick Philipp, Erik Krempel, and Jürgen Beyerer. Privacy and Utility of Private Synthetic Data for Medical Data Analyses.

- Applied Sciences*, 12(23), 2022. ISSN 2076-3417. doi: 10.3390/app122312320. URL <https://www.mdpi.com/2076-3417/12/23/12320>.
- [10] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated Learning with Personalization Layers. *arXiv preprint arXiv:1912.00818*, 2019.
- [11] Sergul Aydore, William Brown, Michael Kearns, Krishnaram Kenthapadi, Luca Melis, Aaron Roth, and Ankit Siva. Differentially private query release through adaptive projection. URL <https://arxiv.org/abs/2103.06641v2>.
- [12] Zahra Azizi, Simon Lindner, Yumika Shiba, Valeria Raparelli, Colleen M. Norris, Karolina Kublickiene, Maria Trinidad Herrero, Alexandra Kautzky-Willer, Peter Klimek, Teresa Gisinger, Louise Pilote, and Khaled El Emam. A comparison of synthetic data generation and federated analysis for enabling international evaluations of cardiovascular health. 13(1):11540. ISSN 2045-2322. doi: 10.1038/s41598-023-38457-3. URL <https://www.nature.com/articles/s41598-023-38457-3>. Publisher: Nature Publishing Group.
- [13] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How To Backdoor Federated Learning. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2938–2948. PMLR, August 2020. URL <https://proceedings.mlr.press/v108/bagdasaryan20a.html>.
- [14] Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. Data mining for wearable sensors in health monitoring systems: A review of recent trends and challenges. 13(12):17472–17500. ISSN 1424-8220. doi: 10.3390/s131217472. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3892855/>.
- [15] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It’s compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. In *Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.
- [16] Barry Becker and Ronny Kohavi. Adult, 1996. Published: UCI Machine Learning Repository.
- [17] James Henry Bell, Kallista A. Bonawitz, Adrià Gascón, Tancrède Lepoint, and Mariana Raykova. Secure Single-Server Aggregation with (Poly)Logarithmic Overhead. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, CCS ’20*, pages 1253–1269, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-7089-9. doi: 10.1145/3372297.3417885. URL <https://doi.org/10.1145/3372297.3417885>. event-place: Virtual Event, USA.
- [18] Steven M Bellovin, Preetam K Dutta, and Nathan Reiter. Privacy and synthetic datasets. *Stan. Tech. L. Rev.*, 22:1, 2019. Publisher: HeinOnline.

- [19] Vincent Bindschaedler, Reza Shokri, and Carl A Gunter. Plausible deniability for privacy-preserving data synthesis. *PVLDB*, 10(5):481–492, 2017.
- [20] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv:1611.04482*, 2016.
- [21] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS '17*, pages 1175–1191, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 978-1-4503-4946-8. doi: 10.1145/3133956.3133982. URL <https://doi.org/10.1145/3133956.3133982>. event-place: Dallas, Texas, USA.
- [22] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards Federated Learning at Scale: System Design. In A. Talwalkar, V. Smith, and M. Zaharia, editors, *Proceedings of Machine Learning and Systems*, volume 1, pages 374–388, 2019. URL <https://arxiv.org/pdf/1902.01046>.
- [23] Antoine Boutet, Carole Frindel, Sébastien Gambis, Théo Jourdan, and Rosin Claude Ngueveu. DySan: Dynamically Sanitizing Motion Sensor Data Against Sensitive Inferences through Adversarial Networks. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security, ASIA CCS '21*, pages 672–686, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8287-8. doi: 10.1145/3433210.3453095. URL <https://doi.org/10.1145/3433210.3453095>. event-place: Virtual Event, Hong Kong.
- [24] Ferdinand Brasser, Urs Müller, Alexandra Dmitrienko, Kari Kostiaainen, Srdjan Capkun, and Ahmad-Reza Sadeghi. Software Grand Exposure: SGX Cache Attacks Are Practical. In *11th USENIX Workshop on Offensive Technologies (WOOT 17)*, Vancouver, BC, August 2017. USENIX Association. URL <https://www.usenix.org/conference/woot17/workshop-program/presentation/brasser>.
- [25] David L. Chaum. Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms. *Commun.*, 24(2):84–90, February 1981. URL <https://dl.acm.org/doi/10.1145/358549.358563>.
- [26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. 16:321–357. ISSN 1076-9757. doi: 10.1613/jair.953. URL <https://www.jair.org/index.php/jair/article/view/10302>.

- [27] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pages 343–362, October 2020. doi: 10.1145/3372297.3417238. URL <http://arxiv.org/abs/1909.03935>. arXiv:1909.03935 [cs].
- [28] Feng Chen, Shuang Wang, Xiaoqian Jiang, Sijie Ding, Yao Lu, Jihoon Kim, S Cenk Sahinalp, Chisato Shimizu, Jane C Burns, Victoria J Wright, and others. Princess: Privacy-protecting rare disease international network collaboration via encryption through software guard extensions. Bioinformatics, 33(6):871–878, 2016. Publisher: Oxford University Press.
- [29] Feng Chen, Chenghong Wang, Wenrui Dai, Xiaoqian Jiang, Noman Mohammed, Md Momin Al Aziz, Md Nazmus Sadat, Cenk Sahinalp, Kristin Lauter, and Shuang Wang. PRESAGE: Privacy-preserving genetic testing via software guard extension. BMC medical genomics, 10(2):48, 2017. Publisher: BioMed Central.
- [30] Jing Chen and Yang Liu. Locally linear embedding: A survey. Artif. Intell. Rev., 36:29–48, June 2011. doi: 10.1007/s10462-010-9200-z.
- [31] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, 5(6):493–497, 2021. Publisher: Nature Publishing Group UK London.
- [32] Olivia Choudhury, Aris Gkoulalas-Divanis, Theodoros Salonidis, Issa Sylla, Yoonyoung Park, Grace Hsu, and Amar Das. Differential privacy-enabled federated learning for sensitive health data. URL <http://arxiv.org/abs/1910.02578>.
- [33] CNIL. ARTICLE 29 DATA PROTECTION WORKING PARTY - opinion 05/2014 on anonymisation techniques, . URL https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.
- [34] CNIL. Recherche scientifique (hors santé) : focus sur certaines catégories de données personnelles, . URL <https://www.cnil.fr/fr/recherche-scientifique-hors-sante/focus-certaines-categories-donnees-personnelles>.
- [35] Joel W Cohen, Steven B Cohen, and Jessica S Banthin. The medical expenditure panel survey: a national information resource to support healthcare cost research and inform policy and practice. Medical care, 47(7_Supplement_1):S44–S50, 2009. Publisher: LWW.
- [36] Victor Costan and Srinivas Devadas. Intel SGX Explained. IACR Cryptology ePrint Archive, 2016(086):1–118, 2016.
- [37] Emiliano De Cristofaro. An Overview of Privacy in Machine Learning. arXiv preprint arXiv:2005.08679, 2020.

- [38] Jessamyn Dahmen and Diane Cook. SynSys: A synthetic data generation system for healthcare applications. 19(5):1181. ISSN 1424-8220. doi: 10.3390/s19051181. URL <https://www.mdpi.com/1424-8220/19/5/1181>. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.
- [39] Georgios Damaskinos, Rachid Guerraoui, Anne-Marie Kermarrec, Vlad Nitu, Richeek Patra, and Francois Taiani. FLeet: Online Federated Learning via Staleness Awareness and Performance Prediction. arXiv preprint arXiv:2006.07273, 2020.
- [40] G. Danezis, R. Dingledine, and N. Mathewson. Mixminion: design of a type III anonymous remailer protocol. In 2003 Symposium on Security and Privacy, 2003., pages 2–15, 2003. doi: 10.1109/SECPRI.2003.1199323. URL <https://www.mixminion.net/minion-design.pdf>.
- [41] Fida K Dankar, Mahmoud K Ibrahim, and Leila Ismail. A multi-dimensional evaluation of synthetic data generators. IEEE Access, 10:11147–11158, 2022. URL https://www.researchgate.net/publication/357961715_A_Multi-Dimensional_Evaluation_of_Synthetic_Data_Generators. Publisher: IEEE.
- [42] Fida Kamal Dankar, Khaled El Emam, Angelica Neisa, and Tyson Roffey. Estimating the re-identification risk of clinical data sets. BMC medical informatics and decision making, 12:1–15, 2012. Publisher: Springer.
- [43] Erfan Darzidehkalani, Mohammad Ghasemi-rad, and P. M. A. van Ooijen. Federated Learning in Medical Imaging: Part I: Toward Multicentral Health Care Ecosystems. Journal of the American College of Radiology, 19(8):969–974, 2022. ISSN 1546-1440. doi: <https://doi.org/10.1016/j.jacr.2022.03.015>. URL <https://www.sciencedirect.com/science/article/pii/S1546144022002800>.
- [44] Ilias Driouich, Chuan Xu, Giovanni Neglia, Frederic Giroire, and Eoin Thomas. A novel model-based attribute inference attack in federated learning. URL <https://openreview.net/forum?id=jJx00vsVVSF>.
- [45] Jiacheng Du, Jiahui Hu, Zhibo Wang, Peng Sun, Neil Zhenqiang Gong, and Kui Ren. SoK: Gradient leakage in federated learning. URL <http://arxiv.org/abs/2404.05403>.
- [46] Cynthia Dwork. Differential privacy: A survey of results. In TAMC, pages 1–19, 2008.
- [47] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. 9(3):211–407. ISSN 1551-305X, 1551-3068. doi: 10.1561/04000000042. URL <http://www.nowpublishers.com/articles/foundations-and-trends-in-theoretical-computer-science/TCS-042>.
- [48] Khaled El Emam. Seven ways to evaluate the utility of synthetic data. IEEE Security & Privacy, 18(4):56–59, 2020. Publisher: IEEE.

- [49] Gokberk Elmas, Salman UH Dar, Yilmaz Korkmaz, Emir Ceyani, Burak Susam, Muzaffer Özbey, Salman Avestimehr, and Tolga Çukur. Federated Learning of Generative Image Priors for MRI Reconstruction, 2022. URL <https://arxiv.org/abs/2202.04175>. _eprint: 2202.04175.
- [50] Linwei Fang, Liming Wang, and Hongjia Li. Iterative and mixed-spaces image gradient inversion attack in federated learning. 7(1):35. ISSN 2523-3246. doi: 10.1186/s42400-024-00227-7. URL <https://doi.org/10.1186/s42400-024-00227-7>.
- [51] Mei Ling Fang, Devendra Singh Dhama, and Kristian Kersting. Dp-ctgan: Differentially private medical data generation using ctgans. In International Conference on Artificial Intelligence in Medicine, pages 178–188. Springer, 2022.
- [52] Dyke Ferber, Omar S. M. El Nahhas, Georg Wölflein, Isabella C. Wiest, Jan Clusmann, Marie-Elisabeth Leßman, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, Manuel Salto-Tellez, Nikolaus Schultz, Daniel Truhn, and Jakob Nikolas Kather. Autonomous artificial intelligence agents for clinical decision making in oncology. URL <http://arxiv.org/abs/2404.04667>.
- [53] André Ferreira, Ricardo Magalhães, Sébastien Mériaux, and Victor Alves. Generation of Synthetic Rat Brain MRI scans with a 3D Enhanced Alpha-GAN, 2022. URL <https://arxiv.org/abs/2112.13626>. _eprint: 2112.13626.
- [54] 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, CCS '15, pages 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 978-1-4503-3832-5. doi: 10.1145/2810103.2813677. URL <https://doi.org/10.1145/2810103.2813677>. event-place: Denver, Colorado, USA.
- [55] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. Property Inference Attacks on Fully Connected Neural Networks Using Permutation Invariant Representations. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS '18, pages 619–633, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 978-1-4503-5693-0. doi: 10.1145/3243734.3243834. URL <https://doi.org/10.1145/3243734.3243834>. event-place: Toronto, Canada.
- [56] Sayantari Ghosh and Saumik Bhattacharya. A data-driven understanding of COVID-19 dynamics using sequential genetic algorithm based probabilistic cellular automata. URL <http://arxiv.org/abs/2008.12020>.
- [57] Matteo Gioni, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. A Unified Framework for Quantifying Privacy Risk in Synthetic Data, 2023. URL <https://petsymposium.org/popets/2023/popets-2023-0055.php>. Publication Title: Proceedings of Privacy Enhancing Technologies Symposium.

- [58] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks, 2014. URL <https://arxiv.org/abs/1406.2661>. _eprint: 1406.2661.
- [59] J. C. Gower. A general coefficient of similarity and some of its properties. 27 (4):857. ISSN 0006341X. doi: 10.2307/2528823. URL <https://www.jstor.org/stable/2528823?origin=crossref>.
- [60] Andreas Grammenos, Rodrigo Mendoza-Smith, Jon Crowcroft, and Cecilia Mascolo. Federated principal component analysis. URL <https://arxiv.org/abs/1907.08059v3>.
- [61] Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. A method for generating synthetic electronic medical record text. 18(1):173–182. ISSN 1557-9964. doi: 10.1109/TCBB.2019.2948985.
- [62] Morgan Guillaudeux, Olivia Rousseau, Julien Petot, Zineb Bennis, Charles-Axel Dein, Thomas Goronflot, Nicolas Vince, Sophie Limou, Matilde Karakachoff, Matthieu Wargny, and Pierre-Antoine Gourraud. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *npj Digital Medicine*, 6(1):37, March 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00771-5. URL <https://doi.org/10.1038/s41746-023-00771-5>.
- [63] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein GANs. URL <http://arxiv.org/abs/1704.00028>.
- [64] Yaron Gurovich, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M. Krawitz, Susanne B. Kamphausen, Martin Zenker, Lynne M. Bird, and Karen W. Gripp. Identifying facial phenotypes of genetic disorders using deep learning. 25(1):60–64. ISSN 1546-170X. doi: 10.1038/s41591-018-0279-0. URL <https://www.nature.com/articles/s41591-018-0279-0>. Publisher: Nature Publishing Group.
- [65] Scott M Hammer, David A Katzenstein, Michael D Hughes, Holly Gundacker, Robert T Schooley, Richard H Haubrich, W Keith Henry, Michael M Lederman, John P Phair, Manette Niu, and others. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15):1081–1090, 1996. Publisher: Mass Medical Soc.
- [66] Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A. Horowitz, and William J. Dally. EIE: Efficient Inference Engine on Compressed Deep Neural Network. *SIGARCH Comput. Archit. News*, 44(3):243–254, June 2016. ISSN 0163-5964. doi: 10.1145/3007787.3001163. URL <https://doi.org/10.1145/3007787.3001163>. Place: New York, NY, USA Publisher: Association for Computing Machinery.
- [67] Anika Hannemann, Jan Ewald, Leo Seeger, and Erik Buchmann. Federated learning on transcriptomic data: Model quality and performance trade-offs. URL <http://arxiv.org/abs/2402.14527>.

- [68] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.
- [69] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Membership Inference Attacks Against Generative Models, August 2018. URL <http://arxiv.org/abs/1705.07663>. arXiv:1705.07663 [cs].
- [70] Geon Heo, Junseok Seo, and Steven Euijong Whang. Personalized DP-SGD using sampling mechanisms. URL <http://arxiv.org/abs/2305.15165>.
- [71] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. 493: 28–45. ISSN 0925-2312. doi: 10.1016/j.neucom.2022.04.053. URL <https://www.sciencedirect.com/science/article/pii/S0925231222004349>.
- [72] Tony Hey and Anne Trefethen. The fourth paradigm 10 years on. 42(6):441–447. ISSN 1432-122X. doi: 10.1007/s00287-019-01215-9. URL <https://doi.org/10.1007/s00287-019-01215-9>.
- [73] Wouter Heyndrickx, Lewis Mervin, Tobias Morawietz, Noé Sturm, Lukas Friedrich, Adam Zalewski, Anastasia Pentina, Lina Humbeck, Martijn Oldenhof, Ritsuya Niwayama, Peter Schmidtke, Nikolas Fechner, Jaak Simm, Adam Arany, Nicolas Drizard, Rama Jabal, Arina Afanasyeva, Regis Loeb, Shlok Verma, Simon Harnqvist, Matthew Holmes, Balazs Pejo, Maria Telenczuk, Nicholas Holway, Arne Dieckmann, Nicola Rieke, Friederike Zumsande, Djork-Arné Clevert, Michael Krug, Christopher Luscombe, Darren Green, Peter Ertl, Peter Antal, David Marcus, Nicolas Do Huu, Hideyoshi Fuji, Stephen Pickett, Gergely Acs, Eric Boniface, Bernd Beck, Yax Sun, Arnaud Gohier, Friedrich Rippmann, Ola Engkvist, Andreas H. Göller, Yves Moreau, Mathieu N. Galtier, Ansgar Schuffenhauer, and Hugo Ceulemans. MELLODDY: Cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. 64(7):2331–2344. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c00799. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11005050/>.
- [74] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. *Proceedings on Privacy Enhancing Technologies*, 2019(4):232–249, October 2019. ISSN 2299-0984. doi: 10.2478/popets-2019-0067. URL <https://petsymposium.org/popets/2019/popets-2019-0067.php>.
- [75] Geoffrey E. Hinton and R. Zemel. Autoencoders, minimum description length and helmholtz free energy. URL <https://www.semanticscholar.org/paper/Autoencoders,-Minimum-Description-Length-and-Free-Hinton-Zemel/3dc3a0efe58eaf8564ca1965c0ffd23ec495b83f>.
- [76] Zhuoqiao Hong, Colin G. Magdamo, Yi-han Sheu, Prathamesh Mohite, Ayush Noori, Elissa M. Ye, Wendong Ge, Haoqi Sun, Laura Brenner, Gregory Robbins, Shibani Mukerji, Sahar Zafar, Nicole Benson, Lidia Moura, John Hsu, Bradley T.

- Hyman, Michael B. Westover, Deborah Blacker, and Sudeshna Das. Natural language processing to detect cognitive concerns in electronic health records using deep learning. URL <http://arxiv.org/abs/2011.06489>.
- [77] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.
- [78] Jihyeon Hyeong, Jayoung Kim, Noseong Park, and Sushil Jajodia. An Empirical Study on the Membership Inference Attack against Tabular Data Synthesis Models, August 2022. URL <http://arxiv.org/abs/2208.08114>. arXiv:2208.08114 [cs].
- [79] Mahmoud Ibrahim, Yasmina Al Khalil, Sina Amirrajab, Chang Sun, Marcel Breeuwer, Josien Pluim, Bart Elen, Gokhan Ertaylan, and Michel Dumontier. Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. URL <http://arxiv.org/abs/2407.00116>.
- [80] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing Differentially Private Machine Learning: How Private is Private SGD? arXiv preprint arXiv:2006.07709, 2020.
- [81] Bargav Jayaraman, Lingxiao Wang, David Evans, and Quanquan Gu. Revisiting Membership Inference Under Realistic Assumptions. arXiv preprint arXiv:2005.10881, 2020.
- [82] Bayrem Kaabachi, Jérémie Despraz, Thierry Meurers, Karen Otte, Mehmed Halilovic, Fabian Prasser, and Jean Louis Raisaro. Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics. medRxiv, 2023. doi: 10.1101/2023.11.28.23299124. URL <https://www.medrxiv.org/content/early/2023/11/28/2023.11.28.23299124>. Publisher: Cold Spring Harbor Laboratory Press eprint: https://www.medrxiv.org/content/early/2023/11/28/2023.11.28.23299124.full.pdf.
- [83] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In Proceedings of the 32nd International Conference on Machine Learning, pages 1376–1385. PMLR. URL <https://proceedings.mlr.press/v37/kairouz15.html>. ISSN: 1938-7228.
- [84] Ali Furkan Kalay. Generating Synthetic Data with The Nearest Neighbors Algorithm. 2022. URL <https://api.semanticscholar.org/CorpusID:252683691>.
- [85] Davood Karimi. Diffusion MRI with Machine Learning, 2024. URL <https://arxiv.org/abs/2402.00019>. eprint: 2402.00019.
- [86] Raouf Kerkouche, Gergely Acs, Claude Castelluccia, and Pierre Genevès. Privacy-Preserving and Bandwidth-Efficient Federated Learning: An Application to In-Hospital Mortality Prediction. In CHIL, pages 1–11, 2021.

- [87] Nazish Khalid, Adnan Qayyum, Muhammad Bilal, Ala Al-Fuqaha, and Junaid Qadir. Privacy-preserving artificial intelligence in healthcare: Techniques and applications. 158:106848. ISSN 0010-4825. doi: <https://doi.org/10.1016/j.combiomed.2023.106848>. URL <https://www.sciencedirect.com/science/article/pii/S001048252300313X>.
- [88] Hui Kwon Kim, Seonwoo Min, Myungjae Song, Soobin Jung, Jae Woo Choi, Younggwang Kim, Sangeun Lee, Sungroh Yoon, and Hyongbum (Henry) Kim. Deep learning improves prediction of CRISPR–cpf1 guide RNA activity. 36(3): 239–241. ISSN 1546-1696. doi: 10.1038/nbt.4061. URL <https://www.nature.com/articles/nbt.4061>. Publisher: Nature Publishing Group.
- [89] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings.
- [90] Panagiotis Kitsos, Paraskevi Pappa, Panagiotis Kitsos, and Paraskevi Pappa. Mobile communications privacy. URL <https://www.igi-global.com/gateway/chapter/www.igi-global.com/gateway/chapter/112617>. Archive Location: mobile-communications-privacy ISBN: 9781466658882 Publisher: IGI Global.
- [91] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492, 2016.
- [92] Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D. Shah, Wai Keong Wong, Zina Ibrahim, James T. Teo, and Richard JB Dobson. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. URL <http://arxiv.org/abs/2010.01165>.
- [93] Aditya Kuppa, Lamine Aouad, and Nhien-An Le-Khac. Towards improving privacy of synthetic datasets. In Annual Privacy Forum, pages 106–119. Springer, 2021.
- [94] Albert Kwon, David Lazar, Srinivas Devadas, and Bryan Ford. Riffle: An Efficient Communication System With Strong Anonymity. Proc. Priv. Enhancing Technol., 2016(2):115–134, 2016.
- [95] Lucas Lange, Nils Wenzlitschke, and Erhard Rahm. Generating synthetic health sensor data for privacy-preserving wearable stress detection. 24(10): 3052. ISSN 1424-8220. doi: 10.3390/s24103052. URL <https://www.mdpi.com/1424-8220/24/10/3052>. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.
- [96] Taro Langner. Machine Learning Techniques for MRI Data Processing at Expanding Scale, 2024. URL <https://arxiv.org/abs/2404.14326>. _eprint: 2404.14326.

- [97] Stevens Le Blond, David Choffnes, Wenxuan Zhou, Peter Druschel, Hitesh Ballani, and Paul Francis. Towards Efficient Traffic-Analysis Resistant Anonymity Networks. *SIGCOMM Comput. Commun. Rev.*, 43(4):303–314, August 2013. URL <https://www.freehaven.net/anonbib/papers/sigcomm13-aqua.pdf>. Place: New York, NY, USA Publisher: Association for Computing Machinery.
- [98] Thomas Lebrun, Louis Béziaud, Tristan Allard, Antoine Boutet, Sébastien Gambs, and Mohamed Maouche. Synthetic data: Generate avatar data on demand. URL <https://hal.science/hal-04715055>.
- [99] Thomas Lebrun, Antoine Boutet, Jan Aalmoes, and Adrien Baud. MixNN: protection of federated learning against inference attacks by mixing neural network layers. In *Proceedings of the 23rd ACM/IFIP International Middleware Conference, Middleware '22*. ACM, November 2022. doi: 10.1145/3528535.3565240. URL <http://dx.doi.org/10.1145/3528535.3565240>.
- [100] Heidi Ledford. Google health-data scandal spooks researchers. doi: 10.1038/d41586-019-03574-5. URL <https://www.nature.com/articles/d41586-019-03574-5>. Bandiera_abtest: a Cg_type: News Publisher: Nature Publishing Group Subject_term: Technology, Society, Law, Health care.
- [101] Mariana Lenharo. An AI revolution is brewing in medicine. What will it look like? *Nature*, 622(7984):686–688, October 2023. ISSN 1476-4687 0028-0836. doi: 10.1038/d41586-023-03302-0. Place: England.
- [102] Ninghui Li, Tiancheng Li, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *ICDE*, pages 106–115, 2007.
- [103] Maxwell W. Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. 16(6):321–332. ISSN 1471-0064. doi: 10.1038/nrg3920. URL <https://www.nature.com/articles/nrg3920>. Publisher: Nature Publishing Group.
- [104] Akis Linardos, Kaisar Kushibar, Sean Walsh, Polyxeni Gkontra, and Karim Lekadir. Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. *Scientific Reports*, 12(1):3551, March 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-07186-4. URL <https://doi.org/10.1038/s41598-022-07186-4>.
- [105] Xiao Ling, Tim Menzies, Christopher Hazard, Jack Shu, and Jacob Beel. Trading off scalability, privacy, and performance in data synthesis. URL <https://arxiv.org/abs/2312.05436v1>.
- [106] Shuyu Lu, Ruoyu Chen, Wei Wei, and Xinghua Lu. Understanding heart-failure patients EHR clinical features via SHAP interpretation of tree-based machine learning model predictions. URL <http://arxiv.org/abs/2103.11254>.
- [107] Laurens van der Maaten and Awni Hannun. The Trade-Offs of Private Prediction. *arXiv preprint arXiv:2007.05089*, 2020. URL <https://arxiv.org/abs/2007.05089>.

- [108] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy Beyond k-anonymity. Transactions on Knowledge Discovery from Data, 1(1), 2007.
- [109] Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Mobile sensor data anonymization. In Proceedings of the International Conference on Internet of Things Design and Implementation, IoTDI '19, pages 49–58. ACM, . ISBN 978-1-4503-6283-2. doi: 10.1145/3302505.3310068. URL <http://doi.acm.org/10.1145/3302505.3310068>. event-place: Montreal, Quebec, Canada.
- [110] Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Protecting sensory data against sensitive inferences. In Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems, W-P2DS'18, pages 2:1–2:6. ACM, . ISBN 978-1-4503-5654-1. doi: 10.1145/3195258.3195260. URL <http://doi.acm.org/10.1145/3195258.3195260>. event-place: Porto, Portugal.
- [111] Frank McKeen, Ilya Alexandrovich, Ittai Anati, Dror Caspi, Simon Johnson, Rebekah Leslie-Hurd, and Carlos Rozas. Intel® Software Guard Extensions (Intel® SGX) Support for Dynamic Memory Management Inside an Enclave. In Proceedings of the Hardware and Architectural Support for Security and Privacy, HASP '16, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-4769-3. doi: 10.1145/2948618.2954331. URL <https://doi.org/10.1145/2948618.2954331>. event-place: Seoul, Republic of Korea.
- [112] Ryan McKenna, Daniel Sheldon, and Gerome Miklau. Graphical-model based estimation and inference for differential privacy. In International Conference on Machine Learning, pages 4435–4444. PMLR, 2019.
- [113] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data, 2021. _eprint: 2108.04978.
- [114] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. arXiv preprint arXiv:1602.05629, 2017.
- [115] John R. McNulty, Lee Kho, Alexandria L. Case, Charlie Fornaca, Drew Johnston, David Slater, Joshua M. Abzug, and Sybil A. Russell. Synthetic Medical Imaging Generation with Generative Adversarial Networks For Plain Radiographs, 2024. URL <https://arxiv.org/abs/2403.19107>. _eprint: 2403.19107.
- [116] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting Unintended Feature Leakage in Collaborative Learning. arXiv preprint arXiv:1805.04049, 2018. URL <https://arxiv.org/pdf/1805.04049>.
- [117] Ofer Mendeleevitch and Michael D. Lesh. Fidelity and privacy of synthetic medical data. URL <http://arxiv.org/abs/2101.08658>.

- [118] Usevalad Milasheuski, Luca Barbieri, Bernardo Camajori Tedeschini, Monica Nicoli, and Stefano Savazzi. On the impact of data heterogeneity in federated learning environments with application to healthcare networks. URL <http://arxiv.org/abs/2404.18519>.
- [119] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets, 2014. URL <https://arxiv.org/abs/1411.1784>. _eprint: 1411.1784.
- [120] Nicole Mitchell and Adam Pearce. How federated learning protects privacy. URL <https://pair.withgoogle.com/explorables/federated-learning/>.
- [121] Fan Mo, Hamed Haddadi, Kleomenis Katevas, Eduard Marin, Diego Perino, and Nicolas Kourtellis. PPFL: Privacy-Preserving Federated Learning with Trusted Execution Environments. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys '21*, pages 94–108, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8443-8. URL <https://doi.org/10.1145/3458864.3466628>. event-place: Virtual Event, Wisconsin.
- [122] Varda Mone and Fayazullaeva Shakhlo. Health data on the go: Navigating privacy concerns with wearable technologies. 23(3):179–188. doi: 10.1017/S1472669623000427.
- [123] Javier Montalt-Tordera, Vivek Muthurangu, Andreas Hauptmann, and Jennifer Anne Steeden. Machine Learning in Magnetic Resonance Imaging: Image Reconstruction, 2020. URL <https://arxiv.org/abs/2012.05303>. _eprint: 2012.05303.
- [124] Hajra Murtaza, Musharif Ahmed, Naurin Farooq Khan, Ghulam Murtaza, Saad Zafar, and Ambreen Bano. Synthetic data generation: State of the art in health care domain. 48:100546. ISSN 1574-0137. doi: 10.1016/j.cosrev.2023.100546. URL <https://www.sciencedirect.com/science/article/pii/S1574013723000138>.
- [125] Mohammad Naseri, Jamie Hayes, and Emiliano De Cristofaro. Toward Robustness and Privacy in Federated Learning: Experimenting with Local and Central Differential Privacy. *arXiv preprint arXiv:2009.03561*, 2021.
- [126] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *Symposium on Security and Privacy (SP), S&P '19*, pages 739–753, 2019. doi: 10.1109/SP.2019.00065. URL <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8835245>.
- [127] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. URL <http://arxiv.org/abs/2209.02299>.
- [128] NIST. 2018 Differential Privacy Synthetic Data Challenge, 2018. URL <https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic>.

- [129] Beata Nowok, Gillian M. Raab, and Chris Dibben. synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, 74(11):1–26, 2016. doi: 10.18637/jss.v074.i11. URL <https://www.jstatsoft.org/index.php/jss/article/view/v074i11>.
- [130] Siddhartha Nuthakki, Sunil Neela, Judy W. Gichoya, and Saptarshi Purkayastha. Natural language processing of MIMIC-III clinical notes for identifying diagnosis and procedures with neural networks. URL <http://arxiv.org/abs/1912.12397>.
- [131] Bristena Oprisanu, Georgi Ganey, and Emiliano De Cristofaro. On Utility and Privacy in Synthetic Genomic Data, January 2022. URL <http://arxiv.org/abs/2102.03314>. arXiv:2102.03314 [cs, q-bio].
- [132] Shaoyan Pan, Elham Abouei, Jacob Wynne, Tonghe Wang, Richard L. J. Qiu, Yuheng Li, Chih-Wei Chang, Junbo Peng, Justin Roper, Pretesh Patel, David S. Yu, Hui Mao, and Xiaofeng Yang. Synthetic CT Generation from MRI using 3D Transformer-based Denoising Diffusion Model, 2023. URL <https://arxiv.org/abs/2305.19467>. _eprint: 2305.19467.
- [133] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The Synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, October 2016. doi: 10.1109/DSAA.2016.49.
- [134] Le Peng, Gaoxiang Luo, sicheng zhou, jiangdong chen, Rui Zhang, Ziyue Xu, and Ju Sun. An in-depth evaluation of federated learning on biomedical natural language processing. URL <http://arxiv.org/abs/2307.11254>.
- [135] Haoyue Ping, Julia Stoyanovich, and Bill Howe. Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pages 1–5, 2017.
- [136] Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. Secure and robust machine learning for healthcare: A survey. URL <http://arxiv.org/abs/2001.08103>.
- [137] Daniele Raimondi, Haleh Chizari, Nora Verplaetse, Britt-Sabina Löscher, Andre Franke, and Yves Moreau. Genome interpretation in a federated learning context allows the multi-center exome-based risk prediction of crohn’s disease patients. 13(1):19449. ISSN 2045-2322. doi: 10.1038/s41598-023-46887-2. URL <https://www.nature.com/articles/s41598-023-46887-2>. Publisher: Nature Publishing Group.
- [138] J. L. Reyes-Ortiz. *Smartphone-based human activity recognition*. Springer, 2015.
- [139] Office for Civil Rights (OCR). Health information privacy. URL <https://www.hhs.gov/hipaa/index.html>. Last Modified: 2024-04-19T18:24:31-0400.
- [140] Frank Rosenblatt. *Perceptrons: An Introduction to Computational Geometry*. Cornell Aeronautical Laboratory. URL <https://bpb-us-e2.wpmucdn.com/websites.umass.edu/dist/a/27637/files/2016/03/rosenblatt-1957.pdf>.

- [141] Hossein Mohammadi Rouzbahani and Hadis Karimipour. Application of artificial intelligence in supporting healthcare professionals and caregivers in treatment of autistic children. URL <http://arxiv.org/abs/2407.08902>.
- [142] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. 323:533–536. URL <https://api.semanticscholar.org/CorpusID:205001834>.
- [143] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. [arXiv:1806.01246](https://arxiv.org/abs/1806.01246), 2018.
- [144] Sajin Sasy, Sergey Gorbunov, and Christopher W. Fletcher. ZeroTrace: Oblivious Memory Primitives from Intel SGX. In [Network and Distributed Systems Security Symposium](https://doi.org/10.14722/ndss.2018.23239), NDSS '18, February 2018. URL <http://dx.doi.org/10.14722/ndss.2018.23239>. Place: San Diego, CA, USA.
- [145] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered Federated Learning: Model-Agnostic Distributed Multi-Task Optimization under Privacy Constraints. [arXiv preprint arXiv:1910.01991](https://arxiv.org/abs/1910.01991), 2019.
- [146] Md Mobashir Hasan Shandhi, Karnika Singh, Natasha Janson, Perisa Ashar, Geetika Singh, Baiying Lu, D. Sunshine Hillygus, Jennifer M. Maddocks, and Jessilyn P. Dunn. Assessment of ownership of smart devices and the acceptability of digital health data sharing. 7(1):1–10. ISSN 2398-6352. doi: 10.1038/s41746-024-01030-x. URL <https://www.nature.com/articles/s41746-024-01030-x>. Publisher: Nature Publishing Group.
- [147] Ofir Ben Shoham and Nadav Rappoport. Federated learning of medical concepts embedding using BEHRT. URL <http://arxiv.org/abs/2305.13052>.
- [148] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In [2017 IEEE symposium on security and privacy \(SP\)](https://doi.org/10.1109/SP.2017.8246244), pages 3–18. IEEE, 2017.
- [149] Jason Peres da Silva. Privacy data ethics of wearable digital health technology | center for digital health | engineering | brown university. URL <https://cdh.brown.edu/news/2023-05-04/ethics-wearables>.
- [150] Abir Smiti. A critical overview of outlier detection methods. [Computer Science Review](https://doi.org/10.1016/j.csr.2020.100306), 38:100306, 2020. Publisher: Elsevier.
- [151] Congzheng Song and Vitaly Shmatikov. Overlearning Reveals Sensitive Attributes. [arXiv preprint arXiv:1905.11742](https://arxiv.org/abs/1905.11742), 2020.
- [152] Liwei Song and Prateek Mittal. Systematic Evaluation of Privacy Risks of Machine Learning Models. [arXiv preprint arXiv:2003.10595](https://arxiv.org/abs/2003.10595), 2020. URL <https://arxiv.org/pdf/2003.10595>.
- [153] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic Data – Anonymisation Groundhog Day, January 2022. URL <http://arxiv.org/abs/2011.07018>. [arXiv:2011.07018](https://arxiv.org/abs/2011.07018) [cs].

- [154] Latanya Sweeney. Weaving technology and policy together to maintain confidentiality. 25(2):98–110. doi: 10.1111/j.1748-720x.1997.tb01885.x.
- [155] Latanya Sweeney. k-anonymity: A model for protecting privacy. International journal of uncertainty, fuzziness and knowledge-based systems, 10(05):557–570, 2002. Publisher: World Scientific.
- [156] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 1701–1708, 2014. doi: 10.1109/CVPR.2014.220.
- [157] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of LLMs help clinical text mining? URL <http://arxiv.org/abs/2303.04360>.
- [158] Florian Tramèr and Dan Boneh. Differentially Private Learning Needs Better Features (or Much More Data). arXiv preprint arXiv:2011.11660, 2020.
- [159] DataRes at UCLA. Make-a-monet: Image style transfer with cycle GANs. URL <https://ucladatares.medium.com/make-a-monet-image-style-transfer-with-cycle-gans-5475dcb525b8>.
- [160] Vibeke Binz Vallevik, Aleksandar Babic, Serena E. Marshall, Severin Elvatun, Helga M.B. Brøgger, Sharmini Alagaratnam, Bjørn Edwin, Narasimha R. Veeraragavan, Anne Kjersti Befring, and Jan F. Nygård. Can I trust my fake data – A comprehensive quality assessment framework for synthetic tabular data in healthcare. International Journal of Medical Informatics, 185:105413, May 2024. ISSN 1386-5056. doi: 10.1016/j.ijmedinf.2024.105413. URL <http://dx.doi.org/10.1016/j.ijmedinf.2024.105413>. Publisher: Elsevier BV.
- [161] Stephan van Schaik, Andrew Kwong, Daniel Genkin, and Yuval Yarom. SGAXe: How SGX Fails in Practice, 2020. URL <https://sgaxeattack.com/>.
- [162] Paul Vanhaesebrouck, Aurélien Bellet, and Marc Tommasi. Decentralized Collaborative Learning of Personalized Models over Networks. In AISTATS, 2017. URL <https://arxiv.org/pdf/1610.05202>.
- [163] George Vavoulas, Charikleia Chatzaki, Thodoris Malliotakis, Matthew Pedititis, and Manolis Tsiknakis. The MobiAct Dataset: Recognition of Activities of Daily Living using Smartphones. In Proceedings of the International Conference on Information and Communication Technologies for Ageing Well and e-Health - ICT4AWE, (ICT4AGEINGWELL 2016), pages 143–151. SciTePress, 2016. ISBN 978-989-758-180-9. doi: 10.5220/0005792401430151. Backup Publisher: INSTICC ISSN: 2184-4984.
- [164] Vini Vijayan, James P. Connolly, Joan Condell, Nigel McKelvey, and Philip Gardiner. Review of wearable devices and data collection considerations for connected health. 21(16):5589. ISSN 1424-8220. doi: 10.3390/s21165589. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8402237/>.

- [165] Isabel Wagner and David Eckhoff. Technical Privacy Metrics: A Systematic Survey. *ACM Comput. Surv.*, 51(3), June 2018. ISSN 0360-0300. doi: 10.1145/3168389. URL <https://doi.org/10.1145/3168389>. Place: New York, NY, USA Publisher: Association for Computing Machinery.
- [166] Alan Westin. Privacy and freedom. 25(1):166. ISSN 0043-0463. URL <https://scholarlycommons.law.wlu.edu/wlulr/vol125/iss1/20>.
- [167] Linda F Wightman. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. 1998. Publisher: ERIC.
- [168] William Wolberg. Breast Cancer Wisconsin (Original), 1992. Published: UCI Machine Learning Repository.
- [169] Jing Wu, Munawar Hayat, Mingyi Zhou, and Mehrtash Harandi. Concealing sensitive samples against gradient leakage in federated learning. URL <http://arxiv.org/abs/2209.05724>.
- [170] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling Tabular data using Conditional GAN, October 2019. URL <http://arxiv.org/abs/1907.00503>. arXiv:1907.00503 [cs, stat].
- [171] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Generation and evaluation of privacy preserving synthetic health data. 416:244–255. ISSN 0925-2312. doi: 10.1016/j.neucom.2019.12.136. URL <https://www.sciencedirect.com/science/article/pii/S0925231220305117>.
- [172] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst. Appl.*, 36:2473–2480, 2009.
- [173] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *31st computer security foundations symposium*, CSF '18, pages 268–282. IEEE, 2018. URL <https://arxiv.org/pdf/1709.01604>.
- [174] Shuo Yu, Qing Qing, Chen Zhang, Ahsan Shehzad, Giles Oatley, and Feng Xia. Data-driven decision making in COVID-19 response: A survey. URL <http://arxiv.org/abs/2202.11435>.
- [175] Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging Federated Learning by Local Adaptation. *arXiv preprint arXiv:2002.04758*, 2020.
- [176] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayses: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):1–41, 2017. URL <https://dl.acm.org/doi/pdf/10.1145/3134428>. Publisher: ACM New York, NY, USA.
- [177] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M. Padilla, Jeffrey Caterino, Ping Zhang, and Dakuo

- Wang. Rethinking human-AI collaboration in complex medical decision making: A case study in sepsis diagnosis. In Proceedings of the CHI Conference on Human Factors in Computing Systems, pages 1–18. doi: 10.1145/3613904.3642343. URL <http://arxiv.org/abs/2309.12368>.
- [178] Wanrong Zhang, Shruti Tople, and Olga Ohrimenko. Dataset-Level Attribute Leakage in Collaborative Learning. arXiv preprint arXiv:2006.07267, 2020.
- [179] Benjamin Zi Hao Zhao, Aviral Agrawal, Catisha Coburn, Hassan Jameel Asghar, Raghav Bhaskar, Mohamed Ali Kaafar, Darren Webb, and Peter Dickinson. On the (In)Feasibility of Attribute Inference Attacks on Machine Learning Models. arXiv preprint arXiv:2103.07101, 2021.
- [180] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. CTAB-GAN+: Enhancing Tabular Data Synthesis, April 2022. URL <http://arxiv.org/abs/2204.00401>. arXiv:2204.00401 [cs].
- [181] Hao Zhong and Kaifeng Bu. Privacy-utility trade-off. URL <http://arxiv.org/abs/2204.12057>.
- [182] Ligeng Zhu, Zhijian Liu, and Song Han. Deep Leakage from Gradients. arXiv preprint arXiv:1906.08935, 2019. URL <https://arxiv.org/pdf/1906.08935>.

Appendix A

Annexes

	Orig.Data	Avatar	SAIPH	M-Avatar	CT-GAN	SynthPop	MST	K-anon.
Task Acc.	.542-.997	.560-.947	.512-.932	.609-.952	.467-.723	.546-.950	.503-.930	.500-.576
SDV Score	1.	.706-.917	.554-.789	.653-.885	.653-.841	.728-.935	.610-.878	.377-.705
Linkability	.111-.999	.011-.306	.003-.114	.005-.207	.002-.020	.002-.068	.003-.024	.001-.017
Singling Out	.991-.993	.020-.190	.004-.063	.007-.177	.009-.021	.012-.032	.006-.021	.006-.023
MIA	.735-.881	.542-.656	.499-.553	.494-.570	.492-.512	.489-.509	.499-.526	.499-.512
AIA Risks								
Gender	.378-.996	.071-.322	.008-.127	.011-.160	.013-.050	.010-.091	.015-.059	.020-.049
Race	.360-.994	.065-.260	.000-.088	.016-.057	.013-.065	.022-.048	.014-.062	.012-.032

Table A.1: Min-Max Values of Utility and Privacy Metrics Across Datasets

	Orig.Data	Avatar	SAIPH	M-Avatar	CT-GAN	SynthPop	MST	K-anon.
Task Acc.	.672	.662	.603	.685	.571	.651	.608	.518
SDV Score	1.	.803	.627	.777	.700	.804	.748	.560
Linkability	.974	.191	.006	.006	.004	.007	.005	.003
Singling Out	.992	.042	.007	.013	.015	.014	.013	.010
MIA	.751	.558	.502	.501	.501	.498	.504	.501
AIA Risks								
Gender	.970	.205	.019	.022	.021	.026	.022	.035
Race	.975	.191	.029	.028	.030	.037	.028	.020

Table A.2: Median Values of Utility and Privacy Metrics Across Datasets

	Orig.Data	Avatar	SAIPH	M-Avatar	CT-GAN	SynthPop	MST	K-anon.
Task Acc.	.997	.820	.640	.727	.501	.637	.930	.533
SDV Score	1.	.917	.789	.853	.830	.935	.836	.584
Linkability	.987	.306	.036	.048	.010	.027	.024	.015
Singling Out	.992	.020	.007	.021	.017	.032	.009	.006
MIA	.751	.593	.551	.536	.492	.498	.502	.501
AIA Risks								
Gender	.935	.322	.127	.160	.038	.091	.059	.045
Race	.975	.231	.088	.052	.065	.037	.062	.026
Homosexuality	.957	.295	.075	.107	.023	.074	.042	.043
Drug Use	.959	.270	.076	.085	.055	.083	.048	.059

Table A.3: AIDS - Utility and privacy metrics comparison between the different baselines.

	Orig.Data	Avatar	SAIPH	M-Avatar	CT-GAN	SynthPop	MST	K-anon.
Task Acc.	.954	.947	.932	.952	.467	.950	.914	.576
SDV Score	1.	.908	.706	.818	.653	.883	.822	.377
Linkability	.528	.194	.114	.207	.020	.068	.024	.017
Singling Out	.993	.190	.063	.177	.012	.013	.011	.023
MIA	.735	.656	.553	.570	.503	.489	.526	.512
AIA Risks								
Attributes	.756	.231	.111	.235	.036	.061	.061	.049

Table A.4: WBCD - Utility and privacy metrics comparison between the different baselines.

	Orig.Data	Avatar	SAIPH	M-Avatar	CT-GAN	SynthPop	MST	K-anon.
Task Acc.	.542	.560	.512	.609	.575	.546	.505	.500
SDV Score	1.	.719	.594	.653	.664	.728	.706	.616
Linkability	.999	.191	.003	.005	.002	.007	.003	.003
Singling Out	.992	.027	.010	.012	.019	.012	.015	.008
MIA	.770	.558	.505	.494	.498	.509	.509	.500
AIA Risks								
Gender	.996	.188	.014	.011	.013	.033	.019	.032
Race	.992	.191	.038	.057	.030	.048	.028	.015

Table A.5: LAWS - Utility and privacy metrics comparison between the different baselines.

	Orig.Data	Avatar	SAIPH	M-Avatar	CT-GAN	SynthPop	MST	K-anon.
Task Acc.	.759	.729	.602	.723	.723	.765	.608	.569
SDV Score	1.	.774	.585	.777	.700	.792	.748	.513
Linkability	.680	.114	.004	.006	.005	.007	.004	.001
Singling Out	.992	.095	.004	.013	.009	.015	.016	.012
MIA	.881	.570	.500	.507	.502	.496	.499	.501
AIA Risks								
Gender	.945	.222	.024	.039	.050	.028	.026	.049
Race	.951	.182	.000	.022	.025	.042	.032	.012

Table A.6: FEWADULT - Utility and privacy metrics comparison between the different baselines.

	Orig.Data	Avatar	SAIPH	M-Avatar	CT-GAN	SynthPop	MST	K-anon.
Task Acc.	.654	.644	.603	.619	.571	.651	.631	.508
SDV Score	1.	.706	.554	.692	.680	.738	.610	.551
Linkability	.974	.181	.011	.006	.004	.006	.006	.003
MIA	.758	.542	.502	.501	.501	.506	.504	.501
AIA Risks								
Gender	.995	.176	.029	.022	.014	.010	.016	.025

Table A.7: CREDIT - Utility and privacy metrics comparison between the different baselines.

	Orig.Data	Avatar	SAIPH	M-Avatar	CT-GAN	SynthPop	MST	K-anon.
Task Acc.	.588	.591	.586	.614	.523	.602	.557	.518
SDV Score	1.	.803	.627	.776	.728	.804	.727	.560
Linkability	.111	.011	.003	.006	.003	.002	.005	.003
Singling Out	.992	.032	.007	.008	.021	.023	.021	.009
MIA	.744	.545	.500	.499	.512	.499	.508	.499
AIA Risks								
Gender	.378	.071	.008	.019	.022	.024	.036	.038
Race	.360	.065	.000	.016	.013	.022	.014	.020

Table A.8: COMPAS - Utility and privacy metrics comparison between the different baselines.

	Orig.Data	Avatar	SAIPH	M-Avatar	CT-GAN	SynthPop	MST	K-anon.
Task Acc.	.672	.662	.658	.685	.658	.664	.503	.504
SDV Score	1.	.902	.771	.885	.841	.904	.878	.705
Linkability	.979	.223	.006	.009	.002	.005	.005	.002
Singling Out	.991	.053	.007	.007	.014	.014	.006	.011
MIA	.746	.545	.499	.495	.501	.495	.501	.501
AIA Risks								
Gender	.994	.246	.013	.023	.021	.023	.015	.020
Race	.994	.260	.029	.028	.035	.032	.024	.032

Table A.9: MEPS - Utility and privacy metrics comparison between the different baselines.

TH1178_LEBRUN Thomas_Manuscrit

9%
Suspicious
texts

6% Similarities
< 1% similarities between quotation
marks
1% among the sources mentioned

4% Unrecognized languages

4% Texts potentially generated by AI
(ignored)

Document name: TH1178_LEBRUN Thomas_Manuscrit.pdf Document ID: 2926669f0bc4824a0cf737566df7e4faa2618845 Original document size: 3.02 MB Authors: []	Submitter: Mickael Lallart Submission date: 10/4/2024 Upload type: interface analysis end date: 10/4/2024	Number of words: 48,016 Number of characters: 318,100
--	--	--

Location of similarities in the document:



Main sources detected






















No.	Description	Similarities	Locations	Additional information
1	Manuscrit_KHALFOUN.pdf Manuscrit_KHALFOUN #4879f1 The document is from my document database 94 similar sources	2%		Identical words: 2% (656 words)
2	TH1163_CORNEJO FUENTES Joaquin Eduardo_Manuscrit.pdf TH1163_CO... #cc8007 The document is from my document database 16 similar sources	< 1%		Identical words: < 1% (365 words)
3	TH1022_ELHARRAB Fatima_Manuscrit.pdf TH1022_ELHARRAB Fatima_M... #c2bad6 The document is from my document database 69 similar sources	< 1%		Identical words: < 1% (373 words)
4	inria.hal.science https://inria.hal.science/hal-03795818/file/MixNN_(49).pdf 4 similar sources	< 1%		Identical words: < 1% (441 words)
5	arxiv.org [2109.12550v1] MixNN: Protection of Federated Learning Against Infe... https://arxiv.org/abs/2109.12550v1#:~:text=In this paper, we present MixNN a proxy-based 5 similar sources	< 1%		Identical words: < 1% (351 words)

Sources with incidental similarities






No.	Description	Similarities	Locations	Additional information
1	www.mdpi.com https://www.mdpi.com/1424-8220/22/21/8254/pdf	< 1%		Identical words: < 1% (40 words)
2	liangli-zhen.github.io https://liangli-zhen.github.io/assets/pdf/GIRG.pdf	< 1%		Identical words: < 1% (39 words)
3	arxiv.org http://arxiv.org/pdf/2310.11739	< 1%		Identical words: < 1% (40 words)
4	dx.doi.org Going Haywire: False Friends in Federated Learning and How to Find T... http://dx.doi.org/10.1145/3579856.3595790	< 1%		Identical words: < 1% (37 words)
5	arxiv.org https://arxiv.org/pdf/1910.01991.pdf#:~:text=Felix Sattler, Klaus-Robert Müller*, Member, IEEE, and ...	< 1%		Identical words: < 1% (34 words)

Ignored sources These sources have been excluded by the document owner from the calculation of the similarity percentage.

No.	Description	Similarities	Locations	Additional information
1	hal.inria.fr https://hal.inria.fr/hal-03354724/file/MixNN_(19).pdf	13%		Identical words: 13% (6,532 words)
2	inria.hal.science https://inria.hal.science/hal-03354724/file/MixNN_(19).pdf	13%		Identical words: 13% (6,525 words)
3	arxiv.org http://arxiv.org/pdf/2109.12550	13%		Identical words: 13% (6,075 words)
4	arxiv.org https://arxiv.org/pdf/2109.12550v1	12%		Identical words: 12% (6,063 words)
5	arxiv.org https://arxiv.org/pdf/2109.12550v1#:~:text=In this paper, we present MixNN a proxy-based privacy-p...	12%		Identical words: 12% (5,626 words)
6	hal.science https://hal.science/hal-03354724/file/MixNN_(19).pdf	9%		Identical words: 9% (4,442 words)

No.	Description	Similarities	Locations	Additional information
7	 inria.hal.science https://inria.hal.science/hal-03795818/file/MixNN (49).pdf#:~:text=MixNN receives the model updat...	8%		🔗 Identical words: 8% (4,180 words)
8	 inria.hal.science https://inria.hal.science/hal-03795818v1/preview/MixNN (49).pdf#page=2	7%		🔗 Identical words: 7% (3,732 words)
9	 inria.hal.science https://inria.hal.science/hal-03354724/file/MixNN (19).pdf	7%		🔗 Identical words: 7% (3,415 words)
10	 inria.hal.science https://inria.hal.science/hal-03795818v1/file/MixNN (49).pdf#:~:text=In this paper, we present MixN...	6%		🔗 Identical words: 6% (3,149 words)
11	 inria.hal.science https://inria.hal.science/hal-03795818/file/MixNN (49).pdf#:~:text=proxy-based privacy-preserving s...	6%		🔗 Identical words: 6% (3,152 words)
12	 arxiv.org https://arxiv.org/pdf/2109.12550v1#:~:text=In this paper, we present MixNN a proxy-based	6%		🔗 Identical words: 6% (2,899 words)
13	 inria.hal.science https://inria.hal.science/hal-03354724/file/MixNN (19).pdf#:~:text=More precisely, to prevent traffic ...	5%		🔗 Identical words: 5% (2,593 words)
14	 inria.hal.science https://inria.hal.science/hal-03354724/document#:~:text=MixNN receives the model updates from p...	5%		🔗 Identical words: 5% (2,593 words)
15	 inria.hal.science https://inria.hal.science/hal-03354724/file/MixNN (19).pdf#:~:text=MixNN reçoit les mises à jour du ...	5%		🔗 Identical words: 5% (2,566 words)
16	 inria.hal.science https://inria.hal.science/hal-03354724/document#:~:text=MixNN receives the model updates from p...	5%		🔗 Identical words: 5% (2,559 words)
17	 inria.hal.science https://inria.hal.science/hal-03354724/document#:~:text=MixNN reçoit les mises à jour	5%		🔗 Identical words: 5% (2,559 words)
18	 arxiv.org https://arxiv.org/pdf/2109.12550v1#:~:text=We show that MixNN signifi-cantly limits the attribute	5%		🔗 Identical words: 5% (2,450 words)
19	 arxiv.org https://arxiv.org/pdf/2109.12550v1.pdf#:~:text=In this paper, we present MixNN a proxy-based	5%		🔗 Identical words: 5% (2,450 words)
20	 arxiv.org https://arxiv.org/pdf/2109.12550	5%		🔗 Identical words: 5% (2,450 words)

Referenced sources (without similarities detected) These sources were cited in the paper without finding any similarities.

-  <https://www.edchimie-lyon.fr>
-  <http://e2m2.universite-lyon.fr>
-  <http://ediss.universite-lyon.fr>
-  <http://ed34.universite-lyon.fr>
-  <https://edeea.universite-lyon.fr>



FOLIO ADMINISTRATIF

THESE DE L'INSA LYON, MEMBRE DE L'UNIVERSITE DE LYON

NOM : **LEBRUN**

DATE de SOUTENANCE : **05/12/2024**

Prénoms : **Thomas, Axel, Pierre**

TITRE : **Données de Santé : Exploration des Mécanismes Émergents de Protection de la Vie Privée**

NATURE : **Doctorat**

Numéro d'ordre : **2024ISAL0114**

École Doctorale : **INFORMATIQUE ET MATHÉMATIQUES**

Spécialité : **Informatique**

RÉSUMÉ :

Les données de santé représentent une grande quantité d'informations, générées quotidiennement et sensibles par nature. Cependant, leur partage est essentiel pour l'avancement de la recherche et, en fin de compte, l'amélioration des soins aux patients. L'utilisation des données médicales est confrontée à des limitations dues à leur sensibilité et à la nécessité de garantir la confidentialité, encadrée par les réglementations en vigueur. Cela nécessite une protection renforcée. L'intérêt pour des alternatives au partage de données brutes, telles que la pseudonymisation ou l'anonymisation, augmente avec les besoins d'accès à des données d'apprentissage pour l'utilisation de l'intelligence artificielle, qui requiert de grandes quantités de données pour fonctionner efficacement en tant qu'assistant médical. Dans cette thèse, nous examinons de nouveaux mécanismes respectant la vie privée, rendues possibles par les avancées rapides de l'intelligence artificielle. Plus spécifiquement, mon analyse porte sur l'amélioration d'alternatives à la centralisation de données sensibles : l'apprentissage fédéré, une méthode décentralisée d'entraînement des modèles d'Intelligence Artificielles qui ne nécessitent pas le partage de données, ainsi que de la génération de données synthétiques, qui crée des données artificielles avec des propriétés statistiques similaires aux données réelles. Considérant l'absence de consensus pour l'évaluation de la confidentialité de ces nouvelles approches, nous avons axé notre travail sur la mesure méthodique de la fuite de confidentialité ainsi que la balance avec l'utilité des données synthétiques ou du modèle d'apprentissage fédéré. Mes travaux incluent un mécanisme pour améliorer les propriétés de confidentialité de l'apprentissage fédéré ainsi qu'une nouvelle méthode de génération conditionnelle de données synthétiques. Cette thèse vise à contribuer au développement de cadres plus robustes pour le partage sécurisé des données de santé, en conformité avec les exigences réglementaires, facilitant ainsi des innovations en matière de santé.

MOTS-CLÉS : Données Synthétiques, Apprentissage Automatique, Apprentissage Fédéré, Confidentialité, Ré-identification, Attaques d'Attributs Sensibles, Attaques d'Appartenance, Enclaves Sécurisées, Données Personnelles, Données de Santé

Laboratoire(s) de recherche : **CITI**

Directeur de thèse : **Cunche Mathieu**

Président du Jury :

Composition du Jury :

Mme. BEN MOKHTAR Sonia, Mme. LESTYAN Szilvia, M. DECOUCHANT Jérémie, M. NGUYEN Benjamin et M. VINCENT Emmanuel