FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

AI-Enhanced Anonymization Workflows for Secure Health Data Sharing

Ricardo Almeida Cavalheiro

Mestrado em Engenharia Informática e Computação

Supervisor: Prof. João Correia Lopes

Co-supervisor: Eng. Gonçalo Gonçalves

July 30, 2025

AI-Enhanced Anonymization Workflows for Secure Health Data Sharing

Ricardo Almeida Cavalheiro

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Carla Teixeira Lopes
Referee: Prof. Benedita Malheiro
Referee: Prof. João Correia Lopes

July 30, 2025

Resumo

A crescente necessidade de partilha segura e ética de dados no setor da saúde destaca a importância de técnicas robustas de anonimização capazes de preservar a privacidade dos pacientes sem comprometer a utilidade dos dados. O processo tradicional de anonimização é frequentemente manual, propenso a erros e difícil de escalar, especialmente em conjuntos de dados estruturados e complexos. Este trabalho apresenta uma plataforma de anonimização melhorada com inteligência artificial, com o objetivo de apoiar os fornecedores de dados de saúde na partilha segura de dados sensíveis.

A solução proposta integra modelos de linguagem de grande escala, executados localmente, para automatizar etapas essenciais do processo de anonimização, incluindo a classificação de atributos, a geração de hierarquias de generalização e a configuração de modelos de privacidade. A arquitetura do sistema combina suporte à decisão baseado em IA com técnicas de anonimização consolidadas, através da integração com a ferramenta ARX.

A plataforma foi avaliada com recurso a conjuntos de dados clínicos, tanto sintéticos como reais. Os resultados experimentais mostram que a classificação de atributos baseada em modelos de linguagem de grande escala atinge uma precisão elevada e tempos de resposta reduzidos, oferecendo uma solução prática para reduzir o esforço manual e melhorar a eficiência e fiabilidade das tarefas de anonimização. Métricas de risco e utilidade foram utilizadas para avaliar a eficácia da anonimização, confirmando a capacidade da plataforma para equilibrar a proteção da privacidade com o valor analítico dos dados.

Este trabalho contribui com uma solução prática e orientada à privacidade para a partilha de dados de saúde, demonstrando o potencial da IA na melhoria do processo de anonimização.

Abstract

The growing demand for secure and ethical data sharing in healthcare highlights the importance of robust anonymization techniques that preserve patient privacy while maintaining data utility. Traditional anonymization workflows are often manual, error-prone and difficult to scale, especially in complex and structured datasets. This work presents an AI-enhanced anonymization platform designed to support healthcare data providers in securely sharing sensitive data.

The proposed solution integrates large language models deployed locally to automate key steps in the anonymization process, including attribute classification, generalization hierarchy generation and privacy model configuration. The system architecture combines AI-driven decision support with established anonymization techniques using the ARX tool.

The platform was evaluated using both synthetic and real-world clinical datasets. Experimental results show that LLM-based attribute classification achieves high accuracy and fast response time, offering a practical solution to reduce manual effort and improve the efficiency and reliability of anonymization tasks. Risk and utility metrics were used to assess anonymization effectiveness, confirming the platform's ability to balance privacy protection and analytical value.

This work contributes a practical, privacy-preserving solution for healthcare data sharing and demonstrates the potential of AI to enhance anonymization workflows.

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework for achieving a better and more sustainable future. The 17 goals address critical global challenges such as poverty, inequality, health, education and environmental protection.

This dissertation supports several SDGs by contributing to ethical and secure data sharing in healthcare through the use of artificial intelligence. The development of a privacy-preserving anonymization platform aligns with global efforts to strengthen health systems, foster innovation and promote responsible use of digital technologies.

The specific Sustainable Development Goals addressed in this work are:

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

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

SDG 16 Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels

SDG	**Target**	**Contribution**	**Performance Indicators and Metrics**
3	3.D	Strengthens data infrastructure in healthcare through secure, anonymized data sharing, enabling better disease research and patient care.	Adoption of anonymization tools in clinical data sharing; reduction in re-identification risk
9	9.5	Promotes innovation in healthcare data management by integrating AI into privacy-preserving workflows.	Integration of AI components; efficiency and accuracy in data classification
16	16.6	Supports transparent, ethical data governance aligned with privacy regulations (e.g., GDPR).	Compliance with legal standards; auditability of anonymization workflows

Acknowledgements

I would like to express my gratitude to everyone who supported me on this journey.

To my supervisors, Prof. João Correia Lopes and Gonçalo Gonçalves, thank you for your patience, for always being present and for your unwavering support throughout this journey. Your guidance was fundamental in helping me stay focused and motivated and your feedback consistently pushed me to improve.

A special thanks to Rui Jorge Ramos, with whom I had many enriching conversations. His deep understanding of the field greatly helped me develop the theoretical foundation necessary for this work. His insights were instrumental in shaping my understanding of the key concepts in data anonymization.

I am grateful to INESC TEC for giving me the opportunity to be part of this project and for providing an inspiring research environment. I also extend my thanks to WhyMob and Centro Hospitalar Universitário São João for their collaboration and for providing essential resources and domain expertise.

To all who walked beside me on this path: thank you.

Ricardo Almeida Cavalheiro

"Some men see things as they are and say, Why?
I dream of things that never were and say, Why not?"

George Bernard Shaw

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Motivation . . . 2
- 1.3 Problem . . . 3
- 1.4 Goals and Results . . . 3
- 1.5 Dissertation Structure . . . 3

2 Background and Related Work **5**
- 2.1 Data Anonymization . . . 5
- 2.2 Attribute Types . . . 6
  - 2.2.1 Identifier . . . 6
  - 2.2.2 Quasi-Identifier . . . 6
  - 2.2.3 Sensitive . . . 6
  - 2.2.4 Insensitive . . . 7
- 2.3 Equivalence Class . . . 7
  - 2.3.1 Definition . . . 7
- 2.4 Anonymization Techniques . . . 7
  - 2.4.1 Generalization . . . 8
  - 2.4.2 Data Masking . . . 8
  - 2.4.3 Perturbation . . . 8
- 2.5 Privacy Models . . . 9
  - 2.5.1 K-Anonymity . . . 9
  - 2.5.2 L-Diversity . . . 10
  - 2.5.3 T-Closeness . . . 11
  - 2.5.4 Differential Privacy . . . 11
- 2.6 Role of Artificial Intelligence . . . 12
- 2.7 Evaluating Anonymization: Risk and Utility Metrics . . . 12
  - 2.7.1 Risk Analysis . . . 12
  - 2.7.2 Utility Metrics . . . 14
  - 2.7.3 Balancing Privacy and Utility . . . 16
- 2.8 Conclusion . . . 17

3 State of the Art **19**
- 3.1 Artificial Intelligence in Data Anonymization . . . 19
  - 3.1.1 Sensitive Data Detection . . . 20
  - 3.1.2 Synthetic Data Generation . . . 20
  - 3.1.3 Hybrid and Context-Aware Approaches . . . 20
  - 3.1.4 Large Language Models for De-Identification . . . 20

3.1.5 Adversarial and Feedback-Guided Anonymization . . . 21
3.2 Applications in Healthcare . . . 21
3.2.1 Anonymization of Clinical Text . . . 22
3.2.2 Privacy in Medical Imaging . . . 22
3.2.3 Healthcare Data Platforms . . . 22
3.3 Impact and Challenges . . . 22
3.4 Conclusion . . . 22

4 Architecture and Design **24**
4.1 Solution Overview . . . 24
4.2 Requirements . . . 25
4.2.1 Data Owner . . . 25
4.2.2 User Stories . . . 25
4.3 Architecture . . . 26
4.3.1 System Overview . . . 27
4.3.2 Key Architectural Components . . . 27
4.4 Interaction Workflow . . . 29
4.5 Deployment . . . 31
4.6 Design Principles . . . 32
4.7 Contributions and Relevance . . . 33
4.7.1 Relevance to the Field . . . 33
4.7.2 Conclusion . . . 33

5 Implementation of the AI-Enhanced Anonymization Platform **34**
5.1 Exploration and Selection of Large Language Models . . . 34
5.1.1 Domain-Specific Transformer Models . . . 34
5.1.2 Generative Language Models and Local Deployment . . . 35
5.2 System Architecture and Technologies . . . 35
5.2.1 Backend . . . 35
5.2.2 AI Integration . . . 36
5.2.3 Frontend . . . 36
5.2.4 Data Storage . . . 36
5.2.5 Containerization . . . 36
5.2.6 API Documentation . . . 36
5.3 Implemented Features . . . 37
5.3.1 Generalization Hierarchies . . . 37
5.3.2 Privacy Models . . . 39
5.3.3 Anonymization Logging . . . 40
5.3.4 AI-Driven Assistance . . . 41
5.4 Comparison to Traditional Anonymization Workflows . . . 47
5.5 Conclusion . . . 48

6 Results and Evaluation **49**
6.1 Experimental Setup . . . 49
6.1.1 Dataset Description . . . 50
6.1.2 Labeling and Ground Truth . . . 50
6.1.3 Evaluation Criteria . . . 51
6.2 Evaluation of Attribute Classification Accuracy . . . 51
6.2.1 Overview . . . 51

6.2.2 LLM Selection . . . 51
6.2.3 Model Accuracy and Performance Comparison . . . 52
6.2.4 Discussion and Observations . . . 52
6.3 Case Study: Real-World Application on Clinical Data . . . 53
6.3.1 Case Study Objectives . . . 53
6.3.2 Dataset Description . . . 53
6.3.3 Workflow Execution . . . 54
6.3.4 API Validation and Usability . . . 54
6.3.5 Evaluation Metrics and Methodology . . . 54
6.3.6 Results and Observations . . . 55
6.3.7 Privacy–Utility Trade-off Analysis . . . 56
6.3.8 Discussion . . . 57
6.4 Limitations . . . 60
6.5 Conclusion . . . 60

7 Conclusion **62**
7.1 Summary of Contributions . . . 62
7.2 Key Findings . . . 63
7.3 Future Work . . . 63
7.4 Final Remarks . . . 63

References **64**

A API Documentation **71**
A.1 Anonymization Endpoints . . . 72
A.2 Hierarchy Generation Endpoints . . . 73
A.3 AI-Assisted Endpoints . . . 73
A.4 Logging Endpoints . . . 74
A.5 Schema Definitions . . . 74

List of Figures

4.1 UML component diagram 28
4.2 UML sequence diagram 29
4.3 Deployment architecture of the anonymization platform 32

5.1 Interval-based generalization hierarchy. 37
5.2 Masking-based generalization hierarchy. 38
5.3 Date generalization hierarchy. 39
5.4 Example of an anonymization log showing metadata. 40
5.5 Traditional anonymization workflow. 48

6.1 Prosecutor risk vs t (Fixed $k = 5$) 57
6.2 Information loss vs t (Fixed $k = 5$) 58
6.3 Suppressed records vs t (Fixed $k = 5$) 58
6.4 Average equivalence class size vs t (Fixed $k = 5$) 59

A.1 Endpoints for loading data, anonymizing tables and computing risk 72
A.2 Endpoints for generating interval, masking, date and categorical hierarchies . . . 73
A.3 Endpoints for attribute classification, hierarchy suggestion and privacy parameter tuning using LLMs 73
A.4 Endpoints for retrieving and deleting anonymization logs 74
A.5 List of reusable schema definitions in the OpenAPI specification 74

List of Tables

2.1 Original dataset 8
2.2 Dataset after generalization 8
2.3 Example of data masking 9
2.4 Original dataset 9
2.5 Dataset after applying perturbation 9
2.6 Original dataset 10
2.7 Dataset after achieving k-anonymity (k=3) 10
2.8 Original dataset with sensitive information 11
2.9 Dataset after achieving l-diversity (l=3) 11

4.1 Data owner's user stories 26

5.1 Categorical generalization hierarchy. 39

6.1 Columns from the synthetic healthcare dataset 50
6.2 Accuracy and inference time of DeepSeek models on attribute classification . . . 52
6.3 Re-identification risk for different anonymization configurations 55
6.4 Utility metrics for different anonymization configurations 56

List of Acronyms

AI	Artificial Intelligence
AECS	Average Equivalence Class Size
BERT	Bidirectional Encoder Representations from Transformers
CHUSJ	Centro Hospitalar Universitário São João
DM	Discernibility Metric
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
LLMs	Large Language Models
NER	Named Entity Recognition
NLP	Natural Language Processing
NCP	Normalized Certainty Penalty
OEC	Outlier Equivalence Class
QID	Quasi-identifiers
SA	Sensitive Attributes
UEC	Utility Equivalence Class

Chapter 1

Introduction

The secure sharing of healthcare data is both a technical and ethical challenge, especially in light of growing privacy regulations and the increasing volume of sensitive information. As medical systems evolve and become more data-driven, the need for robust, scalable and privacy-preserving anonymization solutions becomes critical.

This chapter defines the underlying aspects of this dissertation by outlining the context, explaining the motivation behind the research and identifying the problem it seeks to solve.

1.1 Context

The anonymization of health data is an essential step for preserving patient anonymity when dealing with relational health data [52]. Data owners such as hospitals, banks and insurance companies anonymize their users' data before making it available for secondary purposes such as medical research, clinical studies or data analysis [44]. This process enables external stakeholders, including researchers and public health institutions, to gain valuable insights while ensuring individual privacy is protected. However, anonymized data must still retain sufficient utility to support meaningful analysis and research.

With the digitalization of medical records and as the amount of information continues to grow, securing sensitive data and preserving confidentiality are of the utmost importance [65].

Although collaboration between researchers is far from straightforward, facilitating the sharing of health data is essential to support scientific progress. As they rely significantly on having and providing access to health datasets, it is important to have a secure and efficient way to share them.

By converting identifiable health data into de-identified or anonymized forms, data anonymization provides a solution by lowering the risk of re-identification while maintaining the data's analytical usefulness. This procedure to protect patient data privacy involves applying methods including data encryption, generalization and data masking [52] to mitigate the risk of disclosure.

Artificial Intelligence (AI) is emerging as a valuable tool in various domains and data privacy is no exception to the rule. The growing use of AI in the context of healthcare signals a new era in health data systems [41]. The use of innovative techniques has already shown us that the

improvements derived from AI can be game-changing in the field of health data [71]. From the detection of diseases to the prediction and prevention of new ones, AI has the potential to transform the industry [52].

While there are some challenges associated with the use of AI in healthcare, the opportunities that arise are relevant. It becomes possible to enhance existing strategies, providing an environment where data providers can make AI-assisted decisions regarding the anonymization process while considering the specific context and purpose of the data [36].

This work is developed in the scope of the Health from Portugal (HfPT) project, funded by the NextGenerationEU and the Portuguese Recovery and Resilience Plan programmes (Ref. C630926586-00465198), a strategic initiative designed to stimulate innovation and support the digital transformation of critical sectors in Portugal.

This project is the result of a collaborative effort between INESC TEC (*Institute for Systems and Computer Engineering, Technology and Science*) that contributes with its extensive research expertise in data privacy, artificial intelligence and system integration, WhyMob focusing on the design and implementation of web-based platforms that ensure accessibility and usability of the system and Centro Hospitalar Universitário São João (CHUSJ), a healthcare institution that provides invaluable domain expertise and access to real-world healthcare data.

1.2 Motivation

There is an increased need for information to be securely distributed among healthcare providers which motivated the development of new platforms with a focus on data anonymization [22]. As data in the world continues to grow and expand, researchers and professionals need access to vast amounts of data [65]. These data are essential for research and development purposes and will enable the discovery of new insights and knowledge.

However, the sharing of such information is often restricted due to ethical and legal requirements, such as the General Data Protection Regulation (GDPR) in the European Union [15] and the Health Insurance Portability and Accountability Act (HIPAA) in the United States [19]. In this context, it is essential to balance data privacy and utility in medical situations [52].

This research is driven by the need to develop a secure and efficient anonymization platform. The platform is designed for health data sharing purposes and is enhanced by machine learning techniques to improve the anonymization process efficiency and reliability. As the fast-growing advancements in AI continue to revolutionize the industry, it is essential to leverage these technologies. This project aims to ensure both data privacy and utility, enabling the secure and effective sharing of health data among healthcare providers.

This work is motivated by the challenge of reconciling privacy and usability and aims to deliver a practical, AI-enhanced platform to support secure data sharing in healthcare environments.

1.3 Problem

The problem this research addresses is the lack of effective integration of anonymization workflows in health data-sharing platforms. Existing platforms do not offer a comprehensive solution to the anonymization process, requiring researchers to manually perform these tasks, which are time-consuming and also prone to errors [77]. This can lead to either compromised data privacy or diminished data utility.

There is a growing demand for automation and integration of AI-driven tools to enable researchers to share anonymous data and interactively explore anonymization strategies [33].

Additionally, evaluating the risk profile of an approach and comparing it with other methods and parameters is vital, as we need to understand if the anonymization was properly executed and assess the risk of re-identification.

Currently, no integrated solution leverages AI to assist users in performing anonymization effectively. The absence of intuitive, guided tools leaves data providers with limited support when configuring anonymization parameters or assessing privacy risks.

1.4 Goals and Results

The primary goal of this work is to develop a modular and AI-enhanced anonymization platform that assists healthcare data providers in securely sharing sensitive patient data while preserving analytical value. The proposed system integrates large language models to automate attribute classification, hierarchy suggestion and privacy model configuration, thereby streamlining the anonymization process.

By leveraging local AI inference, the system ensures that data never leaves the secure environment, addressing privacy regulations such as GDPR and HIPAA. The platform should support multiple privacy models, allowing data anonymization experts and privacy professionals to configure and evaluate anonymization strategies in real-time. Through intuitive interfaces and risk-utility visualizations, the tool helps users make informed decisions without deep technical expertise.

The expected outcome is a practical and user-friendly tool that streamlines the anonymization workflow by automating complex decisions and providing contextual guidance. This integration of AI aims to reduce the technical burden on users, increase the consistency of the anonymization process and foster data sharing practices that are both secure and effective.

1.5 Dissertation Structure

Besides this Introduction chapter, this document is organized into six additional chapters:

- Chapter 2 – **Background and Related Work:** Presents the foundational concepts of data anonymization, privacy models and the role of AI in privacy-preserving data sharing.

- Chapter 3 – **State of the Art:** Reviews recent advances in AI-assisted anonymization, highlighting gaps in current solutions for structured healthcare data.
- Chapter 4 – **Architecture and Design:** Details the architecture, design philosophy and functional requirements of the anonymization platform.
- Chapter 5 – **AI-Enhanced Anonymization Platform:** Describes the technical development of the platform, including backend architecture, AI integration and implemented features.
- Chapter 6 – **Results and Evaluation:** Presents quantitative and qualitative assessments of the platform using both synthetic and real clinical datasets.
- Chapter 7 – **Conclusion:** Summarizes key findings and outlines future directions for research and system development.

Chapter 2

Background and Related Work

Data plays a crucial role in driving research and innovation in healthcare. As such, it is vital to develop platforms that ensure privacy while preserving data utility [71].

In the era of Big Data, healthcare systems are generating unprecedented amounts of information, which has been amply characterized by the 3 V's: volume (the sheer amount of data), variety (the diversity of data types) and velocity (the rate at which it is produced). This expansion raises the concern of safeguarding patient privacy [1, 72].

At the intersection of scientific advancement and privacy requirements due to ethical and legal concerns lies the challenge of protecting the sensitive information embedded in the data and still retaining enough information to enable meaningful analysis and effective knowledge extraction by data consumers.

This chapter aims to provide the foundational context and review of the related work necessary to understand the current state of data anonymization, particularly in the healthcare domain.

2.1 Data Anonymization

As the amount of digital data grows [73], the need to protect individuals' privacy is increasing, particularly in data-sharing scenarios [44, 52].

Data anonymization is an important step for protecting individuals' private data [52, 70]. It refers to the process of modifying personal sensitive information in such a way that the individuals cannot be re-identified [44]. Unlike pseudonymization, which is the simple replacement of identifiers with pseudonyms [76, 77], anonymization seeks to remove or alter identifiers to make re-identification computationally unfeasible.

With the rise of data-sharing initiatives and also strict privacy regulations such as GDPR [15, 26, 75] and the HIPAA, data anonymization plays a critical role in protecting data privacy as well as ensuring all ethical and legal requirements are met [48].

2.2 Attribute Types

The classification of attributes in a dataset is a foundational step in the anonymization process. It plays a critical role in determining how each attribute should be handled to protect privacy while preserving data utility. The success of any anonymization technique heavily depends on the accuracy and consistency of this initial classification.

Attributes are classified based on the sensitivity of the information they contain and their potential for being used to identify individuals [32]. This classification ensures that each attribute is appropriately handled based on its privacy risk and impact on the overall utility of the data.

The four types of attributes (identifiers, quasi-identifiers, sensitive and insensitive attributes) are described in the following sections.

2.2.1 Identifier

Identifiers are attributes that uniquely identify an individual. They constitute the biggest risk to the privacy of the data as they can be used to directly associate the data with a single record. Examples of identifiers include social security number, driver's license and phone number. Due to their ability for direct identification, these values must be completely removed during the anonymization process [32, 52, 61].

2.2.2 Quasi-Identifier

Quasi-identifiers (QIDs) are attributes that, by themselves, are not able to uniquely identify an individual but, when combined with other quasi-identifiers, may reveal someone's identity. Examples of quasi-identifiers include age, ZIP code and date of birth. These attributes may appear harmless by themselves, but their combination may be a threat to the privacy of the data, especially when the data is sparse [32, 52, 61].

Such risks are frequently exploited in *linkage attacks*, which are discussed in Section 2.7.1.2. These attacks highlight the importance of properly handling quasi-identifiers during anonymization to prevent re-identification through external data sources.

2.2.3 Sensitive

Sensitive attributes (SA) encode information that is considered confidential and private, such as diseases, disabilities and income. These attributes may not directly identify individuals, but when disclosed can lead to serious consequences such as discrimination or financial harm. For instance, revealing a medical condition can result in professional and insurance problems [32]. These are frequently protected using privacy models such as *l-diversity* (see Section 2.5.2) and *t-closeness* (see Section 2.5.3), which ensure that the data is not disclosed in a way that can be used to infer the sensitive information of an individual [32, 52, 61].

2.2.4 Insensitive

Insensitive attributes do not pose a risk to privacy, as they do not contain any information that can be used to identify an individual. Typically they do not need to be altered or anonymized in any way. Examples of insensitive attributes highly depend on the context of the data. In a healthcare dataset, general attributes including the type of hospital equipment used, the season during which a treatment occurred or the generic category of a medical procedure (e.g., diagnostic vs. surgical) may be deemed insensitive [32].

Understanding and categorizing the attributes in a dataset is a crucial step in the data anonymization process. This classification forms the foundation for applying privacy models and anonymization techniques that ensure the data is shared in a privacy-preserving way.

2.3 Equivalence Class

An **equivalence class** is a key concept in data anonymization, especially in privacy models such as k-anonymity, l-diversity and t-closeness. It denotes a collection of entries in a dataset that possess the same values for a defined set of quasi-identifiers (Section 2.2.2). These groups ensure that individual records cannot be distinguished from others within the same class, providing a level of privacy protection against re-identification attacks.

2.3.1 Definition

Given a dataset D and a set of quasi-identifiers Q, an equivalence class is a subset of D where all records have identical values for Q. Formally, if $R_1, R_2, \ldots, R_n$ are records in an equivalence class, then:

$$Q(R_1) = Q(R_2) = \ldots = Q(R_n)$$

where $Q(R)$ represents the projection of the record R on the quasi-identifiers.

2.4 Anonymization Techniques

Anonymization techniques are essential for minimizing the risk of re-identification while complying with data protection regulations. These techniques aim to modify the data in a way that maintains their utility while protecting the patients' privacy. The choice of techniques to be used relies on the dataset's structure, the privacy requirements and the intended use of the data.

The following subsections outline key anonymization techniques: generalization, data masking and perturbation.

2.4.1 Generalization

Generalization is a technique used to replace quasi-identifiers (see Section 2.2.2) with less specific values or broader categories. This technique is one of the most common ways of modifying data to protect privacy, as it reduces the granularity of the data while maintaining its utility.

Consider the dataset shown in Table 2.1 which contains the original values for the age and ZIP code. Both attributes were classified as quasi-identifiers. To reduce the risk of potentially identifying the individuals, the generalization method was applied on both attributes:

1. **Age:** The specific numerical values for age were substituted with an interval [20, 30[, containing all individuals in the dataset within one age interval. This transformation lowers the detail level, making it hard to identify a person's precise age while preserving valuable data for analysis (e.g., trends concerning those in a certain age interval).

2. **ZIP code:** The particular ZIP codes were replaced with larger geographic areas, such as "Aveiro", "Braga" and "Portimão". This guarantees that people are linked to a broader region instead of a particular street or neighborhood, further concealing their identity.

The generalized dataset is shown in Table 2.2.

2.4.2 Data Masking

Data masking involves replacing individuals' information with obfuscated, altered or masked data. Table 2.3 shows how masking can be applied to the attribute "ZIP code".

In this example, we can see that the last three digits of every ZIP code were replaced with a masking character (*), avoiding compromising security but still being able to retain some information about the individual's area.

2.4.3 Perturbation

Perturbation involves adding noise to the data while still maintaining the statistical properties of the original dataset, making it still useful for analysis purposes. Data perturbation is a common technique used in differential privacy, as explained in Section 2.5.4.

Consider the dataset shown in Table 2.4. In this example, the column 'Income' is a sensitive attribute. By adding noise to the income, we protect the individual's privacy while still preserving

Table 2.1: Original dataset

Age	ZIP code
21	3800-110
22	4700-201
23	2900-999
24	4700-721
25	3800-517
26	8500-411

Table 2.2: Dataset after generalization

Age	ZIP code
[20, 30[	Aveiro
[20, 30[	Braga
[20, 30[	Setúbal
[20, 30[	Braga
[20, 30[	Aveiro
[20, 30[	Portimão

Table 2.3: Example of data masking

ZIP code	Masked ZIP code
3800-110	3800-***
4700-201	4700-***
2900-999	2900-***
4700-721	4700-***
3800-517	3800-***
8500-411	8500-***

the general properties of the data as shown in Table 2.5. Common statistical properties preserved through perturbation include metrics such as the *mean*, *variance*, *standard deviation* and *correlation coefficients*.

Perturbation can be effective when we have numerical information. However, we must be cautious to achieve the right balance between data privacy and utility.

2.5 Privacy Models

Anonymization models are structured approaches for achieving anonymized data, each offering a different approach to minimizing the risk of re-identification [77]. This section presents the most relevant models from the literature [16, 44, 52, 75].

2.5.1 *K*-Anonymity

Samarati and Sweeney [59, 64] first presented the notion of *k-anonymity*. This privacy model ensures that each individual in a dataset cannot be distinguished from at least $k-1$ others based on the quasi-identifiers. To achieve k-anonymity, techniques such as *generalization* and *data masking* are applied. A dataset satisfies k-anonymity if, for every combination of quasi-identifiers, there are at least k identical records in the dataset.

Table 2.6 illustrates an example of a dataset with three quasi-identifiers: age, gender and ZIP code.

In its original form, the dataset reveals sensitive information, making it vulnerable to linkage attacks. To achieve k-anonymity with $k = 3$, the dataset was transformed so that each individual record is indistinguishable from at least two others based on the quasi-identifiers. The attributes *age* and *ZIP code* were generalized, ages were grouped into the interval $[20, 30[$ and ZIP codes

Table 2.4: Original dataset

id	Income (€)
1	10 000
2	25 000
3	35 000
4	60 000

Table 2.5: Dataset after applying perturbation

id	Income (€)
1	10 300
2	24 700
3	35 200
4	59 800

Table 2.6: Original dataset

age	gender	ZIP code
21	male	416-120
22	female	416-223
23	male	416-999
24	female	416-721
25	male	416-517
26	female	416-411

were partially masked. This transformation resulted in the creation of two equivalence classes, each containing three records that share identical values for the quasi-identifiers. These equivalence classes, as explained in Section 2.3, are a key concept in enforcing *k*-anonymity, as they ensure that individual records cannot be distinguished from others within the same group. The result of this transformation is shown in Table 2.7.

2.5.2 *L*-Diversity

Machanavajjhala *et al.* [43] introduced an enhancement to *k*-anonymity in 2006. While *k-anonimity* protects against re-identification, it does not prevent *attribute disclosure*. For instance, if all records in an equivalence class (see Section 2.3) share the same value for a sensitive attribute (e.g., disease), an adversary can still infer that value. *l-diversity* is an extension of *k*-anonymity that enforces diversity in sensitive attributes. This privacy model ensures that each equivalence class has at least *L distinct values* for the sensitive attribute. This additional constraint ensures diversity in sensitive attributes, reducing the risk of inference attacks. Table 2.8 presents the same dataset as Table 2.6, but the *disease* column was added as a sensitive attribute.

In its original form, the dataset remains vulnerable to both re-identification and attribute disclosure risks. To mitigate this, we apply both *k*-anonymity with $k = 3$ and *l*-diversity with $l = 3$.

The transformed dataset is shown in Table 2.9. The sensitive attribute (*disease*) has been adjusted to meet *l*=3-diversity. This means that each equivalence class contains at least three distinct values for the sensitive attribute. However, it is important to note that the dataset did not have sufficient records or diversity in the sensitive attribute to fully satisfy both 3-anonymity and

Table 2.7: Dataset after achieving *k*-anonymity (*k*=3)

age	gender	ZIP code
[20,30[	male	416-***
[20,30[	male	416-***
[20,30[	male	416-***
[20,30[	female	416-***
[20,30[	female	416-***
[20,30[	female	416-***

Table 2.8: Original dataset with sensitive information

age	gender	ZIP code	disease
21	male	416-120	flu
22	female	416-223	asthma
23	male	416-999	diabetes
24	female	416-721	heart disease
25	male	416-517	flu
26	female	416-411	diabetes

3-diversity without data loss. As a result, some records were suppressed to ensure compliance with both privacy models.

2.5.3 T-Closeness

Li *et al.* [40] proposed t-Closeness to address l-diversity limitations. While l-diversity ensures the sensitive attribute within each equivalence class is diverse, it does not account for the distribution of this attribute. t-Closeness requires that the distance between the distribution of the sensitive attribute in each equivalence class and the distribution of the sensitive attribute in the overall dataset is less than a threshold t. The smaller the value of t, the more similar the distributions must be.

For instance, suppose we have a dataset where 40 % of the individuals are diabetic, 10 % have asthma and 30 % have a heart condition, each equivalence class after the anonymization process should have a disease distribution that is closely aligned to these proportions (depending on the value of t).

2.5.4 Differential Privacy

Differential privacy was introduced by Cynthia Dwork [18] as a framework for protecting private data, while still allowing the data to be shared and properly analyzed. The key property of differential privacy is that the inclusion or exclusion of a single record from the dataset does not alter the output of the analysis. This guarantee is achieved through randomized algorithms that introduce carefully calibrated noise. The degree of noise is controlled by a parameter ε (epsilon). A smaller

Table 2.9: Dataset after achieving l-diversity (l=3)

age	gender	ZIP code	disease
*	*	*	flu
[20,30[	female	416-***	asthma
*	*	*	diabetes
[20,30[	female	416-***	heart disease
*	*	*	flu
[20,30[	female	416-***	diabetes

value of ε corresponds to stronger privacy because it ensures that the outputs of the algorithm are nearly indistinguishable regardless of whether any single record is included. On the contrary, a larger ε allows for higher data utility but provides weaker privacy guarantees.

Cynthia Dwork laid the foundational groundwork for the concept of differential privacy in numerous domains. Building upon this approach, Bild et al [9] addressed the challenges of truthful data anonymization. Instead of using a perturbative approach, which generates synthetic data, Bild focused on maintaining the integrity of the data by applying generalization and sampling techniques. This method is notably used in anonymization processes, as it achieves strong privacy guarantees while maintaining high data utility when the noise is chosen thoughtfully.

2.6 Role of Artificial Intelligence

The era of big data has made privacy and confidentiality extremely difficult to maintain, as sensitive information is openly and easily available. In the healthcare sector, patient-related data contains fundamentally sensitive information, including medical history, diagnosis reports and genetic information. Therefore, this is an extremely important issue. Maintaining confidentiality of such data is a technological obligation and also a moral and legal one. AI has become an integral tool in these problems by putting forward creative solutions for sensitive data identification, categorization and anonymization [4, 33, 71].

AI is important in the anonymization process for striking a balance between utility and the need for privacy. For instance, healthcare research relies on patient data to find trends to improve diagnostics and come up with new treatments. However, if such data are used or shared without proper anonymization, it might breach regulations such as the HIPAA or the GDPR [15, 26, 75]. AI-driven anonymization methods, when executed in secure and local environments, ensure that the data maintains its value for research while protecting the identities of patients [4]. This paves the way for privacy-preserving analytics and fosters trust among stakeholders.

2.7 Evaluating Anonymization: Risk and Utility Metrics

An effective anonymization process must balance privacy protection with the preservation of data utility. Evaluating both aspects is critical to ensure that data releases are not only legally compliant but also practically useful for analysis, research or AI model development. This section presents established methods for assessing anonymized datasets, focusing on quantifying re-identification risks and measuring analytical utility.

2.7.1 Risk Analysis

Risk analysis refers to the systematic estimation of the likelihood that an individual can be re-identified from an anonymized dataset. A key concept here is that re-identification is rarely random. Adversaries typically operate with structured strategies and external auxiliary data. As such,

modern risk assessments simulate real-world threats through adversary models and statistical metrics.

2.7.1.1 Re-identification Risk

Re-identification risk quantifies the probability that an anonymized record can be matched to an individual. This is not simply about guessing names, it involves combining quasi-identifiers (e.g., age, ZIP code, gender) with external datasets such as voter rolls, public registries or insurance claims.

To simulate real-world threats, Scaiano et al. [60] introduced a set of adversary models:

- The **Prosecutor model** assumes that the adversary knows that a specific individual is included in the dataset. The goal is to re-identify that person by matching quasi-identifiers or other attributes. This is a high-risk scenario, particularly relevant in cases where the attacker has auxiliary information about the target.
- The **Journalist model** represents an attacker who is unsure whether a specific person is present in the dataset. The attacker scans the data for potentially identifiable records, looking for unique or suspicious combinations that may reveal someone's identity. This model reflects investigative scenarios, such as whistleblowing or public record analysis.
- The **Marketer model** simulates an attacker who is not targeting any specific individual but instead tries to re-identify as many records as possible. The goal is to extract personal information at scale, often for commercial purposes such as targeted advertising or profiling.

These models shift the evaluation from theoretical guarantees to scenario-based simulations. Scaiano et al. also criticized over-reliance on standard metrics such as micro-average recall, noting that missing a single sensitive term in clinical text may still lead to complete re-identification. They proposed the more realistic *"all-or-nothing" recall*, where the failure to remove any identifying element from a document results in a 100 % re-identification score for that record.

This probabilistic understanding was extended by Ratra et al. [57], where they evaluated healthcare datasets anonymized using a hybrid of k-anonymity and differential privacy (k-ADP). They found that combining suppression/generalization with noise injection significantly reduced risk: prosecutor-model risk dropped from 100 % to under 0.2 %, showing that multiple techniques in concert are more robust than any single method.

2.7.1.2 Linkage Attacks

Linkage attacks are one of the most common and potent forms of re-identification. They exploit combinations of quasi-identifiers to match anonymized data with identifiable external datasets. For example, knowing a person's ZIP code, birth date and gender can often uniquely identify them in the population.

The canonical example is the re-identification of the Governor of Massachusetts from anonymized hospital discharge records by linking them with voter registration data [21]. Despite removing names and Social Security numbers, records were successfully matched based on demographic quasi-identifiers.

El Emam et al. [21] reviewed numerous such attacks and concluded that the majority relied on publicly available auxiliary datasets. The implication is clear: anonymization must account not only for the internal structure of a dataset but also for what is publicly knowable about the population.

Lapwattanaworakul et al. demonstrated that even minimal generalization of attributes namely the age and marital status on the dataset could leave records vulnerable to linkage. Using tools such as ARX[1] [38, 55], they showed that unless quasi-identifiers are generalized sufficiently, they may still uniquely point to individuals when cross-referenced with external sources.

2.7.1.3 Population Uniqueness

Most anonymization strategies, such as k-anonymity, operate on the assumption that indistinguishability within a dataset implies privacy. However, this can be misleading. If a record is unique not just in the dataset but also in the larger population, it remains re-identifiable regardless of how well it blends internally.

This concept is captured by **population uniqueness**, which estimates how frequently a combination of quasi-identifiers occurs in the general population. Bandara et al. [7] emphasize that attackers often have access to population-level statistics and they introduce metrics such as the *Special Uniqueness Detection Algorithm (SUDA)* to quantify this risk.

SUDA assigns scores to records based on the estimated uniqueness of their quasi-identifier combinations within the external population. Even if a record is not unique in the dataset, it might correspond to a rare demographic slice (e.g., a 97-year-old neurosurgeon in a rural ZIP code), making it vulnerable to re-identification.

Their findings show that re-identification risk remains non-negligible after applying k-anonymity or even l-diversity if population uniqueness is not considered. This highlights the need to augment anonymization strategies with external risk models and population-informed metrics.

2.7.2 Utility Metrics

While privacy is the primary concern of anonymization, preserving utility is equally important. Utility metrics provide quantitative measures of how well anonymized data retains its original structure, semantics and analytical properties.

2.7.2.1 Information Loss

Information loss metrics assess how much detail or resolution has been removed due to generalization and suppression:

[1] https://arx.deidentifier.org/

- **Discernibility Metric (DM)** quantifies information loss by assigning a penalty to each record based on the size of its equivalence class. Larger classes receive higher penalties, reflecting reduced specificity and utility.
- **Average Equivalence Class Size (AECS)** calculates the mean number of records within each equivalence class. While a higher AECS typically indicates stronger privacy, it may also signal a greater loss of data granularity.
- **Normalized Certainty Penalty (NCP)** measures the degree of generalization applied to each attribute. It provides a detailed, attribute-level assessment of how much information was lost during anonymization.

Vural and Aydos proposed distinguishing between Utility Equivalence Classes (UEC) and Outlier Equivalence Classes (OEC) [69]. UECs retain analytical value and can be preserved with minimal generalization. OECs, on the other hand, often represent extreme or rare records that require heavy suppression. Separating these allows for targeted anonymization strategies that preserve the utility of the majority while protecting the few.

2.7.2.2 Data Quality Metrics

These metrics compare statistical features between the original and anonymized datasets:

- **Attribute-level distortion** counts how often and to what extent values are changed or masked.
- **Distribution similarity metrics**, such as Kullback-Leibler (KL) divergence or Earth Mover's Distance (EMD), evaluate whether attribute distributions (e.g., income levels, age) have been preserved.
- **Model fidelity** assesses how well predictive models trained on anonymized data replicate the performance of those trained on the original. Common measures include accuracy, precision and recall.

2.7.2.3 Task-specific Metrics

When data is intended for a known task (e.g., diagnosis prediction, customer segmentation), utility should be evaluated in that context:

- **Classification tasks**: Common metrics include *accuracy*, which measures the proportion of correctly predicted instances; *precision*, which evaluates the proportion of true positives among predicted positives; *recall*, which assesses the proportion of true positives among actual positives; and the *F1-score*, which is the harmonic mean of precision and recall.

- **Regression tasks**: These are typically evaluated using the *Mean Squared Error*, which calculates the average squared difference between predicted and actual values and the *R-squared* score, which indicates the proportion of variance in the dependent variable explained by the model.
- **Clustering tasks**: For unsupervised learning, metrics such as the *Silhouette Coefficient*, which measures how similar a data point is to its own cluster compared to others and the *Adjusted Rand Index (ARI)*, which quantifies the similarity between predicted and true cluster assignments, are commonly used.

2.7.3 Balancing Privacy and Utility

The balance between privacy and utility lies at the heart of all anonymization efforts. Anonymized data must be private enough to protect individuals from re-identification, yet still useful for meaningful analysis. This trade-off is not merely technical, shaping whether data can be ethically and legally shared while still serving scientific, operational or public policy goals.

Why the Trade-Off Exists

Anonymization typically involves modifying data to obscure individual identities. Common techniques include generalization, suppression and noise addition. These techniques are described in detail in Section 2.4 and are critical components of modern anonymization frameworks. Each of these improves privacy by reducing uniqueness, but at a cost: they distort the data and reduce its analytical accuracy.

For example, increasing the parameter k in k-anonymity reduces re-identification risk by ensuring that each record is indistinguishable from at least $k-1$ others. However, doing so often requires heavy generalization, which flattens important distinctions in the data. A $k = 10$ anonymized health dataset may no longer allow researchers to differentiate between patients aged 45 and 65, potentially invalidating age-sensitive studies.

Similarly, differential privacy offers strong mathematical guarantees of privacy by injecting calibrated noise into query results. But with low ε values, the noise may overwhelm weak data signals, impairing tasks such as disease prediction or patient stratification.

Why It Matters in Practice

This tension is especially critical in high-stakes domains such as healthcare, where datasets must protect patient confidentiality but also support evidence-based decisions. Overemphasis on privacy may lead to "safe but useless" data, while insufficient privacy undermines trust and legal compliance.

As highlighted in Ratra et al. [57], hybrid approaches such as combining k-anonymity with differential privacy can mitigate this tension. These methods allow fine-tuned control, adjusting anonymization levels to suit the risk tolerance and utility requirements of specific use cases.

Vural and Aydos [69] further show that separating Utility Equivalence Classes (UEC) from Outlier Equivalence Classes (OEC) helps focus protection efforts where risk is highest, preserving utility elsewhere.

Navigating the Trade-Off

Effective anonymization requires explicit strategies for managing this trade-off:

- **Parameter tuning**: Practitioners can iteratively adjust privacy parameters (k, ε, t) and immediately observe their impact on risk and utility metrics. Tools such as ARX enable this interactively, visualizing suppression levels, re-identification risk and equivalence class size [38].
- **Multi-objective optimization**: Some frameworks use optimization algorithms to find the best solutions that maximize utility and minimize risk. These methods aim to automatically identify trade-offs between conflicting objectives, such as preserving data quality while reducing the likelihood of re-identification, without requiring manual tuning [13].
- **Use-case awareness**: Data should be anonymized in context. For exploratory research, a lower privacy threshold may be acceptable. In contrast, datasets intended for public release demand significantly stronger privacy guarantees. Tailoring anonymization strategies to specific use cases helps avoid both over-protection, which can degrade data utility and under-protection, which can expose individuals to re-identification risk.
- **Iterative evaluation**: Rather than anonymizing once, practitioners should evaluate and refine anonymization in cycles, guided by both quantitative metrics and expert judgment.

The privacy-utility trade-off is not a flaw in anonymization, it is its defining challenge. It demands careful, context-aware design choices grounded in both legal mandates and analytical goals. Ignoring either side of the equation risks producing datasets that are either dangerously revealing or functionally worthless. Striking the right balance is the mark of a mature anonymization strategy.

2.8 Conclusion

This chapter established the conceptual foundation necessary to understand data anonymization within the healthcare context. It introduced all the key concepts which are essential to grasp the theoretical principles that underpin anonymization strategies.

By reviewing related work and practical challenges, especially in healthcare, the chapter highlighted why anonymization is both a legal obligation and a technical challenge. The role of artificial intelligence was also discussed as a catalyst for improving the detection and anonymization of sensitive information at scale. Given the wide range of anonymization techniques, the complexity

of parameter tuning and the inherent trade-offs between privacy and utility, AI can play a crucial role in supporting data providers by suggesting appropriate strategies and helping configure privacy models. This assistance reduces the technical burden and improves both consistency and effectiveness in the anonymization process.

Finally, the discussion on evaluating anonymization through risk and utility metrics reinforces a central theme of this dissertation: achieving a practical and intelligent balance between protecting individual privacy and preserving data utility. This background informs the design choices and evaluation framework proposed in the following chapters.

Chapter 3

State of the Art

The growing complexity and scale of healthcare datasets has increased the importance of anonymization in preserving patient privacy while enabling the utility of data for research and innovation.

Traditional anonymization techniques such as generalization, suppression and masking have formed the foundation of privacy-preserving data practices. However, due to the complexity, diversity and volume of healthcare data, traditional approaches often struggle to meet modern demands. AI emerged as a solution to this challenge, providing scalable and effective methods that address the current limitations of traditional anonymization techniques.

This chapter reviews the recent progress, findings and applications of anonymization in the healthcare industry, focusing on the integration of AI.

3.1 Artificial Intelligence in Data Anonymization

The exponential growth of healthcare data has made it essential to create advanced anonymization techniques to tackle privacy issues while preserving data usefulness. GDPR and HIPAA impose stringent guidelines for managing sensitive patient information [61, 74]. These regulations have led to the use of innovative AI-driven anonymization methods, which provide significant benefits over conventional techniques.

AI's role in anonymization extends beyond de-identification. It can adaptively modify strategies depending on the characteristics of the dataset and the specific use cases [8, 23]. Unlike traditional methods that often sacrifice usability, AI-based solutions achieve a balance between privacy and functionality through approaches such as Natural Language Processing (NLP) [31], generative models [74] and hybrid methods involving large language models [63].

The incorporation of AI into data anonymization has revolutionized the management of sensitive healthcare data, enabling secure data sharing for research and innovation while addressing privacy issues. The subsequent sections explore AI techniques and their applications, emphasizing both progress and the challenges that remain.

3.1.1 Sensitive Data Detection

Sensitive data detection is the cornerstone of anonymization. Modern AI approaches, especially in NLP, have made significant strides in identifying sensitive entities such as names, addresses and medical conditions. Named Entity Recognition (NER) models classify text into predefined categories and have become essential in structured and unstructured datasets.

Jin et al. [30] introduced a hybrid framework that merges rule-based systems with deep learning for de-identifying unstructured clinical texts in Electronic Medical Records, demonstrating AI's adaptability in managing intricate, multilingual datasets.

Kuzina et al. [35] emphasized the utility of transformer-based models, such as Bidirecional Encoder Representations from Transformers (BERT), in identifying sensitive entities. These models excel in capturing contextual nuances, offering higher accuracy and scalability compared to older methods. Additionally, commercial tools such as Google Cloud DLP[1] and Gretel AI[2] employ hybrid techniques to detect sensitive information across formats, such as text, images and audio [8].

3.1.2 Synthetic Data Generation

Synthetic data generation addresses the limitations of traditional anonymization methods by creating data that mirrors real-world datasets without exposing sensitive information. Generative Adversarial Networks (GANs) are at the forefront of this innovation.

For instance, Yoon et al. [74] presented ADS-GAN, a Generative Adversarial Network designed to produce synthetic datasets that maintain statistical characteristics while reducing privacy risks. Jadon and Kumar [29] extended the application of synthetic data for collaborative research. Their work balances privacy with research requirements, enabling secure multi-institutional collaborations.

3.1.3 Hybrid and Context-Aware Approaches

Hybrid approaches combine rule-based systems with AI models to optimize anonymization. Rule-based methods efficiently handle structured data, such as social security numbers, while AI models excel with unstructured formats [8, 23].

Context-aware techniques leveraging transformers such as BERT further enhance performance in sensitive data detection. Kuzina et al. [35] highlighted how these models adapt to diverse data environments, improving both scalability and accuracy.

3.1.4 Large Language Models for De-Identification

The emergence of Large Language Models (LLMs) has reshaped the landscape of text anonymization, particularly for unstructured healthcare data. Unlike traditional machine learning models,

[1] https://cloud.google.com/dlp
[2] https://gretel.ai

LLMs such as GPT-4, LLaMA and ClinicalBERT are capable of identifying complex patterns and sensitive entities through contextual understanding and in-context learning.

Recent studies have shown that LLMs can achieve de-identification accuracy above 98 % without task-specific fine-tuning. Liu et al. [42] proposed DeID-GPT, a prompt-engineered GPT-4 system that removes personally identifiable information in clinical notes using zero-shot learning, matching regulatory frameworks.

Lee et al. [39] evaluated the application of LLMs for de-identifying Chinese-English code-mixed clinical notes. They demonstrated the importance of prompt engineering and found that code-mixed training instances significantly improved the ability of LLMs to detect sensitive information.

Singh et al. [62] extended this work in multilingual healthcare settings. They trained PI-RoBERTa on synthetic summaries generated from LLMs (LLaMA-3, Gemma and Mistral), improving generalization in Indian clinical data with limited annotated resources.

Pissarra et al. [54] proposed new evaluation metrics to assess both privacy and utility in anonymized texts. Their experiments showed that open-source LLMs outperformed rule-based systems such as Presidio[3], maintaining clinical meaning while anonymizing sensitive content.

This shift toward promptable, adaptable and locally deployable LLMs offers new pathways to privacy-preserving health data sharing while avoiding cloud-based inference risks.

3.1.5 Adversarial and Feedback-Guided Anonymization

As the capabilities of LLMs increase, so do the risks of re-identification through indirect inference. Staab et al. [63] addressed this by introducing a novel adversarial anonymization framework. In this approach, one LLM acts as an adversary to infer attributes (e.g., age, location, gender), while another LLM iteratively anonymizes text based on feedback. This feedback-guided loop ensures that quasi-identifiers and subtle cues are removed and not just explicit entities.

Their experiments showed that adversarial LLM anonymization outperforms commercial tools such as Microsoft Azure Presidio in both utility and privacy. Moreover, even small models such as Mixtral or LLaMA-3 deployed locally can achieve strong anonymization results, providing a scalable, privacy-preserving alternative to cloud APIs.

This new class of anonymization systems aligns with GDPR's requirement that no personal attributes can be "reasonably inferred", marking a paradigm shift from detection-based to inference-aware anonymization.

3.2 Applications in Healthcare

AI-driven anonymization has transformed healthcare data management, enabling secure sharing of structured datasets, clinical text and medical imaging. Below are highlighted key applications.

[3] https://microsoft.github.io/presidio/

3.2.1 Anonymization of Clinical Text

Clinical text, such as Electronic Medical Records and clinical notes, often contains sensitive information requiring advanced anonymization. Jin et al. [30] demonstrated how hybrid models improve accuracy in multilingual datasets. Transformer-based NER models [35] further streamline text anonymization, ensuring regulatory compliance and data utility.

3.2.2 Privacy in Medical Imaging

Medical imaging files pose unique challenges due to embedded metadata. Hafsa and Sevrani [24] reviewed AI techniques for anonymizing medical images while retaining diagnostic value. Synthetic image generation represents another promising direction, enabling privacy-preserving datasets for research [74].

3.2.3 Healthcare Data Platforms

Frameworks such as FHIR-DIET[4] integrate AI to anonymize healthcare datasets compliant with standards such as FHIR [27]. Raso et al. [56] showcased configurable pipelines that balance privacy and utility, facilitating secure data sharing. Test-Driven Anonymization (TDA), proposed by Augusto et al. [5] guarantees data privacy while retaining usefulness by continuously assessing the anonymized data's compatibility for AI applications. These examples underscore how AI-driven techniques are transforming data anonymization practices within healthcare.

3.3 Impact and Challenges

AI-driven anonymization techniques have revolutionized healthcare data sharing by providing scalable, robust solutions that address privacy concerns. However, challenges remain, including:

- **Scalability:** Adapting methods to large and diverse datasets.
- **Bias:** Mitigating biases introduced by AI models.
- **Standardization:** Developing unified frameworks for anonymization.

Future research should focus on integrating explainable AI techniques, improving scalability and addressing ethical concerns to ensure the continued success of AI in data anonymization.

3.4 Conclusion

The landscape of data anonymization in healthcare has evolved significantly, driven by the increasing complexity of data and the emergence of AI technologies. Traditional approaches, while still relevant in specific contexts, are increasingly being supplemented or replaced by AI-driven

[4] https://www.hosmartai.eu/fhir-diet/3256/

methods that offer greater adaptability, context awareness and scalability. In particular, LLMs have revolutionized de-identification of unstructured clinical text, enabling high accuracy through zero-shot learning, prompt-based strategies and adversarial anonymization techniques.

However, a critical gap remains: the application of LLMs to structured healthcare datasets has received comparatively little attention. While unstructured data has benefitted from sophisticated NLP and NER techniques, structured formats such as tabular electronic health records still rely heavily on rule-based systems or manually configured pipelines. This lack of solutions for structured data presents a clear limitation in the current state of the art.

This work aims to address this gap by integrating LLMs into the anonymization process for structured datasets. Through attribute classification and strategy recommendation powered by locally deployed language models, the current work bridges the strengths of LLMs with the practical needs of structured data anonymization in healthcare settings. This novel approach extends the reach of modern AI beyond free-text processing, offering a path forward for more comprehensive, intelligent anonymization workflows.

Chapter 4

Architecture and Design

The proposed solution aims to address the challenges associated with securely sharing anonymized health data while preserving its utility for research and analysis. This approach integrates AI into an anonymization workflow tailored for health data-sharing environments, enabling data providers to achieve optimal anonymization strategies efficiently and effectively. The following sections detail the system's design, architecture and key components supporting this approach.

4.1 Solution Overview

The key features of the proposed solution include:

1. **AI-Driven Anonymization Workflow:**
    - Integrates locally deployed LLMs to assist data providers in automatically classifying dataset attributes (e.g., identifiers, quasi-identifiers, sensitive attributes).
    - Provides recommendations for privacy models and generalization strategies based on the semantic context of the dataset, enhancing both the accuracy and efficiency of the anonymization process.
    - Ensures that all AI inference is performed locally, preserving data privacy and complying with regulatory frameworks such as GDPR and HIPAA.
    - Enables dynamic decision-making and real-time evaluation of anonymization strategies through semantic understanding of the data.
2. **Intuitive Interface:**
    - Provides intuitive interfaces that allow users to interactively explore anonymization strategies and their effects on data utility and privacy.
    - Offers guidance to data providers through visualizations and actionable insights, including AI-suggested classifications and risk reports.
3. **Modular Architecture:**

- Comprises well-defined components for data anonymization, risk analysis, AI-driven classification and hierarchy creation.
- Ensures scalability, adaptability and seamless integration into existing health data-sharing environments through containerized services.

4. **Privacy Requirements:**
   - Includes workflows for risk assessment to ensure adherence to regulations such as GDPR.
   - Facilitates auditing and traceability of anonymization processes, including AI-driven decisions.
5. **Real-World Applicability:**
   - Designed to address practical constraints and requirements identified in collaboration with healthcare stakeholders.
   - Ensures the solution is robust, efficient and applicable to diverse healthcare scenarios while preserving analytical value.

4.2 Requirements

This section defines the system requirements focusing on the primary actor involved in the proposed solution and presenting user stories that describe the desired functionalities from that actor's perspective. The main actor in the system is the *Data Owner*.

4.2.1 Data Owner

The **Data Owner** represents a healthcare institution or organization responsible for managing sensitive patient datasets. This actor is tasked with uploading data, classifying attributes, configuring anonymization settings and overseeing the entire anonymization process. The Data Owner ensures that data privacy is maintained and that all procedures comply with applicable legal and regulatory frameworks, such as GDPR and HIPAA.

4.2.2 User Stories

The following user stories outline the desired system functionalities from the perspective of the Data Owner. Each user story is identified by a unique ID, a descriptive name, a priority level and a detailed description.

Table 4.1: Data owner's user stories

ID	Name	Priority	Description
US01	Upload Dataset	High	As a *Data Owner*, I want to upload sensitive healthcare datasets so that I can prepare them for anonymization and secure sharing.
US02	Classify Attributes	High	As a *Data Owner*, I want to classify data attributes manually or with AI assistance to guide the application of appropriate anonymization techniques.
US03	Generate Hierarchies	High	As a *Data Owner*, I want to be able to create or validate AI-generated hierarchies to support generalization techniques that will be used in the anonymization process.
US04	Configure Anonymization	High	As a *Data Owner*, I want to select privacy models and configure anonymization settings, optionally guided by AI-generated recommendations.
US05	Review Risk Analysis	Medium	As a *Data Owner*, I want to review risk analysis reports so that I can assess the effectiveness of the anonymization process.
US06	Maintain Anonymization Logs	Medium	As a *Data Owner*, I want the system to record a detailed log of all anonymization operations so that I can review and verify changes made to the data for transparency, auditing and compliance purposes.

The user stories provided for the Data Owner capture the functional requirements necessary for the successful operation of the proposed system. These requirements ensure that the Data Owner can effectively manage and anonymize sensitive datasets. By addressing the needs of the Data Owner, the system aims to provide a secure, privacy-compliant and user-friendly environment for healthcare data sharing.

4.3 Architecture

This section presents the system architecture of the AI-enhanced anonymization platform. It describes the main components, their interactions and the technologies employed to ensure secure,

privacy-compliant processing of sensitive healthcare data. The architecture is designed with modularity, scalability and data protection in mind, supporting both manual and AI-assisted anonymization workflows.

4.3.1 System Overview

The architecture integrates three complementary open-source technologies to support secure, utility-preserving anonymization of health data: **Opal**[1], a widely adopted data management and governance platform specifically designed for biomedical research environments, the **ARX Data Anonymization Tool**[2], a powerful framework for implementing privacy-preserving data transformations and **Ollama**[3], a lightweight runtime for executing LLMs locally.

The system architecture supports a modular, privacy-centric workflow that facilitates the seamless flow of data from ingestion to anonymization and secure sharing, incorporating both manual controls and intelligent AI-assisted recommendations.

4.3.2 Key Architectural Components

The proposed system architecture is designed to securely and efficiently handle the anonymization of sensitive health data while maintaining data utility and ensuring compliance with privacy regulations. The architecture integrates modular components that collectively support data ingestion, anonymization, risk analysis and hierarchical data processing. These components are interconnected through well-defined interfaces to facilitate seamless data flow, scalability and security.

Figure 4.1 illustrates the overall structure and interaction of the system's key components. The modular, service-oriented backend design ensures scalability and adaptability to future requirements, enabling efficient data processing while maintaining strong privacy protections and compliance with data regulations. Each of the components is described in detail below.

4.3.2.1 Web Application

The **Web Application** serves as the primary interface for system users. It facilitates the uploading of sensitive health data, configuration of anonymization settings and retrieval of anonymized datasets. All communication between the Web Application and backend services is handled through secure RESTful API calls.

4.3.2.2 Backend Services

The **Backend** implements the system's core functionality and is organized into modular components to ensure scalability, maintainability and clear separation of concerns. These include:

[1] https://www.obiba.org/pages/products/opal/
[2] https://arx.deidentifier.org/
[3] https://ollama.com/

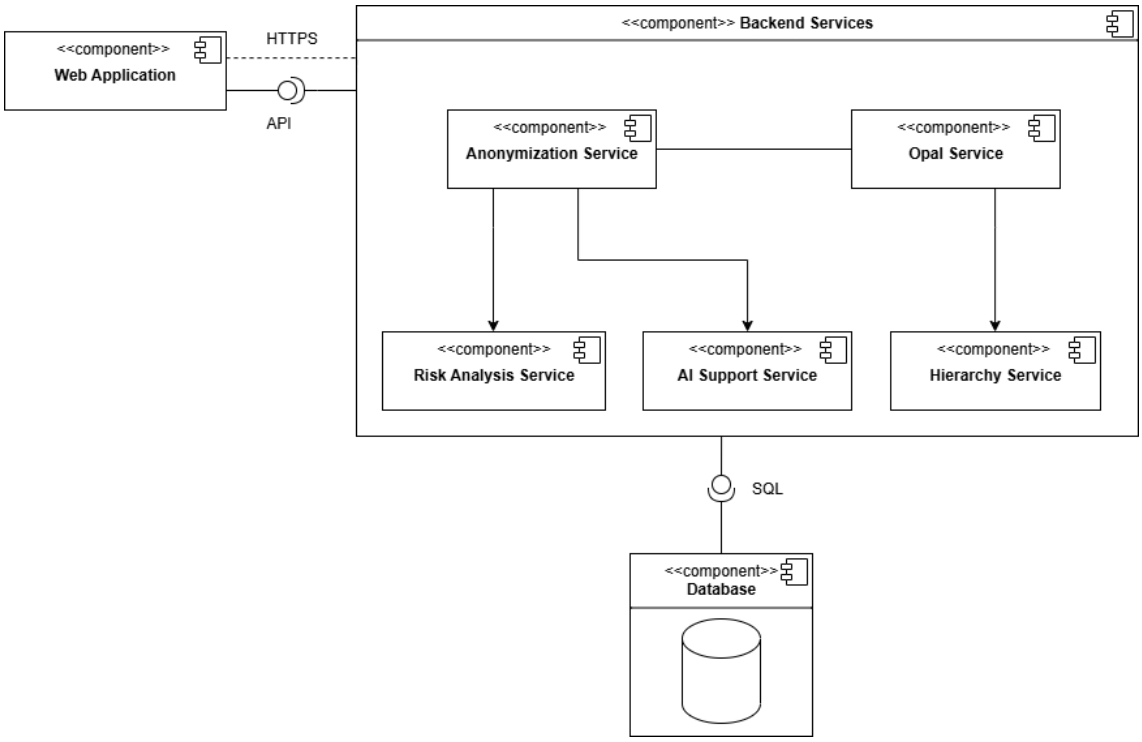

Figure 4.1: UML component diagram

- **Anonymization Service:** This service handles the anonymization of sensitive data using various privacy models, such as k-anonymity, l-diversity and t-closeness. It applies data transformation techniques including generalization, suppression, masking and perturbation to minimize re-identification risks. It is built using the **ARX Data Anonymization Tool** as a library, which provides a comprehensive and well-tested framework for implementing these privacy-preserving transformations.

- **Risk Analysis Service:** This service evaluates the effectiveness of anonymization by calculating both privacy and data utility metrics. It assesses the likelihood of re-identification and the impact of anonymization on data quality, providing real-time feedback to help users fine-tune their anonymization strategies.

- **Hierarchy Service:** This service generates data generalization hierarchies to support anonymization. It implements techniques such as masking, interval-based generalization for numerical data and temporal generalization for date values.

- **AI Support Service:** This service leverages locally executed large language models through the **Ollama** framework to assist in the anonymization process. It provides intelligent, context-aware support for tasks such as attribute classification, hierarchy generation and privacy model selection. By incorporating AI-driven suggestions into the workflow, the system enhances decision-making while maintaining full data locality to comply with privacy regulations.

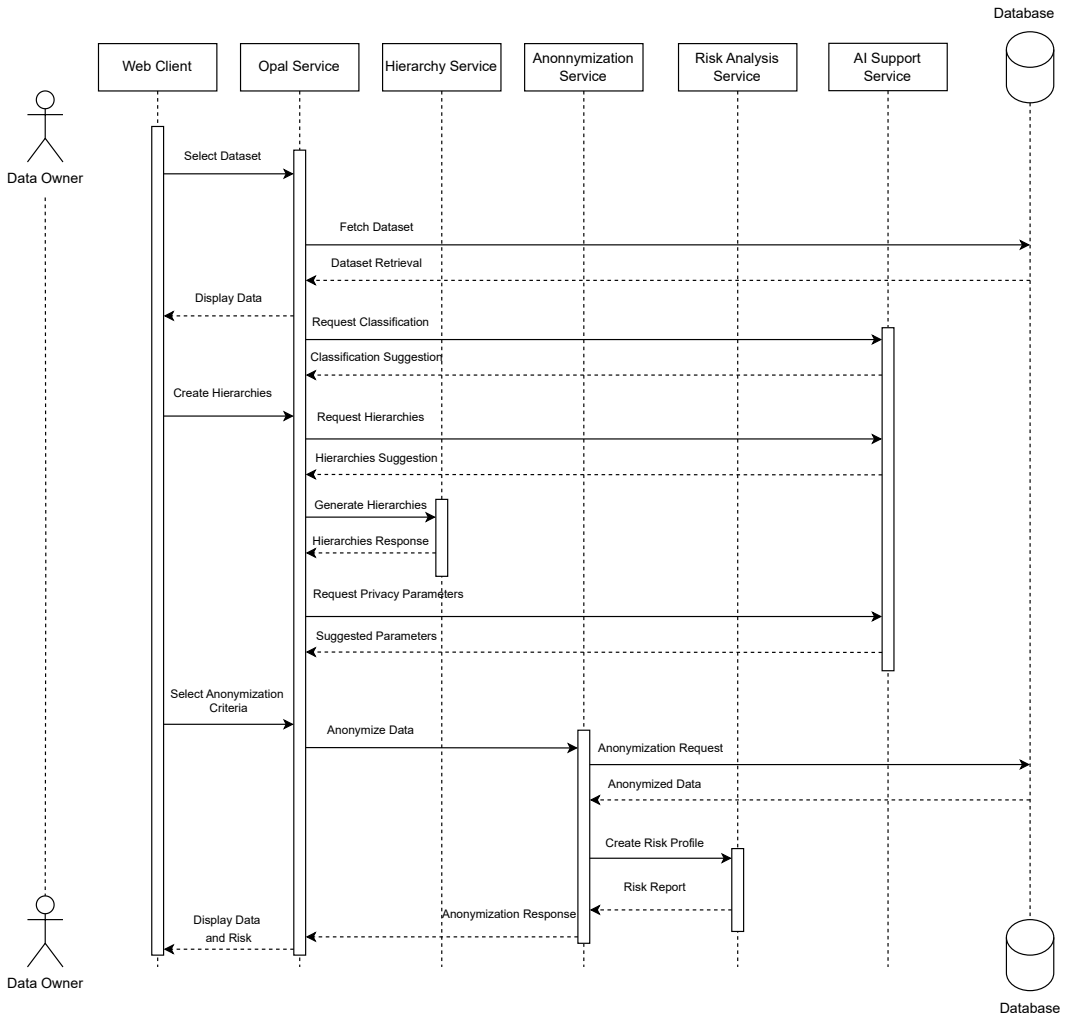

Figure 4.2: UML sequence diagram

4.3.2.3 Database

The **Database** securely stores both raw and anonymized datasets, along with user profiles, configuration settings, risk reports and audit logs. Access to the database is strictly controlled through backend services, ensuring data integrity and privacy compliance.

4.4 Interaction Workflow

Figure 4.2 illustrates the complete sequence of interactions among the core components of the proposed anonymization platform. This diagram includes the user-facing Web Client, backend orchestration, modular services and the AI Support Service responsible for intelligent automation.

The anonymization process unfolds through the following key steps:

1. **User Data Upload**
   The Data Owner initiates the workflow by either uploading a new sensitive dataset or selecting a previously uploaded one through the Web Client. This interface allows the user to configure anonymization preferences and provides real-time feedback throughout the process.

2. **Backend Service**
   Once the data is submitted, the Backend service orchestrates the anonymization process by coordinating communication between internal services. It validates the uploaded dataset and forwards it for attribute classification, hierarchy generation and anonymization.

3. **Attribute Classification**
   The Backend invokes the AI Support Service to classify each attribute as an identifier, quasi-identifier, sensitive or insensitive. The service returns the classification and a justification for each decision. Once the classifications are received, they must be reviewed and validated by the user.

4. **Hierarchy Suggestion**
   For attributes identified as quasi-identifiers, the AI Support Service provides recommended generalization hierarchies. These are based on attribute semantics and sample values.

5. **Privacy Model Configuration**
   The AI Support Service may also be queried to recommend suitable privacy parameters such as the value of k for k-anonymity or t for t-closeness. These recommendations take into account the distribution of quasi-identifiers and the distribution of sensitive attributes.

6. **Hierarchy Service**
   The suggested hierarchies are passed to the Hierarchy Service, which formalizes them into tree-based generalization structures. These hierarchies are required for downstream anonymization using generalization and masking techniques.

7. **Anonymization Service**
   The dataset is anonymized according to the chosen privacy models and hierarchies. Supported models include k-anonymity, l-diversity, t-closeness and differential privacy. Techniques such as suppression, masking and interval generalization are applied to transform the data.

8. **Risk Analysis Service**
   The anonymized data is then evaluated for re-identification risk and utility degradation. The Risk Analysis Service computes key metrics such as prosecutor risk, discernibility, NCP and AECS.

9. **Output and Feedback**
   The anonymized dataset, along with its associated risk and utility metrics, is returned to the

Web Client. The user can then visualize the final result or iteratively adjust the configuration to refine the anonymization process, supporting both usability and informed decision-making.

This interaction workflow ensures that sensitive healthcare data is anonymized efficiently and securely, with strong guarantees of privacy and utility. The integration of AI-powered support and interactive feedback provides a practical and intelligent framework for privacy-preserving data sharing.

4.5 Deployment

The deployment architecture of the proposed anonymization platform is structured around a containerized model using Docker[4], which promotes modularity, scalability and isolation between components. This approach allows each service to operate independently while maintaining seamless integration within the overall system. Figure 4.3 presents the architecture, detailing the containers and their respective roles.

The main container hosts both the Web Application and the Opal Service, which together form the core of the system's logic. The Web Application provides the graphical interface through which users can upload data, configure anonymization settings and trigger workflows. Internally, it communicates directly with the Opal Service, which is responsible for orchestrating the anonymization pipeline, coordinating interactions with auxiliary services such as classification, risk analysis and hierarchy generation.

An Apache Web Server is deployed in a dedicated container to handle incoming HTTPS requests and securely route them to Opal. This layer ensures secure external access to the platform while enforcing network boundaries between frontend and backend components.

A separate container runs the Ollama framework, which serves as the local runtime environment for executing large language models. This container supports models such as DeepSeek-R1[5], Mistral AI[6], and Gemma[7], enabling AI-assisted tasks like attribute classification, hierarchy suggestion, and privacy parameter recommendation. Since all AI inference occurs locally within the same deployment, the system ensures that sensitive data remains local, fully adhering to privacy-by-design principles and relevant data protection regulations.

Persistent storage is handled through a database container. Access to the database is strictly mediated by backend services to ensure security, auditability and compliance with legal frameworks such as the GDPR and HIPAA.

[4] https://www.docker.com/
[5] https://ollama.com/library/deepseek-r1
[6] https://ollama.com/library/mistral
[7] https://ollama.com/library/gemma3

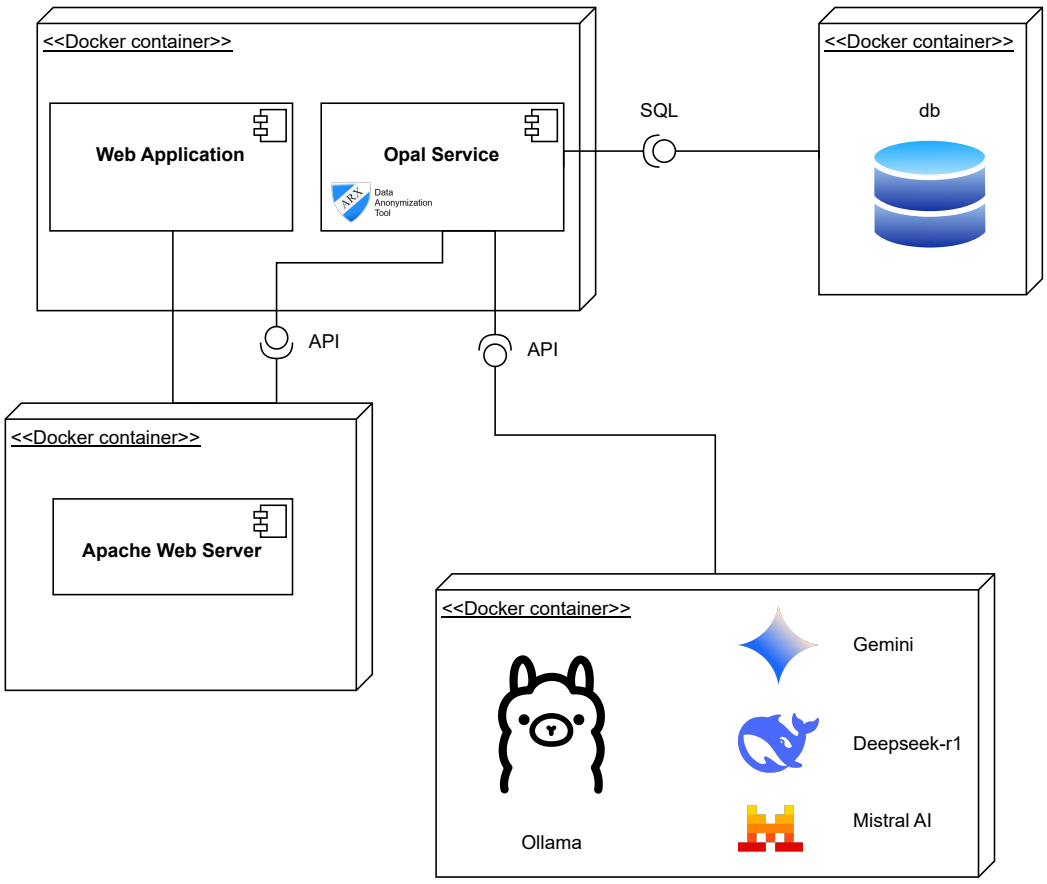

Figure 4.3: Deployment architecture of the anonymization platform

Overall, this architecture leverages the benefits of containerization to create a robust, maintainable and reproducible deployment environment. It facilitates secure data processing, simplifies system updates and provides the flexibility needed to scale and adapt the platform to future healthcare scenarios.

4.6 Design Principles

The system is built on the principles of privacy-by-design, explainability and modularity. All AI inferences are executed locally using the Ollama framework, ensuring that no sensitive data leaves the deployment environment. Anonymization operations are managed through the ARX library, which provides robust support for established privacy models.

By combining AI-driven intelligence with user control, the platform upholds stringent privacy standards while improving usability and adaptability in real-world healthcare data-sharing scenarios.

4.7 Contributions and Relevance

Based on the literature review and the theoretical framework established in this research, the proposed solution is designed to meet the growing need for privacy-preserving data-sharing mechanisms in healthcare, while maintaining the utility of data for research and analysis. By integrating the Opal secure data management platform with the ARX data anonymization tool and incorporating AI-driven decision support, the system provides a robust, scalable, and regulation-compliant approach to anonymizing and securely sharing sensitive healthcare data.

4.7.1 Relevance to the Field

The successful implementation and validation of the proposed system will have significant implications for both the academic field and practical applications. In the academic domain, this research contributes to the advancement of privacy-preserving data-sharing frameworks, particularly in the health domain, where data sensitivity and privacy concerns are paramount. The system bridges a critical gap between data privacy and utility, offering a practical solution that aligns with theoretical advancements in data anonymization.

In practical terms, the solution offers healthcare institutions a scalable and compliant method for sharing sensitive data to support research and innovation. By integrating secure data management with advanced anonymization techniques, the system addresses real-world challenges in data governance, privacy compliance and secure data sharing. This fosters collaboration between healthcare providers and researchers, enabling innovation while maintaining strict privacy standards. The system's modularity and adaptability also make it suitable for adoption in other data-sensitive industries beyond healthcare, such as finance and government.

4.7.2 Conclusion

The proposed system offers a secure, scalable and effective framework for anonymizing and sharing sensitive healthcare data. Its successful deployment can contribute both to academic research in privacy-preserving data sharing and to real-world healthcare innovation.

Chapter 5

Implementation of the AI-Enhanced Anonymization Platform

This chapter details how the proposed architecture was translated into a working software platform. It focuses on the engineering aspects of the anonymization workflow, including module development, integration of local LLMs via the Ollama runtime, audit and logging capabilities and the technologies used across the system.

While Chapter 4 introduced the high-level system architecture and design rationale, this chapter emphasizes the engineering and development aspects. It begins with the exploration and selection of Large Language Models, which played a central role in AI-driven decision support throughout the anonymization workflow.

5.1 Exploration and Selection of Large Language Models

As part of the platform's AI integration, several LLMs were evaluated to determine their suitability for supporting privacy-preserving tasks such as attribute classification, hierarchy suggestion and privacy parameter recommendation. The selection process focused on balancing inference quality, performance and local deployability to ensure full data locality and compliance with privacy regulations.

The evaluation began with domain-specific transformer models and later expanded to generative models that offer greater flexibility and interactivity.

5.1.1 Domain-Specific Transformer Models

The initial experimentation focused on transformer models pre-trained on medical data. *ClinicalBERT*[1], a variant of the BERT architecture trained on clinical notes and biomedical text, was selected due to its domain specialization. The model showed strong performance in tasks involving medical terminology understanding and named entity recognition.

[1] https://huggingface.co/medicalai/ClinicalBERT

However, transformer encoders like ClinicalBERT are primarily designed for token-level or sentence-level classification tasks and are optimized for unstructured data such as free-text clinical notes. As such, they lack the flexibility required for dynamic user-driven workflows, such as reasoning about attribute sensitivity based on context or providing open-ended recommendations.

5.1.2 Generative Language Models and Local Deployment

To address the limitations of transformer encoders, the evaluation next considered generative LLMs, which are better suited for tasks involving free-form reasoning, summarization and user interaction. These models offered the ability to dynamically assess attributes and suggest classifications based on semantic context, an essential feature for a usable anonymization platform.

Given the sensitivity of the data involved, a critical constraint was that all inference must occur locally, without transmitting any data to third-party services. As such, high-performance cloud-based models such as xAI's Grok[2] or OpenAI's GPT-4[3] were excluded.

Among the local deployment options explored, *GPT4All*[4] was initially tested due to its open-source availability and offline capabilities. However, it suffered from slow response times and high memory usage, which made it impractical for responsive use within the platform.

Ollama, by contrast, provided a streamlined and actively maintained framework for deploying LLMs locally. It supports several performant models such as LLaMA, Mistral and Deepseek and was designed for lightweight execution and ease of integration. Its modular architecture allowed the platform to experiment with different models quickly and under realistic hardware constraints.

5.2 System Architecture and Technologies

The implementation of the platform required a combination of frameworks, engines and runtime environments that align with both privacy guarantees and performance constraints. This section details the technologies selected for each core component.

5.2.1 Backend

The backend is developed in **Java** using the **Spring Framework**, providing a robust environment for RESTful APIs, data processing workflows and anonymization orchestration.

A key contribution of this work was the integration of the **ARX anonymization engine** as a Java library within the backend. This integration was developed independently and is not part of the original Opal platform. It allows the system to invoke privacy-preserving transformations and dynamically configure anonymization parameters such as k, l, t and ε (see Section 2.5).

[2] https://x.ai/blog/grok
[3] https://openai.com/gpt-4
[4] https://github.com/nomic-ai/gpt4all

5.2.2 AI Integration

Another central contribution of this work is the integration of local Large Language Models into the anonymization workflow. The platform uses the **Ollama** runtime to execute these models locally, deployed on a dedicated high-performance machine on the same network to reduce computational overhead on the main application and improve inference responsiveness.

This AI integration was implemented and is external to both Opal and ARX. The platform supports multiple models, with **DeepSeek**[5] as the default configuration. Users can select the desired model via the application interface. LLMs assist with attribute classification, generalization hierarchy suggestions and privacy parameter recommendations. All communication with the models is handled through structured prompts and RESTful interfaces designed specifically for this system.

5.2.3 Frontend

The graphical interface is built using **Vue.js**, offering a responsive and user-friendly web interface for managing data upload, workflow configuration and visualization of anonymization outcomes. The frontend communicates with the backend through REST APIs, supporting a clear separation of concerns and enabling modular deployment.

5.2.4 Data Storage

The platform uses **MySQL** as its primary database for persisting datasets, anonymization configurations, user-selected parameters and audit logs. This ensures traceability of operations and alignment with data governance requirements, including support for reproducibility and auditing.

5.2.5 Containerization

All backend services are containerized using **Docker**, ensuring reproducibility across environments and simplifying development and deployment workflows. The system uses the base image `maven:3.8-eclipse-temurin-21` for Java services. Containerization also supports future scalability and facilitates CI/CD integration.

5.2.6 API Documentation

To facilitate communication between the frontend and backend services, the platform provides a RESTful API developed using the Spring Boot framework and documented via the OpenAPI 3.0 specification[6]. This API exposes all core functionality of the platform, including data ingestion, attribute classification, hierarchy generation, privacy model configuration, anonymization execution, risk analysis and audit log management.

[5] https://ollama.com/library/deepseek-r1
[6] https://swagger.io/specification/

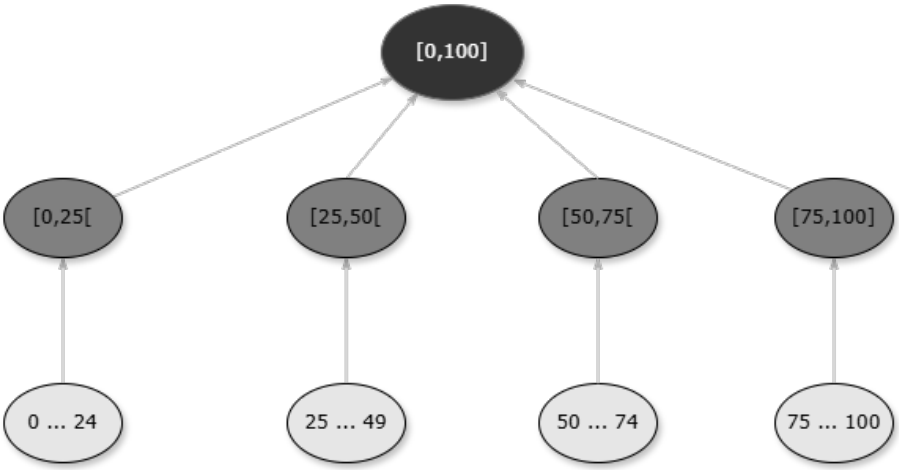

Figure 5.1: Interval-based generalization hierarchy.

To improve accessibility and enable rapid testing, the API is self-documented using Swagger UI[7]. This interface automatically generates a live, interactive API documentation page where users can inspect request parameters, view schema definitions and examine server responses. A complete visual overview of the API endpoints and their groupings is presented in Appendix A.

This API-centric design ensures that all anonymization workflows can be invoked programmatically and supports integration with third-party systems or data pipelines. It also enhances modularity and testability, allowing for the independent validation of system components.

5.3 Implemented Features

The implemented platform translates the architecture outlined in Chapter 4 into a working prototype that supports the secure and efficient anonymization of structured healthcare datasets. The system is composed of modular components that collectively address the challenges of privacy preservation, data utility and usability in real-world health data-sharing scenarios.

The following subsections detail the core functionalities developed to support the anonymization workflow.

5.3.1 Generalization Hierarchies

Four types of data transformation hierarchies were implemented to support generalization operations:

- **Interval-Based Hierarchies:** Used for numerical attributes such as `age` or `income`, where raw values are grouped into predefined ranges. Figure 5.1 illustrates an interval-based hierarchy, showing how values (e.g. `25`) are progressively abstracted into broader intervals.

[7] https://swagger.io/tools/swagger-ui/

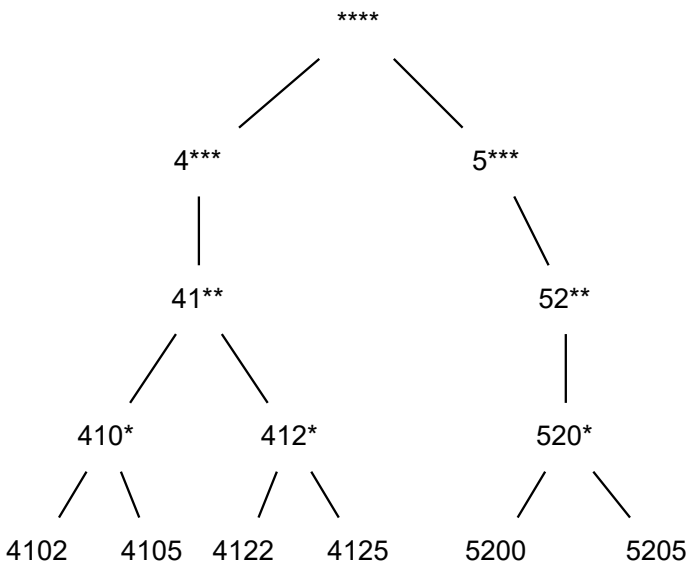

Figure 5.2: Masking-based generalization hierarchy.

- **Masking Hierarchies:** This approach is commonly applied to identifiers such as `ZIP codes` or `phone numbers`. It consists of replacing a growing number of characters, either from right to left or left to right, with a masking symbol (e.g. $\star$), thereby reducing the risk of re-identification.

  Figure 5.2 shows six ZIP code examples and how they are generalized through successive masking levels. This enables partial obfuscation while preserving enough structure for meaningful analysis.

- **Date Hierarchies:** Used for temporal fields such as `birth dates` or `admission dates`. Generalization is achieved by reducing the granularity of the date. Figure 5.3 demonstrates how individual dates can be generalized across multiple levels, preserving temporal structure while improving privacy.

- **Categorical Hierarchies:** These are used to generalize nominal attributes such as `occupation` or `diagnosis`, where values are grouped into semantically related categories using domain-specific taxonomies.

  Table 5.1 presents a categorical hierarchy for the `Diagnosis` attribute, where common medical conditions are grouped into diagnostic categories and further abstracted into general medical classifications.

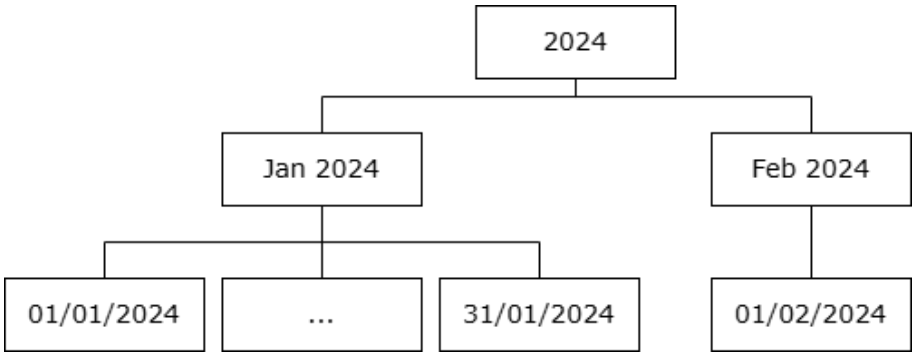

Figure 5.3: Date generalization hierarchy.

5.3.2 Privacy Models

The platform supports multiple privacy models, enabling users to select the one that best aligns with their privacy requirements and acceptable utility trade-offs. These models were introduced and discussed in detail in Section 2.5, which outlines their theoretical foundations and practical implications. The implemented models include:

- **k-Anonymity:** Ensures that each record is indistinguishable from at least $k-1$ others.
- **l-Diversity:** Requires that sensitive attributes have at least l distinct values within each equivalence class.
- **t-Closeness:** Ensures the distribution of sensitive attributes in any equivalence class is close to the overall dataset.
- **Differential Privacy:** Provides ε-bounded noise injection to mitigate membership inference attacks.

Table 5.1: Categorical generalization hierarchy.

Diagnosis	Disease Group	General Type
Flu	Respiratory Infection	Infectious Disease
Pneumonia	Respiratory Infection	Infectious Disease
COVID-19	Viral Respiratory Disease	Infectious Disease
Hypertension	Cardiovascular Condition	Chronic Disease
Heart Attack	Cardiovascular Condition	Chronic Disease
Type 2 Diabetes	Metabolic Disorder	Chronic Disease
Migraine	Neurological Condition	Chronic Disease

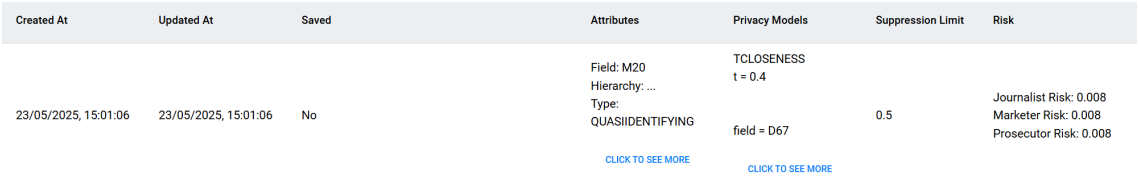

Figure 5.4: Example of an anonymization log showing metadata.

5.3.3 Anonymization Logging

Every anonymization operation is logged with its parameters, selected hierarchies and resulting risk metrics (e.g. re-identification risk). This ensures full traceability and auditability, which are essential for privacy compliance and governance.

Figure 5.4 presents an anonymization log, capturing key metadata such as applied privacy models, selected parameters, hierarchy strategies and corresponding risk scores. This log format underpins the system's ability to support transparent, auditable workflows.

Comprehensive logging is fundamental to ensuring transparency, accountability and regulatory compliance in anonymization workflows. In the context of health data, where the risk of re-identification must be carefully managed, maintaining detailed logs serves several key purposes:

- **Auditing and Compliance:** Logs provide an auditable trail of actions taken during the anonymization process, including which privacy models and parameters were used (e.g. k, l, t, or ε) and the rationale for those choices. This supports both internal governance and external regulatory audits.

- **Reproducibility and Traceability:** Each anonymization task may involve multiple parameters and transformations. Logging allows users and stakeholders to reproduce the anonymization process exactly, ensuring that results can be validated or repeated on new datasets with similar characteristics. This is especially important in scientific research, where reproducibility is a cornerstone of integrity.

- **Risk Monitoring and Analysis:** By capturing detailed risk metrics—such as prosecutor, journalist and marketer model scores—logs help stakeholders assess the effectiveness of anonymization strategies over time. For instance, a sudden increase in re-identification risk across datasets may indicate a need to adjust model parameters or data handling procedures.

- **Troubleshooting and Decision Justification:** If a data utility issue arises (e.g. a dataset becomes too generalized or loses analytical value), logs can help identify which transformations caused the degradation. This also enables users to understand and justify why certain fields were masked or generalized, facilitating informed discussions between data engineers, legal teams and clinical stakeholders.

5.3.4 AI-Driven Assistance

The platform integrates local Large Language Models to assist with key decision points in the anonymization workflow. All models are executed locally via the Ollama framework to maintain strict privacy constraints. The following subsections describe the AI contributions in detail.

5.3.4.1 Attribute Classification

The first and arguably most critical step in any anonymization workflow is the accurate classification of dataset attributes. Misclassification at this stage can lead to improper anonymization strategies either exposing sensitive data or overly degrading data utility. This component classifies each attribute into one of four predefined privacy categories, as detailed in Section 2.2: *identifying*, *quasi-identifying*, *sensitive* and *insensitive*. The classification is performed by a locally hosted LLM such as DeepSeek, using structured prompts that include contextual and semantic cues.

Methodology

- The classification process begins by extracting relevant context for each attribute, including:
  - The attribute's **name**;
  - Its associated **label**, if available;
  - A sample of **non-NULL values**;
- This information is embedded into a structured prompt sent to an LLM. The prompt includes:
  - A concise definition of each privacy category (*identifying*, *quasi-identifying*, *sensitive*, *insensitive*) with illustrative examples;
  - The attribute name, label and sample values;
  - A direct instruction to classify the attribute and justify the decision;
- The model returns a JSON-formatted response containing:
  - The predicted `classification`;
  - A `justification` outlining the reasoning behind the decision;

Example prompt (simplified):

> You are an expert in data anonymization. Classify the following attributes based on the definitions below:
>
> **Identifying** – Directly reveals identity (e.g. name, SSN)
> **Quasi-identifying** – Can indirectly identify someone when combined with other attributes (e.g. age, ZIP code)

Sensitive – Contains confidential or private information (e.g. diagnosis, income)
Insensitive – Neutral information posing minimal or no privacy risk (e.g. procedure type)

Attribute name: `age`
Label: `Patient Age (Years)`
Sample values: 22, 35, 41, 56

Return a JSON object with two fields: `"classification"` and `"justification"`.

Example output:

Listing 5.1: Example of AI-driven attribute classification

```
{
  "age": {
    "classification": "QUASIIDENTIFYING",
    "justification": "Age is not unique by itself but can help re-identify
        a person when combined with other fields."
  },
  // Other attributes
}
```

Benefits

- Automates a critical step, reducing reliance on manual review.
- Produces explainable, auditable decisions via justifications.
- Ensures consistent attribute handling across datasets.

Limitations

- Ambiguous or generic attributes (e.g. `Notes`, `Comments`) may require additional context for accurate classification.
- Final decisions remain subject to user validation as part of the human-in-the-loop process.

5.3.4.2 Hierarchy Suggestions

Once attributes are classified, the next step is to define how quasi-identifying attributes can be generalized. This is essential for applying privacy models such as k-anonymity and l-diversity. To assist data providers, the system uses a locally hosted LLM to propose suitable generalization hierarchies based on the attribute's semantics, data type and value distribution.

Methodology

The hierarchy suggestion workflow operates as follows:

- For each quasi-identifying attribute, the system extracts:
  - The attribute's **name** and **label**, if available;
  - A sample of **non-NULL values**;
- A structured prompt is created and sent to the LLM. It includes:
  - A summary of the generalization strategies supported by the platform (as described in Section 5.3.1), including:
    - **Interval-based** for numeric attributes;
    - **Date-based** hierarchies for temporal attributes;
    - **Masking-based** for strings and identifiers (e.g., ZIP codes);
    - **Categorical grouping** for domain-specific terms (e.g., diagnoses);
  - The attribute name, label and sample values;
  - A direct request to output a hierarchy in JSON format, along with a justification and the recommended generalization type;
- The output is parsed into a hierarchical structure compatible with the anonymization engine.

Example prompt (simplified):

> You are an expert in data anonymization. Suggest an appropriate generalization hierarchy for the attribute below using one of the following supported strategies:
>
> - **Interval generalization** – for numeric values, group values into ranges (e.g., 20–30);
> - **Date generalization** – reduce precision from day → month → year;
> - **Masking** – replace part of a string with a masking symbol (e.g., 1234→12**);
> - **Categorical grouping** – group categories into semantically similar sets;
>
> **Attribute name**: `age`
> **Label**: `Patient Age (Years)`
> **Sample values**: 22, 35, 41, 56
>
> Return a JSON object with the recommended hierarchy structure, the type of generalization and a justification.

Example output:

Listing 5.2: Example output of AI-suggested generalization hierarchy

```
1 {
```

```
  "age": {
    "start": 0,
    "end": 100,
    "step": [
      10,
      20
    ],
    "justification": "Interval generalization in steps of 10 and 20 years
       protects individual ages while allowing demographic analysis.",
    "type": "INTERVAL"
  }
   // Other attributes
}
```

The model recommended an **interval-based generalization hierarchy** for the attribute `age`, with values ranging from 0 to 100. The `step` array specifies two levels of generalization granularity: intervals of 10 years and 20 years. This results in hierarchical groupings such as `[0-10[`, `[10-20[`, `[20-30[` for the finer level and broader intervals such as `[0-20[`, `[20-40[` and so on for the broader level.

Benefits

- Automates the creation of structured, context-aware generalization hierarchies tailored to each attribute.
- Reduces manual workload, especially for data providers without expertise in privacy-preserving transformations.
- Encourages consistent and semantically meaningful generalization across different datasets.

Limitations

- The model may struggle to suggest appropriate hierarchies for attributes that are ambiguous, highly domain-specific, or have unclear value semantics.
- Final hierarchies must be validated by the user to ensure alignment with the dataset's context and anonymization goals.

5.3.4.3 Privacy Model Parameter Suggestions

After classifying attributes and defining generalization hierarchies, users must configure the parameters of their chosen privacy models. These parameters such as k for k-anonymity, l for l-diversity and t for t-closeness control the strength of privacy guarantees. However, determining appropriate values is non-trivial and context-dependent.

To support this decision, the platform integrates an AI-powered module that suggests suitable parameter values using local LLMs. Prompts are dynamically generated based on the selected privacy model and dataset statistics.

Model-Specific Prompting

Each supported privacy model is associated with a tailored prompt template that includes an explanation of the model, relevant dataset features and formatting instructions. The LLM returns a structured JSON response containing the suggested parameter and reasoning.

- **k-Anonymity**: Suggests an integer k based on the dataset size and the generalization hierarchies of quasi-identifiers.
- **l-Diversity**: Analyzes the frequency distribution of a selected sensitive attribute and proposes an appropriate l value.
- **t-Closeness**: Computes the global frequency distribution of a sensitive attribute and asks the model to suggest a suitable t threshold based on statistical variation.

Example: k-Anonymity Prompt

> You are a privacy expert.
>
> k-Anonymity is a privacy model that ensures that each record in a dataset is indistinguishable from at least $k-1$ others based on quasi-identifiers. Higher values of k provide stronger privacy but may reduce data utility.
>
> Based on the following quasi-identifying attributes and their generalization hierarchies, suggest an appropriate value for k. Explain your reasoning.
>
> **Dataset size:** 1000
>
> **Generalization hierarchies (example paths):**
>
> - age: 25 → [20,30[→ [0,40[
> - zipcode: 12345 → 1234* → 123**
>
> Return only a JSON object:

Listing 5.3: Example output of AI-suggested k-anonymity parameters

```
{
  "k": 6,
  "reasoning": "With two quasi-identifiers and 1000 records, k=6 ensures
      moderate anonymity while preserving demographic granularity."
}
```

Example: l-Diversity Prompt

You are a privacy expert.

l-Diversity enhances k-anonymity by requiring at least l distinct sensitive values per equivalence class. Higher l increases privacy by diversifying sensitive attributes.

Based on the following sensitive attribute value distribution, suggest a suitable l. Explain your reasoning.

Attribute: Diagnosis
Value distribution:

- Flu: 320
- COVID-19: 220
- Hypertension: 160
- Diabetes: 100

Return only a JSON object.

Listing 5.4: Example output of AI-suggested l-diversity parameters

```
{
  "field": "Diagnosis",
  "l": 4,
  "reasoning": "The attribute has enough diversity across values. Setting
      l=4 reduces inference risk while maintaining analytical use."
}
```

Example: t-Closeness Prompt

You are a data privacy expert.

Equal-distance t-Closeness ensures that the distribution of sensitive attributes within each group is close to the overall distribution. The threshold t controls this closeness. Lower values increase privacy.

Attribute: Diagnosis
Global frequency distribution:

- Flu: 400
- Diabetes: 250
- Heart Attack: 200
- Cancer: 150

Dataset size: 1000

Return only a JSON object.

Listing 5.5: Example output of AI-suggested t-closeness parameters

```
{
  "field": "Diagnosis",
  "t": 0.15,
  "reasoning": "This value of t reflects moderate sensitivity, allowing
      some distribution shift while preventing inference of rare conditions
      ."
}
```

Benefits

- Model-aware prompts ensure recommendations are aligned with theoretical definitions and practical implications.
- LLMs use real dataset distributions and hierarchies to contextualize suggestions.
- Structured JSON output with justifications supports transparency and user oversight.

Limitations

- Final responsibility remains with the user to approve or revise the suggested parameters.

By leveraging LLMs, the platform transitions anonymization from a fully manual process to a guided, semi-automated experience. The AI acts as a *knowledgeable assistant*, improving speed, consistency and quality of decisions while preserving expert oversight and final validation.

5.4 Comparison to Traditional Anonymization Workflows

Traditional anonymization workflows often require domain experts to manually inspect each attribute, define generalization hierarchies, select appropriate privacy models and iteratively fine-tune parameters. This manual approach can be slow, error-prone and inconsistent particularly when applied to large, complex healthcare datasets with many variables.

Figure 5.5 illustrates the traditional anonymization process as a sequential pipeline consisting of four major stages: attribute classification, hierarchy creation, privacy model selection and risk analysis.
In contrast, the AI-assisted workflow implemented in this platform integrates local LLMs to automate or augment key decision points. Attribute types are automatically classified based on contextual analysis; generalization strategies are suggested based on data type and semantics; and privacy parameters such as k, l and t are recommended according to the sensitivity and distribution of the data.

This hybrid approach not only accelerates the anonymization process but also improves the consistency and quality of decisions, while preserving the ability for expert oversight and final validation.

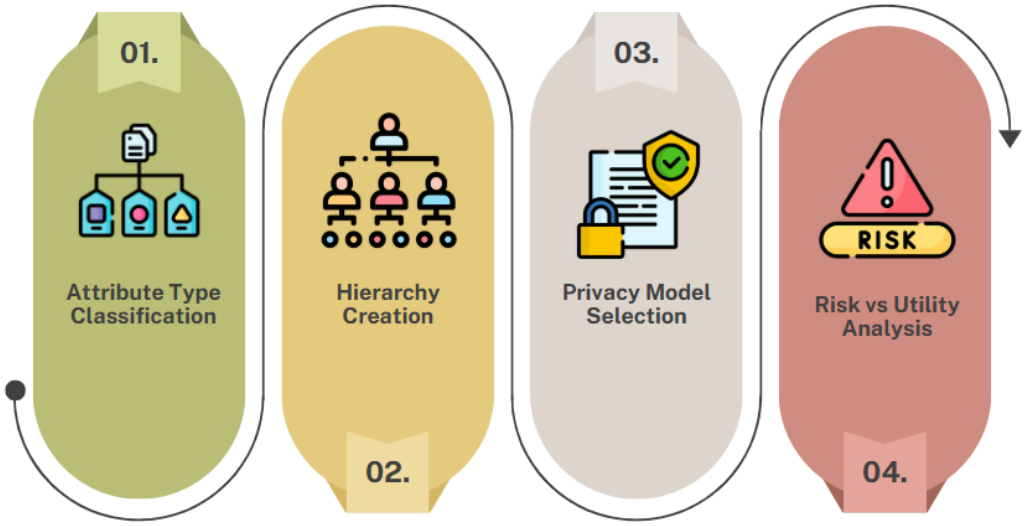

Figure 5.5: Traditional anonymization workflow.

5.5 Conclusion

This chapter detailed the practical implementation of the AI-enhanced anonymization platform, focusing on the translation of architectural concepts into an operational system. Through the integration of modular services, local LLMs and robust privacy-preserving technologies, the platform addresses the core challenges of privacy, usability and auditability in healthcare data anonymization.

Key contributions include the seamless integration of the ARX, the deployment of local LLMs via the Ollama runtime for AI-driven assistance and the implementation of a user-friendly web interface supported by a scalable backend and documented API. These elements collectively enable end-to-end anonymization workflows, from attribute classification to privacy model execution, with explainable and auditable outputs.

By leveraging AI for decision support and automation, the platform enhances both the efficiency and consistency of anonymization tasks, significantly reducing the manual burden traditionally required in such workflows. The system's architecture is designed for extensibility, ensuring compatibility with evolving privacy standards and future model integrations.

Together, these implementation efforts establish a solid foundation for the platform's evaluation, which is explored in the following chapter through synthetic benchmarks and a real-world case study.

Chapter 6

Results and Evaluation

Following the implementation of the proposed anonymization platform, a series of experiments were conducted to evaluate its functionality, effectiveness and applicability in realistic healthcare data-sharing scenarios. This chapter focuses on assessing the platform's ability to meet key objectives, namely the preservation of patient privacy, the maintenance of data utility and the overall usability of the system from a data provider's perspective.

A central aspect of this evaluation involved exploring the use of Artificial Intelligence to support and automate key components of the anonymization process. Given the sensitivity of healthcare data, particular attention was paid to ensuring that all processing could occur in local environments, thereby avoiding reliance on external cloud-based services. Various tools and models were explored under this constraint, leading to the integration and testing of LLMs within the system.

In addition to technical validation, a case study using real clinical data provided by CHUSJ was carried out. This enabled a comprehensive assessment of the platform in a realistic context, using a dataset with a complex structure and sensitive content. The evaluation focused on the platform's ability to support attribute classification, apply anonymization techniques effectively and compute privacy risk while preserving the analytical value of the data.

The results obtained from this evaluation are presented and discussed in this chapter, providing insights into the platform's strengths, limitations and areas for future improvement.

6.1 Experimental Setup

To evaluate the attribute classification performance of AI models within the anonymization platform, a publicly available synthetic healthcare dataset from Kaggle[1] was selected. This dataset was designed to emulate realistic healthcare records providing a safe and representative proxy for real patient data while avoiding privacy concerns.

[1] https://www.kaggle.com/datasets/prasad22/healthcare-dataset

Table 6.1: Columns from the synthetic healthcare dataset

Attribute	**Description**
Name	Full name of the patient (e.g. "John Smith")
Age	Patient age in years
Gender	Biological sex ("Male" or "Female")
Blood Type	Patient blood group (e.g. "A+", "O-")
Medical Condition	Primary diagnosis (e.g. "Diabetes", "Hypertension")
Date of Admission	Admission date
Doctor	Attending physician name
Hospital	Name of healthcare facility
Insurance Provider	Insurance company covering the patient
Billing Amount	Total billed amount for care, in USD
Room Number	Patient's room number during admission
Admission Type	Nature of admission ("Emergency", "Elective" or "Urgent")
Discharge Date	Date of discharge from the hospital
Medication	Medication prescribed (e.g. "Paracetamol", "Lipitor")
Test Results	Outcome of lab results

6.1.1 Dataset Description

The dataset contains multiple patient records, each characterized by demographic, medical and administrative features. The columns reflect attributes commonly found in electronic health records. A summary of the columns is shown in Table 6.1.

6.1.2 Labeling and Ground Truth

Each attribute was manually classified into one of the following four privacy categories:

- **Identifying (ID)**: Directly reveals a patient's identity (e.g. `Name`, `Doctor`).
- **Quasi-identifying (QID)**: Can indirectly identify individuals when combined with other data (e.g. `Age`, `Room Number`, `Date of Admission`, `Discharge Date`, `Hospital`).
- **Sensitive (SA)**: Encodes confidential clinical or financial details (e.g. `Medical Condition`, `Billing Amount`, `Test Results`, `Medication`, `Insurance Provider`, `Blood Type`).
- **Insensitive**: Contains neutral information that poses negligible re-identification risk (e.g. `Gender`, `Admission Type`).

A total of **15 attribute-label pairs** were defined as ground truth for evaluating classification accuracy. These labels were manually assigned based on established guidelines in the data anonymization literature, reflecting commonly accepted classifications of identifying, quasi-identifying, sensitive and insensitive attributes. They served as the reference for comparing model predictions across all variants tested.

6.1.3 Evaluation Criteria

To assess model performance, two key metrics were used:

- **Classification Accuracy**: The proportion of correctly predicted labels compared to a predefined ground-truth set of attributes.
- **Average Inference Time**: The mean time (in seconds) required for the model to classify all atributes.

To ensure reliability, each model was executed five times and both classification accuracy and inference time were averaged across these runs.

6.2 Evaluation of Attribute Classification Accuracy

An essential goal of the platform is to assist users in reliably classifying dataset attributes into privacy-relevant categories with minimal manual effort. This section evaluates the accuracy and performance of local large language models when used for this task. By testing various DeepSeek model variants, we examine how model size influences classification quality and inference time. The experiment provides a data-driven basis for selecting an optimal model for practical deployment in anonymization workflows.

6.2.1 Overview

This section evaluates the accuracy and performance of different DeepSeek language model variants when applied to the task of attribute classification in structured healthcare datasets. Since accurate classification of dataset attributes into *identifying*, *quasi-identifying*, *sensitive* and *insensitive* is critical for applying the correct anonymization strategies, this experiment helps determine which model provides the most reliable support under local deployment constraints.

6.2.2 LLM Selection

The DeepSeek family of models was selected for this evaluation due to its strong performance-to-efficiency ratio and availability in open-source, locally deployable formats. Unlike proprietary models such as OpenAI's GPT-4, DeepSeek models offer full transparency and offline usability, an essential criteria for our privacy-focused context, where no data should leave the local environment.

Table 6.2: Accuracy and inference time of DeepSeek models on attribute classification

Model	**Accuracy (%)**	**Avg. Inference Time (s)**
DeepSeek-r1 32B	81.7	17.67
DeepSeek-r1 14B	78.3	12.36
DeepSeek-r1 8B	61.7	10.22
DeepSeek-r1 7B	58.3	7.71
DeepSeek-r1 1.5B	41.7	7.21

DeepSeek also supports a wide range of model sizes, making it ideal for comparing trade-offs between inference efficiency and classification accuracy. Additionally, the models demonstrated robust instruction-following capabilities in early testing, making them suitable for structured prompt-based classification tasks without requiring fine-tuning.

The combination of open licensing, local deployment via the Ollama runtime and favorable reasoning performance made DeepSeek a practical and scalable choice for AI-assisted anonymization workflows in healthcare contexts.

Five DeepSeek model variants were tested in this evaluation to compare performance across different model sizes: **DeepSeek-r1 1.5B**, **DeepSeek-r1 7B**, **DeepSeek-r1 8B**, **DeepSeek-r1 14B** and **DeepSeek-r1 32B**.

All models were executed locally using the Ollama inference framework on a high-performance server equipped with the following hardware: a 28-core CPU, 62 GB of RAM and an NVIDIA GH100 (H200 SXM, 141 GB) GPU. This configuration ensured consistent runtime performance and enabled the deployment of large-scale language models without reliance on external cloud services.

6.2.3 Model Accuracy and Performance Comparison

Table 6.2 presents the evaluation of each DeepSeek model based on the criteria described in Section 6.1.3. Accuracy was calculated as the proportion of attributes correctly classified into one of the four categories.

6.2.4 Discussion and Observations

The results reveal a strong correlation between model size and classification accuracy. The best-performing model, DeepSeek-r1 32B, achieved an accuracy of 81.7 %, but required the highest average inference time of 17.67 s. While its classification quality was highest, the latency and resource requirements make it suitable primarily for high-performance environments.

DeepSeek-r1 14B achieved a comparable accuracy of 78.3 % with significantly lower latency (12.36 s), offering a favorable trade-off between performance and efficiency. This makes it a strong candidate for integration into local anonymization workflows where time and hardware constraints are moderate.

Mid-sized models such as DeepSeek-r1 8B and 7B offered intermediate accuracy levels of 61.7 % and 58.3 %, respectively. While faster, they showed a tendency to misclassify quasi-identifying and sensitive attributes and may require manual oversight in real-world use.

The smallest model, DeepSeek-r1 1.5B, demonstrated limited capability with an accuracy of 41.7 %. Although it was among the fastest (7.21 s), its reduced accuracy limits its usefulness for nuanced attribute classification tasks.

These findings highlight the importance of model selection in AI-assisted anonymization. While higher-accuracy models offer improved classification, they also demand more resources and introduce latency. DeepSeek-r1 14B stands out as a balanced option for most deployment scenarios. Importantly, these results also demonstrate and confirm that accurate, AI-assisted attribute classification is achievable using large language models, even under local inference constraints, supporting the integration of LLMs into privacy-preserving workflows in sensitive domains such as healthcare.

6.3 Case Study: Real-World Application on Clinical Data

To validate the practical applicability of the proposed platform, a real-world case study was conducted using structured clinical data. This evaluation aimed to demonstrate the system's ability to assist users in anonymizing sensitive health records through AI-supported workflows while maintaining data utility. The case study also served to illustrate how different privacy configurations affect key risk and utility metrics in a realistic healthcare setting.

6.3.1 Case Study Objectives

The goal of this case study was to validate the platform's functionality in a real-world setting using actual clinical data. Specifically, the aim was to assess the effectiveness of AI-assisted anonymization workflows in supporting privacy protection without compromising data utility and to explore the practical implications of applying different privacy configurations in a healthcare context.

6.3.2 Dataset Description

The dataset used in this case study was provided by CHUSJ and consisted of structured electronic health records related to breast cancer. It contained **138 attributes** and **475 patient records**. Attributes included demographic variables (e.g. `Sex`, `Age`), clinical dates (e.g. `Date of Mammogram`, `Date of First Breast Operation`), diagnostic outcomes (e.g. `Grade of Invasive Cancer`) and outcome variables such as `Cause of Death`.

Due to the dataset's complexity and sensitivity, maintaining privacy while preserving analytical utility posed a significant challenge.

6.3.3 Workflow Execution

The anonymization process was executed using the proposed platform which consisted of the following steps:

1. **Attribute Classification:** The DeepSeek-14B model was used to perform a preliminary classification of all attributes into four privacy categories: *identifying*, *quasi-identifying*, *sensitive* and *insensitive*. Due to LLM token limitations, the attributes were processed in batches of 20. The resulting classifications were then reviewed and validated by the research team. This model was selected based on its strong balance between classification accuracy and inference time in earlier evaluations.

2. **Hierarchy Generation:** All date-related attributes were generalized using date hierarchies (day/month/year → month/year → year) and `Age` was generalized using interval-based hierarchies.

3. **Anonymization:** The validated attribute classifications and hierarchies were used to apply k-anonymity and t-closeness across all sensitive attributes. Multiple combinations of parameters were explored to assess the privacy-utility trade-off.

6.3.4 API Validation and Usability

In addition to evaluating AI accuracy and anonymization performance, the underlying RESTful API was tested for robustness, clarity and usability. The API exposes all platform features, including dataset upload, attribute classification, hierarchy generation, risk analysis and log retrieval, via a modular interface documented using the OpenAPI 3.0 specification.

Endpoints were tested manually using the integrated Swagger UI, which allowed researchers to simulate common usage scenarios and inspect request/response behavior. Example endpoint groups include `/arx/anonymizeTable`, `/arx/suggestKAnonymity` and `/hierarchy/masking`, all of which were exercised during the case study execution.

The full list of endpoints and example screenshots of the Swagger interface are provided in Appendix A. These visual references demonstrate the platform's extensibility and reinforce the modular design that enables integration into external tools or clinical data pipelines. This API-centric approach also ensured that the anonymization logic could be evaluated independently of the web interface.

6.3.5 Evaluation Metrics and Methodology

The anonymized datasets were evaluated along two primary axes, using the risk and utility metrics described in Section 2.7:

- **Privacy Risk:** Measured using ARX's attacker models — *prosecutor*, *journalist* and *marketer*. Since all three models produced identical re-identification probabilities under our configurations, only the prosecutor model is reported in the graphical analysis.

Table 6.3: Re-identification risk for different anonymization configurations

Configuration	Prosecutor Risk (%)	Journalist Risk (%)	Marketer Risk (%)
Original Dataset	99.16	99.16	99.16
$k = 5, t = 0.8$	4.93	4.93	4.93
$k = 10, t = 0.5$	2.07	2.07	2.07
$k = 10, t = 0.4$	0.89	0.89	0.89

- **Utility and Information Loss:** Assessed using ARX's statistical tools, including:
  - Average Equivalence Class Size;
  - Discernibility Metric;
  - Suppressed Record Count;
  - Normalized Certainty Penalty;

6.3.6 Results and Observations

To evaluate the trade-off between privacy protection and data utility, multiple anonymization configurations were applied using different combinations of k-anonymity and t-closeness. Tables 6.3 and 6.4 summarize the results across three selected configurations.

Privacy Risk

Table 6.3 presents the estimated re-identification risks as reported by ARX's attacker models. All three attacker profiles (prosecutor, journalist, marketer) yielded identical success probabilities for each configuration. Compared to the original dataset, which had a near-certain re-identification risk ((99.16 %)), every anonymization configuration led to a substantial reduction in attacker success rates. Notably, the strictest setting ($k = 10, t = 0.4$) reduced risk to as low as (0.89 %).

Utility Metrics

Table 6.4 summarizes utility indicators as discussed in Section 2.7, including:

- **AECS** – average group size after anonymization;
- **Suppressed Records** – number of entries removed to meet privacy criteria;
- **DM** – a penalty score for indistinguishability due to generalization;
- **NCP** – how much information was lost during anonymization;

The results highlight the expected privacy–utility trade-off:

Table 6.4: Utility metrics for different anonymization configurations

Configuration	**AECS**	**Suppressed Records**	**Discernibility**	**NCP**
$k = 5, t = 0.8$	20.28	110	0.709	0.469
$k = 10, t = 0.5$	48.40	233	0.448	0.284
$k = 10, t = 0.4$	112.50	250	0.353	0.263

- The configuration $k = 5$, $t = 0.8$ had the least suppression (110 records) and the smallest equivalence class sizes (AECS = 20.28), both of which support higher data usability. However, it also showed the highest information loss according to the Discernibility Metric (0.709) and Normalized Certainty Penalty (0.469) and resulted in the highest re-identification risk (4.93 %).
- The configuration $k = 10, t = 0.5$ achieved a lower re-identification risk (2.07 %) and exhibited a trade-off between increased suppression (233 records) and improved utility metrics, such as a lower NCP (0.284) and Discernibility Metric (0.448), compared to the more relaxed configuration.
- The strictest configuration ($k = 10, t = 0.4$) achieved the strongest privacy guarantees (risk = 0.89 %), but at the cost of suppressing 250 records (more than half the dataset) and producing very large equivalence classes (AECS = 112.5), which improves anonymity but reduces data granularity for fine-grained analyses. Interestingly, it also yielded the lowest NCP (0.263) and Discernibility Metric (0.353), likely due to aggressive suppression and efficient generalization.

These results suggest that choosing the appropriate anonymization strategy depends on the acceptable level of privacy risk and the analytical needs of the data consumer. The platform effectively supports such decision-making by quantifying the effects of each configuration across a spectrum of metrics.

6.3.7 Privacy–Utility Trade-off Analysis

To better understand the impact of varying anonymization parameters, we analyzed how privacy and utility metrics evolve as the t-closeness constraint is adjusted under a fixed $k = 5$. The following figures illustrate key trends and help interpret the trade-offs involved.

Figure 6.1 shows that as t increases, prosecutor risk also increases. Specifically, re-identification risk rises from 0.89 % at $t = 0.4$ to 4.93 % at $t = 0.8$, reflecting a relaxation of the privacy constraint. This highlights the direct effect of t-closeness on the adversary's inference success.

Figure 6.2 illustrates how information loss metrics respond to varying t. In this case, both the Normalized Certainty Penalty (NCP) and Discernibility decrease as t decreases, indicating that stricter privacy constraints were associated with improved utility. Notably, NCP drops from 0.469

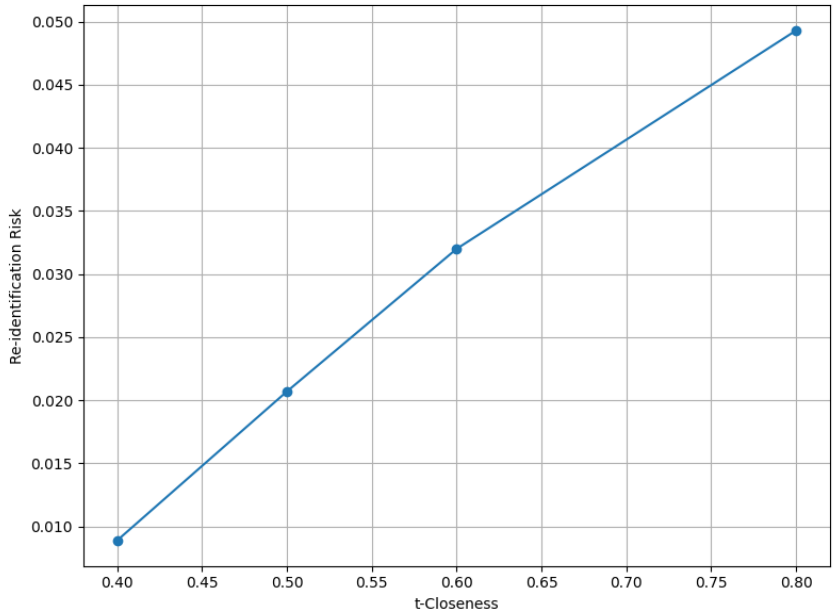

Figure 6.1: Prosecutor risk vs t (Fixed $k = 5$)

to 0.263 as t decreases, suggesting that the utility gain may result from effective generalization and aggressive suppression rather than excessive distortion of quasi-identifiers.

Figure 6.3 shows a decline in suppressed records with increasing t. At $t = 0.4$, 250 records are suppressed, compared to only 110 at $t = 0.8$. This reinforces that stricter privacy guarantees (lower t) require more aggressive data suppression to enforce indistinguishability.

Finally, Figure 6.4 depicts the average equivalence class size (AECS) as a function of t. As privacy constraints loosen, AECS decreases, from 112.5 at $t = 0.4$ to 20.28 at $t = 0.8$,indicating finer-grained groupings and better data granularity.

Together, these results confirm the intuitive trade-off: lower t values strengthen privacy (Figure 6.1) and, somewhat counterintuitively, were also associated with improved utility metrics (Figures 6.2–6.4). This suggests that in this dataset, generalization and suppression were applied in a way that minimized information loss more effectively under stricter privacy constraints.

6.3.8 Discussion

The experimental results presented in this chapter demonstrate the effectiveness and practical viability of the proposed AI-enhanced anonymization platform in both synthetic and real-world healthcare contexts. The integration of large language models for attribute classification, generalization hierarchy suggestion and privacy parameter tuning proved critical in reducing manual workload and improving the consistency of anonymization workflows.

The key findings and insights are summarized as follows:

- **Strengths:** While DeepSeek-r1 32B achieved the highest overall classification accuracy, the DeepSeek-r1 14B model offered the best trade-off between accuracy and inference time.

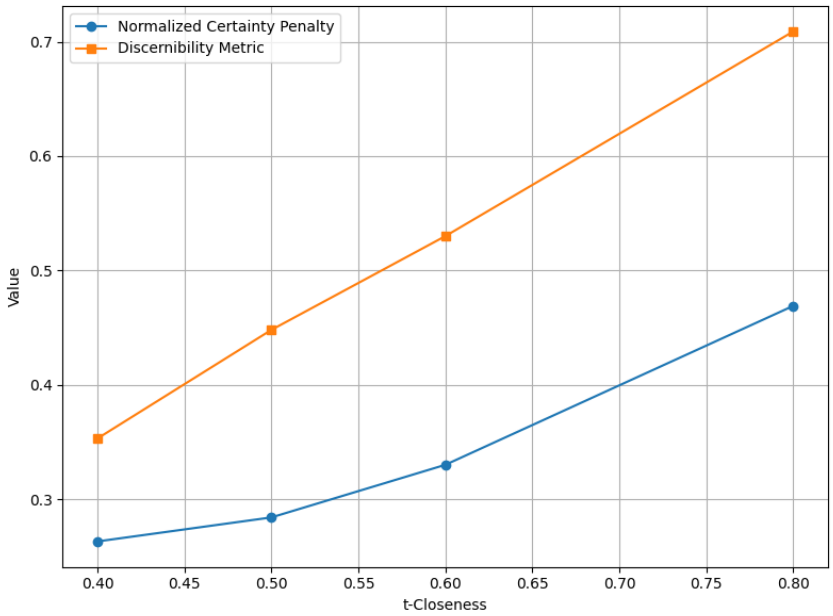

Figure 6.2: Information loss vs t (Fixed $k = 5$)

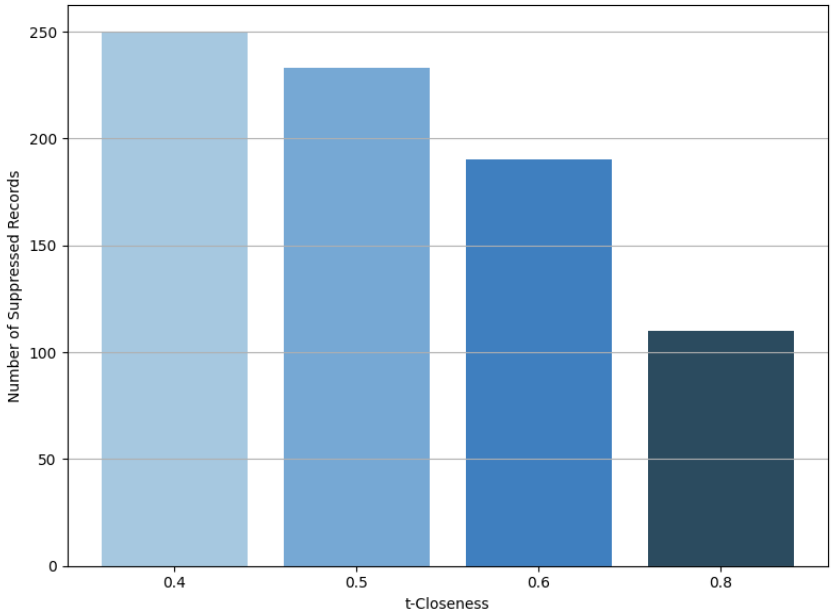

Figure 6.3: Suppressed records vs t (Fixed $k = 5$)

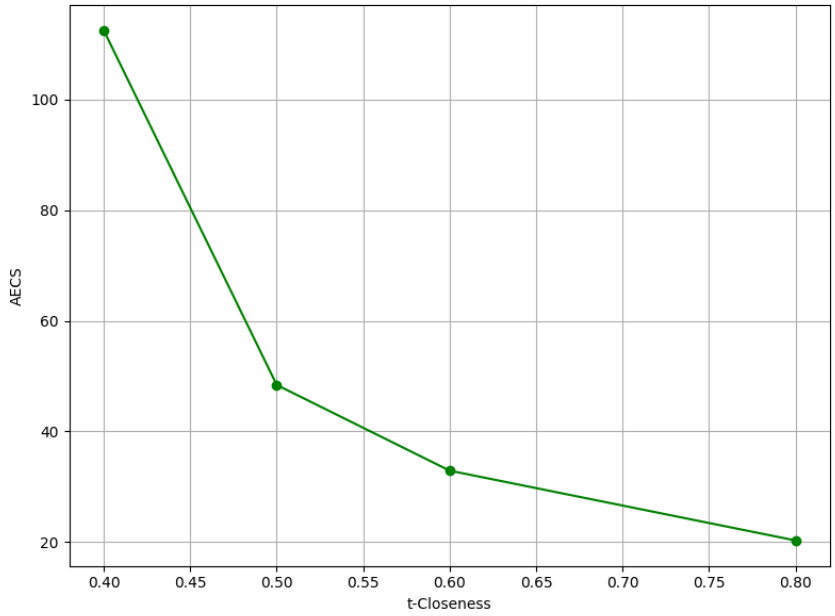

Figure 6.4: Average equivalence class size vs t (Fixed $k = 5$)

This made it the most suitable option for local deployment in structured healthcare datasets. The platform's modular architecture and visual feedback mechanisms also facilitated rapid experimentation with anonymization configurations and risk–utility adjustments.

- **Challenges:** The synthetic dataset experiments revealed that smaller LLMs (e.g. 1.5B, 7B) struggled with nuanced attribute classification, while larger models introduced latency and resource constraints. In the clinical case study, high dimensionality and limited records complicated the application of stricter privacy models, requiring suppression and careful tuning.

- **Implications:** The results validate the feasibility of embedding AI into anonymization pipelines to support non-expert users in sensitive data environments. Local inference ensured regulatory compliance while maintaining operational flexibility. However, effective deployment requires careful model selection, hardware planning and clear interpretability of AI-generated recommendations.

- **Future Work:** Promising directions include fine-tuning language models on structured medical metadata, improving attribute batching and prompt engineering and developing active-learning workflows that adapt privacy strategies based on ongoing utility feedback.

Taken together, these findings highlight how AI can assist in navigating the complex trade-offs between data privacy and utility in healthcare data-sharing environments.

6.4 Limitations

While the platform demonstrated promising results in controlled and real-world evaluations, several limitations must be acknowledged:

- **Lack of Fine-Tuning for AI Tasks:** The large language models used in the platform were not fine-tuned for the specific tasks of attribute classification, hierarchy generation or privacy model parameter suggestion. While prompt engineering provided acceptable results, the absence of domain-specific fine-tuning may limit accuracy and consistency, especially when handling ambiguous or complex healthcare attributes.
- **Token and Batch Constraints:** Due to token limitations in local LLM inference, large datasets required manual or automated batching of attributes. This adds complexity and may introduce inconsistencies or classification drift across batches.
- **Limited Dataset Size:** The real-world dataset used in the case study, while valuable, had a limited number of records (475) compared to its dimensionality (138 attributes). This imbalance constrained the applicability of high-strength privacy models and increased the risk of information loss or excessive record suppression.
- **Hardware Constraints:** Running large language models locally introduces substantial hardware demands. Models such as DeepSeek require high-performance GPUs and large memory footprints to operate efficiently. This limits accessibility for organizations without advanced computing infrastructure and may hinder real-time responsiveness in resource-constrained environments.

These limitations provide avenues for future research and platform enhancement, particularly in supporting broader dataset types, scaling LLM inference more effectively and validating usability across diverse clinical settings.

6.5 Conclusion

This chapter presented a comprehensive evaluation of the proposed AI-assisted anonymization platform through both synthetic benchmarks and a real-world case study. The experiments demonstrated that large language models, particularly DeepSeek-r1 14B, are capable of accurately supporting key privacy engineering tasks such as attribute classification within structured healthcare datasets while maintaining acceptable inference times under local deployment.

The synthetic dataset experiment revealed the trade-offs between model size, accuracy and inference latency. While DeepSeek-r1 32B achieved the highest classification accuracy, it required substantially more computational resources. In contrast, DeepSeek-r1 14B provided the best overall balance between performance and efficiency, making it the most practical candidate for integration into privacy-sensitive environments with limited hardware availability.

The real-world case study using clinical breast cancer data from CHUSJ further validated the platform's applicability in operational settings. Through the use of parameterized anonymization strategies (e.g., k-anonymity and t-closeness), the platform enabled users to tune privacy protection levels against data utility needs. The analysis of privacy risk and utility metrics across multiple configurations confirmed the expected trade-offs and showed that aggressive generalization and suppression can, in some cases, reduce uncertainty and improve utility, particularly in high-dimensional and small sample datasets.

Overall, the results confirm that accurate, AI-assisted attribute classification is achievable using large language models, even when operating entirely within local environments. Embedding LLMs into end-to-end anonymization workflows reduces manual burden, enhances reproducibility and enables non-expert users to manage sensitive data more confidently. The evaluation also highlighted areas for further development, including domain-specific fine-tuning, scalability enhancemen

Chapter 7

Conclusion

This work developed an AI-enhanced anonymization platform designed to support secure, compliant and efficient sharing of structured healthcare data. By integrating locally deployed large language models with a modular anonymization pipeline, the platform enables data providers to apply privacy-preserving transformations while maintaining analytical utility. Through a combination of architectural design, AI-assisted classification and real-time privacy–utility evaluation, the system addresses critical challenges in balancing data privacy with usability in clinical research and data-sharing environments.

7.1 Summary of Contributions

The key contributions of this work are as follows:

- **Design and Implementation of a Modular Architecture:** A scalable, containerized platform integrating Opal, ARX and Ollama was developed to support end-to-end anonymization workflows for structured health datasets.
- **Integration of Local LLMs:** The system uses locally deployed large language models to automate attribute classification, generalization hierarchy suggestion and privacy model configuration enabling privacy-preserving processing without external cloud dependencies.
- **Evaluation on Synthetic and Real Data:** The platform was evaluated using both a synthetic healthcare dataset and a real-world dataset from CHUSJ. Results demonstrated high classification accuracy, significant reductions in re-identification risk and the ability to manage privacy–utility trade-offs in real-world scenarios.
- **Risk–Utility Visualizations and Feedback:** The platform incorporates explainable metrics such as Prosecutor Risk, AECS, NCP and Number of Suppressed Records to support informed decision-making by non-expert users.

7.2 Key Findings

Experimental results validated that LLMs can meaningfully assist in automating complex anonymization decisions. DeepSeek-14B emerged as a strong balance between performance and resource usage, offering reliable results across various anonymization tasks.

The modular architecture, visual feedback and AI-guided support reduced the manual burden of privacy configuration, demonstrating the feasibility of integrating AI into sensitive health data workflows.

7.3 Future Work

Future work may address the previously identified limitations in Section 6.4 by:

- Fine-tuning lightweight domain-specific language models for structured health data to improve recommendations.
- Optimizing model inference pipelines and batching strategies to reduce latency and streamline workflows.
- Expanding evaluations across a wider range of datasets and healthcare contexts to test generalizability and scalability.

7.4 Final Remarks

As healthcare data becomes increasingly essential for research and innovation, ensuring privacy without compromising analytical value is a critical challenge. This work shows that by combining AI-driven insights with privacy-by-design principles, it is possible to build anonymization tools that enable data providers to share sensitive information responsibly and effectively. The platform lays a foundation for future research and deployment of privacy-aware data ecosystems in healthcare and beyond.

References

[1] Bilal Abu-Salih et al. “Introduction to Big Data Technology”. In: Mar. 2021, pp. 15–59. ISBN: 978-981-33-6651-0. DOI: 10.1007/978-981-33-6652-7_2.

[2] Ahmad Alzu’bi et al. “A Review of Privacy and Security of Edge Computing in Smart Healthcare Systems: Issues, Challenges, and Research Directions”. In: *Tsinghua Science and Technology* 29 (4 Aug. 2024), pp. 1152–1180. ISSN: 1007-0214. DOI: 10.26599/TST.2023.9010080.

[3] Dimitris Asimopoulos et al. “Benchmarking Advanced Text Anonymisation Methods: A Comparative Study on Novel and Traditional Approaches”. In: *2024 13th International Conference on Modern Circuits and Systems Technologies (MOCAST)*. IEEE, June 2024, pp. 1–6. ISBN: 979-8-3503-8542-7. DOI: 10.1109/MOCAST61810.2024.10615642.

[4] Cristian Augusto et al. “Test-Driven Anonymization for Artificial Intelligence”. In: *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*. IEEE, Apr. 2019, pp. 103–110. ISBN: 978-1-7281-0492-8. DOI: 10.1109/AITest.2019.00011.

[5] Cristian Augusto et al. “Test-Driven Anonymization in Health Data: A Case Study on Assistive Reproduction”. In: *2020 IEEE International Conference On Artificial Intelligence Testing (AITest)*. IEEE, Aug. 2020, pp. 81–82. ISBN: 978-1-7281-6984-2. DOI: 10.1109/AITEST49225.2020.00019.

[6] Amsaveni Avinashiappan and Bharathi Mayilsamy. “Internet of Medical Things: Security Threats, Security Challenges, and Potential Solutions”. In: *Internet of Medical Things: Remote Healthcare Systems and Applications*. Ed. by D. Jude Hemanth, J. Anitha, and George A. Tsihrintzis. Cham: Springer International Publishing, 2021, pp. 1–16. ISBN: 978-3-030-63937-2. DOI: 10.1007/978-3-030-63937-2_1. URL: https://doi.org/10.1007/978-3-030-63937-2_1.

[7] P.L.M Kelani Bandara, HMN Dilum Bandara, and Shantha Fernando. “Evaluation of Re-identification Risks in Data Anonymization Techniques Based on Population Uniqueness”. In: *2020 5th International Conference on Information Technology Research (ICITR)*. IEEE, Dec. 2020, pp. 1–5. ISBN: 978-1-6654-1475-3. DOI: 10.1109/ICITR51448.2020.9310884.

[8] Bruce A Beckwith et al. “Development and evaluation of an open source software tool for deidentification of pathology reports”. In: *BMC Medical Informatics and Decision Making* 6 (1 Dec. 2006), p. 12. ISSN: 1472-6947. DOI: 10.1186/1472-6947-6-12.

[9] Raffael Bild, Klaus A. Kuhn, and Fabian Prasser. “SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees”. In: *Proceedings on Privacy Enhancing Technologies* 2018 (1 Jan. 2018), pp. 67–87. ISSN: 2299-0984. DOI: 10.1515/popets-2018-0004.

[10] Rihab Boussada, Mariem Bouchaala, and Leila Azouz Saidane. "Privacy and Tracking in the Emerging Mobile Applications: A Survey". In: *2023 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, June 2023, pp. 1491–1496. ISBN: 979-8-3503-3339-8. DOI: 10.1109/IWCMC58020.2023.10182497.

[11] Bianca Buff, Joschka Kersting, and Michaela Geierhos. "Detection of Privacy Disclosure in the Medical Domain: A Survey". In: *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*. SCITEPRESS - Science and Technology Publications, 2020, pp. 630–637. ISBN: 978-989-758-397-1. DOI: 10.5220/0009347506300637.

[12] Fer Carmona, Jordi Conesa, and Jordi Casas-Roma. "Towards the Analysis of How Anonymization Affects Usefulness of Health Data in the Context of Machine Learning". In: *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, June 2019, pp. 604–608. ISBN: 978-1-7281-2286-1. DOI: 10.1109/CBMS.2019.00126.

[13] Loredana Caruccio et al. "A decision-support framework for data anonymization with application to machine learning processes". In: *Information Sciences* 613 (Sept. 2022). DOI: 10.1016/j.ins.2022.09.004.

[14] Raphaël Chevrier et al. "Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review". In: *Journal of Medical Internet Research* 21 (5 May 2019), e13484. ISSN: 1438-8871. DOI: 10.2196/13484.

[15] J. Cusick. *The General Data Protection Regulation (GDPR): What Organizations Need to Know*. CT Corporation Resource Center. 2018. URL: http://refhub.elsevier.com/S2772-4425(23)00059-X/sb8.

[16] Paula Delgado-Santos et al. "A Survey of Privacy Vulnerabilities of Mobile Device Sensors". In: *ACM Computing Surveys* 54 (11s Jan. 2022), pp. 1–30. ISSN: 0360-0300. DOI: 10.1145/3510579.

[17] Gaurav Dhiman et al. "Federated Learning Approach to Protect Healthcare Data over Big Data Scenario". In: *Sustainability* 14 (5 Feb. 2022), p. 2500. ISSN: 2071-1050. DOI: 10.3390/su14052500.

[18] Cynthia Dwork. "Differential Privacy". In: *Automata, Languages and Programming*. Ed. by Michele Bugliesi et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12. ISBN: 978-3-540-35908-1.

[19] Peter F. Edemekong et al. *Health Insurance Portability and Accountability Act (HIPAA) Compliance*. StatPearls [Internet]. Updated 2024 Nov 24. Available from: https://www.ncbi.nlm.nih.gov/books/NBK500019/. Treasure Island (FL): StatPearls Publishing, 2025.

[20] Femi.A. Elegbeleye et al. "Data Privacy on using Four Models- A Review". In: *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*. IEEE, July 2022, pp. 1–9. ISBN: 978-1-6654-7087-2. DOI: 10.1109/ICECET55527.2022.9872999.

[21] Khaled El Emam et al. "A Systematic Review of Re-Identification Attacks on Health Data". In: *PLoS ONE* 6 (12 Dec. 2011), e28071. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0028071.

[22] Joseph Ficek et al. “Differential privacy in health research: A scoping review”. In: *Journal of the American Medical Informatics Association* 28 (10 Sept. 2021), pp. 2269–2276. ISSN: 1527-974X. DOI: 10.1093/jamia/ocab135.

[23] F. J. Friedlin and C. J. McDonald. “A Software Tool for Removing Patient Identifying Information from Clinical Documents”. In: *Journal of the American Medical Informatics Association* 15 (5 Sept. 2008), pp. 601–610. ISSN: 1067-5027. DOI: 10.1197/jamia.M2702.

[24] Laçi Hafsa and Sevrani Kozeta. “Preserving privacy in medical images while still enabling AI-driven research: A comprehensive review”. In: *2024 13th Mediterranean Conference on Embedded Computing (MECO)*. IEEE, June 2024, pp. 1–5. ISBN: 979-8-3503-8756-8. DOI: 10.1109/MECO62516.2024.10577795.

[25] Shilan S. Hameed et al. “A systematic review of security and privacy issues in the internet of medical things; the role of machine learning approaches”. In: *PeerJ Computer Science* 7 (Mar. 2021), e414. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.414.

[26] Madhuri Hiwale et al. “A systematic review of privacy-preserving methods deployed with blockchain and federated learning for the telemedicine”. In: *Healthcare Analytics* 3 (Nov. 2023), p. 100192. ISSN: 27724425. DOI: 10.1016/j.health.2023.100192.

[27] HL7 International. *FHIR Release 5: Fast Healthcare Interoperability Resources*. https://www.hl7.org/fhir/. Accessed: 2025-06-27. 2023.

[28] Muhammad Akbar Husnoo et al. “Differential Privacy for IoT-Enabled Critical Infrastructure: A Comprehensive Survey”. In: *IEEE Access* 9 (2021), pp. 153276–153304. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3124309.

[29] Aryan Jadon and Shashank Kumar. “Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy”. In: *2023 International Conference on Smart Applications, Communications and Networking (SmartNets)*. IEEE, July 2023, pp. 1–4. ISBN: 979-8-3503-0252-3. DOI: 10.1109/SmartNets58706.2023.10215825.

[30] Meng Jin et al. “A Hybrid Machine Learning Method for the De-identification of Un-Structured Narrative Clinical Text in Multi-center Chinese Electronic Medical Records Data”. In: *2019 IEEE International Conference on Big Knowledge (ICBK)*. IEEE, Nov. 2019, pp. 115–121. ISBN: 978-1-7281-4607-2. DOI: 10.1109/ICBK.2019.00023.

[31] Alistair E.W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. “Deidentification of free-text medical records using pre-trained bidirectional transformers”. In: *ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning*. Association for Computing Machinery, Inc, Feb. 2020, pp. 214–221. ISBN: 9781450370462. DOI: 10.1145/3368555.3384455.

[32] Jipmin Jung et al. “A Determination Scheme for Quasi-Identifiers Using Uniqueness and Influence for De-Identification of Clinical Data”. In: *Journal of Medical Imaging and Health Informatics* 10 (2 Feb. 2020), pp. 295–303. ISSN: 2156-7018. DOI: 10.1166/jmihi.2020.2966.

[33] Nazish Khalid et al. “Privacy-preserving artificial intelligence in healthcare: Techniques and applications”. In: *Computers in Biology and Medicine* 158 (May 2023), p. 106848. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2023.106848.

[34] Clete A. Kushida et al. “Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies”. In: *Medical Care* 50 (July 2012), S82–S101. ISSN: 0025-7079. DOI: 10.1097/MLR.0b013e3182585355.

[35] Vjeko Kuzina, Eugen Vusak, and Alan Jovic. “Methods for Automatic Sensitive Data Detection in Large Datasets: a Review”. In: *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*. IEEE, Sept. 2021, pp. 187–192. ISBN: 978-953-233-101-1. DOI: 10.23919/MIPRO52101.2021.9596735.

[36] Vjeko Kužina et al. “CASSED: Context-based Approach for Structured Sensitive Data Detection”. In: *Expert Systems with Applications* 223 (Aug. 2023), p. 119924. ISSN: 09574174. DOI: 10.1016/j.eswa.2023.119924.

[37] Vjeko Kužina et al. “CASSED: Context-based Approach for Structured Sensitive Data Detection”. In: *Expert Systems with Applications* 223 (Aug. 2023), p. 119924. ISSN: 09574174. DOI: 10.1016/j.eswa.2023.119924.

[38] Jiraphat Lapwattanaworakul, Chetneti Srisa-An, and Supanit Angsirikul. “Guideline for Data Anonymization for Data Privacy in Thailand”. In: *2022 6th International Conference on Information Technology (InCIT)*. IEEE, Nov. 2022, pp. 211–215. ISBN: 978-1-6654-8912-6. DOI: 10.1109/InCIT56086.2022.10067859.

[39] You Qian Lee et al. “Unlocking the Secrets Behind Advanced Artificial Intelligence Language Models in Deidentifying Chinese-English Mixed Clinical Text: Development and Validation Study”. In: *Journal of Medical Internet Research* 26 (1 2024). ISSN: 14388871. DOI: 10.2196/48443.

[40] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity”. In: *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, Apr. 2007, pp. 106–115. ISBN: 1-4244-0802-4. DOI: 10.1109/ICDE.2007.367856.

[41] WeiKang Liu et al. “A Survey on Differential Privacy for Medical Data Analysis”. In: *Annals of Data Science* 11 (2 Apr. 2024), pp. 733–747. ISSN: 2198-5804. DOI: 10.1007/s40745-023-00475-3.

[42] Zhengliang Liu et al. “DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4”. In: (Mar. 2023). URL: http://arxiv.org/abs/2303.11032.

[43] A. Machanavajjhala et al. “L-diversity: privacy beyond k-anonymity”. In: *22nd International Conference on Data Engineering (ICDE’06)*. IEEE, 2006, pp. 24–24. ISBN: 0-7695-2570-9. DOI: 10.1109/ICDE.2006.1.

[44] Abdul Majeed and Sungchang Lee. “Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey”. In: *IEEE Access* 9 (2021), pp. 8512–8545. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3045700.

[45] Blaz Meden et al. “Privacy–Enhancing Face Biometrics: A Comprehensive Survey”. In: *IEEE Transactions on Information Forensics and Security* 16 (2021), pp. 4147–4183. ISSN: 1556-6013. DOI: 10.1109/TIFS.2021.3096024.

[46] Fatemeh Mosaiyebzadeh et al. “Privacy-Enhancing Technologies in Federated Learning for the Internet of Healthcare Things: A Survey”. In: *Electronics* 12 (12 June 2023), p. 2703. ISSN: 2079-9292. DOI: 10.3390/electronics12122703.

[47] Mary Nankya et al. “Security and Privacy in E-Health Systems: A Review of AI and Machine Learning Techniques”. In: *IEEE Access* 12 (2024), pp. 148796–148816. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3469215.

[48] Gary S. Nelson. "Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification". In: *SAS Global Forum Proceedings*. Apr. 2015, pp. 1–23.

[49] Gregory S Nelson. *Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification*.

[50] Gregory S Nelson. "Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification". In: *SAS global forum proceedings*. 2015, pp. 1–23.

[51] Leysan Nurgalieva, David O'Callaghan, and Gavin Doherty. "Security and Privacy of mHealth Applications: A Scoping Review". In: *IEEE Access* 8 (2020), pp. 104247–104268. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2999934.

[52] Iyiola E. Olatunji et al. "A Review of Anonymization for Healthcare Data". In: *Big Data* (Mar. 2022). ISSN: 2167-6461. DOI: 10.1089/big.2021.0169.

[53] J. Andrew Onesimu et al. "Privacy Preserving Attribute-Focused Anonymization Scheme for Healthcare Data Publishing". In: *IEEE Access* 10 (2022), pp. 86979–86997. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3199433.

[54] David Pissarra et al. "Unlocking the Potential of Large Language Models for Clinical Text Anonymization: A Comparative Study". In: (May 2024). URL: http://arxiv.org/abs/2406.00062.

[55] Fabian Prasser et al. "ARX–A Comprehensive Tool for Anonymizing Biomedical Data". In: *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2014 (Nov. 2014), pp. 984–93.

[56] Emanuele Raso et al. "Anonymization and Pseudonymization of FHIR Resources for Secondary Use of Healthcare Data". In: *IEEE Access* 12 (2024), pp. 44929–44939. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3381034.

[57] Ritu Ratra, Preeti Gulia, and Nasib Singh Gill. "Evaluation of Re-identification Risk using Anonymization and Differential Privacy in Healthcare". In: *International Journal of Advanced Computer Science and Applications* 13 (2 2022). ISSN: 21565570. DOI: 10.14569/IJACSA.2022.0130266.

[58] Jared Romeo et al. "Privacy-Preserving Machine Learning for E-Health Applications: A Survey". In: *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*. IEEE, Apr. 2024, pp. 1–6. ISBN: 979-8-3503-7297-7. DOI: 10.1109/ICMI60790.2024.10586115.

[59] Pierangela Samarati and Latanya Sweeney. "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression". In: 1998. URL: https://api.semanticscholar.org/CorpusID:2181340.

[60] Martin Scaiano et al. "A unified framework for evaluating the risk of re-identification of text de-identification tools". In: *Journal of Biomedical Informatics* 63 (Oct. 2016), pp. 174–183. ISSN: 15320464. DOI: 10.1016/j.jbi.2016.07.015.

[61] Navoda Senavirathne and Vicenç Torra. "On the Role of Data Anonymization in Machine Learning Privacy". In: *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, Dec. 2020, pp. 664–675. ISBN: 978-1-6654-0392-4. DOI: 10.1109/TrustCom50675.2020.00093.

[62] Sanjeet Singh et al. "Generation and De-Identification of Indian Clinical Discharge Summaries using LLMs". In: (July 2024). URL: http://arxiv.org/abs/2407.05887.

[63] Robin Staab et al. “Large Language Models are Advanced Anonymizers”. In: (Feb. 2024). URL: http://arxiv.org/abs/2402.13846.

[64] Latanya Sweeney. “k-anonymity: a model for protecting privacy”. In: *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10.5 (Oct. 2002), pp. 557–570. ISSN: 0218-4885. DOI: 10.1142/S0218488502001648. URL: https://doi.org/10.1142/S0218488502001648.

[65] Rodrigo Tertulino, Nuno Antunes, and Higor Morais. “Privacy in electronic health records: a systematic mapping study”. In: *Journal of Public Health* (Jan. 2023). ISSN: 2198-1833. DOI: 10.1007/s10389-022-01795-z.

[66] Chandra Thapa and Seyit Camtepe. “Precision health data: Requirements, challenges and existing techniques for data security and privacy”. In: *Computers in Biology and Medicine* 129 (Feb. 2021), p. 104130. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2020.104130.

[67] Olga Vovk, Gunnar Piho, and Peeter Ross. “Methods and tools for healthcare data anonymization: a literature review”. In: *International Journal of General Systems* 52 (3 Apr. 2023), pp. 326–342. ISSN: 0308-1079. DOI: 10.1080/03081079.2023.2173749.

[68] Yilmaz Vural and Murat Aydos. “A New Approach to Utility-Based Privacy Preserving in Data Publishing”. In: *2017 IEEE International Conference on Computer and Information Technology (CIT)*. IEEE, Aug. 2017, pp. 204–209. ISBN: 978-1-5386-0958-3. DOI: 10.1109/CIT.2017.27.

[69] Yilmaz Vural and Murat Aydos. “A New Approach to Utility-Based Privacy Preserving in Data Publishing”. In: *2017 IEEE International Conference on Computer and Information Technology (CIT)*. IEEE, Aug. 2017, pp. 204–209. ISBN: 978-1-5386-0958-3. DOI: 10.1109/CIT.2017.27.

[70] Gaetan Kamdje Wabo et al. “Data Quality– and Utility-Compliant Anonymization of Common Data Model–Harmonized Electronic Health Record Data: Protocol for a Scoping Review”. In: *JMIR Research Protocols* 12 (Aug. 2023), e46471. ISSN: 1929-0748. DOI: 10.2196/46471.

[71] Steven M. Williamson and Victor Prybutok. “Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare”. In: *Applied Sciences* 14 (2 Jan. 2024), p. 675. ISSN: 2076-3417. DOI: 10.3390/app14020675.

[72] Caesar Wu, Rajkumar Buyya, and Kotagiri Ramamohanarao. “Big Data Analytics = Machine Learning + Cloud Computing”. In: (Jan. 2016). DOI: 10.48550/arXiv.1601.03115.

[73] Dingyi Xiang and Wei Cai. “Privacy Protection and Secondary Use of Health Data: Strategies and Methods”. In: *BioMed Research International* 2021 (Oct. 2021), pp. 1–11. ISSN: 2314-6141. DOI: 10.1155/2021/6967166.

[74] Jinsung Yoon, Lydia N. Drumright, and Mihaela van der Schaar. “Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN)”. In: *IEEE Journal of Biomedical and Health Informatics* 24 (8 Aug. 2020), pp. 2378–2388. ISSN: 2168-2194. DOI: 10.1109/JBHI.2020.2980262.

[75] Athanasios Zigomitros et al. “A Survey on Privacy Properties for Data Publishing of Relational Data”. In: *IEEE Access* 8 (2020), pp. 51071–51099. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2980235.

[76] Mishall Al-Zubaidie, Zhongwei Zhang, and Ji Zhang. "PAX: Using Pseudonymization and Anonymization to Protect Patients' Identities and Data in the Healthcare System". In: *International Journal of Environmental Research and Public Health* 16 (9 Apr. 2019), p. 1490. ISSN: 1660-4601. DOI: 10.3390/ijerph16091490.

[77] Zheming Zuo et al. "Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study". In: *JMIR Medical Informatics* 9 (10 Oct. 2021), e29871. ISSN: 2291-9694. DOI: 10.2196/29871.

Appendix A

API Documentation

A.1 Anonymization Endpoints

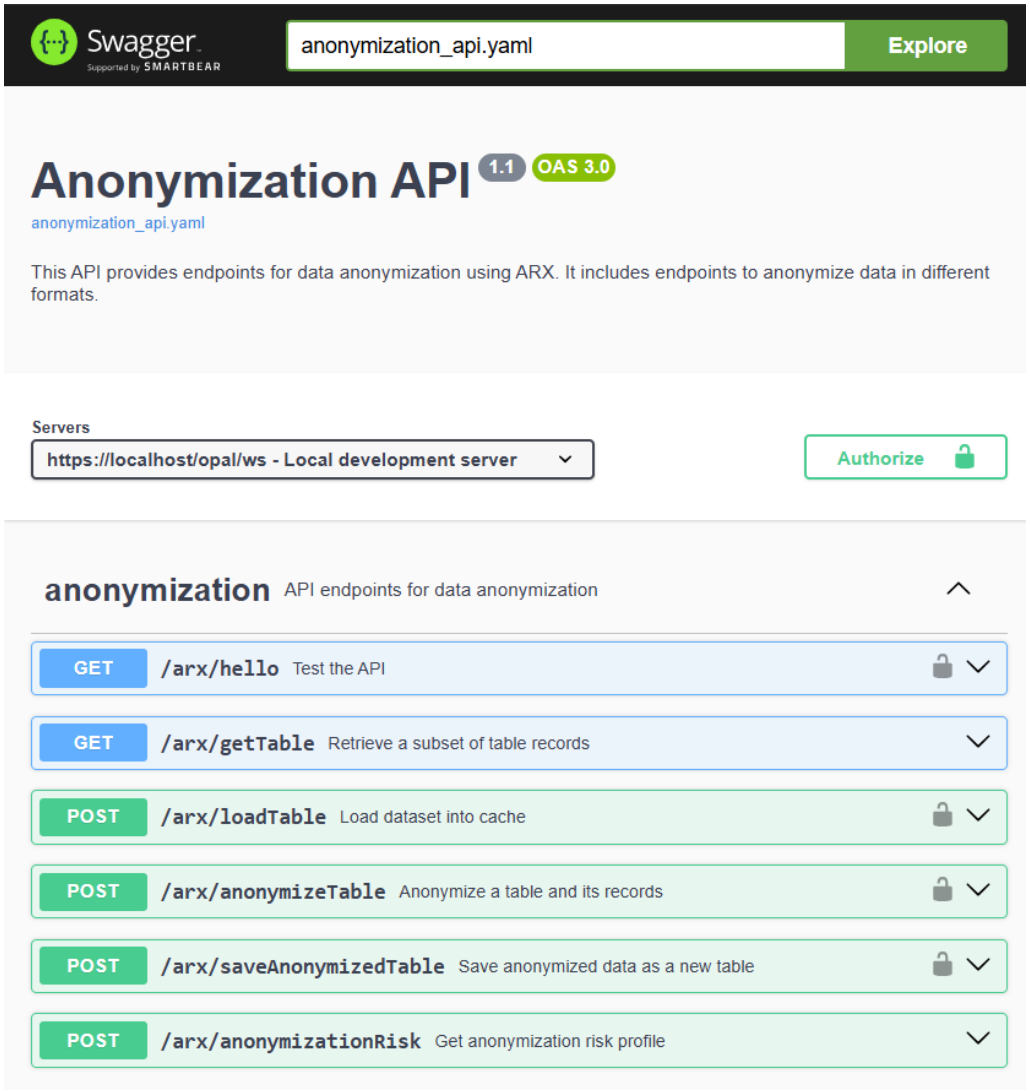

Figure A.1: Endpoints for loading data, anonymizing tables and computing risk

A.2 Hierarchy Generation Endpoints

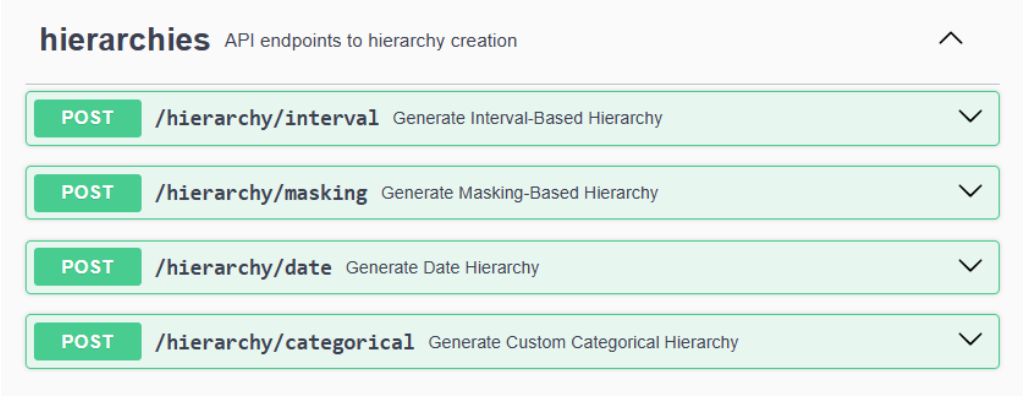

Figure A.2: Endpoints for generating interval, masking, date and categorical hierarchies

A.3 AI-Assisted Endpoints

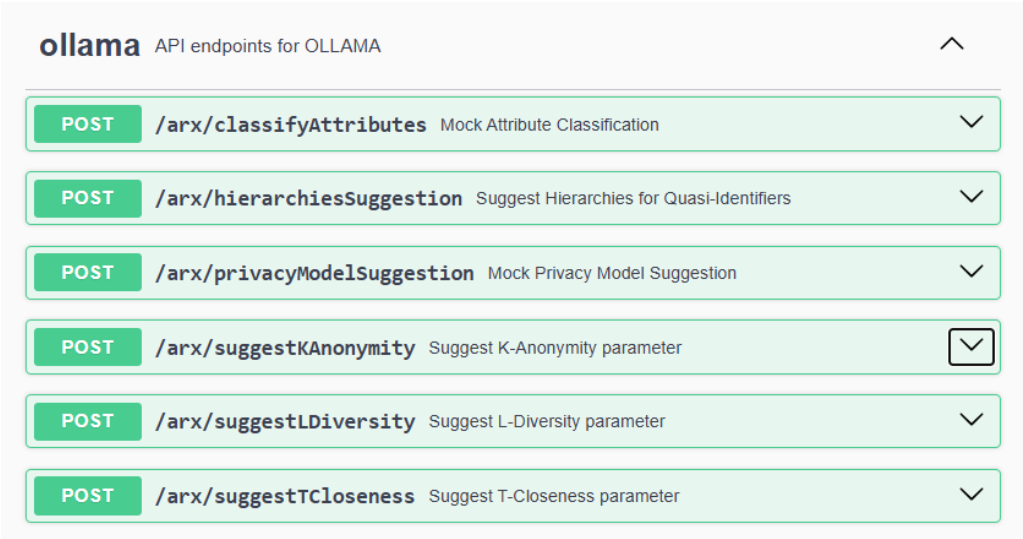

Figure A.3: Endpoints for attribute classification, hierarchy suggestion and privacy parameter tuning using LLMs

A.4 Logging Endpoints

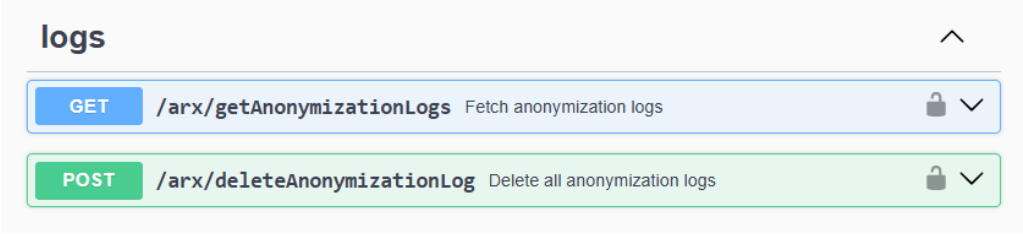

Figure A.4: Endpoints for retrieving and deleting anonymization logs

A.5 Schema Definitions

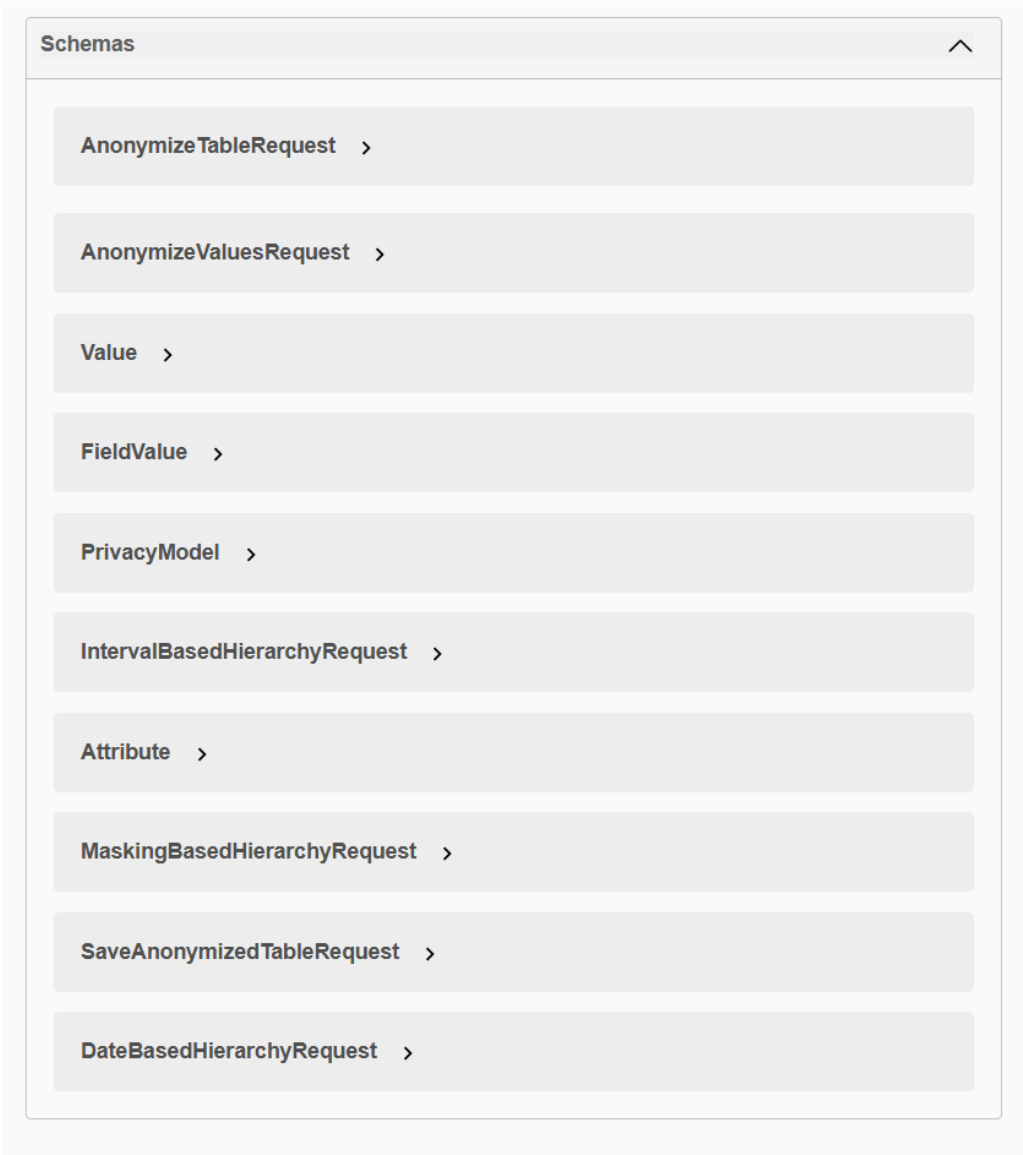

Figure A.5: List of reusable schema definitions in the OpenAPI specification