Enhancing Healthcare Data Privacy with Client-Side Anonymization: A Lightweight, Rule-Based Architecture

Adeeba Imtiyaz[1], Dr Pertik Garg[2]
[1]*Student, Department of Computer Science and Engineering, Swami Vivekanand Institute of Engineering & Technology, Ramnagar, Banur, Punjab, India*
[2]*Professor, Department of Computer Science and Engineering, Swami Vivekanand Institute of Engineering & Technology, Ramnagar, Banur, Punjab, India*

***Abstract*—As more health-related information is moved to cloud, so are concerns about patient-privacy, even with traditional encryption and access control methods, as sensitive patient details are still at risk of being leaked. Current anonymization methods often involve complicated infrastructure or a lack of visibility into whether privacy is being ensured. This paper proposes a lightweight, rule-based, client-side anonymization system for structured healthcare datasets that anonymizes data locally before it is stored or shared in cloud environments. The system uses keyword-based attribute classification to identify sensitive fields and applies masking, generalization, temporal reduction, and suppression accordingly, leaving non-sensitive attributes unchanged. Anonymization effectiveness is validated using k-anonymity, l-diversity, a custom information-loss metric, and analytically modeled record-linkage attacks under varying attacker knowledge levels. Experiments on synthetic healthcare datasets of varying sizes (2000, 5000, and 10000 records) assessed privacy protection, utility preservation, and scalability. Results showed minimum equivalence class size of k=9-45 and Identity Disclosure Risk of 0.64%-1.60%respectively, implying an effective privacy-utility balance with robust resistance to record-linkage attacks even as attacker capability increases, demonstrating that measurable privacy preservation is achievable without computationally intensive infrastructure.**

***Index Terms*—Data anonymization, healthcare data privacy, k-anonymity, l-diversity, Information Loss, re-identification risk, client-side privacy, cloud data sharing.**

I. INTRODUCTION

The rapid growth of digital technologies has sharply increased the volume of sensitive data generated across sectors, with healthcare producing some of the most sensitive information, including patient records, diagnoses, and personal identifiers [1]. Healthcare organizations increasing rely on cloud computing due to its scalability and cost-effectiveness for managing this data [2], but storing and sharing it in cloud environments exposes sensitive information to unauthorized access, breaches, and misuse [3]. Encryption and access control secure data in transit and storage, but offer limited protection once data is decrypted for analysis, leaving it susceptible to identity disclosure [4].

Healthcare information combines personally identifiable information with confidential medical details such as diagnoses, treatment history, laboratory reports, and prescribed medications [5]. Unauthorized disclosure of this information can lead to identity theft, financial fraud, discrimination, and a loss of patient trust [6]. The widespread adoption of Electronic Health Records (EHRs) has further increased the amount of healthcare data being collected and exchanged, making privacy protection an even greater challenge [7].

In response, regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) establish strict requirements for handling sensitive healthcare information [8]. However, issues such as cross-system data sharing, insider threats, cloud misconfigurations, and disclosure attacks continue to expose healthcare data to privacy risks [9].

One of the most challenging privacy issues is the risk of re-identification. One can still be identified with combination of quasi-identifiers (age, gender, and geographic location) with other publically available datasets, even if direct identifiers are anonymized or

removed. This restriction highlights the need of anonymization, achieved using traditional security measures as masking techniques, generalization, and suppression [12] [13].

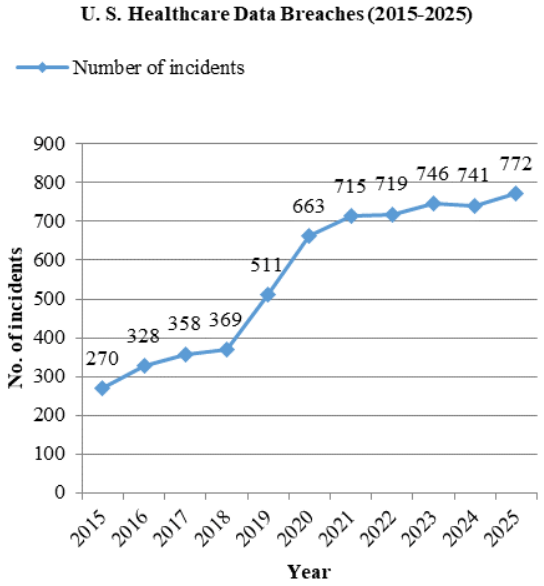

Fig. 1: U.S. Healthcare Data Breaches Reported to HHS OCR (2015-2025). [14]

As the data in Fig. 1 shows, the number of healthcare data breaches has been increasing and measures to protect privacy need to be strengthened.

According to HIPAA Journal, in 2025, over 112 million healthcare records were compromised as 772 incidents of healthcare breaches were reported, as compared to 270 in 2015. Moreover, healthcare has been the most costly industry for past 15 years, with average cost of a data breach reaching $7.42 million in 2025 [15]. These statistics show that being on guard to protect against breaches isn't sufficient. It is also important to minimize the risk of re-identification of an individual when accessing, sharing, and exposure [16] [17].

Apart from these concerns, there are other privacy issues in cloud-based healthcare systems as organizations lose their control over data when sharing it with cloud systems [18]. The risk of privacy breach has further worsened in multi-tenant environments, recurrent cyber-attacks, and temporary access to data during processing [19] [20]. If data is anonymized on the client-side prior to being sent outside of client's infrastructure, only anonymized information will be sent to cloud and therefore have a diminished impact if stolen [21].

To protect sensitive information, a few anonymization techniques have been proposed including masking, generalization, and suppression [22]. There are also some other methods like pseudonymization, perturbation, and differential privacy applied and used for purpose of privacy as per the requirements of intended application [23]. Some of the existing anonymization methods, however, are computationally very intensive, relying on special hardware or lack evidence of anonymity [24] [25]. Data must be anonymized with techniques as such to be re-identification resistant and to achieve data utility along with anonymization.

As a solution to these shortcomings, this paper proposes a lightweight, rule-based client-side anonymization system for healthcare data to anonymize sensitive information before storage and sharing through cloud data services, by masking, generalizing, and suppressing the data, as well as using temporal reduction. There are several existing solutions [26] that do not provide quantitative assessment of effectiveness, but the proposed method quantitatively measures the effectiveness of anonymization techniques using measures of privacy, such as k-anonymity [27], l-diversity [28], Information Loss [29], and Identity Disclosure Risk [30]. The anonymization pipeline is simple, easy to follow along with the inclusion of standardized privacy assessment, providing a practical and transparent way of sharing and keeping healthcare data private.

This work has the following main contributions:

1. A simple and transparent anonymization pipeline without requiring specialized infrastructure.
2. Integration of existing privacy metrics for evaluating anonymization, beyond data transformations techniques.
3. A thorough assessment process, including functional assessment, privacy assessment, and re-identification analysis.

II. LITERATURE SURVEY

Over the years, numerous research studies have been dedicated to the issue of privacy preserving data publishing, with the objective of safeguarding

sensitive data while releasing data that is still useful. While much of the previous literature has centered on how to prevent disclosure of an identity via anonymization, the most recent literature has focused on how to make the data more useful, how to minimize the amount of information lost as a result of anonymization, and how to deal with the new and emerging privacy concerns associated with sharing of healthcare data.

K-anonymity is one of the very first models that were suggested for privacy preserving. It proposed a lattice-based full domain generalization algorithm to generate k-anonymous datasets in an efficient way to lower computational complexity [31]. This approach provided effective protection against identity disclosure but was vulnerable to significant loss of information from full domain generalization. Later, multi-dimensional portioning techniques were developed to provide more flexibility in how data is anonymized while still meets privacy requirements, for improved data utility [32].

K-anonymity was however not sufficient to prevent attribute disclosure in homogeneous equivalence class, therefore (α, k)-anonymity was introduced to limit dominance of sensitive attributes in an equivalence class [33]. The, t-closeness was calculated for each equivalence class, based on the similarity between frequency distribution of sensitive attributes in an equivalence class and in entire dataset, using Earth Mover's Distance metric [34]. This has been found to be more robust against disclosure attacks, but more complicated to compute and has higher information loss.

Further research was carried out on privacy and usefulness of data. To meet privacy requirements while minimizing information loss, several privacy-preserving anonymization techniques based on optimization were developed [35]. Some studies have incorporated non-homogeneous generalization to make generalizations at multiple levels, preventing the needless distortion [36].

Later comparative surveys focused on the lack of a standard method for evaluating different anonymization techniques [37]. Similar comments were noted in the healthcare studies, with concerns related to availability of complimentary evaluation strategies and implementation methods [38].
When it comes to privacy guarantees, comparative studies have always shown that t-closeness provides more guarantees of privacy than k-anonymity and l-diversity, while, it results in higher information loss [39] [40]. For practical implementation, anonymization frameworks have begun to incorporate multiple privacy models, in addition to measures to manage disclosure risk and information loss, however, requiring increased levels of sophistication in their configuration and deployment [41].

The other significant field of interest is risk-based de-identification, where level of anonymization is decided based on the likelihood of re-identification, rather than a set of pre-defined transformation rules [42] [43]. Some works has also compared traditional anonymization techniques to differential privacy, demonstrating that these techniques are easy to implement and reduce complexity but still vulnerable to linkage attacks if information is present. Differential privacy, by contrast, offers more mathematical guarantees of privacy, but comes with an extra computational burden and can have negative impacts on utility of data, especially for data analytics and machine learning applications [44]. However, healthcare data also has additional challenges other than tackled by traditional anonymization models.

The privacy requirements in medical data are generally more complicated than in typical single sensitive attribute data, where there are multiple sensitive attributes which are correlated in the dataset [45]. Additionally, data in healthcare applications are often not static and are evolving over time when stored in the cloud. It has been shown that anonymization rules for an already anonymized set can be violated when new records are appended to the set, and that a full re-anonymization of that set is required, a fact that has led to incremental anonymization approaches that anonymize new records when added to the set [46].

There three trends that are important, overall, from the literature. First, the most prevalent privacy models for structured healthcare data are k-anonymity, l-diversity, and t-closeness, and are commonly evaluated using Identity Disclosure Risk, and Information Loss. Second, differential privacy offers

more convincing theoretical privacy assurances, but implementing it is more complex and in some applications, it might be better to have more powerful assurances.

Third, using several anonymization techniques together is more likely to yield more privacy protection than using one of them. Many current solutions, however, do this by optimizing some algorithms, automating attribute detection, or in-depth anonymization solutions, but introduces implementation complexity, and diminishes transparency [47] [48]. Interestingly, a few studies take anonymization as obligatory at the client-side prior to transmission [49] and, few only, introduce systematic privacy validation in anonymization framework [50].

Based on these observations, it is no doubt that there is a definite gap in research. While masking, generalization, suppression, and temporal reduction are effective anonymization techniques, little research has been dedicated to a lightweight, transparent, and client-side architecture that can be used together to implement any of the techniques or their combinations. In addition, anonymization and privacy evaluation are often considered as two distinct processes, instead of parts of a comprehensive privacy-preserving approach [51].

III. PROPOSED SYSTEM

This section presents the design and implementation of a lightweight, rule-based anonymization system for structured healthcare datasets, performing anonymization locally before any data is stored or transferred.

A. System Overview

Protecting healthcare information requires more than securing communication channels or restricting access to cloud resources. Once data leaves the user's environment, the data owner has limited control over how it is processed or stored. Motivated by this challenge, this work proposes a lightweight client-side anonymization framework that transforms sensitive healthcare data before it is transferred to external storage. Unlike conventional approaches that rely on optimization-based anonymization models, the proposed framework performs privacy protection through predefined transformation rules while using established privacy metrics only to assess the quality of anonymization.

The framework is designed around three principles:

1. Pre-processing of data before uploading to cloud, reducing the exposure of sensitive healthcare information during transmission and cloud storage.
2. The framework's use of simple rule-based transformations instead of optimization-based algorithm reduces system complexity and eases interpretation.
3. The proposed framework is modularly organized, thus making it easy to add additional anonymization techniques and privacy models without re-designing the framework.

B. System Architecture

The proposed framework is a multi-layered framework, independently working on data processing, storage, and user interaction. This separation makes maintenance, extensibility, and future modifications easy without impacting the overall workflow.

The system architecture is organized in the following three layers:

1. The presentation layer provides access to user interface to import healthcare data, process, and review it.
2. The application logic layer carries out all the operations related to anonymization such as identification of attributes, data transformations based on rules, etc.
3. The data storage layer is responsible for persistent storage and only anonymized data is stored in the cloud.

This separation of layers helps to achieve a clear boundary between valuable raw data and externally stored data. The detailed architecture of proposed anonymization system is given in Fig. 2.

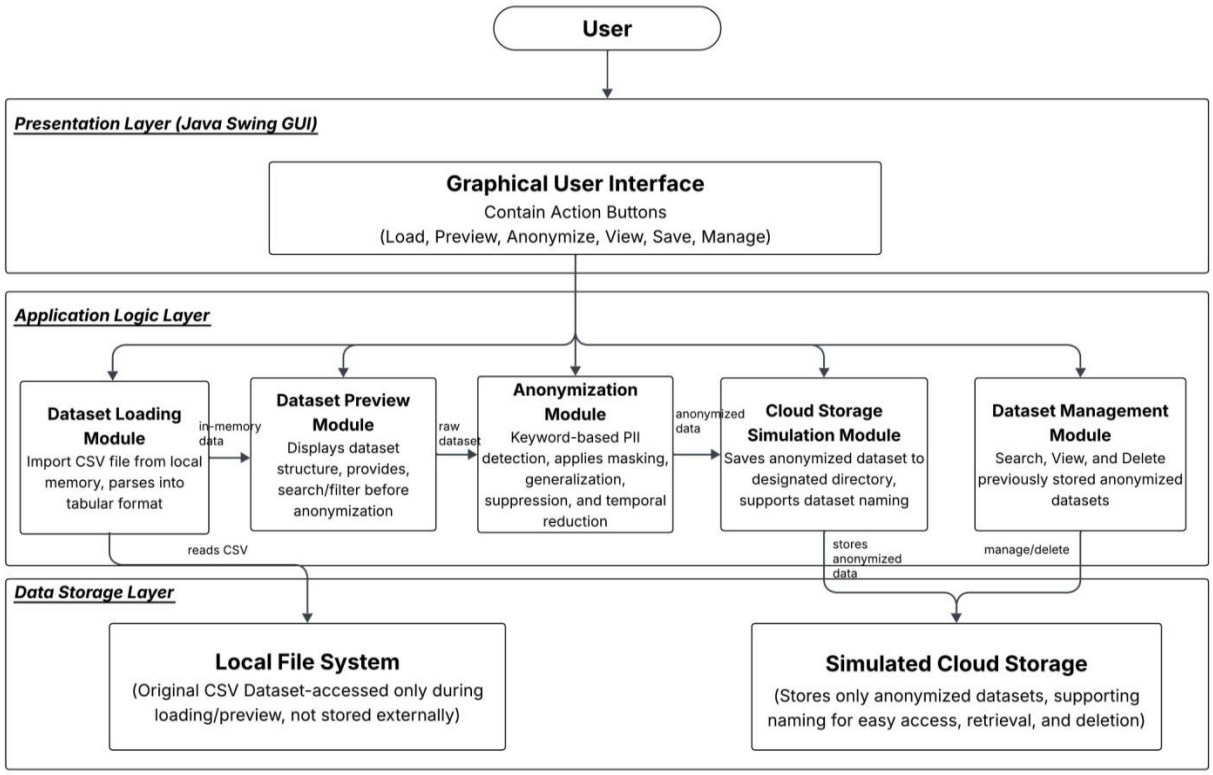

Fig. 2: Multi-Layer Architecture and Component Interaction of Privacy-Preserving Anonymization System.

C. Client-Side Anonymization Framework

The proposed approach processes structured healthcare information in CSV format and computations are performed locally. The dataset is stored in memory and all transformations are performed prior to any storage operations. As a result, there is no patient information that can be identified in the simulated cloud environment, in its original form. Sensitive attributes are detected using simple data-pattern recognition and rule-based inspection of column names.

The framework utilizes simple transformation rules which are easy to understand, replicate, and modify, instead of suing computationally costly statistical models or machine learning methods.

The main part of the proposed framework is based on 4 anonymization techniques:

1. Masking:

It partially hides sensitive information, while leaving a small amount of the original information, so that the information is not directly identifiable without removing the entire attribute.

2. Generalization:

Exact attribute values are generalized, so that the records are not as distinct, but still useful for statistical analysis.

3. Temporal Reduction:

Date and time attributes are abstracted into higher level temporal representations, reduce the amount of temporal disclosure without losing chronological information.

4. Suppression:

Attributes with little analytical value that are very sensitive are removed from the data and filled in with some sort of “placeholder” value, such as “REDACTED”.

D. Privacy-Preserving System Workflow

The proposed framework’s operation is a sequential process as shown in Fig. 3, which is to ensure that privacy protection is enforced prior to external storage.

1. The first step is to upload a healthcare dataset into the application and make it available for preview inspection.

2. Upon user's trigger for anonymization, the framework identifies the sensitive attributes and applies the rules of anonymization.
3. After processing, the anonymized dataset is previewed for verification and archived in the stimulated cloud repository.
4. Only anonymized data is processed, not the original dataset. The original data is only stored during the processing and is deleted after the data is anonymized.

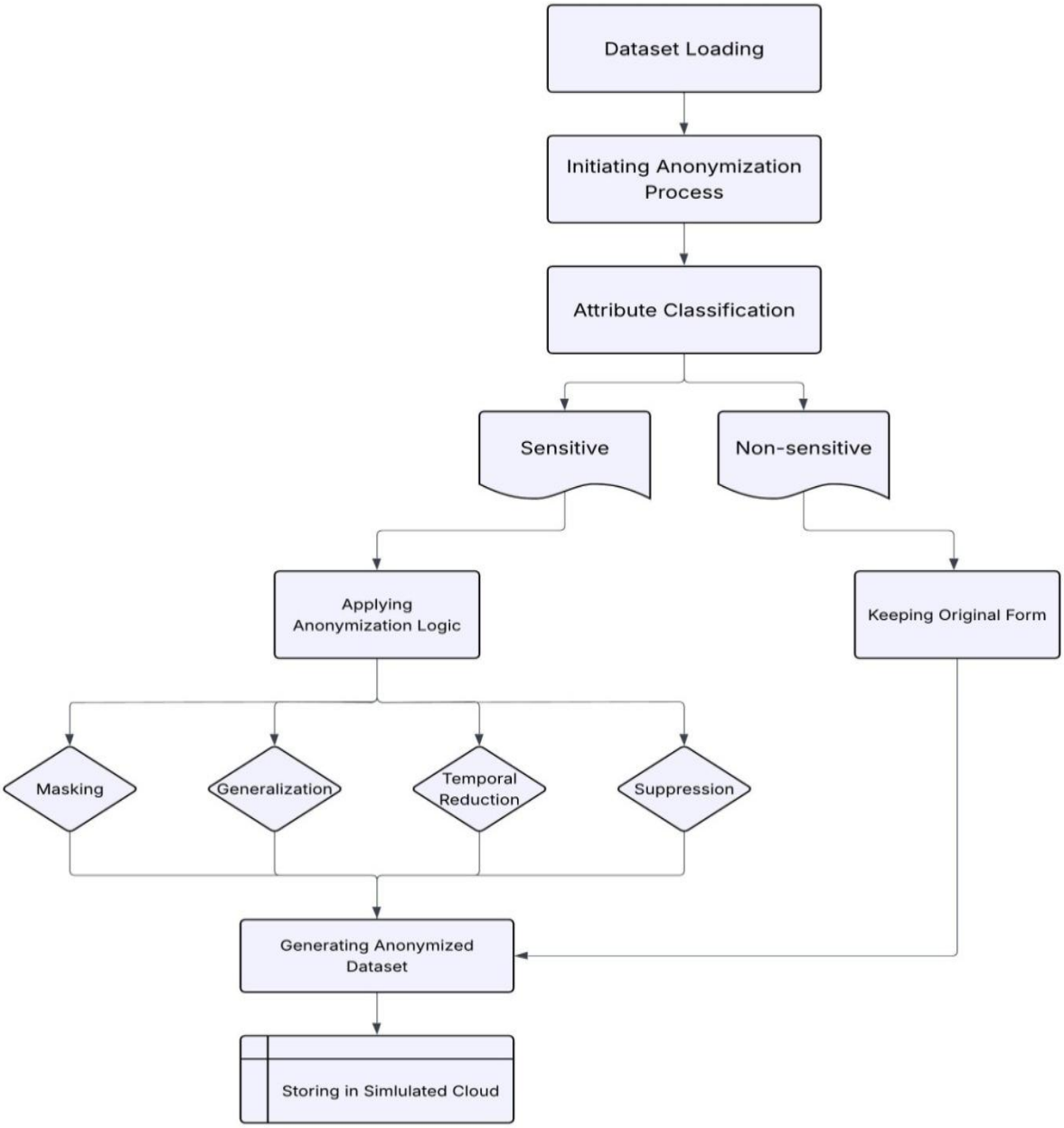

Fig. 3: Workflow of the Proposed Privacy-Preserving Data Anonymization Approach.

IV. EVALUATION DESIGN AND STRATEGY

The proposed anonymization framework was tested to evaluate the system's functional correctness, privacy preserving capability, and effect on the utility of data. The main evaluation criteria focused on determining if the framework could provide reasonable protection for sensitive healthcare data while at the same time being useful to secondary applications. Experiments were carried out in a controlled environment on different

sized datasets and the privacy performance and scalability were measured.

A. Experimental Setup

The following environmental dependencies and tool chains have been employed in the development of this standalone desktop application:

1. Development Setup
a) Programming Language: Java (JDK 8).
b) Integrated Development Environment (IDE): Net Beans IDE 8.0.2.
c) User Interface Framework: Java Swing (with AWT components where required).

2. System Requirements
a) Hardware Requirements: Minimum 4GB RAM.
b) Software Requirements:
1) Java Runtime Environment (JRE 8 or higher).
2) Operating System: Windows 10 (64-bit).

3. Data Requirements
a) File Format Supported: CSV (Comma-Separated Values).
b) Storage Mechanism: Simulate cloud directory (local file-based storage).

B. Experimental Dataset

Experiments were carried out using synthetically generated healthcare datasets containing commonly used patient attributes, including patient name, patient ID, date of birth, gender, phone number, email address, address, diagnosis, etc. Synthetic data was selected to provide a controlled evaluation environment while avoiding privacy concerns associated with real patient records.

Each attribute was assigned an anonymization strategy according to its role within the dataset.

1. Direct identifiers, such as names, phone numbers, and email addresses, were protected through masking.
2. Quasi-identifiers, including age and location, were generalized to reduce re-identification risk.
3. Date and time information was processed using temporal reduction.
4. Highly sensitive attributes with limited analytical significance were suppressed.
5. Attributes that did not require privacy protection were retained without modification.

To examine the scalability of proposed framework, experiments were repeated using datasets containing:

1. 2000 records.
2. 5000 records.
3. 10000 records.

C. Evaluation Strategy and Metrics

The performance of proposed framework was assessed using following complementary evaluation stages:

1. Unit testing:
This was conducted to verify the correctness of individual functional modules, including dataset loading, dataset preview, anonymization, storage simulation, and dataset management.

2. Integration testing:
This evaluated the interaction between these modules by validating the complete processing pipeline, beginning with dataset import and ending with storage of the anonymized dataset.

3. Privacy evaluation:
These tests measured the effectiveness of anonymization process using established privacy and utility metrics, enabling quantitative assessment of disclosure protection and information preservation.

4. Re-identification testing:
These tests modeled record-linkage attacks under different attacker knowledge assumptions to evaluate the framework's resistance to identity disclosure.

The effectiveness of proposed framework was evaluated using the following privacy and utility measures:

1. K-anonymity
K-anonymity measures whether every unique quasi-identifier combination appears in at least k records, thereby reducing the likelihood of identifying an individual record.

$$min_{\left(E \in \frac{D}{QI}\right)}(|E|) \geq k \qquad (1)$$

Where,

D/QI represents the set of equivalence classes formed using quasi-identifiers.
E denotes an individual equivalence class.

2. L-diversity
L-diversity extends k-anonymity by requiring each equivalence class to contain at least l distinct sensitive attribute values, reducing the possibility of attribute disclosure.

$$min_{\left(E \in \frac{D}{QI}\right)}(|S_E|) \geq l \quad (2)$$

Where,
S_E represents the set of unique sensitive values contained within equivalence class *E*.
l = user-defined positive integer ($l \leq k$).

3. Information Loss
Information loss quantifies the reduction in data utility resulting from anonymization. Information Loss in this study is quantified using weighted retention-based metric, conceptually adapted from the Precision metric proposed in [52]. Since the formulation [52] assumes a multi-level generalization hierarchy for each quasi-identifier, whereas the anonymization techniques implemented in this study apply a single-step transformation per attribute, a simplified weighted retention formula was adopted instead as follows:

$$IL = 100 - Retained \quad (3)$$

Where:

$$Retained = \frac{(N_{unchanged} \cdot 1.0 + N_{generalized} \cdot 0.7 + N_{masked} \cdot 0.5 + N_{suppressed} \cdot 0.0)}{N_{total}} \times 100 \quad (4)$$

Where:
N_{total}= total fields (total rows × total columns).
$N_{unchanged}$= unchanged fields (data matches exactly).
$N_{generalized}$= generalized fields (includes generalized and temporal columns).
N_{masked}= masked fields.
$N_{suppressed}$= suppressed fields.

4. Identity Disclosure Risk
Identity disclosure risk measures the probability that an individual record in an anonymized dataset can be linked back to a specific person using available quasi-identifier information.
The probability of re-identifying an anonymized record is estimated as

$$R_{risk} = \frac{1}{K} \quad (5)$$

Where:
K represents the number of records sharing the same quasi-identifier combination.

5. Precision, Recall, and F1-score
Resistance to record-linkage attacks was evaluated using precision, recall, and F1-score. These measures quantify the success of simulated re-identification attempts by measuring the accuracy, completeness, and overall effectiveness of predicted record matches.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (8)$$

Where:
TP (expected true positives) = the attacker's expected probability-weighted count of correctly isolating truly unique individuals within attackable equivalence classes (size ≤ k), computed as sum of 1/group size.

FP (expected false positives) = the remaining share of attempted claims within attackable classes that do not correspond to a correct match.

FN (expected false negatives) = truly unique individuals whose equivalence class exceeds the threshold k, and are therefore never attempted.

In addition to standard privacy metrics, effectiveness of anonymization techniques and column verification across all anonymization techniques was also computed to verify correctness of applied transformations.

V. RESULTS AND DISCUSSION

This section presents the evaluation findings of proposed privacy-preserving anonymization model. The evaluation is assessed across three dimensions; functional correctness, privacy protection, and resistance to re-identification, to examine effectiveness and scalability of the system.

A. System Validation
Functional correctness was verified through unit and integration testing across all dataset sizes, with representative results shown in Table I and Table II.

Table I: Representative Unit Test Cases and Results across Core Anonymization Techniques.

Technique	Test Cases	Input Example-1	Output
Masking	13	John Smith	J**n S***h
		9876543210	*******210
Generalization	6	Age: 93	90+
		DOB: 2005-01-10	2000-2009
Temporal Reduction	6	Date: 2024-03-15	2024
		Date: (Empty)	YEAR
Suppression	4	Insurance: Blue Cross	[ANONYMIZED]
		Policy: POL12345	[ANONYMIZED]

Table II: Representative Column-Level Verification Results across Dataset Sizes.

Data set Size	Example-1	Technique Applied	Output
2000	William Rodriguez (name)	Masking	W*****m R*******z
	Age: 19	Generalization	0-29
5000	Jaipur (city)	Generalization	South Asia
	Consultant (occupation)	Generalization	Professional
10000	9059984674 (phone_number)	Masking	*******674
	Insurance_Co_27 (insurance_provider)	Suppression	[ANONYMIZED]

Unit and integration testing verified correctness across all anonymization techniques and dataset sizes, with 100% pass rates across 29-unit tests and full structural verification at n = 2000, 5000, and 10000.

B. Privacy and Utility Metrics

The model's privacy protection and utility preservation were further assessed using different evaluation and technique-level effectiveness metrics. Minimum equivalence class size and average Identity Disclosure Risk (1/k) across dataset sizes, and minimum and average l-diversity in are shown in Fig. 4 and Fig. 5 respectively.

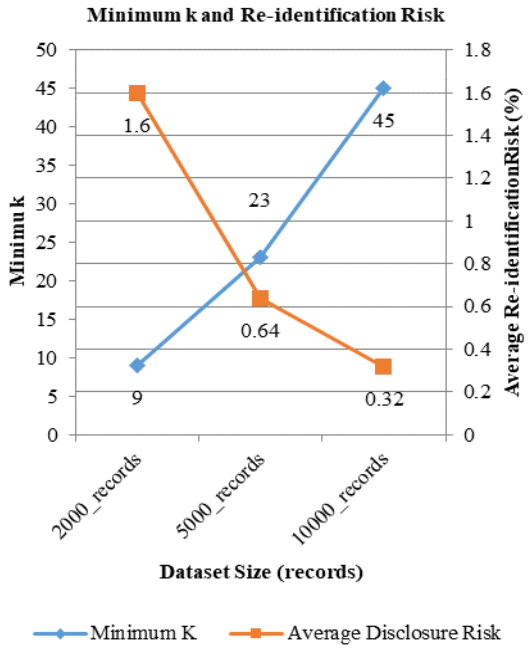

Fig. 4: Minimum Equivalence Class Size (k) and Average Re-identification Risk (%) across dataset sizes.

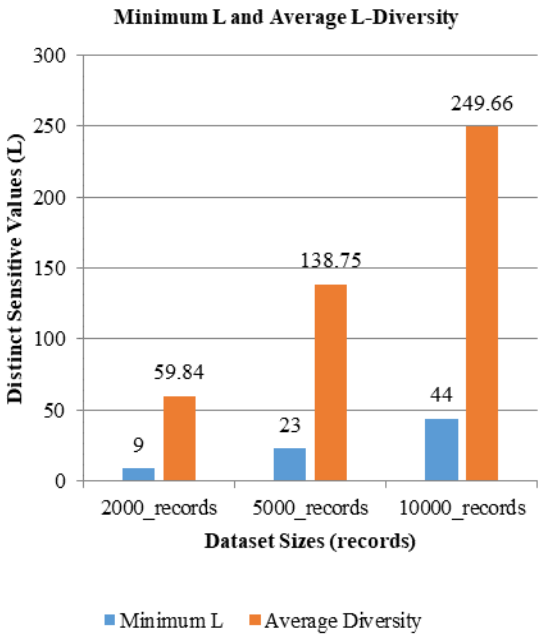

Fig. 5: Minimum and Average l-diversity across Datasets Sizes.

Table III and Table IV report the Information Loss and effectiveness of anonymization techniques across all dataset sizes, respectively.

Table III: Information Loss across Dataset sizes

Dataset size	Unchanged Fields	Generalize/Temporal Fields	Masked Fields	Suppressed Fields	Information Retained
2000	12000	18000	12000	4000	66.52 %
5000	30000	45000	30000	10000	66.52 %
10000	60000	90000	60000	20000	66.52 %

Table IV: Anonymization Technique Effectiveness by Dataset Size

Dataset Size	Suppression (Fields checked)	Generalization (Fields checked)	T R (fields checked)	Effectiveness (%)
2000	4000	8000	6000	100%
5000	10000	20000	15000	100%
10000	20000	40000	30000	100%

The column-wise verification results across all dataset sizes are presented in Table V.

Table V: Column-Wise Verification across Anonymization Techniques

Category	Fields Checked	Result
Masking	600	Successful
Generalization	400	Successful
Suppression	200	Successful
Temporal Reduction	300	Successful
Keep-As-Is	600	Successful

The size of minimum equivalence classes (k), Identity Disclosure Risk (1/k), and the average l-diversity grew with the size of dataset and with the increasing scale of datasets, better anonymity was achieved.

The Information Loss remained same for all dataset sizes (66.52%) because of per-column transformation mapping in the rule-based design. In all three transformation techniques; masking, generalization, and temporal reduction, the effectiveness of transformation was 100% in all targeted fields.

C. Re-identification Evaluation

Re-identification risk was assessed by considering the threshold sensitivity analysis across six fixed thresholds for each of the three attacker assumptions of weak, medium, and strong in all datasets. The number of attempted and expected correct claims is shown in Fig. 6 and Fig. 7, respectively.

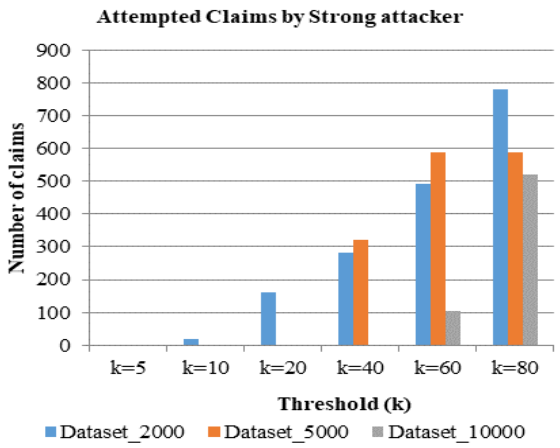

Fig. 6: Total Number of Attempted Claims

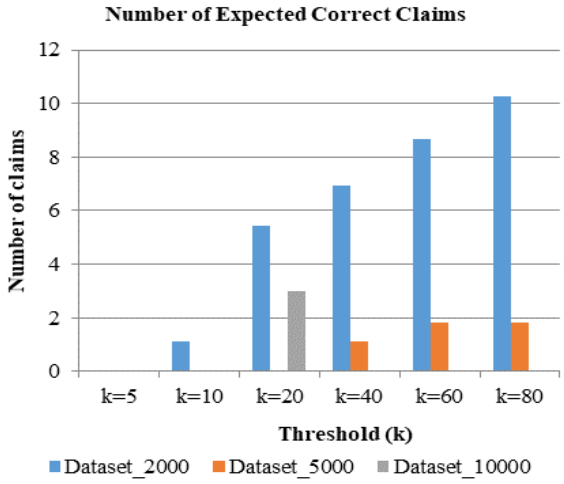

Fig. 7: Total Number of Expected Correct Claims

The Precision, Recall, and F1-score across all thresholds are shown in Fig. 8, Fig. 9, and Fig. 10 respectively.

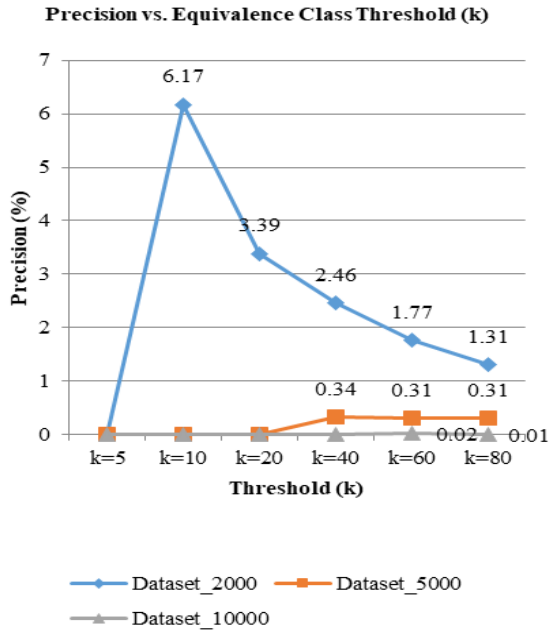

Fig. 8: Precision across Datasets Sizes at Different Threshold Levels.

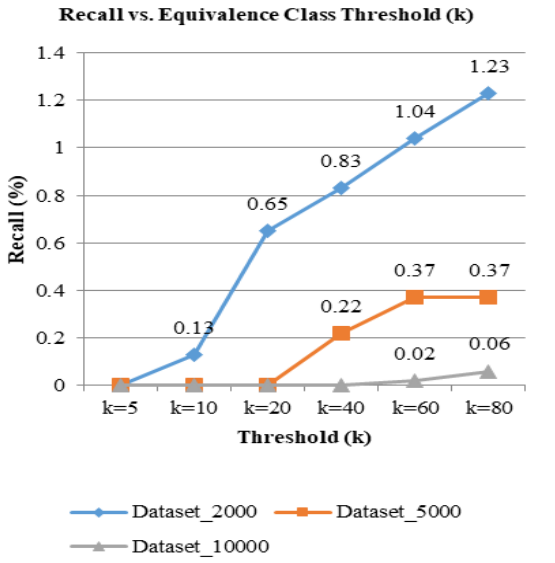

Fig. 9: Recall as a Function of Equivalence-Class Threshold (k) across Dataset Sizes.

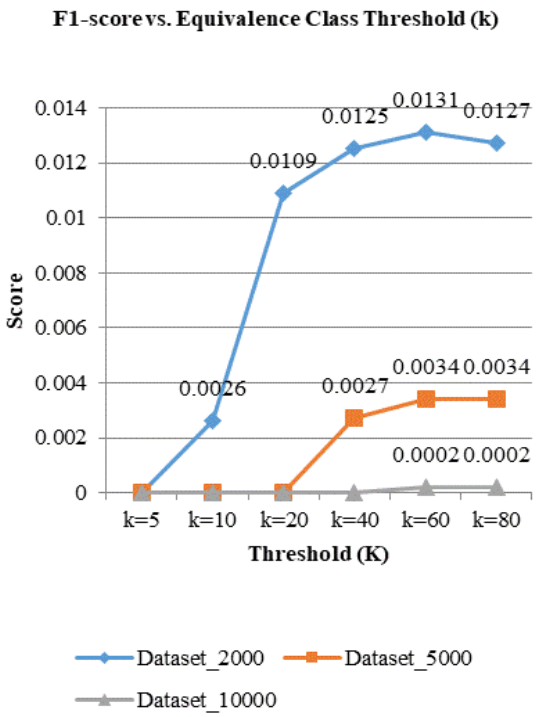

Fig. 10: F1-score across Dataset Sizes at Different Threshold Values.

0% vulnerability was observed for all datasets for the standard setting of $k \geq 5$, and in the case of strong attacker, vulnerability declined with increasing k (0% to 1.23% for n=2000, k=80).

Weak and medium scenarios were constant at 0% as their attribute combinations were limited and produced group sizes far exceedingly even the largest tested threshold. Further, F1-score did not exceed 0.014 in any tested condition.

The values reported for the metrics of re-identification under scaled threshold of 2% of dataset size are shown in Table VI, compared with fixed threshold k=40.

Table VI: Re-identification Metrics under Threshold scaled to 2% of Dataset size, compared against fixed k=40 condition.

Dataset Size (N)	K used (2% of N)	TP	Claims	Truly Unique	Precision	Recall	F1-score	Recall (k=40)
2000	40	6.94	282	832	2.46%	0.83%	0.0125	0.83%
5000	100	1.98	683	496	0.29%	0.40%	0.0034	0.22%
10000	200	0.17	1265	105	0.01%	0.16%	0.0003	0.00%

An attacker whose effort grows with the dataset size, yields modest additional benefits for larger datasets (recall=0.83% for n=2000, 0.40% for n=5000, and 0.16% for n=10000), for a scaled threshold of 2% of the dataset.

D. Observations and Insights

In the analysis of the proposed model, the following observations were made:

1. Client-side anonymization ensured that raw identifiable data is not ever stored in the simulated cloud, giving an extra layer of privacy protection beyond access control mechanisms.
2. Keyword-based attribute classification reliably routed sensitive fields to the correct technique (masking, generalization, temporal reduction, suppression) across all dataset sizes, without requiring statistical or ML-based detection.
3. Information loss remained constant at 66.52% across all sizes, confirming stable, predictable utility retention regardless of scale.
4. Re-identification vulnerability stayed at 0% at the standard $k \geq 5$ threshold across all datasets, consistent with accepted privacy standards.
5. Larger datasets showed progressively lower risk at any fixed threshold, as more records filled the same finite set of generalized quasi-identifier combinations.
6. Vulnerability appeared only under the strong attacker scenario (age + gender + city); weak and medium scenarios remained at 0%, confirming risk concentrates in higher-dimensional quasi-identifier combinations.

E. Limitations

Although the proposed system yielded encouraging results, it has several limitations:

1. Sensitive attribute detection relies on keyword-based column matching, which may need adjustment for unconventional naming conventions.
2. Only three quasi-identifiers (age, gender, city) were evaluated; a broader set could reveal higher risk.
3. Generalization uses fixed, coarse category boundaries rather than adaptive or multi-level hierarchies.
4. The system lacks formal privacy guarantees such as differential privacy, relying instead on rule-based transformation.

REFERENCES

[1] K. Abouelmehdi, A. B. Hssane, and H. Khaloufi, “Big Healthcare Data: Preserving Security and Privacy,” *Journal of Big Data*, vol. 5, 2018. [Online]. Available: https://doi.org/10.1186/s40537-017-0110-7

[2] T. P. Dr and K. Amin, “A Study on k-Anonymity, l-Diversity, and t-Closeness Techniques of Privacy Preservation Data Publishing,” 2019.

[3] S. Chenthara, K. Ahmed, H. Wang, and F. Whittaker, “Security and Privacy-Preserving Challenges of e-Health Solutions in Cloud Computing,” IEEE Access, vol. 7, pp. 74361–74382, 2019. [Online]. Available: https://doi.org/10.1109/access.2019.2919982

[4] J. A. Onesimu, K. J, J. Eunice, M. Pomplun, and H. Dang, “Privacy Preserving Attribute-Focused Anonymization Scheme for Healthcare Data Publishing,” IEEE Access, vol. 10, pp. 86979–86997, 2022. [Online]. Available: https://doi.org/10.1109/access.2022.3199433

[5] M. Kayaalp, “Patient Privacy in the Era of Big Data,” Balkan Medical Journal, vol. 35, pp. 8–17, 2017. [Online]. Available: https://doi.org/10.4274/balkanmedj.2017.0966

[6] Z. Zandesh et al., “Privacy, Security, and Legal Issues in the Health Cloud,” 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10897789/

[7] R. Nowrozy, K. Ahmed, A. Kayes, H. Wang, and T. R. McIntosh, “Privacy Preservation of Electronic Health Records in the Modern Era: A Systematic Survey,” ACM Computing Surveys, vol. 56, pp. 1–37, 2024. [Online]. Available: https://doi.org/10.1145/3653297

[8] H. Khan, L. Menten, and L. M. Peeters, “An Algorithmic Pipeline for GDPR-Compliant Healthcare Data Anonymisation: Moving Toward Standardisation,” arXiv, abs/2506.02942, 2025. [Online]. Available: https://doi.org/10.48550/arxiv.2506.02942

[9] L. N. Zlatolas, T. Welzer, and L. Lhotská, “Data Breaches in Healthcare: Security Mechanisms for Attack Mitigation,” Cluster Computing, vol. 27, pp. 8639–8654, 2024. [Online]. Available: https://doi.org/10.1007/s10586-024-04507-2

[10] Y. J. Lee and K. Lee, "Re-Identification of Medical Records by Optimum Quasi-Identifiers," 2017 19th International Conference on Advanced Communication Technology (ICACT), pp. 428–435, 2017. [Online]. Available: https://doi.org/10.23919/icact.2017.7890125

[11] Z. Wang, E. Khatibi, F. Firouzi, S. R. Mousavi, K. Chakrabarty, and A. M. Rahmani, "Linkage Attacks Expose Identity Risks in Public ECG Data Sharing," 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1–7, 2025. [Online]. Available: https://doi.org/10.1109/embc58623.2025.11253028

[12] D. Kalinin, V. Severyn, and M. Bezmenov, "Privacy Models and Anonymization Techniques for Tabular Healthcare Data," Bulletin of National Technical University 'KhPI'. Series: System Analysis, Control and Information Technologies, 2024. [Online]. Available: https://doi.org/10.20998/2079-0023.2024.02.12

[13] S. Smadi, N. A. Karim, H. Kanaker, and W. K. Abdulraheem, "Anonymization Techniques for Privacy-Preserving Data Publishing: A Comprehensive Survey," Bulletin of Electrical Engineering and Informatics, 2026. [Online]. Available: https://doi.org/10.11591/eei.v15i2.11224

[14] HIPAA Journal, "Healthcare Data Breach Statistics – Updated for 2026," 2026. [Online]. Available: https://www.hipaajournal.com/healthcare-data-breach-statistics/

[15] IBM Security, "Cost of a Data Breach Report 2025," IBM Corporation, 2025. [Online]. Available: https://www.ibm.com/reports/data-breach

[16] S. S. Jayakumar, K. Meher, U. Rout, G. Srinath, S. Khurana, S. Ghumman, and S. Singh, "Risk Analysis of Data Privacy Violations in Digital Health Records and Patient Confidentiality," Seminars in Medical Writing and Education, 2024. [Online]. Available: https://doi.org/10.56294/mw2024498

[17] M. Kantarcioglu et al., "A Novel Analysis Methodology for Assessment of Re-Identification Risks for the National Cancer Institute Cancer Registry Privacy Preserving Record Linkage Technique," Journal of the American Medical Informatics Association: JAMIA, 2025. [Online]. Available: https://doi.org/10.1093/jamia/ocaf172

[18] I. Khalil et al., "Security and Privacy Requirements for Cloud Computing in Healthcare: A Systematic Mapping Study," 2020. [Online]. Available: https://doi.org/10.1145/3386160

[19] A. Mishra et al., "Enhancing Privacy-Preserving Mechanisms in Cloud Data Storage," 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/cpe.7831

[20] A. Abbasi and B. Mohammadi, "A Clustering-Based Anonymization Approach for Privacy-Preserving in the Healthcare Cloud," Concurrency and Computation: Practice and Experience, vol. 34, 2021. [Online]. Available: https://doi.org/10.1002/cpe.6487

[21] S. Jeon, J. Seo, S. Kim, J. Lee, J.-H. Kim, J. Sohn, J. Moon, and H. J. Joo, "Proposal and Assessment of a De-Identification Strategy to Enhance Anonymity of the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) in a Public Cloud-Computing Environment: Anonymization of Medical Data Using Privacy Models," Journal of Medical Internet Research, vol. 22, 2020. [Online]. Available: https://doi.org/10.2196/19597

[22] P. Samarati and L. Sweeney, "Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression," 1998. [Online]. Available: https://dataprivacylab.org/projects/kanonymity/paper3.html

[23] R. Aufschläger et al., "Anonymization Procedures for Tabular Data," Information, vol. 14, no. 9, p. 487, 2023. [Online]. Available: https://doi.org/10.3390/info14090487

[24] F. Wirth, T. Meurers, M. Johns, and F. Prasser, "Privacy-Preserving Data Sharing Infrastructures for Medical Research: Systematization and Comparison," BMC Medical Informatics and Decision Making, vol. 21, 2021. [Online]. Available: https://doi.org/10.1186/s12911-021-01602-x

[25] S. M. Narayan, N. Kohli, and M. M. Martin, "Addressing Contemporary Threats in Anonymised Healthcare Data Using Privacy

Engineering," NPJ Digital Medicine, vol. 8, 2025. [Online]. Available: https://doi.org/10.1038/s41746-025-01520-6

[26] F. G. H. Hammer, M. Buglowski, and A. Stollenwerk, "Semi-Local Time-Sensitive Anonymization of Clinical Data," Scientific Data, 2024/2025. [Online]. Available: https://doi.org/10.1038/s41597-024-04192-1

[27] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, pp. 557–570, 2002. [Online]. Available: https://doi.org/10.1142/s0218488502001648

[28] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy Beyond k-Anonymity," 22nd International Conference on Data Engineering (ICDE'06), p. 24, 2006. [Online]. Available: https://doi.org/10.1109/icde.2006.1

[29] F. Kohlmayer, F. Prasser, and K. A. Kuhn, "The Cost of Quality: Implementing Generalization and Suppression for Anonymizing Biomedical Data with Minimal Information Loss," Journal of Biomedical Informatics, vol. 58, pp. 37–48, 2015. [Online]. Available: https://doi.org/10.1016/j.jbi.2015.09.007

[30] M. Scaiano, K. Middleton, and K. El Emam, "A Unified Framework for Evaluating the Risk of Re-identification of Health Data," Journal of Biomedical Informatics, 2016. [Online]. Available: https://doi.org/10.1016/j.jbi.2016.07.015

[31] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-Domain k-Anonymity," 2005. [Online]. Available: https://doi.org/10.1145/1066157.1066164

[32] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity," 2006. [Online]. Available: https://dblp.org/rec/conf/icde/LefevreDR06.html

[33] R. C. Wong, J. Li, A. Fu, and K. Wang, "(α, k)-Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing," Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06), 2006. [Online]. Available: https://doi.org/10.1145/1150402.1150499

[34] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," 2007 IEEE 23rd International Conference on Data Engineering, pp. 106–115, 2007. [Online]. Available: https://doi.org/10.1109/icde.2007.367856

[35] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "A Framework for Efficient Data Anonymization under Privacy and Accuracy Constraints," 2009. [Online]. Available: https://doi.org/10.1145/1538909.1538911

[36] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "Non-Homogeneous Generalization in Privacy Preserving Data Publishing," 2010. [Online]. Available: https://doi.org/10.1145/1807167.1807248

[37] A. Zigomitros, F. Casino, A. Solanas, and C. Patsakis, "A Survey on Privacy Properties for Data Publishing of Relational Data," IEEE Access, vol. 8, pp. 51071–51099, 2020. [Online]. Available: https://doi.org/10.1109/access.2020.2980235

[38] A. Sepas, A. Bangash, O. Alraoui, K. Emam, and A. El-Hussuna, "Algorithms to Anonymize Structured Medical and Healthcare Data: A Systematic Review," Frontiers in Bioinformatics, vol. 2, 2022. [Online]. Available: https://doi.org/10.3389/fbinf.2022.984807

[39] K. Rajendran, M. Jayabalan, and M. Ehsan, "A Study on k-Anonymity, l-Diversity, and t-Closeness Techniques Focusing Medical Data," 2018.

[40] I. E. Olatunji, J. Rauch, M. Katzensteiner, and M. Khosla, "A Review of Anonymization for Healthcare Data," Big Data, vol. 12, pp. 538–555, 2021. [Online]. Available: https://doi.org/10.1089/big.2021.0169

[41] F. Prasser, F. Kohlmayer, R. Lautenschläger, and K. A. Kuhn, "ARX — A Comprehensive Tool for Anonymizing Biomedical Data," AMIA Annual Symposium Proceedings, 2014.

[42] K. El Emam, "Risk-Based De-Identification of Health Data," IEEE Security & Privacy, vol. 8, pp. 64–67, 2010. [Online]. Available: https://doi.org/10.1109/msp.2010.103

[43] K. El Emam and L. Arbuckle, Anonymizing Health Data: Case Studies and Methods to Get You Started, 2013. [Online]. Available:

https://www.oreilly.com/library/view/anonymizing-health-data/9781449363073/

[44] R. Ratra, P. Gulia, and N. S. Gill, "Evaluation of Re-identification Risk Using Anonymization and Differential Privacy in Healthcare," International Journal of Advanced Computer Science and Applications, 2022. [Online]. Available: https://doi.org/10.14569/ijacsa.2022.0130266

[45] T. Kanwal, A. Anjum, S. Malik, H. Sajjad, A. Khan, U. Manzoor, and A. Asheralieva, "A Robust Privacy Preserving Approach for Electronic Health Records Using Multiple Dataset with Multiple Sensitive Attributes," Computers & Security, vol. 105, p. 102224, 2021. [Online]. Available: https://doi.org/10.1016/j.cose.2021.102224

[46] Y. A. A. S. Aldeen, M. Salleh, and Y. Aljeroudi, "An Innovative Privacy Preserving Technique for Incremental Datasets on Cloud Computing," Journal of Biomedical Informatics, vol. 62, pp. 107–116, 2016. [Online]. Available: https://doi.org/10.1016/j.jbi.2016.06.011

[47] Y. Tao et al., "Privacy Preserving Publication: Anonymization Frameworks and Algorithms." [Online]. Available: https://doi.org/10.1007/978-0-387-48533-1_20

[48] F. Prasser, J. Eicher, R. Bild, H. Spengler, and K. A. Kuhn, "Flexible Data Anonymization Using ARX — Current Status and Challenges Ahead," Software: Practice and Experience, vol. 50, pp. 1277–1304, 2020. [Online]. Available: https://doi.org/10.1002/spe.2812

[49] Z. Zuo, M. Watson, D. Budgen, R. Hall, C. Kennelly, and A. N. Moubayed, "Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study," JMIR Medical Informatics, vol. 9, 2021. [Online]. Available: https://doi.org/10.2196/29871

[50] S. Pitoglou, A. Filntisi, A. Anastasiou, G. Matsopoulos, and D. Koutsouris, "Measuring the Impact of Anonymization on Real-World Consolidated Health Datasets Engineered for Secondary Research Use: Experiments in the Context of MODELHealth Project," Frontiers in Digital Health, vol. 4, 2022. [Online]. Available: https://doi.org/10.3389/fdgth.2022.841853

[51] S. Jahan, Y.-F. Ge, E. Kabir, and K. Wang, "Analysis and Multi-Objective Protection of Public Medical Datasets from Privacy and Utility Perspectives," Data Science and Engineering, vol. 10, pp. 362–375, 2025. [Online]. Available: https://doi.org/10.1007/s41019-025-00283-0

[52] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 571–588, 2002.