ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 11, No. 1, 51-68 (2026)
www.astesj.com
Special Issue on Multidisciplinary Frontiers in Engineering, Computing and Applied Sciences

ASTES Journal
ISSN: 2415-6698

Multi Attribute Stratified Sampling: An Automated Framework for Privacy-Preserving Healthcare Data Publishing with Multiple Sensitive Attributes

Vikas Thammanna Gowda*, Landis Humphrey, Aiden Kadoch, YinBo Chen, Olivia Roberts

Division of Information Technology and Sciences, Champlain College, Burlington, 05401, USA

*Corresponding Author: Vikas Thammanna Gowda, Division of Information Technology and Sciences, Champlain College, Burlington, 05401, USA, +18023836605 & Email: vthammannagowda@champlain.edu

ARTICLE INFO

Article history:
Received: 31 December, 2025
Revised: 26 January, 2026
Accepted: 29 January, 2026
Online: 9 February, 2026

Keywords:
Healthcare data anonymization
k-anonymity
Privacy-utility tradeoff
Multiple sensitive attributes
Automated Parameter Optimization
Machine Learning Utility

ABSTRACT

The accumulation and analysis of large-scale patient data have led to breakthrough discoveries in potential flags for diseases based on pattern recognition, highlight medication efficacy, and local population health trends that would be impossible with traditional paper-based records. However, these benefits come with unique challenges posed by the application of data sharing for research and analysis, and mandatory requirements that require careful balance between privacy protection and usefulness of data especially when the data contains several sensitive information. We propose a framework, Multi Attribute Stratified Sampling (MASS), to achieve automatic parameter optimization by separating the sanitization process from manual privacy parameter configuration. Most traditional privacy-preserving techniques require experts to specify privacy parameters such as k, l, and t values for k-anonymity, l-diversity, and t-closeness respectively based on intuition or trail and error resulting in sub-optimal privacy-utility-trade-offs. In contrast our framework employs a self-tuning paradigm which uses GetAnonymized, CandidateBuilder, and Optimizer modules. CandidateBuilder produces multiple anonymized versions of the original preprocessed data by iteratively calling GetAnonymized on a range of anonymization levels creating a solution space. The Optimizer then scans through the solution space using an objective function to determine the optimal anonymized version. The objective function utilizes privacy and information losses along with the classification recall to discover the privacy parameters that yield the best balance between privacy protection and data utility, eliminating the burden of manually fine tuning the privacy parameters and ensuring reproducible outcomes across various healthcare datasets and analytical contexts. Experimental validation on four datasets (1k, 10k, 100k and 10M) demonstrates that MASS achieves strong privacy protection across datasets with a privacy loss < 0.25 while maintaining >95% recall retention on datasets exceeding 10k records. Given that the computational complexity to generate anonymized dataset is NP-hard, MASS presents a polynomial-time heuristic solution which validates practical implementation and scalability for real-world deployment.

1. Introduction

In the current digital era, healthcare organizations across the world have welcomed Electronic Health Records (EHRs) as a fundamental building block to support modern healthcare practices and clinical trails. EHR systems have transformed healthcare access and delivery by providing healthcare professionals with instant access to patient information, minimizing errors, and enabling data-driven decision making. Apart from clinical care, the non-volatile health data repositories stored in EHR data warehouses hold extraordinary potential in healthcare policy development, public health surveillance, medical research, and government initiatives.

The transition to EHR has altered how data is collected, stored, and shared. Healthcare data publishing serves various objectives that extend beyond archival preservation. While research and analysis remain the primary focus, the application and landscape of data publishing have grown through government regulatory mandates, ethical obligations, and funding requirements that compel organizations to make data publicly accessible.

https://dx.doi.org/10.25046/aj110106

The American Food and Drug Administration Amendments Act of 2007 (FDAAA 801) [1] established mandatory registration and result reporting for clinical trails by the sponsors. Noncompliance carries civil penalties reaching 10,000 USD per day. The European Union (EU) Clinical trails Regulation 536/2014 [2] implemented transparency requirements by mandating that trail protocols, results, and clinical study reports be publicly available through EU Clinical trail Register. Effective January 25, 2023, all National Institutes of Health funded research are required to share data management and sharing policies regardless of funding level. These mandates create knotty commitments that must be balanced against privacy protection by the Health Insurance Probability and Accountability Act (HIPPA) in the USA and the General Data Protection Regulation (GDPR) in the EU. This very tension between mandate disclosure requirement and privacy obligations drives the privacy engineers to innovate techniques in the field of Privacy Preserving Data Publishing (PPDP). Figure 1 illustrates the pathway of data publication from Data Owners to Data Users.

Organizations and agencies (Data Holders/Custodians) that collect patients' data (Data Owners–whose personal health information must be protected) with the patients' consent are responsible for security, compliance (HIPPA, GDPR) and access control. Data publishers (a team of privacy engineers) who may be within the same organization as Data Holders or a separate entity, transform and anonymize data using privacy preserving approaches before publishing that data. Data Publishers face the privacy-utility tradeoff challenge that differentiates healthcare data from social media or commercial data.

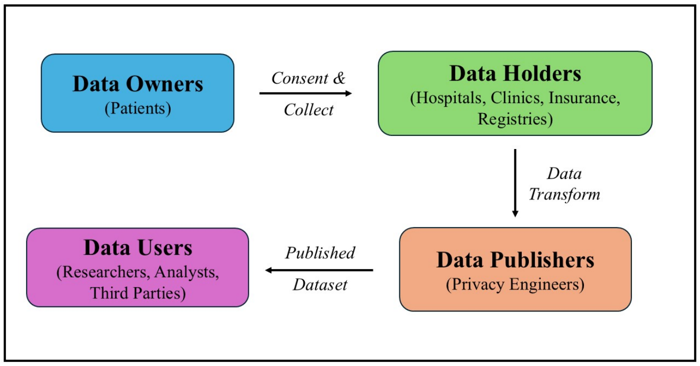


Figure 1: Data Publishing Flow

Healthcare data can be categorized into:

- **Administrative Attributes (AA):** serves as the backbone of healthcare operations that capture logistic and financial aspects of patient information. These attributes hold Personal Identifiable Information (PII) (PatientID, Patient Name, emailID, and SSN), Sensitive Information (SI) (charges, payments, and reimbursements), Quasi Identifiers (QI) (length of stay and scheduling data), and Operational Information (OI) (Department/Unit, Institution ID, and VisitID).

- **Demographic Attributes (DA):** describes the characteristics of the patient population (age/DOB, gender, race, ethnicity, etc.) and usually contains QI (information that does not identify the patient but a combination of them might).

- **Clinical Attributes (CA):** constitutes the core medical content of healthcare data that holds diagnoses, laboratory results, medications, vital signs, and allergies to list a few. These usually hold SI that must be protected.

- **Socio-Economic Attributes (SEA):** provides context about patients' social and economic circumstances that may impact health outcomes. They include, but are not limited to education level, employment status, income level, marital status, neighborhood characteristics, and behavioral and lifestyle data. They hold both SI and QI.

PPDP provides techniques and approaches to safeguard SI while restoring high data usage. The k-anonymity privacy model [3, 4] set up the elementary foundations for PPDP by requiring each record to be like $k-1$ other records in terms of QIs. While groundbreaking, researchers identified significant limitations since k-anonymity does not take SIs into consideration. Extensions like l-diversity [5] (requiring minimum distinct sensitive values in each equivalence class) and t-closeness [6] (preserves SI distribution in each equivalence class that is close to the SI distribution in pre-anonymized data) were proposed to k-anonymity that specifically considers SIs.

This work makes the following key contributions to the field of privacy preserving healthcare data publishing:

- We propose a novel framework that presents an end-to-end solution that generates publishable healthcare data containing multiple SI attributes.

- Our methodology automatically discovers the optimal privacy configurations through systematic solution exploration coupled with data-driven optimization there by eliminating the requirement to specify the privacy parameters.

- We address the NP-hard problem of forming optimal equivalence classes with minimum information loss through a similarity based heuristic that operates in polynomial time.

The rest of the paper is organized as follows, Section 2 provides a brief literature review on the existing work in publishing medical data. Section 3 gives the framework of our approach, and Section 4 lays out the experiments and discusses the findings. The paper concludes in Section 5.

2. Related Work

The challenge of releasing healthcare data while safeguarding patient privacy has directed large-scale exploration on various dimensions. The challenge enhances when handling data that contains several private medical information. This section reviews existing approaches focused on healthcare context to identify potential research gaps.

In [7], a method was proposed to prevent over-generalization in medical data using k-anonymity and h-ceiling. The method preserves data utility by inserting counterfeit records to the EHR that contains demographics and diagnosis. While the approach substantially prevents unreasonable information loss that occurs with k-anonymity based full-domain generalization, it does not extend to l-diversity or t-closeness. The insertion of counterfeit records may introduce artificial patterns that impact the results of medical data analysis especially when handling a rare disease or condition.

An investigation in [8] addressed a problem in PPDP where traditional anonymization techniques used in published datasets frequently underrepresent minority groups. Existing methods, such as k-anonymity and l-diversity, can mask or delete minority patterns when data is widely generalized, potentially compromising the quality of decisions made based on this data, and subsequently affecting the benefits derived from it. Their method automatically identifies "equity-vulnerable" QI attributes and prevents them from being erased, while still maintaining considerable privacy. Preserving this information is crucial for ensuring fair representation in healthcare data and serves as a solution to the privacy-equity trade-off (PET) often observed in data publishing scenarios.

A new privacy-preserving technique that partitions data both horizontally and vertically known as slicing was proposed in [9]. This new approach overcomes problems found in popular anonymization techniques (generalization and bucketization), as it can handle high-dimensional data while preserving attribute correlations and l-diversity through random permutation of values across columns within each bucket. By leveraging the strengths of both techniques simultaneously, slicing is demonstrated to be particularly suitable for complex healthcare data, where a clear separation between QI and SI is normally required.

An enhanced k-anonymity algorithm, KAPP, that integrates t-closeness and an improved African vultures optimization algorithm (AVOA) was introduced in [10] to counter challenges found in common k-anonymity methods. Traditional methods can lead to clustering inconsistencies and are still vulnerable to skewness and similarity attacks. Additionally, other models, like l-diversity, have specific weaknesses when it comes to protecting SIs. KAPP enhances clustering accuracy by featuring a multi-dimensional SI clustering algorithm, while ensuring that equivalence classes remain distinct.

A comparative evaluation of three anonymization algorithms was reported in [11] highlighting the Multi-Objective Optimization-Based Anonymization Model (MO-OBAM) as the most effective for maintaining data utility while protecting against homogeneity and linkage attacks. When compared to the basic k-anonymity model, MO-OBAM significantly decreases vulnerability in groups at a higher risk of linkage attacks. By pointing out the need for more advanced anonymization methods that protect vulnerable populations without compromising model performance, this analysis supports the increased use of advanced anonymization models, such as MO-OBAM, to balance the privacy-utility trade-off in published data.

In [12], the authors discuss a flaw with k-anonymity and similar metrics like t-closeness or l-diversity, being that they are static and only account for a limited time span. This means that dynamic trajectory data fails to be fully obscured as time moves forward. To solve this problem, they propose a method which involves partitioning the data by time, then scanning their trajectories for efficient clustering and privacy protection under (k, δ) security constraints. They have tested their method on various real-world trajectory data examples, and have concluded that it successfully preserves privacy and the data's utility and integrity.

A new way of understanding k-anonymity within the context of moving objects' data was identified in [13]. The authors state that because a moving object has an inherent uncertainty due to sampling and positioning systems, a new version of k-anonymity using co-localization can be utilized for privacy preservation. The trajectory of a moving object is not a line in three-dimensional space, but a cylindrical volume containing its potentiality; this means any object within that same cylinder is indistinguishable from all others. Due to this quality, they define (k, δ)-anonymity, where δ represents the possible location imprecision, then develop a method tailored to space translation called Never Walk Alone (NWA). This method achieves (k, δ)-anonymity by first minimally distorting the spatial data, then utilizing a greedy clustering algorithm based on clustering and ad hoc pre-processing and outlier removal. Rethinking privacy algorithms based on the field knowledge and context of the data itself leads to valuable innovations in data privacy.

An anonymization algorithm which includes dimensionality reduction by applying feature selection, then uses feature and attribute suppression to satisfy k-anonymity was proposed in [14]. This k-anonymous pattern-based multidimensional suppression algorithm (kPB-MS) was tested by comparing the performance of five anonymized datasets against the originals using four distinct classifiers. In the majority of test cases, the accuracy of the classifiers was not diminished by using the anonymized datasets, proving the algorithm's viability for maintaining the usability of the data. The primary shortcomings of the research are the reliance of the algorithm on the manual definition of k values and the potential to miss interacting features with low main effects due to its use of a forward selection method.

A k-Anonymity based on Center Point Clustering (KACPC) was introduced in [15] to improve data utility while guaranteeing k-anonymity in EHR. The algorithm defines new similarity metrics between healthcare records and clustering technique decrease information loss as compared to traditional generalization methods. Although the algorithm achieves higher data utility than the Samarati's [4] baseline while maintaining polynomial time complexity it requires a predefined k value and does not apply to EHR with more than one sensitive attribute.

A framework was developed in [16] to achieve (k, l)-diversity in healthcare records with multiple sensitive attributes. It first vertically slices the data into QI buckets and SI buckets, later applies hierarchical classification to retain effective data utility since it does not treat all attributes uniformly. Even though their framework allows various levels of privacy for different attributes providing flexibility in safeguarding several sensitive attributes, the vertical slicing breaks the correlation between QI and SI that are critical for medical research analysis especially in epidemiological studies.

3. Multi Attribute Stratified Sampling (MASS)

In this section, we present a comprehensive end-to-end framework for MASS where we transform a raw healthcare data into privacy protected version of the original data. This anonymized data is suitable for publication and sharing with high analytical usefulness. Our methodology starts by preprocessing the raw data which involves cleaning the data and mapping the healthcare attributes to privacy classes. We then pass this preprocessed data through three modules: the first module is the core anonymization engine (*GetAnonymized*) that generate a sanitized version, using a stratification-based approach that partitions records by SI combinations before forming equivalence classes, for an anonymity level and stores the privacy and information losses. This anonymized version achieves implicit l-diversity and t-closeness without explicitly specifying the l and t parameters. The second module (*CandidateBuilder*) generates a solution space of anonymized data variants for a range of anonymity levels. In this module we test for the data utility (i.e,, the ability of the anonymized healthcare data to perform on predictive analytics) of all the variants by capturing the classification recall metric. The final module (*Optimizer*) uses an objective function to scan through the solution space and select the best anonymized version to publish and share. For each of the modules we perform time-space analysis. This approach eliminates the need for manual parameter tuning and provides a practical solution in a polynomial time, for a NP-hard problem.

3.1. Data Preprocessing and Feature Engineering

Data preprocessing is a crucial step to ensure data integrity for privacy modeling and successive analytical utility. We begin by cleaning the raw data to address the common quality issues such as:

- Missing values (Null/NaN, empty strings, placeholders): for example, unrecorded/omitted vital signs, lab results, or diagnosis codes in EHRs.
- Duplicate records: for example, multiple records for the same patient caused by system migrations between healthcare providers, slight variation in patient information, or re-enrollments.
- Data inconsistencies (capitalization, spelling errors, mixed data formats, units of measurement, invalid/incorrect data, encoding/formatting, structural errors).
- Outliers.

Once the quality issues are resolved we proceed to mapping healthcare attributes to privacy classes. We define a function $\varphi : H \rightarrow P$ that associates healthcare-related attributes with standard privacy categories. Let $H = \{AA, DA, CA, SEA\}$ represent the healthcare attributes and $P = \{PII, QI, SI, OI\}$ represent the privacy attributes where AA, DA, CA, and SEA are mapped to PII, QI, SI, and OI respectively. We then drop the PII attributes since they directly identify the patient.

3.2. Data Representation

We denote the preprocessed healthcare data, $\mathcal{H}^C$, as a set that relates each patient record to a tuple h_i,

$$\mathcal{H}^C = \{h_1, h_2, h_3, ..., h_n\} \tag{1}$$

$$h_i = \{a_{i,1}, a_{i,2}, a_{i,3}, ..., a_{i,m}\} \tag{2}$$

where n is the total number of records, m is the total number of attributes after preprocessing, $a_{i,j}$ is the value of j^{th} attribute in the i^{th} record. The m attributes represent the three privacy classes.

Table 1: Attribute Categorization

Privacy Classes	Attributes	Cardinality
QI	$\{a_1, a_2, \ldots, a_q\}$	q
SI	$\left\{a_{q+1}, a_{q+2}, \ldots, a_s\right\}$	s - q
OI	$\{a_{s+1}, a_{s+2}, \ldots, a_m\}$	m - s

3.3. GetAnonymized

The *GetAnonymized* module serves as the core sanitization engine to transform the preprocessed healthcare dataset into a privacy-preserved version that balances the necessities of patient privacy and data utility preservation. The algorithm deploys an applied k-anonymity principle that indirectly achieves l-diversity and t-closeness within the context of data anonymization. This module accepts the $\mathcal{H}^C$, QI and SI column names, and anonymity level (privacy parameter k) to generate an anonymized version of the preprocessed dataset, $\mathcal{H}^{C^*}$. During the sanitization process, *GetAnonymized* computes and records two key metrics:

- Privacy Loss (PL): quantifies the level of privacy protection achieved on the SI through the t-closeness principle using Earth Mover's Distance [17], also known as Wasserstein distance. It measures the distributional divergence between SI values in the original and anonymized datasets by estimating the amount of work done in successfully converting one distribution into another. This gives us the privacy parameter t.
- Information Loss (IL): measures how much the data has been altered from its original form through generalizing the QIs within each equivalence class using General Loss Metric [18].

GetAnonymized module is a three-step process, where we first cluster $\mathcal{H}^C$ into non overlapping subsets such that each subset contains patient records that share the same combination of SI values. This approach ensures that the subsequent QI generalization operation do not internationally create equivalence classes that homogenize SIs, which would ironically increase privacy risks by making SIs more inferrable. For example, consider a scenario where all records in an equivalence class share the same diagnosis, an attacker knowing that a patient belonging to this equivalence class will immediately infer their medical condition with certainty, even if their QI characteristics remain hidden.

The second step involves the anonymization mechanism that operates independently on each subset produced in the first step. For each subset, we use a similarity-based clustering approach to identify records whose QI values are in comparable range there by naturally minimizing the information loss required to achieve the desired anonymity, i.e., for each group satisfying the minimum size constraint ($|group| \geq k$), we select $|group|/e$ records, where e is the size of equivalence class calculated as n/k. This selective sampling strategy is vital for optimizing privacy preservation through t-closeness since every equivalence class gets equal portion of records from each subset which retains the semantic closeness of SI values and similar QI values thereby reducing the privacy loss and information loss respectively. The final step implements metric computation to quantify the privacy protection achieved through *t*-closeness principle and the information utility sacrificed during the generalization of QI. These metrics act as key components in decision making process for the third module, *Optimizer*.

3.3.1. Implicit l-diversity, t-closeness and Low Information loss Achievement

The *GetAnonymized* module offers reliable privacy guarantees beyond *k*-anonymity through its architectural design implementation to satisfies *l*-diversity and approximates *t*-closeness without requiring explicit privacy parameter specifications. During the module's first two steps where we cluster the patient's records by unique SI combinations into non-overlapping subsets and form equivalence classes by selecting equal portions of records inherently prevents homogeneity attacks, i.e., the records within each SI-homogeneous subset share identical sensitive values, and the final equivalence classes combine records from multiple different subsets, the resulting anonymized dataset naturally exhibits SI diversity. An equivalence class in this anonymized version contains records originating from different SI subsets which guarantees that SI are not homogeneous within the equivalence class. This design approach achieves high SI diversity as an emergent property rather than through explicit constraint enforcement there by avoiding the additional information loss and computational overhead required by traditional *l*-diversity implementations, i.e., after forming each equivalence class the implementations must verify the diversity constraint for each iteration. As stated earlier, during equivalence class creation selecting records whose QI values are close results in low cardinal generalization which reduces the information loss

Also, the module addresses *t*-closeness objectives without requiring specification of the *t* parameter. It is a critical advantage given that determining appropriate *t* values lacks principled methodology. Unlike, *l*-diversity that requires minimum *l* distinct SI values within each equivalence class, the *t* value is subjective i.e., distinguishing whether $t = 0.15$ versus $t = 0.18$ is adequate. To address this distribute SI combinations equally among equivalence classes during clustering by ensuring each SI-homogeneous subset contributes proportionally to equivalence classes. Now the module minimizes distributional distance between individual equivalence classes. This equal distribution strategy reduces the overall distance between the original SI distribution and the distributions within each equivalence class, effectively approximating *t*-closeness by making all equivalence class distributions similar to each other.

Key Advantages:

- Eliminates the need to determine appropriate *l* or *t* values
- Privacy properties emerge from SI-based clustering design rather than explicit checking
- Achieves diversity through subset combination rather than forced mixing
- Equal SI distribution minimizes distances between subset distributions and the original distribution
- No runtime diversity or closeness constraint checking required

3.3.2. Algorithm

Algorithm 1: GetAnonymized: Core Sanitization Engine

```
Input: $\mathcal{H}^C$ (preprocessed dataset), QI columns, SI
       columns, $k$ (anonymity level)
Result: $\mathcal{H}^{C^*}$ (anonymized dataset), PL (Privacy Loss), IL
        (Information Loss)
// Step 1: Cluster by Sensitive Information
$Subsets \leftarrow \emptyset$;
foreach unique combination of SI values do
    $subset \leftarrow$ all records in $\mathcal{H}^C$ with this SI combination;
    $Subsets \leftarrow Subsets \cup \{subset\}$;
end
// Step 2: Form Equivalence Classes
$\mathcal{H}^{C^*} \leftarrow \emptyset$;
foreach $subset \in Subsets$ do
    Group records in $subset$ by similar QI values;
    foreach group of size $\geq k$ do
        Select $|group|/e$ records with similar QI values;
        Generalize QI values for selected records;
        Add generalized records to $\mathcal{H}^{C^*}$;
    end
end
// Step 3: Compute Metrics
$PL \leftarrow$ ComputePrivacyLoss($\mathcal{H}^C$, $\mathcal{H}^{C^*}$) using Earth
  Mover's Distance;
$IL \leftarrow$ ComputeInformationLoss($\mathcal{H}^C$, $\mathcal{H}^{C^*}$) using General
  Loss Metric;
return $\mathcal{H}^{C^*}$, PL, IL;
```

3.3.3. Time Complexity Analysis

Creating subsets based on unique SI combinations requires scanning all n records and grouping them by s sensitive attributes. This operation takes $O(n \cdot s)$ time. For each subset, grouping records by similar QI values and generalizing them involves: sorting or comparing QI values: $O(n \log n)$ in the worst case, and electing and generalizing n/k records per group: $O(n \cdot q)$ where q is the number of QI attributes. Computing the Privacy Loss (Earth Mover's Distance) takes $O(n^2 \cdot m)$ in worst case, and Information Loss (General Loss Metric) takes $O(n \cdot m)$

Overall Time Complexity: $O(n^2 \cdot m + n \log n)$, dominated by the Earth Mover's Distance computation.

3.3.4. Space Complexity Analysis

- **Input Storage:** Original dataset $\mathcal{H}^C$ requires $O(n \cdot m)$ space.
- **Subsets Storage:** In worst case, each record could have unique SI values, requiring $O(n \cdot m)$ space for all subsets.
- **Output Storage:** Anonymized dataset $\mathcal{H}^{C^*}$ requires $O(n \cdot m)$ space.
- **Metrics:** Privacy Loss and Information Loss values require $O(1)$ space.

Overall Space Complexity: $O(n \cdot m)$

3.4. *CandidateBuilder*

The *CandidateBuilder* module systematically generates a comprehensive solution space Ω to explore the tradeoff present in PPDP. Our strategy is to let go of complexity constraints to avoid sub-optimal parameter selection that could result in week data analytics or lack of better privacy protection. To this end, we build a lattice of anonymized versions of the preprocessed data by iteratively invoking the *GetAnonymized* module for a range of anonymity levels (extracted from the dataset itself and not a user defined range) rather than producing a single anonymized data variant. Each point in this solution space represents a unique privacy-utility-trade-off point. This enables our framework to identify best configuration that satisfy an objective function in the later stage.

This module initially computes the maximum cardinality l_v across all SI columns in the original dataset, defined as

$$l_v = max(l_{1,}, l_2, l_3, ..., l_s) \tag{3}$$

where, l_x is the total number of unique instances of sensitive values in the x^{th} SI column, where $(q + 1 \leq x \leq s)$. This maximum cardinality serves as a vital benchmark to define the upper bound of the anonymity parameter exploration range. We set $4l_v$ as the upper bound. The rationale for using this as the upper bound reflects two important considerations on the practical limits of anonymization:

- This upper bound guarantees that the module explores k values well beyond the diversity of SI attributes, testing configurations where the groupings are considerably larger than the number of distinct SI values. This is essential because k-anonymity alone do not guarantee diverse SI.
- As k approaches as and exceeds $4l_v$, the equivalence classes become large that generalization typically destroys the QI information leading to high information loss.

The fundamental computational loop of *CandidateBuilder* cycles through k values from 2 to $4l_v$, generating $4l_v - 1$ unique anonymized dataset by calling the *GetAnonymized* module as shown in Figure 2. The choice of $k = 2$ as the lower bound indicates the minimum meaningful k-anonymity guarantee, $k = 1$ would provide no anonymization as each record would belongs to its own equivalence class and returns the original data as anonymized data.

The organized iteration produces a solution space Ω containing $4l_v - 1$ candidate solutions, each representing a heterogeneous point in the utility-privacy tradeoff ecosystem. For each candidate solution ω_i, the module computes a comprehensive collection of metrics to capture the anonymization quality. The privacy loss and information loss metrics are inherited directly from the *GetAnonymized* module. However, recognizing that the terminal merit of anonymized healthcare data is situated in its capacity to support valid inferential and analytical tasks, we introduce an additional practical utility metric through classification recall (R) computation, which quantifies the proportion of relevant information retained after sanitization. (The choice of recall as a metric rather than classification accuracy or precision demonstrates the domain-specific priorities). Recall is a significant measure in healthcare data analysis and research as it measures model's sensitivity in detecting positive cases.The goal is to reduce the False Negatives (FN) in the actual prediction. FN represents the number instances where a patient is having a disease and the model predicted not having disease. The module trains a classifier on each anonymized variant $\mathcal{H}^{C^*}$ and computes its recall performance.

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_{4l_v-1}\} \tag{4}$$

where $|\Omega| = 4l_v - 1$. Each candidate solution $\omega_i \in \Omega$ is a tuple

$$\omega_i = (\mathcal{H}^{C^*}, PL, IL, R)_i \tag{5}$$

containing the anonymized dataset along with its respective privacy loss, information loss, and recall. These tuples together populate the solution space which serves as input to subsequent optimization module that select the most appropriate candidate.

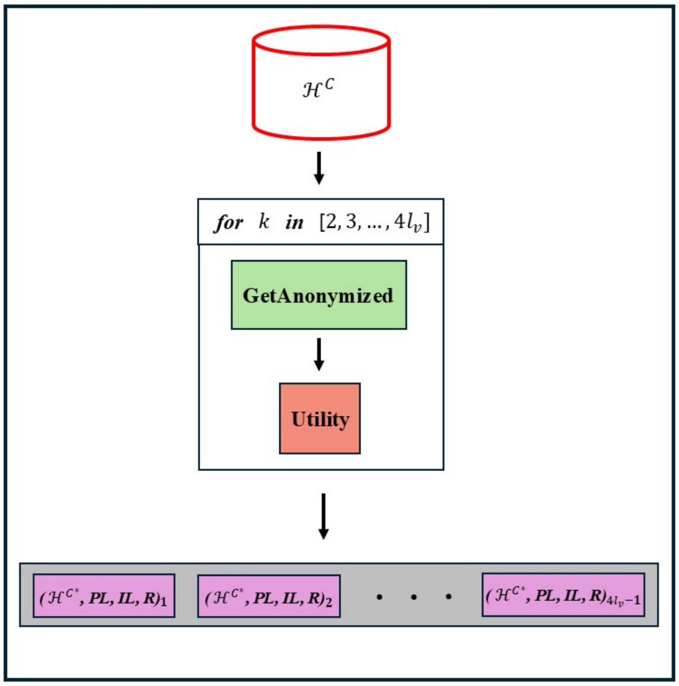

$\mathcal{H}^C$
$for\ \ k\ \ in\ \ [2, 3, \ldots, 4l_v]$
GetAnonymized
Utility
$(\mathcal{H}^{C^*}, PL, IL, R)_1$
$(\mathcal{H}^{C^*}, PL, IL, R)_2$
• • •
$(\mathcal{H}^{C^*}, PL, IL, R)_{4l_v-1}$

Figure 2: CandidateBuilder flow diagram

3.4.1. Recall Computation Methodology

Before invoking the *GetAnonymized* to create a solution space and soon after the preprocessing we set up a benchmark to ensure that the classification recall measurements accurately capture the practical analytical value of the anonymized datasets by training multiple classification algorithms on the $\mathcal{H}^C$ to identify and eliminate algorithms exhibiting overfitting, i.e., we set a threshold of 15% as the accuracy gap between the training and testing data. The algorithms demonstrating stable performance across the classification accuracy form the evaluation ensemble. We use this ensemble algorithms during the *CandidateBuilder*'s iteration process, to get the classification recall across each anonymized version. The recorded recall R for the each candidate solution ω_i is the average recall across all classifiers in the ensemble

$$R_i = \frac{1}{m} \sum_{j=1}^{m} \text{Recall}_{ij} \tag{6}$$

where m is the number of classifier algorithms in the valid ensemble and Recall_{ij} is the recall value of the j-th algorithm on the i-th anonymized dataset calculated as:

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegatives}} \tag{7}$$

By averaging the recall rather than depending on a the behavior of the single model, we reduce the bias that the classification algorithms introduce.

3.4.2. Solution Space Properties

The solution space Ω possesses several key structural properties that underpin its interpretation and practical use:

- The candidates are tacitly organized in increasing order of the privacy parameter k that corresponding to decreasing information loss and increasing privacy protection. The ordering naturally facilitates efficient visualization of the privacy-information tradeoff as seen in Figure 4(A) .
- The solution space captures non-linear relationships between k and the metrics that enables to identify *sweet spots* where privacy and utility are optimally balanced as seen in Figure 4(B).
- Although the losses captures valuable theoretical insight into the anonymization process, they do not directly answer the pragmatic question: “Can this anonymized dataset still support clinical predictions we care about?” By contrast, recall explicitly addresses this question directly. Incorporating and tracking recall as an additional utility measure bridges this gap between the measure of privacy loss and information loss and their practical implications.

3.4.3. Algorithm

Algorithm 2: CandidateBuilder: Solution Space Generator

```
Input: $\mathcal{H}^C$ (preprocessed dataset), QI columns, SI
    columns
Result: Ω (solution space of sanitized dataset variants)
// Compute maximum cardinality
$l_v \leftarrow \max(l_1, l_2, l_3, \ldots, l_s)$;
// where $l_x$ is the number of unique instances
    in $x^{th}$ SI column
$\Omega \leftarrow \emptyset$;
// Generate candidate solutions
for $k \leftarrow 2$ to $4l_v$ do
    $(\mathcal{H}^{C^*}, PL, IL) \leftarrow$ GetAnonymized($\mathcal{H}^C$, QI, SI, $k$);
    // Compute data utility through recall
    Train classifier on $\mathcal{H}^{C^*}$;
    $R \leftarrow$ ComputeRecall(classifier);
    // Store candidate solution
    $\omega_{k-1} \leftarrow (\mathcal{H}^{C^*}, PL, IL, R)$;
    $\Omega \leftarrow \Omega \cup \{\omega_{k-1}\}$;
end
return Ω;
```

3.4.4. Time Complexity Analysis

Finding the maximum cardinality l_v among s SI columns requires scanning each SI column which takes $O(n \cdot s)$ time. Generating each solution candidate involves calling GetAnonymized algorithm iteratively from $k = 2$ to $k = 4l_v$, creating $(4l_v - 1)$ iterations. Each GetAnonymized iteration takes $O(n^2 \cdot m + n \log n)$. Training ensemble classifiers and computing the aggregate recall takes $O(n \cdot m \cdot d)$ where d is the ensemble classifier complexity.

Overall Time Complexity: $O(l_v \cdot (n^2 \cdot m + n \cdot m \cdot d))$

3.4.5. Space Complexity Analysis

- **Input Storage:** Original dataset $\mathcal{H}^C$ requires $O(n \cdot m)$ space.
- **Solution Space Ω:** Stores $(4l_v - 1)$ candidate solutions, where each ω_i contains:
  - Anonymized dataset $\mathcal{H}^{C^*}$: $O(n \cdot m)$
  - Metrics (PL, IL, R): $O(1)$

  Total for Ω: $O(l_v \cdot n \cdot m)$
- **Classifier Storage:** Temporary space for training classifier: $O(n \cdot m)$

Overall Space Complexity: $O(l_v \cdot n \cdot m)$. This is the dominant space requirement as it stores all candidate anonymized datasets.

3.5. Optimizer

The *Optimizer* module implements the decision-making logic to select the optimal sanitized dataset from the solution space Ω generated by *CandidateBuilder*. This module employs a composite objective function $\Im(\omega)$ that consolidates the multidimensional metrics into a single scalar optimization criterion to systematically evaluate every candidate and returns the best balance across all three metrics. The optimization problem is formally defined as:

$$\omega^* = \arg\min_{\omega \in \Omega} \Im(\omega) \tag{8}$$

where ω^* represents the solution that minimizes the objective function,

$$\Im(\omega_i) = \frac{PL_i + IL_i}{R_i} \tag{9}$$

The rationale behind the objective function is to achieve maximum analytical utility while incurring minimum privacy and information costs. The numerator $PL_i + IL_i$ represents the total cost of anonymization, where lower PL indicates stronger privacy protection and lower IL indicates better data quality preservation. The function treats both losses as equally important costs to minimize. The denominator R_i represents the analytical utility in the anonymized dataset through classification recall. Higher recall means the model is better at identifying actual positives. Placing recall in the denominator ensures that the configuration in Ω maintain high utility. The ratio structure creates an inherent tradeoff mechanism. Configurations with low total cost and high recall achieve low objective function values and are preferred.

The *Optimizer* uses an exhaustive search strategy that evaluates every candidate solution in Ω against the objective function to identify the global minimum. This exhaustive linear search is computationally feasible because it requires simple arithmetic operations on pre-computed metrics.

3.5.1. Algorithm

```
Algorithm 3: Optimizer: Optimal Solution Selector
Input: $\Omega = \{\omega_1, \omega_2, \ldots, \omega_{4l_v-1}\}$ (solution space)
Result: $\omega^*$ (optimal sanitized dataset with metrics)
$\Im_{min} \leftarrow \infty$;
$\omega^* \leftarrow$ null;
// Evaluate each candidate solution
foreach $\omega_i \in \Omega$ do
    Extract $(\mathcal{H}^{C^*}{}_i, PL_i, IL_i, R_i)$ from $\omega_i$;
    // Compute objective function
    $\Im(\omega_i) \leftarrow \frac{PL_i+IL_i}{R_i}$;
    // Update optimal solution if better
    if $\Im(\omega_i) < \Im_{min}$ then
        $\Im_{min} \leftarrow \Im(\omega_i)$;
        $\omega^* \leftarrow \omega_i$;
    end
end
// Extract optimal components
$(\mathcal{H}^{C^*}{}_{optimal}, PL^*, IL^*, R^*) \leftarrow \omega^*$;
return $\mathcal{H}^{C^*}{}_{optimal}$, $PL^*$, $IL^*$, $R^*$;
```

3.5.2. Time Complexity Analysis

The optimizer module evaluates $(4l_v - 1)$ candidates in the solution space and for each candidate it computes the objective function in a constant time of $O(1)$.

Overall Time Complexity: $O(l_v)$

3.5.3. Space Complexity Analysis

- **Input Storage:** Solution space Ω requires $O(l_v \cdot n \cdot m)$ space.
- **Working Variables:** Minimum objective value $\Im_{min}$ and optimal solution ω^* require $O(n \cdot m)$ space.

Overall Space Complexity: $O(l_v \cdot n \cdot m)$

Unlike traditional approaches where practitioners or data engineers must pre-define privacy constraints (k or t) through trail and error, MASS's integrated system of *GetAnonymized*, *CandidateBuilder*, and *Optimizer* operates as a self-tuning framework that automatically identifies the anonymized dataset with optimal privacy-utility balance.

3.6. Solving the NP-Hard Problem

Give a dataset with n records, QI, privacy parameter k, and an information loss metric, finding the anonymized version that minimizes the information loss while upholding k-anonymity requires exploring an exponentially large solution space which fundamentally is a NP-Hard [19, 20]. When explicit constraints for l-diversity and t-closeness are introduced, the computational complexity dramatically increases [5, 6]. The algorithms must verify these additional constrains by checking at each partition step which further expands the exponential search space. Enforcing explicit l and t constrains results in a nested optimization loops, where the algorithms first partition for k constraint, check diversity constraints, then measure SI distributional distance, and backtracking when the constraints are not satisfied [21, 22]. When handling datasets with several SI columns, finding the combination of SI values that satisfy both l and t constraints across all attributes concurrently compounds the difficulty, making optimal solutions computationally infeasible even for moderately sized datasets.

The MASS framework addresses this intractability through a decomposition-based polynomial-time heuristic that achieves near-optimal information loss, without claiming to compute the global optimum. The time complexity of $\boldsymbol{O(l_v \cdot (n^2 \cdot m + n \cdot m \cdot d))}$ implicitly satisfies l-diversity and t-closeness without additional computational overhead.The partitioning of $\mathcal{H}^C$ into SI-homogeneous subsets transforms the global optimization problem into several independent smaller problems. Then we employ a similarity based greedy clustering within each subset which reduces generalization distance and thus the information loss. Critically our framework achieves l-diversity and t-closeness as emergent properties of the decomposition design and not through explicit constraint checking. The SI-based subset creation ensures that equivalence classes naturally contain records from multiple different subsets that ensures diversity in SI columns (implicit l-diversity). The identical allocation strategy dispenses SI combinations proportionally across all equivalence classes which leads to minimizing distributional distances (implicit t-closeness). These privacy properties are achieved without any additional computational cost beyond the base k-anonymity algorithm, i.e.,

- no diversity verification loops,
- no SI distribution distance verification loops,
- no constraint-violation backtracking.

MASS's polynomial-time complexity enables practical anonymization of large datasets where optimal algorithms are infeasible, i.e., $\boldsymbol{O(l_v \cdot (n^2 \cdot m + n \cdot m \cdot d))}$ time complexity remains unchanged regardless of whether we consider only k-anonymity or the combination of k-anonymity, l-diversity, and t-closeness which is a significant notable advantage over methods that must explicitly enforce these three constraints, positioning the framework as a scalable heuristic rather than an exact solver. The time and space complexities are summarized in Table 2.

Table 2: Time and Space complexities

Algorithm	Time Complexity	Space Complexity
GetAnonymized	$O(n^2 \cdot m + n \log n)$	$O(n \cdot m)$
CandidateBuilder	$O(l_v \cdot (n^2 \cdot m + n \cdot m \cdot d))$	$O(l_v \cdot n \cdot m)$
Optimizer	$O(l_v)$	$O(l_v \cdot n \cdot m)$
Overall	$\boldsymbol{O(l_v \cdot (n^2 \cdot m + n \cdot m \cdot d))}$	$\boldsymbol{O(l_v \cdot n \cdot m)}$

4. Experimental Evaluation and Results

We conduct wide-range experiments on synthetic healthcare datasets of varying sizes to validate and study the effectiveness and scalability of the MASS framework. The experimental design systematically examines MASS's performance across settings ranging from small-scale pilot studies to large-scale institutional data repositories. Since publicly available healthcare datasets are typically pre-anonymized, which would confound our analysis of the anonymization process, we use synthetically generated datasets that allow us to control the original data characteristics and systematically evaluate the impact of our anonymization framework.

4.1. Setup

We generate four synthetic healthcare datasets with realistic attribute distributions and correlation patterns representative of electronic health records. The datasets contain identical schema with multiple QIs (age, gender, zip code, occupation, race), multiple SIs (diagnosis, treatment, symptoms, total bill, high risk), and OI (visit date, blood pressure systolic, blood pressure diastolic, heart rate, bmi, temperature) but vary in size:

- **Small-scale (1K records):** Representative of specialized clinical studies or rare disease registries
- **Medium-scale (10K records):** Typical of single-institution research datasets
- **Large-scale (100K records):** Representative of multi-institutional collaborative studies
- **Very large-scale (10M records):** Approaching national-level healthcare databases

This scale progression enables evaluation of MASS across realistic deployment scenarios while assessing computational scalability as dataset size increases by four orders of magnitude. Each feature is generated from empirically validated distributions to ensure clinical realism.

All experiments were run on Dell XPS series laptop (2.60 GHz 13th gen i9 processor with 32 GB RAM and 8 GB graphics card) using Python 3.13 to implement the framework.

4.2. Evaluation Objectives

Our experiments investigate three key research questions as we apply MASS to varying data sizes:

- We examine how different classification algorithms behave on anonymized data compared to original data. While recall drives our optimization function, we also track classification accuracy to understand the broader impact of anonymization on predictive modeling. This reveals whether accuracy and recall degrade similarly or exhibit different sensitivity patterns as anonymization strength increases. We expect the anonymized datasets to perform slightly better due to the fact that we introduce additional data points while generalizing the QIs.
- We analyze the multidimensional tradeoff between privacy loss, information loss, and utility as k increases which is expected in PPDP. This analysis helps us to understand if the privacy gains justify utility costs at various k levels and examine how information loss correlates with actual utility degradation.
- We validate the computational scalability of our approach by measuring how execution time scales with dataset size and k value to confirm that our polynomial-time heuristic achieves practical runtime performance. We expect the time per-k execution to show a slight decrease as k increases for smaller dataset because fewer equivalence classes need to be formed and should execute in seconds. As the data size increases the execution should increase.

For each dataset, we follow a systematic protocol beginning with classifier validation, proceeding through comprehensive anonymization testing, and concluding with boundary validation.

4.3. Classifier Ensemble Validation

We begin by identifying a valid ensemble of classifiers that demonstrate dependable generalization on $\mathcal{H}^C$ to ensure robust and meaningful utility evaluation. We train seven widely used classification models–Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), k-Nearest Neighbors, Gradient Boosting (XGBoost), and Naive Bayes– on $\mathcal{H}^C$. Each model is assessed on held-out test data where both training and test accuracy are recorded. The generalization behavior of each model is quantified via an accuracy gap measured as the absolute difference between their training and test accuracy. Those models that exhibit an accuracy gap of 15% are deemed to overfit the data and are excluded from subsequent evaluation. This screening ensures that our utility measurements are driven by true predictive capability rather than artifacts of model memorization. The validation results are summarized in Table 3, showing training and test metrics for each algorithm along with overfitting status. The accuracy gap is visualized in Figure 3 with the 15% threshold marked in black dashed line.

Decision Trees and Random Forests display pronounced overfitting across all four dataset sizes, achieving perfect training accuracy (1.000) while test accuracy remains around 55-66%. In both cases, they consistently show accuracy gaps exceeding 34% which are well above the established 15% threshold. Consequently, these classifiers are removed from the validated ensemble. Gradient Boosting similarly overfits on the small-scale dataset, exhibiting an accuracy gap of 18.08%. However, its performs stabilizes on larger datasets, indicating that limited training data exacerbates its tendency to memorize patterns.

We note that the decision tree-based models can often be stabilized through parameter tuning like pruning or depth constraints. However, we intentionally avoid classifier-specific optimization in this study since the objective in this study is to evalaute how anonymization generalizer under a uniform and minimal-configuration setting. Applying classifier-specific optimization would introduce additional degrees of freedom.

In contrast, the remaining four classifiers, i.e., Logistic Regression, SVM, K-Nearest Neighbors, and Naive Bayes, exhibit consistent generalization across all dataset sizes with accuracy gaps below 15%. Logistic Regression demonstrates remarkable consistency with gaps of under 1% across all scales. SVM and Naive Bayes maintain gaps under 7% and 3% respectively. K-Nearest Neighbors exhibits slightly higher gap but still acceptable around 13%. On the medium, large, and very large datasets, Gradient Boosting also joins the validated ensemble with gaps under 5%. Interestingly, test accuracy remains relatively stable across dataset sizes for most classifiers, typically ranging from 60-66% suggesting the synthetic data's inherent complexity rather than sample size determines predictive capacity. The validated ensemble of 4-5 classifiers (depending on dataset) provides a robust foundation for evaluating anonymization impact, with ensemble-averaged baseline test accuracy around 63-65% across all scales.

Table 3: Baseline classification performance on original (non-anonymized) data

Dataset	Classifier	Training Accuracy	Test Accuracy	Accuracy Gap (%)	Overfit?
Small-scale	Logistic Regression	0.670	0.665	0.42	No
	Decision Tree	1.000	0.606	39.33	**Yes**
	Random Forest	1.000	0.660	34.00	**Yes**
	SVM	0.791	0.663	12.83	No
	K-Nearest Neighbors	0.736	0.606	13.00	No
	Gradient Boosting	0.830	0.650	18.08	**Yes**
	Naive Bayes	0.633	0.660	2.67	No
Medium-scale	Logistic Regression	0.660	0.659	0.05	No
	Decision Tree	1.000	0.553	44.65	**Yes**
	Random Forest	1.000	0.643	35.70	**Yes**
	SVM	0.721	0.653	6.80	No
	K-Nearest Neighbors	0.728	0.603	12.52	No
	Gradient Boosting	0.695	0.653	4.24	No
	Naive Bayes	0.639	0.621	1.80	No
Large-scale	Logistic Regression	0.658	0.649	0.83	No
	Decision Tree	1.000	0.555	44.42	**Yes**
	Random Forest	1.000	0.638	36.20	**Yes**
	SVM	0.699	0.646	5.32	No
	K-Nearest Neighbors	0.735	0.596	13.94	No
	Gradient Boosting	0.672	0.650	2.16	No
	Naive Bayes	0.645	0.630	1.50	No
Very large-scale	Logistic Regression	0.6510	0.645	0.56	No
	Decision Tree	1.000	0.559	44.01	**Yes**
	Random Forest	1.000	0.635	36.42	**Yes**
	SVM	0.682	0.644	3.80	No
	K-Nearest Neighbors	0.731	0.593	13.78	No
	Gradient Boosting	0.657	0.645	1.25	No
	Naive Bayes	0.640	0.628	1.24	No

By excluding the Decision Trees and Random Forests from the majority of datasets, our utility assessment focuses on algorithms with demonstrated generalization capability. This is appropriate since we aim to measure how anonymization affects genuine predictive patterns rather than memorization. To format the results into tables and visualize the patterns efficiently, we exclude Gradient Boosting too. The validated ensemble spans diverse learning paradigms from linear (Logistic Regression), to distance-based (SVM, K-Nearest Neighbors), and probabilistic (Naive Bayes) which enables comprehensive and balanced assessment of utility across diverse modeling approaches.

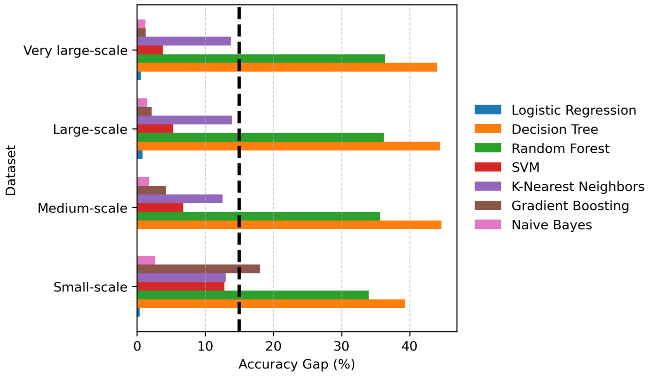


Figure 3: Classification accuracy gap across dataset

4.4. Solution Space Generation and Evaluation

After establishing the validated ensemble, we find each dataset's SI structure to determine the range of k. Table 4 presents the maximum cardinality for each dataset, which we use to determining the upper bound of $4l_v - 1$. While we execute the anonymization process for every value of k within the exploration range, presenting all the results is space extensive and obscure the key tends (for example: the Medium scale data has 43 set of results, instead we prune it to 15). Therefore, we also show the representative k values for which the results are recorded in the subsequent utility measurement. These values are chosen after completing all the experiments and identifying the optimum anonymized version for each dataset. This presentation strategy ensures that we are not unintentionally biased by selectively displaying only convenient values. We provide transparency about the solution quality by including dense sampling around the data-driven optimal k.

We present the five core anonymization metrics (privacy loss, information loss, computational time, ensemble-averaged recall, and the objective function) in Table 5 across the representative anonymity levels for all four datasets. These aggregate metrics are visualized in Figure 4(A) revealing the fundamental privacy-information tradeoffs present in PPDP (i.e., how they evolve inversely as anonymization strength increases). Figure 4(B) tracks the execution time in seconds across the datasets and Figure 5 displays the objective function $\mathfrak{I}(\omega)$ showing where optimal configurations emerge.

MASS upholds the classic privacy-information loss trade off, i.e., as k increases, privacy loss consistently decreases which indicates stronger privacy protection and information loss monotonically increases which indicating the effect of anonymization through generalization. The tradeoff pattern is consistent across all four datasets, for small-scale data, privacy loss drops from 0.367 at $k = 2$ to 0.078 at $k = 27$ which accounts to a 79% reduction, while information loss rises from 0.076 to 0.462 a 508% increase. On medium-scale, large-scale, and very large-scale datasets privacy loss declining from 0.402 to 0.051 (87% reduction), 0.412 to 0.045 (89% reduction), and 0.506 to 0.011 (98% reduction), and information loss increases from 0.074 to 0.436 (489% increase), 0.072 to 0.515 (615% increase), and 0.073 to 0.512 (601% increase) respectively.

Table 4: Dataset Characteristics and k-Parameter Exploration

Dataset	l_v	$4l_v - 1$	**Representative k**
Small-scale	7	27	2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 27
Medium-scale	11	43	2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 43
Large-scale	16	63	2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 45, 55, 63
Very large-scale	19	75	2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 45, 55, 65, 75

Despite the persistent growth of information loss with k, Table 5 shows that ensemble-averaged recall remains remarkably stable, fluctuating within narrow ranges: 0.631-0.749 for small-scale, 0.750-0.778 for medium-scale, 0.780-0.791 for large-scale, and 0.781-0.792 for very large-scale.

This stability is achieved due to the addition on a data point while generalizing the QI columns and allows the objective function to identify meaningful optima where the combined burden of privacy and information loss is minimized relative to preserved utility. Figure 5 visualizes these U-shaped objective function curves, with minima occurring at $k = 4$ (small), $k = 9$ (medium), $k = 8$ (large), and $k = 9$ (very large). These optimal points correspond to configurations in Table 5 where objective values reach their minima: 0.533, 0.559, 0.479, and 0.547 respectively. These optimal configurations achieve privacy loss values below 0.25 while maintaining recall above 0.70.

The computation time on small and medium scale datasets decrease with increase in k due to fewer equivalence classes.Small-scale shows clear descent from 0.507s to 0.231s and medium-scale drops from 5.66s to 2.44s. Large scale dataset shows modest decrease from 493.5s to 421.1s. The very large scale shows relatively stable times fluctuating around 3000s to 3500s across k values rather than a strictly monotonic trend. This behavior is seen due to the interaction between MASS's stratification-based processing and system-level effects. At this scale, the algorithmic variations are further amplified by Python runtime overhead, memory access patterns, and cache misses. These variations dominate execution time once the dataset no longer fits entirely in fast memory reflecting data-dependent partitioning and hardware-level effects rather than instability in the MASS algorithm itself.

Table 6 provides detailed per-classifier performance data showing accuracy, recall, accuracy retention percentage, and recall retention percentage at each representative k value. The detailed breakdown enables us to understand the effect of anonymization effectively. The retention patterns are visualized in Figure 6(A) for accuracy and Figure 6(B) for recall across all four datasets. These retention patterns are re-framed as percentage gains or losses from the baseline in Figure 7 highlighting where anonymization provides unexpected benefits (values above 0%) versus expected degradation (values below 0%).

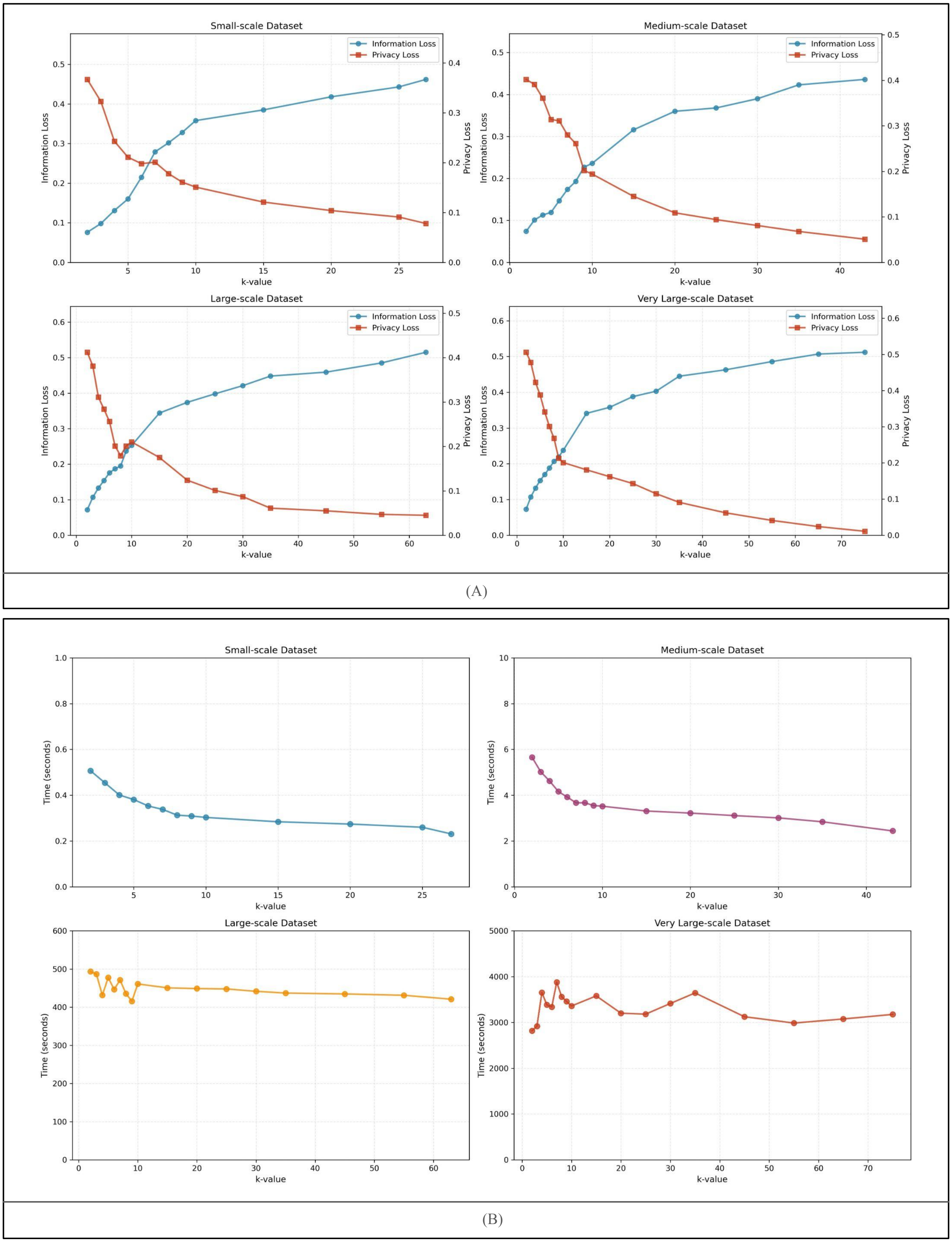
The retention percentages are computed as:

$$\text{Recall Retention (\%)} = \frac{\text{Recall}_{\text{anonymized}}}{\text{Recall}_{\text{original}}} \times 100 \qquad (10)$$

$$\text{Accuracy Retention (\%)} = \frac{\text{Accuracy}_{\text{anonymized}}}{\text{Accuracy}_{\text{original}}} \times 100 \qquad (11)$$

Table 5: Metrics across dataset

Dataset	k	**Info Loss**	**Privacy Loss**	**Time (s)**	**Recall (avg)**	$\mathfrak{I}(\omega)$
Small-scale	2	0.076	0.367	0.507	0.730	0.607
	3	0.098	0.323	0.454	0.749	0.562
	4	0.131	0.243	0.401	0.702	0.533
	5	0.160	0.211	0.381	0.663	0.560
	6	0.215	0.198	0.353	0.731	0.565
	7	0.279	0.201	0.338	0.729	0.658
	8	0.302	0.178	0.313	0.711	0.675
	9	0.328	0.161	0.309	0.694	0.705
	10	0.358	0.151	0.303	0.676	0.753
	15	0.385	0.121	0.284	0.652	0.776
	20	0.418	0.104	0.274	0.721	0.724
	25	0.443	0.091	0.260	0.642	0.832
	27	0.462	0.078	0.231	0.631	0.856
Medium-scale	2	0.074	0.402	5.66	0.778	0.612
	3	0.101	0.391	5.02	0.773	0.637
	4	0.113	0.361	4.62	0.750	0.632
	5	0.119	0.314	4.16	0.772	0.561
	6	0.147	0.311	3.92	0.775	0.591
	7	0.174	0.280	3.67	0.757	0.600
	8	0.193	0.261	3.67	0.768	0.591
	9	0.227	0.202	3.55	0.768	0.559
	10	0.236	0.194	3.52	0.767	0.561
	15	0.316	0.145	3.31	0.770	0.599
	20	0.360	0.109	3.22	0.763	0.615
	25	0.368	0.094	3.11	0.764	0.605
	30	0.390	0.081	3.01	0.765	0.616
	35	0.423	0.068	2.84	0.766	0.641
	43	0.436	0.051	2.44	0.766	0.636
Large-scale	2	0.072	0.412	493.5	0.787	0.615
	3	0.107	0.381	486.8	0.786	0.621
	4	0.133	0.311	431.6	0.784	0.566
	5	0.154	0.284	477.6	0.782	0.560
	6	0.176	0.256	446.7	0.784	0.551
	7	0.187	0.201	471.2	0.784	0.495
	8	0.195	0.179	435.7	0.780	0.479
	9	0.237	0.201	415.7	0.786	0.558
	10	0.253	0.210	461.0	0.791	0.585
	15	0.344	0.175	450.8	0.786	0.661
	20	0.374	0.124	448.9	0.783	0.636
	25	0.398	0.101	447.9	0.781	0.639
	30	0.421	0.087	441.7	0.781	0.651
	35	0.448	0.061	437.1	0.781	0.652
	45	0.459	0.055	434.8	0.789	0.652
	55	0.485	0.047	431.3	0.785	0.678
	63	0.515	0.045	421.1	0.782	0.716
Very Large-scale	2	0.073	0.506	2819.97	0.788	0.735
	3	0.107	0.478	2919.72	0.786	0.744
	4	0.132	0.423	3653.78	0.784	0.708
	5	0.153	0.388	3385.69	0.789	0.686
	6	0.170	0.341	3340.55	0.791	0.646
	7	0.188	0.301	3880.88	0.791	0.618
	8	0.207	0.268	3562.65	0.791	0.601
	9	0.219	0.213	3459.25	0.790	0.547
	10	0.238	0.201	3360.35	0.789	0.557
	15	0.341	0.181	3580.39	0.789	0.662
	20	0.358	0.162	3202.45	0.792	0.657
	25	0.388	0.143	3182.46	0.781	0.680
	30	0.403	0.115	3415.12	0.790	0.656
	35	0.445	0.091	3645.22	0.791	0.678
	45	0.463	0.062	3125.54	0.790	0.665
	55	0.486	0.041	2987.71	0.789	0.668
	65	0.507	0.024	3076.84	0.787	0.675
	75	0.512	0.011	3176.52	0.792	0.660



(A)

(B)

Figure 4: Anonymization metrics: privacy loss and information loss trade-off (A), computational time (B)

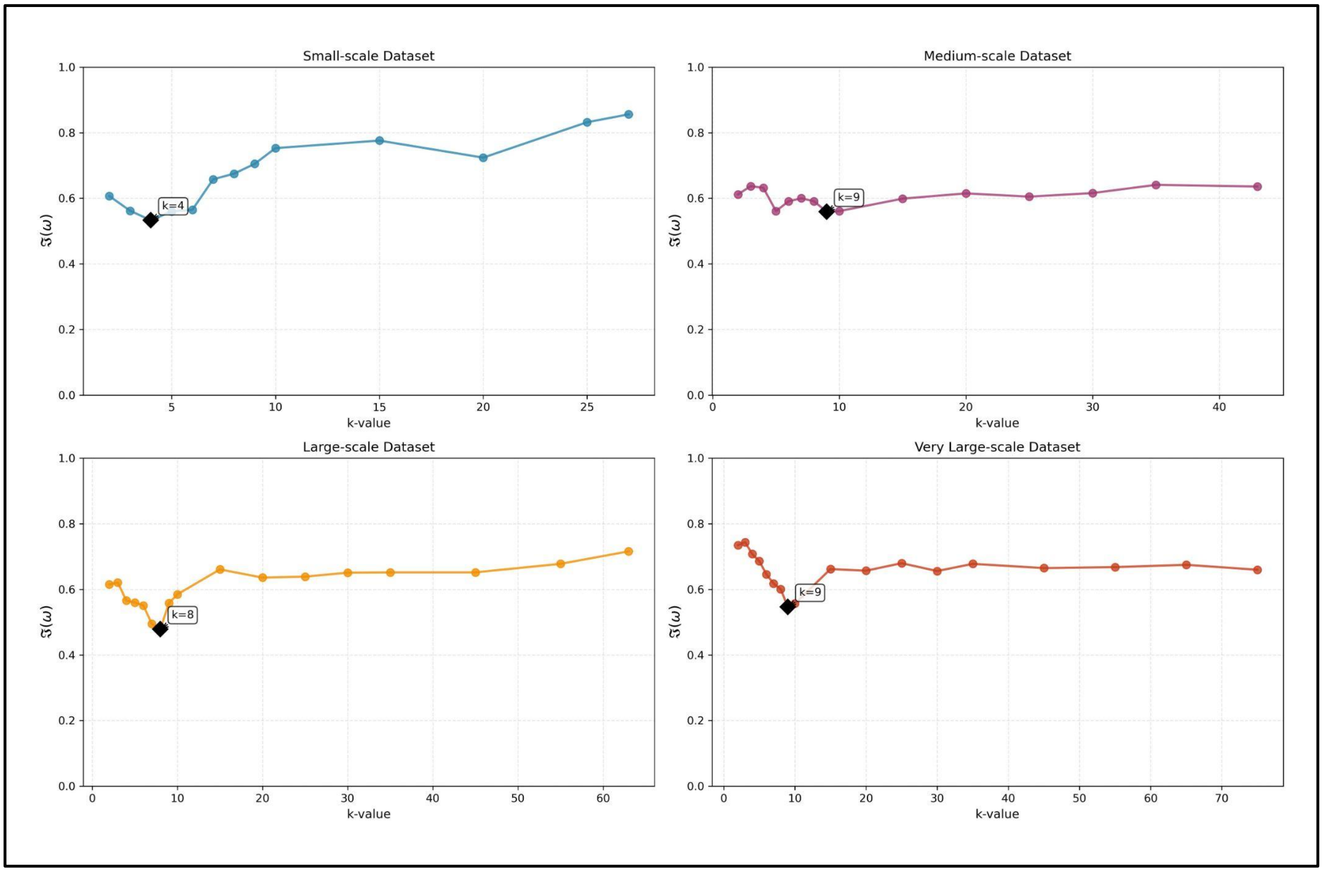
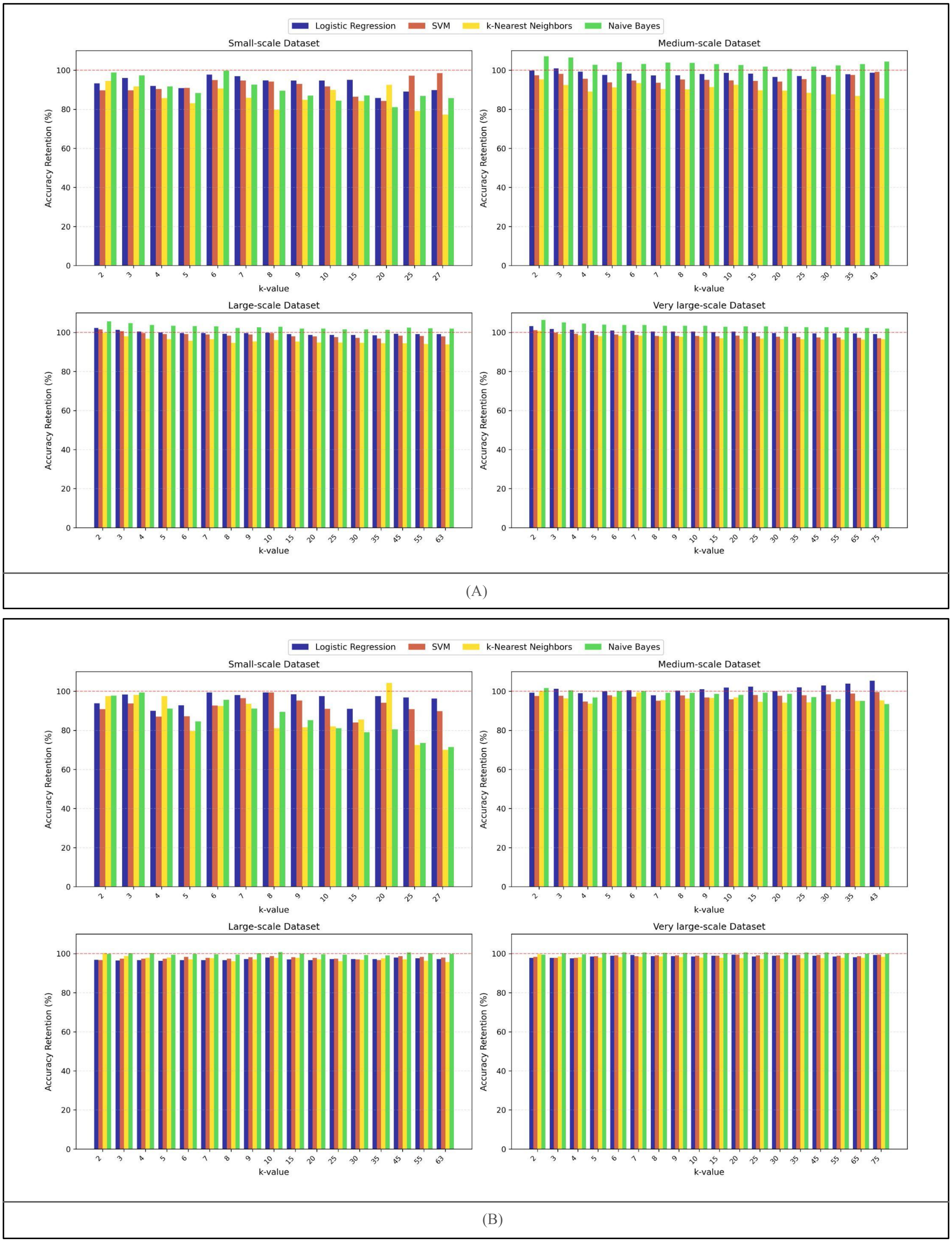


Figure 5: Objective function indicating optimal *k* across datasets

Table 6: Accuracy and Recall Across k Values for All Dataset Scales

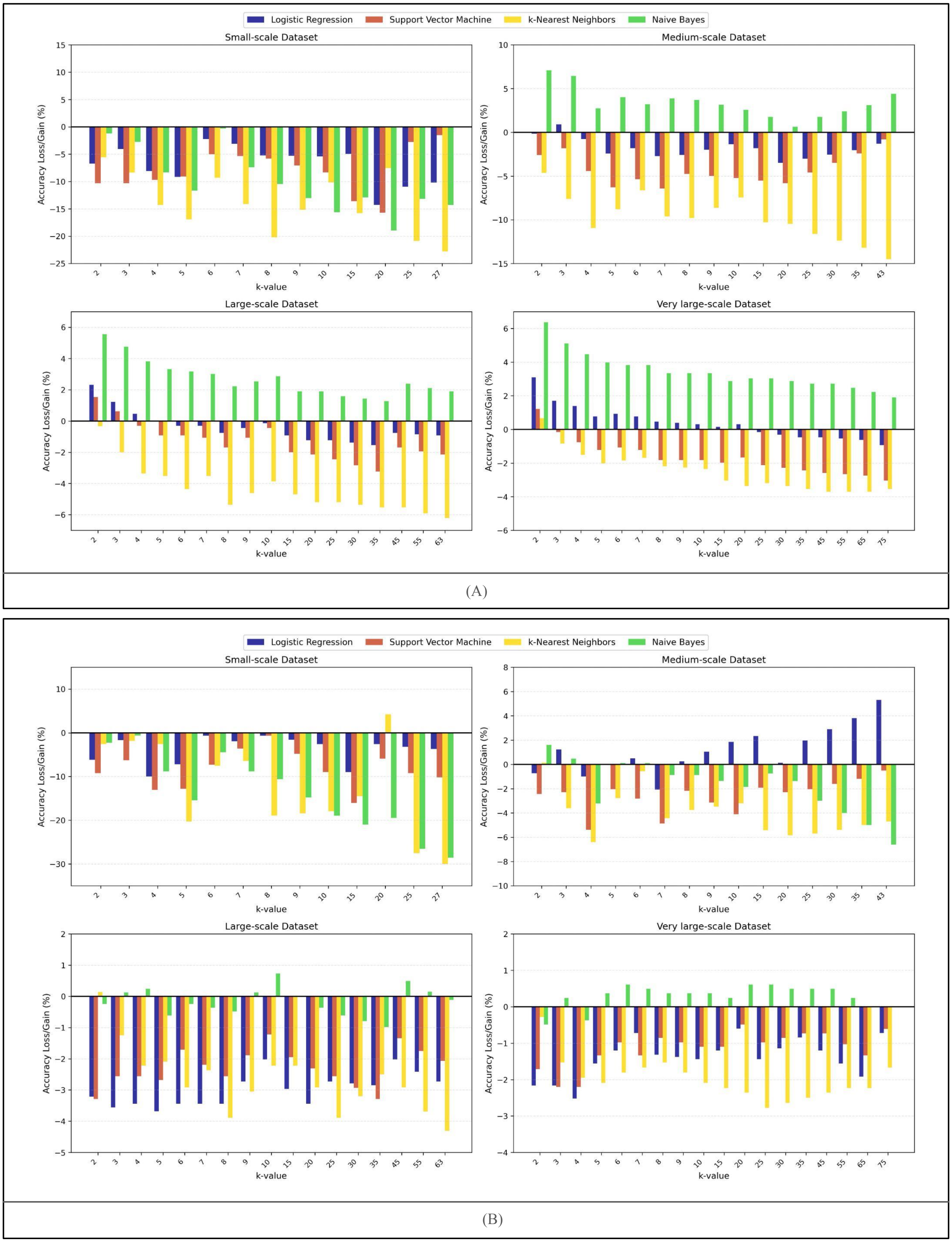
Dataset	k	**Logistic Regression**				**SVM**				**k-NN**				**Naive Bayes**			
		Acc	Rec	Acc_{Ret}	Rec_{Ret}	Acc	Rec	Acc_{Ret}	Rec_{Ret}	Acc	Rec	Acc_{Ret}	Rec_{Ret}	Acc	Rec	Acc_{Ret}	Rec_{Ret}
Small-scale	2	0.621	0.733	93.27	93.85	0.597	0.709	89.71	90.78	0.573	0.703	94.45	97.44	0.652	0.773	98.79	97.72
	3	0.639	0.768	95.94	98.34	0.597	0.732	89.71	93.73	0.556	0.708	91.64	98.13	0.642	0.786	97.27	99.37
	4	0.612	0.703	91.92	90.01	0.601	0.679	90.31	86.94	0.520	0.703	85.71	97.44	0.605	0.721	91.67	91.15
	5	0.605	0.725	90.86	92.83	0.605	0.681	90.91	87.20	0.504	0.575	83.09	79.70	0.583	0.669	88.33	84.58
	6	0.651	0.776	97.78	99.36	0.632	0.724	94.97	92.70	0.550	0.667	90.67	92.45	0.658	0.756	99.70	95.57
	7	0.645	0.766	96.88	98.08	0.630	0.753	94.67	96.41	0.521	0.675	85.88	93.56	0.611	0.721	92.58	91.15
	8	0.631	0.776	94.77	99.36	0.627	0.776	94.21	99.36	0.484	0.585	79.78	81.08	0.591	0.707	89.55	89.38
	9	0.631	0.769	94.70	98.40	0.619	0.744	92.94	95.20	0.515	0.589	84.82	81.57	0.574	0.674	86.97	85.21
	10	0.630	0.761	94.62	97.44	0.610	0.711	91.66	91.04	0.545	0.592	89.85	82.05	0.557	0.641	84.39	81.04
	15	0.633	0.711	95.04	91.04	0.575	0.656	86.40	83.99	0.511	0.617	84.22	85.52	0.575	0.625	87.12	79.01
	20	0.571	0.761	85.75	97.44	0.561	0.735	84.30	94.11	0.561	0.752	92.47	104.23	0.535	0.637	81.06	80.53
	25	0.593	0.756	89.06	96.80	0.647	0.709	97.22	90.78	0.480	0.523	79.13	72.49	0.573	0.581	86.82	73.45
	27	0.598	0.752	89.80	96.30	0.655	0.701	98.50	89.80	0.468	0.505	77.20	70.00	0.566	0.565	85.70	71.40
Medium-scale	2	0.658	0.806	99.85	99.26	0.636	0.762	97.40	97.57	0.575	0.720	95.36	100.14	0.665	0.822	107.09	101.59
	3	0.665	0.822	100.91	101.23	0.641	0.763	98.16	97.70	0.557	0.693	92.37	96.38	0.661	0.813	106.44	100.48
	4	0.654	0.804	99.24	99.01	0.624	0.739	95.56	94.62	0.537	0.673	89.05	93.60	0.638	0.783	102.74	96.78
	5	0.643	0.812	97.57	100.00	0.612	0.765	93.72	97.95	0.550	0.699	91.21	97.22	0.646	0.810	104.03	100.11
	6	0.647	0.816	98.18	100.49	0.618	0.759	94.64	97.18	0.563	0.715	93.37	99.44	0.641	0.810	103.22	100.11
	7	0.641	0.795	97.27	97.91	0.611	0.743	93.57	95.13	0.545	0.687	90.38	95.55	0.645	0.802	103.87	99.13
	8	0.642	0.814	97.42	100.25	0.622	0.764	95.25	97.82	0.544	0.692	90.21	96.24	0.644	0.802	103.71	99.13
	9	0.646	0.821	98.02	101.05	0.621	0.757	95.02	96.86	0.551	0.694	91.38	96.52	0.641	0.798	103.14	98.64
	10	0.650	0.827	98.63	101.85	0.619	0.749	94.79	95.90	0.558	0.696	92.54	96.80	0.637	0.794	102.58	98.15
	15	0.647	0.831	98.18	102.34	0.617	0.766	94.49	98.08	0.541	0.680	89.72	94.58	0.632	0.803	101.77	99.25
	20	0.636	0.813	96.51	100.12	0.615	0.763	94.18	97.70	0.540	0.677	89.55	94.16	0.625	0.798	100.64	98.63
	25	0.639	0.828	96.97	101.97	0.623	0.765	95.41	97.95	0.533	0.678	88.39	94.30	0.632	0.785	101.77	97.02
	30	0.641	0.835	97.45	102.90	0.629	0.767	96.50	98.40	0.528	0.679	87.60	94.60	0.636	0.777	102.40	96.00
	35	0.643	0.842	97.95	103.80	0.635	0.769	97.60	98.80	0.523	0.680	86.80	95.00	0.640	0.769	103.10	95.00
	43	0.647	0.853	98.70	105.30	0.645	0.773	99.20	99.50	0.515	0.682	85.50	95.30	0.648	0.755	104.40	93.40
Large-scale	2	0.664	0.815	102.31	96.79	0.660	0.794	101.54	96.71	0.594	0.720	99.66	100.14	0.665	0.817	105.56	99.76
	3	0.657	0.812	101.23	96.44	0.654	0.800	100.62	97.44	0.584	0.710	97.99	98.75	0.660	0.820	104.76	100.12
	4	0.652	0.813	100.46	96.56	0.648	0.800	99.69	97.44	0.576	0.703	96.64	97.78	0.654	0.821	103.81	100.24
	5	0.649	0.811	100.00	96.32	0.644	0.799	99.08	97.32	0.575	0.704	96.48	97.91	0.651	0.814	103.33	99.39
	6	0.647	0.814	99.69	96.56	0.644	0.807	99.08	98.29	0.570	0.698	95.64	97.08	0.650	0.817	103.17	99.76
	7	0.647	0.814	99.69	96.56	0.643	0.803	98.92	97.81	0.575	0.702	96.48	97.63	0.649	0.816	103.02	99.63
	8	0.644	0.814	99.23	96.56	0.639	0.800	98.31	97.44	0.564	0.691	94.63	96.11	0.644	0.815	102.22	99.51
	9	0.646	0.820	99.54	97.27	0.643	0.806	98.93	98.11	0.568	0.697	95.38	96.95	0.646	0.820	102.54	100.12
	10	0.648	0.825	99.85	97.98	0.647	0.811	99.54	98.78	0.573	0.703	96.14	97.78	0.648	0.825	102.86	100.73
	15	0.643	0.817	99.08	97.03	0.637	0.805	98.00	98.05	0.568	0.703	95.30	97.78	0.642	0.819	101.90	100.00
	20	0.641	0.814	98.77	96.56	0.636	0.802	97.85	97.69	0.565	0.698	94.80	97.08	0.642	0.816	101.90	99.63
	25	0.641	0.819	98.77	97.27	0.634	0.800	97.54	97.44	0.565	0.691	94.80	96.11	0.640	0.814	101.59	99.39
	30	0.640	0.818	98.62	97.21	0.632	0.797	97.16	97.07	0.564	0.696	94.63	96.80	0.639	0.812	101.43	99.20
	35	0.639	0.818	98.46	97.15	0.629	0.794	96.77	96.71	0.563	0.701	94.46	97.50	0.638	0.811	101.27	99.02
	45	0.644	0.825	99.23	97.98	0.639	0.810	98.31	98.66	0.563	0.698	94.46	97.08	0.645	0.823	102.38	100.49
	55	0.643	0.822	99.15	97.59	0.637	0.807	98.05	98.25	0.561	0.692	94.09	96.31	0.643	0.820	102.11	100.15
	63	0.643	0.819	99.08	97.27	0.636	0.804	97.85	97.93	0.559	0.688	93.79	95.69	0.642	0.818	101.90	99.88
Very large-scale	2	0.665	0.814	103.10	97.84	0.665	0.805	101.22	98.29	0.597	0.717	100.67	99.72	0.668	0.816	106.37	99.51
	3	0.656	0.814	101.71	97.84	0.656	0.801	99.85	97.80	0.588	0.708	99.16	98.47	0.660	0.822	105.10	100.24
	4	0.654	0.811	101.40	97.48	0.652	0.801	99.24	97.80	0.584	0.705	98.48	98.05	0.656	0.817	104.46	99.63
	5	0.650	0.819	100.78	98.44	0.649	0.808	98.78	98.66	0.581	0.704	97.98	97.91	0.653	0.823	103.98	100.37
	6	0.651	0.822	100.93	98.80	0.650	0.811	98.93	99.02	0.582	0.706	98.15	98.19	0.652	0.825	103.82	100.61
	7	0.650	0.826	100.78	99.28	0.649	0.808	98.78	98.66	0.583	0.707	98.31	98.33	0.652	0.824	103.82	100.49
	8	0.648	0.821	100.47	98.68	0.645	0.812	98.17	99.15	0.580	0.708	97.81	98.47	0.649	0.823	103.34	100.37
	9	0.648	0.821	100.39	98.62	0.645	0.811	98.17	99.02	0.580	0.706	97.73	98.19	0.649	0.823	103.34	100.37
	10	0.647	0.820	100.31	98.56	0.645	0.810	98.17	98.90	0.579	0.704	97.64	97.91	0.649	0.823	103.34	100.37
	15	0.646	0.822	100.16	98.80	0.644	0.810	98.02	98.90	0.575	0.703	96.96	97.77	0.646	0.822	102.87	100.24
	20	0.647	0.827	100.31	99.40	0.646	0.815	98.33	99.51	0.573	0.702	96.63	97.64	0.647	0.825	103.03	100.61
	25	0.644	0.820	99.84	98.56	0.643	0.811	97.87	99.02	0.574	0.699	96.80	97.22	0.647	0.825	103.03	100.61
	30	0.643	0.823	99.69	98.86	0.642	0.812	97.72	99.15	0.573	0.700	96.63	97.36	0.646	0.824	102.87	100.49
	35	0.642	0.825	99.53	99.16	0.641	0.813	97.56	99.27	0.572	0.701	96.46	97.50	0.645	0.824	102.71	100.49
	45	0.642	0.822	99.53	98.80	0.640	0.813	97.41	99.27	0.571	0.702	96.29	97.64	0.645	0.824	102.71	100.49
	55	0.642	0.819	99.46	98.44	0.640	0.811	97.34	98.97	0.571	0.703	96.29	97.77	0.644	0.822	102.47	100.24
	65	0.641	0.816	99.38	98.08	0.639	0.808	97.26	98.66	0.571	0.703	96.29	97.77	0.642	0.820	102.23	100.00
	75	0.639	0.826	99.07	99.28	0.637	0.814	96.96	99.39	0.572	0.707	96.46	98.33	0.640	0.820	101.91	100.00



(A)

(B)

Figure 6: Classification performance study: (A) % accuracy retention, (B) % recall retention



(A)

(B)

Figure 7: Classification performance study: (A) Accuracy gain/loss, (B) Recall gain/loss from the baseline

5. Conclusion and Future Work

This study addresses fundamental challenges in PPDP for healthcare datasets containing multiple sensitive attributes. We have presented MASS, a comprehensive framework that eliminates the need for manual privacy parameter specification, provides polynomial-time solution to the NP-hard problem of optimal anonymization, and implicitly achieves *l*-diversity and *t*-closeness guarantees through architectural design rather than explicit constraint enforcement.

Our experimental validation across four dataset scales demonstrates that MASS successfully navigates the privacy-utility-information loss tradeoff space. The framework achieves strong privacy protection (privacy loss <0.25 at optimal configurations) while maintaining excellent utility preservation, with recall retention exceeding 95% on datasets with 10,000 or more records. The remarkable retention stability observed in larger datasets where accuracy and recall retention remain within 93-106% and 95-101% ranges respectively throughout the entire *k* parameter space validates that the stratification approach enables aggressive anonymization without catastrophic utility degradation. The occasional performance gains exceeding baseline, where anonymized data outperforms original data, reveal an unexpected benefit: generalization can act as beneficial regularization that improves classifier performance rather than degrading it.

MASS represents a significant advancement in making privacy-preserving healthcare data publishing both theoretically sound and practically deployable. By addressing the dual challenges of multiple sensitive attributes and automated parameter selection, while maintaining computational tractability through our polynomial-time heuristic, we enable organizations to publish valuable health data that advances medical research, supports clinical decision-making, and informs public health policy all while rigorously protecting patient privacy. The framework's open architecture supports future extensions including integration of differential privacy mechanisms, incorporation of alternative anonymization techniques beyond generalization, and adaptation to other domains beyond healthcare where multiple sensitive attributes require protection. Future work will focus on three directions: parallelization to achieve further scalability improvements for very large datasets, integration of multiple anonymization techniques (microaggregation, anatomization, permutation, noise addition) with adaptive selection mechanisms, and extension to dynamic data publishing scenarios where datasets evolve over time. The comprehensive framework presented here provides a solid foundation for these enhancements while already delivering production-ready capabilities for static healthcare data publication. Privacy-preserving healthcare data publishing for multiple sensitive attributes is no longer a theoretical challenge requiring expert intervention and MASS makes it a practical, automated reality where the framework is be applicable to:

- Multi-institutional research databases require standardized de-identification across diverse healthcare systems and data warehouses to track multiple health conditions over time.
- Public health surveillance needs to publish patterns and trends while protecting individual patient privacy.
- Clinical trial requirement where patient characteristics and findings must be released to the public.
- Health information exchanges facilitate data sharing among healthcare providers.

References

[1] "Food and Drug Administration Amendments Act of 2007," Pub. L. No. 110-85, § 801, 121 Stat. 823, 2007, codified at 42 U.S.C. § 282(j).

[2] "Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC," Official Journal of the European Union, 2014, available at: https://eur-lex.europa.eu/eli/reg/2014/536/oj.

[3] L. Sweeney, "k-anonymity: a model for protecting privacy," Int. J. Uncertain. Fuzziness Knowl.-Based Syst., **10**(5), 557–570, 2002, doi:10.1142/S0218488502001648.

[4] P. Samarati, L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," Technical Report SRI-CSL-98-04, SRI International, Menlo Park, CA, 1998.

[5] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," ACM Trans. Knowl. Discov. Data, **1**(1), 3–es, 2007, doi:10.1145/1217299.1217302.

[6] N. Li, T. Li, S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in 2007 IEEE 23rd International Conference on Data Engineering, 106–115, 2007, doi:10.1109/ICDE.2007.367856.

[7] H. Lee, S. Kim, J. Kim, Y. Chung, "Utility-preserving anonymization for health data publishing," BMC Medical Informatics and Decision Making, **17**, 2017, doi:10.1186/s12911-017-0499-0.

[8] A. Majeed, S. O. Hwang, "Solving the Privacy-Equity Trade-off in Data Sharing By Using Homophily, Diversity, and t-Closeness Based Anonymity Algorithm," IEEE Access, **12**, 181953–181974, 2024, doi:10.1109/ACCESS.2024.10772434.

[9] T. Li, N. Li, J. Zhang, I. Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing," IEEE Transactions on Knowledge and Data Engineering, **24**(3), 561–574, 2012, doi:10.1109/TKDE.2010.236.

[10] B. Su, J. Huang, K. Miao, Z. Wang, X. Zhang, Y. Chen, "K-Anonymity Privacy Protection Algorithm for Multi-Dimensional Data against Skewness and Similarity Attacks," Sensors, **23**(3), 1554, 2023, doi:10.3390/s23031554.

[11] Y. Wei, H. Y. Benson, M. Capan, "An Analytical Approach to Privacy and Performance Trade-Offs in Healthcare Data Sharing," arXiv preprint arXiv:2508.18513, 2025, doi:10.48550/arXiv.2508.18513.

[12] S. Li, H. Shen, Y. Sang, H. Tian, "An efficient method for privacy-preserving trajectory data publishing based on data partitioning," The Journal of Supercomputing, **76**(7), 5276–5300, 2020, doi:10.1007/s11227-019-02906-6.

[13] O. Abul, F. Bonchi, M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," in 2008 IEEE 24th International Conference on Data Engineering, 376–385, 2008, doi:10.1109/ICDE.2008.4497446.

[14] A. Aristodimou, A. Antoniades, C. S. Pattichis, "Privacy preserving data publishing of categorical data through k-anonymity and feature selection," Healthcare Technology Letters, **3**(1), 16–21, 2016, doi:10.1049/htl.2015.0050.

[15] H. Wang, J. He, N. Zhu, "Improving Data Utilization of K-anonymity through Clustering Optimization," in Transactions on Data Privacy, volume 15, 177–192, 2022, available at: https://www.tdp.cat/issues21/tdp.a441a21.pdf.

[16] J. Jayaram, P. Manickam, "An efficient privacy-preserving data publishing in health care records with multiple sensitive attributes," in 2021 6th International Conference on Inventive Computation Technologies (ICICT), 623–629, IEEE, Coimbatore, India, 2021, doi:10.1109/ICICT50816.2021.9358639.

[17] Y. Rubner, C. Tomasi, L. J. Guibas, "The earth mover's distance as a metric for image retrieval," International Journal of Computer Vision, **40**(2), 99–121, 2000, doi:10.1023/A:1026543900054.

[18] J. Cao, B. Carminati, E. Ferrari, K.-L. Tan, "CASTLE: Continuously Anonymizing Data Streams," IEEE Transactions on Dependable and Secure Computing, **8**(3), 337–352, 2011, doi:10.1109/TDSC.2009.47.

[19] A. Meyerson, R. Williams, "On the complexity of optimal k-anonymity," in Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '04, 223–228, ACM, Paris, France, 2004, doi:10.1145/1055558.1055591.

[20] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A. Zhu, "Approximation algorithms for k-anonymity," in Journal of Privacy Technology, 20051120001, 2005.

[21] R. Khan, X. Tao, A. Anjum, H. Sajjad, S. u. R. Malik, A. Khan, F. Amiri, "Wireless Communications and Mobile Computing," in Privacy preserving for multiple sensitive attributes against fingerprint correlation attack satisfying c-diversity, volume 2020, 1–18, 2020, doi:10.1155/2020/8416823.

[22] J. Jayapradha, M. Prakash, Y. Alotaibi, O. I. Khalaf, S. A. Alghamdi, "IEEE Access," in Heap bucketization anonymity—An efficient privacy-preserving data publishing model for multiple sensitive attributes, volume 10, 28773–28791, 2022, doi:10.1109/ACCESS.2022.3158312.