ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Data Privacy and Utility Trade-Off: An Efficient K-Anonymization Algorithm with Low Information Loss

Charles R. Haruna, Maame G. Asante-Mensah, Festus S. Doe* and Sandro K. Amofa

Department of Computer Science and Information Technology, University of Cape Coast, Cape Coast, Ghana

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150400063

Received: 11 April 2026; Accepted: 16 April 2026; Published: 08 May 2026

ABSTRACT

Privacy Preserving Data Publishing (PPDP) remains a critical challenge in the era of large-scale data sharing, where the need to balance data utility and individual privacy is inherently conflicting. Among existing models, k-anonymity continues to be widely adopted due to its simplicity and interpretability; however, traditional k-anonymization algorithms suffer from key limitations, including distribution-agnostic partitioning and inadequate handling of outliers, which lead to excessive information loss and reduced data utility.

This paper proposes RAYDEN, a novel hybrid k-anonymization algorithm that integrates distribution-aware VP-tree partitioning with Connectivity-based Outlier Factor (COF) detection to address these limitations. The algorithm employs Gower distance to support mixed-type datasets and introduces a statistically adaptive threshold for robust outlier identification. Unlike existing approaches, RAYDEN incorporates a recursive outlier recovery mechanism that re-partitions detected outliers, maximizing data retention before applying suppression as a last resort.

Experimental evaluation on the UCI Adult dataset demonstrates that RAYDEN consistently outperforms compared algorithms across key utility metrics utilized in the study. The outlier recovery mechanism achieves a mean recovery rate exceeding 90% across all k values, substantially reducing suppression-related information loss compared to Mondrian with COF. While incurring higher computational cost, the algorithm achieves practical execution times and significantly improves the privacy–utility trade-off, particularly at commonly used k values. These results establish RAYDEN as a robust and effective framework for privacy-preserving data publishing in mixed-type datasets.

Keywords: k-anonymity, Privacy-preserving data publishing, Data utility, Outlier detection, VP-tree

INTRODUCTION

The exponential growth of data collection in healthcare, finance, social media, and government sectors has created unprecedented opportunities for data-driven research and decision-making. However, the publication of such data raises significant privacy concerns, as individual identities can often be inferred through linkage attacks even after removal of direct identifiers [1]. Privacy Preserving Data Publishing (PPDP) has emerged as a critical research area aimed at enabling data sharing while protecting individual privacy [2], [3].

The fundamental challenge in PPDP is achieving an optimal balance between privacy protection and data utility, a problem proven to be NP-hard [4], [5]. Excessive privacy protection renders data useless for analysis, while insufficient protection exposes individuals to re-identification risks. This trade-off becomes particularly acute in high dimensional datasets with complex distributions and outlier data points [6]. The legislative response including the European Union's General Data Protection Regulation (GDPR) [7] and Ghana's Data Protection Act 843 [8] has imposed binding obligations on data controllers to implement technical anonymization measures before publication.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Real-world microdata often includes outliers that deviate significantly from the general population pattern, and such irregular observations are commonly found in many datasets that have been made publicly available [9]. Beyond their well-documented impact on data utility, outliers also pose significant risks to data privacy. In response, scholarly work has approached this challenge in two broad ways: some researchers opt to remove outliers entirely from the dataset, while others seek to extract value from them rather than discard them.

Anonymization is a privacy-preserving technique that conceals the identities of individuals in a dataset while still maintaining the usefulness of the data for intended purposes [10]. k-anonymity, introduced by Sweeney [10], remains the most widely adopted formal privacy model for tabular data publication. It requires that every record in a published dataset be indistinguishable from at least k−1 other records with respect to a defined set of quasi-identifiers (QIDs) attributes, that in combination, could identify an individual. Although existing k-anonymity algorithms have gained broad adoption, inherent structural limitations continue to constrain their overall effectiveness. Specifically, current k-anonymization approaches exhibit several notable shortcomings: (i) Algorithms like Mondrian [11] use KD-tree structures that partition data based on frequency distributions rather than actual data similarity, leading to suboptimal groupings and reduced utility. (ii) Most k-anonymity algorithms do not incorporate dedicated mechanisms for handling outliers which can significantly degrade both privacy and utility. Outliers are either blindly included in partitions (causing loose generalization) or suppressed entirely (causing unnecessary information loss) [9].

In the current landscape, safeguarding personal data has emerged as an essential obligation, aimed at preserving the highest possible level of data usability while simultaneously upholding individuals' right to privacy [12]. This study presents RAYDEN, a novel hybrid k anonymization algorithm developed to address the shortcomings of existing approaches mentioned above. RAYDEN seeks to strike an effective balance between data privacy and data utility within a unified framework by employing a distribution-aware anonymization that partitions the data space based on the underlying distribution of data points. This creates equivalence classes by grouping similar records. Simultaneously, it integrates a COF-based outlier detection technique, which identifies outliers using connectivity patterns and statistical thresholds within each partition, and intelligently mitigates the adverse effects of these outliers.

The contributions of this study are as follows:

1. A novel application of VP-Tree partitioning to k-anonymity using Gower distance, which natively handles mixed-type QID spaces without attribute encoding or preprocessing.

2. A statistically adaptive COF threshold that calibrates outlier sensitivity to the density distribution within each partition, eliminating the need for dataset-specific tuning.

3. A recursive outlier recovery strategy that applies VP-Tree re-partitioning to the outlier group at the partition level, recovering equivalence classes before resorting to suppression.

4. Ability to effectively handle both categorical and numerical data types, making it more versatile and applicable to real-world datasets with mixed attributes.

5. The effectiveness of the proposed approach is validated through a comprehensive experimental evaluation on the Adult Dataset, using multiple performance metrics including CAVG, DM, and NCP, where it consistently demonstrates superior performance across all measures.

The remainder of this paper is structured as follows. Section II reviews foundational concepts in data privacy, k-anonymity-based anonymization models, and existing methods. Section III presents the RAYDEN algorithm. Section IV describes the dataset and evaluation metrics employed in the study, reports the experimental results, and discusses the performance of the algorithms. Section V concludes the paper and offers directions for future research.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

REVIEW OF EXISTING LITERATURE

Privacy Models in Data Publication

Several anonymization models have been developed that allow data publishers to release sensitive data in a manner that safeguards individual privacy. Among the various privacy-preserving models applied in data publishing k-anonymity[10], l-diversity[13], t-closeness[14] and differential privacy [15] are the most widely recognized and frequently adopted. Each of these models is briefly described below.

k-anonymity [10] defines the baseline requirement that each published record must be identical to at least k−1 others on all QID attributes. Anonymization's techniques such as masking and generalization are subsequently applied to the QIDs, ensuring that no individual within a group can be uniquely distinguished from others [11]. This ensures that each record within an equivalence class is indistinguishable from at least (k - 1) other records in the same class, thereby providing protection against record linkage attacks. This model has been extensively studied and extended. Machanavajjhala et al. [13] extended this with l-diversity, requiring that the sensitive attribute within each equivalence class contain at least l distinct values, thereby guarding against attribute-disclosure attacks. Li et al. [14] further strengthened the model with t-closeness, demanding that the distribution of sensitive values within each class closely approximates the global distribution. Differential privacy [15], introduced by Dwork, provides a probabilistic framework with rigorous worst-case privacy guarantees but introduces utility trade-offs that make it unsuitable for direct tabular data release in many domains. While these models provide stronger privacy guarantees, k-anonymity remains widely adopted due to its simplicity, computational efficiency, and intuitive interpretation [16], [17]. Despite appearing straightforward on the surface, achieving optimum k-anonymity has been formally established as an NP-hard problem[18]. Recent work by Karagiannis et al. [19] demonstrates k-anonymity's continued relevance in healthcare data sharing, while Yuan et al. [20] show its effectiveness in big data contexts.

Related work

Despite having been first introduced by Sweeney in 2002, k-anonymity continues to feature prominently in contemporary research. Kacha et al.[21] proposed KAB, a k-anonymity algorithm that integrates the Black Hole Algorithm (BHA), a nature-inspired metaheuristic, to address the NP-hard challenge of optimal k-anonymization. KAB combines BHA with an NCP-based clustering technique to find an optimal k-anonymous solution that minimizes information loss while maximizing data utility. Validated on the UCI Adult dataset, the algorithm demonstrates superior performance in balancing privacy protection and data utility. Andrew and Karthikeyan [22] proposed (K, L) Anonymity, a hybrid privacy-preserving model designed for big data publication. Motivated by the limitations of conventional anonymization techniques, which remain vulnerable to re-identification attacks despite their widespread use, the study introduces an approach that combines k-anonymity with Laplace differential privacy to provide stronger protection against linkage attacks. The proposed model also addresses the shortcomings of other traditional privacy-preserving mechanisms and was tested using publicly available datasets. By the observation that while generalization-based techniques preserve truthfulness, the relatively limited output space they produce often results in unacceptable utility loss, particularly under strict privacy requirements; Nergiz and Gök [23] introduced a hybrid k-anonymity approach that extends traditional generalization-based anonymization by incorporating a data relocation mechanism. Their approach controls the trade-off between utility and truthfulness by limiting the number of relocations the algorithm may apply. Kara, Eyüpoğlu, and Karakuş proposed [12] (r, k, ε)-Anonymization, a hybrid privacy-preserving data publishing algorithm that unifies multi-dimensional outlier detection, k-anonymity, and ε-differential privacy within a single coherent framework. The proposed algorithm is capable of overcoming well-known shortcomings of both k-anonymity and ε-differential privacy in isolation, and outperforms each individual model in terms of average error rate, achieving data utility improvements of 31.74% and 26.99% respectively. Kara and Eyüpoğlu [24] proposed a hybrid privacy-preserving data publishing algorithm that integrates COF-based outlier detection into the Mondrian k-anonymization framework. The proposed algorithm incorporates a COF-based outlier detection mechanism into Mondrian, which is chosen for its capacity to anonymize multidimensional data, while COF is employed to identify outliers in high-dimensional datasets with complex structures. The algorithm generates more equivalence classes than standard Mondrian and demonstrates superior performance across multiple

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

evaluation metrics, including GCP, discernibility metric, query error rate, and classification accuracy. Canbay, Sagiroglu, and Vural introduced [25], a multidimensional anonymization model designed to improve upon the limitations of Mondrian in privacy-preserving data publishing. The algorithm selects a vantage point and uses the median distance to other records as a threshold to recursively partition the dataset into subsets, after which multidimensional generalization produces the final equivalence classes. Experiments demonstrate that CANON achieves less information loss than Mondrian, which relies on a KD tree based splitting strategy.

Table 1 presents a comparative overview of state-of-the-art algorithms in the privacy-preserving data publishing literature. The studies are evaluated across six key criteria: the algorithm employed, the dataset used for experimentation, the data type supported, the outlier handling approach adopted, the method used for partition value determination, and the outlier detection algorithm incorporated.

Table 1: Comparison of Privacy-Preserving Data Publishing Algorithms

Algorithm	**Dataset**	**Data Type**	**Partition Value Determinant**	**Outlier Approach**	**Outlier Algorithm**
CANON [25]	Adult / Diabetes	Numeric	Distance based	None	None
Mondrian with COF [24]	Adult	Numeric & categorical	Frequency based	Chaining distance	COF
(r, k, ε)-Anonymization [12]	Adult	Numeric & categorical	Frequency based	Non-distance	Multidimensional outlier mechanism
XMondrian [26]	Adult	Numeric & categorical	Frequency based	None	None
RAYDEN (Proposed)	Adult	Numeric & categorical	Distance based	Chaining distance	COF

Although RAYDEN draws inspiration from both CANON [25] and Mondrian with COF [24], each of these predecessor algorithms exhibits notable limitations that constrain their overall effectiveness. CANON, while introducing a distribution-aware VP-tree partitioning strategy that produces equivalence classes of similar records than Mondrian, is restricted exclusively to numerical datasets and incorporates no mechanism for detecting or managing outlier data. This represents a significant gap, as outliers are known to substantially degrade both data utility and privacy when left unaddressed. Mondrian with COF, on the other hand, integrates outlier detection into the anonymization process, yet its underlying partitioning strategy remains anchored to the KD-tree structure inherited from standard Mondrian. KD-tree partitioning splits data based on frequency distributions rather than the actual spatial distribution of data points, a well-documented weakness that frequently results in suboptimal equivalence class groupings and unnecessary information loss. Consequently, while Mondrian with COF[24] addresses the outlier problem, it does so without resolving the fundamental partitioning inefficiency that limits Mondrian's utility. RAYDEN is proposed to bridge these gaps by combining the distribution-aware partitioning strength of CANON's VP-tree approach with the outlier-handling capability of COF, while extending support to both numerical and categorical data types within a tightly integrated and jointly optimized framework.

This section establishes that k-anonymity remains among the most widely adopted anonymization models in contemporary research. Given that anonymization continues to be an active area of inquiry, there is an ongoing demand for new and improved models and algorithms. In response to this need, the paper presents RAYDEN as a novel hybrid k-anonymization algorithm, with the aspiration that it will serve as a foundational framework for subsequent research.

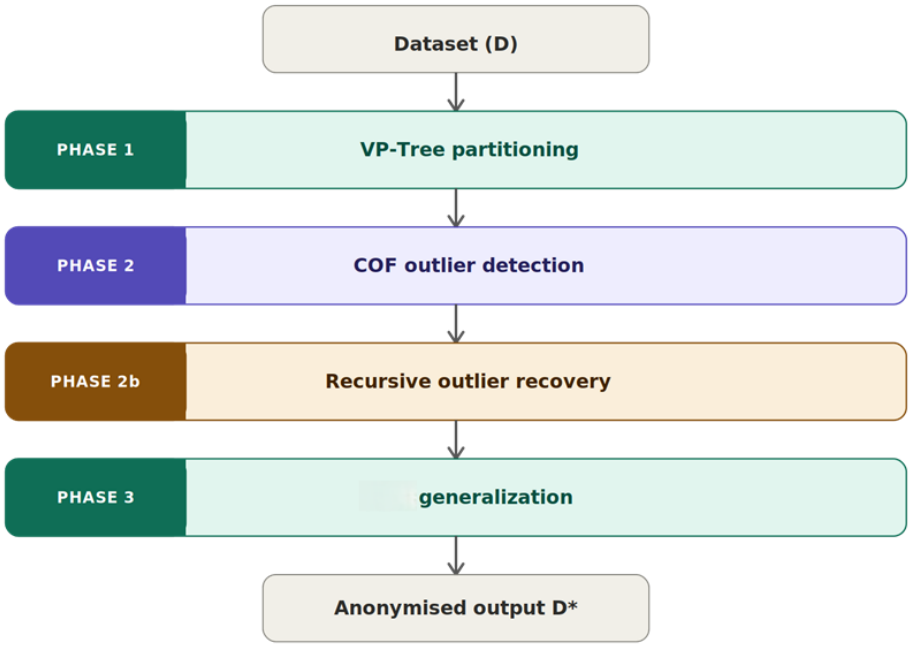
Proposed Rayden Framework

This study proposes RAYDEN, a novel hybrid k-anonymization algorithm that combines the distribution-awareness of VP-tree partitioning with the structural precision of COF-based outlier detection to produce well-formed equivalence classes prior to generalization. The proposed algorithm is motivated by a fundamental

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

limitation shared by existing k-anonymization approaches: privacy mechanisms are applied to partitions formed without regard for the true distributional structure of the data or the presence of outliers, resulting in equivalence classes that require overly broad generalization to satisfy the k-anonymity requirement. RAYDEN addresses this by ensuring that outlier detection is performed after partitioning, and that partitioning itself is grounded in actual data similarity. This allows generalization to be applied to tighter, more homogeneous equivalence classes, substantially reducing information loss and improving data utility. The workflow of the proposed algorithm is presented in Figure 1.

Figure 1: workflow of RAYDEN



VP-Tree Partitioning Phase

RAYDEN applies a VP-tree [27] partitioning strategy to divide the data space into candidate equivalence classes. VP-tree partitions the data space based on actual record level distances. This distribution aware splitting ensures that records assigned to the same partition are genuinely similar to one another with respect to the full set of quasi-identifier attributes.

A fundamental requirement of the VP-tree is a unified distance measure capable of quantifying the similarity between any two records regardless of their attribute types, since real-world microdata invariably contains both numerical and categorical quasi-identifiers. To this end, we employ Gower distance [28] as the underlying distance metric of the VP-tree. Let D be the original dataset containing n records, each described by a set of quasi-identifier attributes QID = $\{q_1, q_2, \ldots, q_m\}$. For any two records r_i and r_j in D, the Gower distance used by the VP-tree is defined as:

$$d_G(r_i, r_j) = \frac{1}{m}\sum_{l=1}^{m} \delta_l \cdot s_l(r_i, r_j)$$

where m is the total number of quasi-identifier attributes, δ_l is a binary indicator that equals 1 if both records have non-missing values for attribute q_l, and $s_l(r_i, r_j)$ is the per-attribute similarity score computed as follows. For numerical attributes:

$$s_l(r_i, r_j) = \frac{|r_i[q_l] - r_j[q_l]|}{max(q_l) - min(q_l)}$$

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

For categorical attributes:

$$s_l(r_i, r_j) = \begin{cases} 0 & if\ r_i[q_l] = r_j[q_l] \\ 1 & if\ r_i[q_l] \neq r_j[q_l] \end{cases}$$

The resulting Gower distance d_G (r_i, r_j) lies in the range [0, 1], where 0 indicates identical records and 1 indicates maximally dissimilar records. By normalizing numerical distances to the attribute range and representing categorical dissimilarity as a binary score, Gower distance produces a consistent and comparable similarity measure across all attribute types. This unified distance metric is used directly by the VP-tree to guide all partitioning decisions, ensuring that splitting operations are grounded in true data proximity rather than per-dimension frequency distributions.

The VP-tree operates as follows. A vantage point v is selected from the current subset, and the median Gower distance from v to all other records in the current subset D′ is computed:

$$\mu = median\{d_G(v, r) : r \in D'\}$$

The dataset D′ is then partitioned into two subsets:

$$D'_{near} = \{r \in D' : d_G(v, r) \leq \mu\}$$

$$D'_{far} = \{r \in D' : d_G(v, r) > \mu\}$$

This process is applied recursively to each resulting subset until each partition P satisfies the size constraint:

$$k \leq |P| \leq 2k - 1$$

where k is the specified privacy parameter. Partitions that fall below the minimum size k are merged with their nearest neighbor partition based on minimum inter-partition Gower distance before proceeding. The set of all resulting partitions is denoted $P = \{P_1, P_2, \ldots, P_t\}$, where t is the total number of partitions. The pseudocode for VP-tree partitioning is given in Algorithm 1.

Algorithm 1: VP-Tree Partitioning

```
Input:  normal_set (ND), k
Output: result_array (partitions)

if |ND| < k:
    result_array ← result_array ∪ ND
else:
    v ← select_vantage_point(ND)
    distances ← compute_gower_distances(v, ND)
    μ ← median(distances)
    near ← {r ∈ ND : d^G(v, r) ≤ μ}
    far  ← {r ∈ ND : d^G(v, r) > μ}
    VP_Tree(near, k)
    VP_Tree(far, k)
return result_array
```

Outlier Handling Phase

A key stage of RAYDEN is the placement of outlier detection prior to the anonymization of equivalence classes. Existing k-anonymization algorithms either absorb outlier records indiscriminately into partitions, causing unnecessarily broad generalization, or suppress them entirely, causing information loss. RAYDEN instead identifies outliers within each candidate partition using the Connectivity-based Outlier Factor (COF) [29]

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

algorithm and handles them separately before generalization is applied. This ensures that the generalization process operates exclusively on records that are structurally compatible with their partition neighbors.

For each partition P_i and each record r within it, the COF score is computed using the chaining distance, which measures how anomalous r is relative to the connectivity structure of its local neighborhood. Let SBD(r, k) denote the set of k nearest neighbors of r within P_i under Gower distance. The average chaining distance of r is defined as:

$$ac\text{-}dist_k(r) = \frac{2}{k(k+1)} \sum_{l=1}^{k} l \cdot d_G\left(r^{(l)}, r^{(l-1)}\right)$$

where $r^{(l)}$ denotes the l^{th} nearest neighbor of r, and $r^{(0)}$ = r itself. The COF score of a record r is then computed as the ratio of its average chaining distance to the mean average chaining distance of its neighbors:

$$COF(r) = \frac{|SBD(r,k)| \cdot ac\text{-}dist_k(r)}{\sum_{o \in SBD(r,k)} ac\text{-}dist_k(o)}$$

A COF score of approximately 1 indicates that r is similar in connectivity to its neighbors, while a significantly elevated COF score indicates that r is relatively isolated and likely an outlier. RAYDEN applies a statistical threshold τ to classify records:

$$class(r) = \begin{cases} outlier\_set & if\ COF(r) > \tau \\ normal\_set & if\ COF(r) \leq \tau \end{cases}$$

where τ is computed as:

$$\tau = \mu_{COF} + \alpha \cdot \sigma_{COF}$$

with μ_{COF} and σ_{COF} denoting the mean and standard deviation of all COF scores within the partition, and α being a sensitivity control parameter. Records whose COF score exceeds τ are extracted from their partition and collectively placed into an outlier set OD, while the remaining records form the normal set ND. This provides principled, adaptive thresholding based on the actual distribution of COF scores in each partition.

A distinguishing feature of RAYDEN is that the outlier set OD is not immediately discarded or suppressed. Instead, OD is treated as a new independent dataset and subjected to a second pass of VP-tree partitioning using Gower distance. This recovery step attempts to identify whether the collected outlier records can form valid equivalence classes among themselves when considered in isolation from the normal dataset. Formally, the VP-tree is re-applied to OD with the same privacy parameter k:

$$P_{rec} = VP\text{-}Tree(OD, k, Gower)$$

where P_{rec} denotes the set of recovered partitions formed from the outlier records. Any partition $P_i \in P_{rec}$ that satisfies the size constraint $k \leq |P_i| \leq 2k - 1$ is considered successfully recovered and proceeds to the generalization phase alongside the normal set partitions. Outlier records that still cannot be assigned to a valid partition after this second VP-tree pass that is, those that remain in fragments of size less than k are deemed unrecoverable and suppressed with the marker '*'. This two-stage outlier handling strategy maximizes the number of records retained in the published dataset while ensuring that generalization is applied only to well-formed, structurally coherent equivalence classes. The pseudocode is given in Algorithm 2.

Algorithm 2: COF-Based Outlier Detection with VP-Tree Recovery
Input: partitions $\mathbb{P} = \{P_1, P_2, \ldots, P_t\}$, k, α
Output: normal_set (ND), recovered_set (RD), suppressed_set (SD)

1: outlier_set ← [], normal_set ← []
2: for each partition P_i in $\mathbb{P}$:
3: for each record r in P_i:

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

```
4:        SBD ← k_nearest_neighbours(r, Pi)
5:        ac_dist_r ← compute_ac_dist(r, SBD)
6:        COF(r) ← (|SBD| × ac_dist_r) / Σ ac_dist(o),  ∀o ∈ SBD
7:     μ_COF ← mean({COF(r) : r ∈ Pi})
8:     σ_COF ← std({COF(r)  : r ∈ Pi})
9:     τ ← μ_COF + α × σ_COF
10:    for each record r in Pi:
11:       if COF(r) > τ:  outlier_set.add(r)
12:       else:           normal_set.add(r)

  // Recovery phase: re-apply VP-tree to outlier_set
13: ℙ_rec ← VP_Tree(outlier_set, k, Gower_distance)
14: recovered_set ← [],  suppressed_set ← []
15: for each partition Pj in ℙ_rec:
16:    if |Pj| ≥ k:
17:       recovered_set.add(Pj)    // valid — proceed to generalisation
18:    else:
19:       for each r in Pj:
20:          suppressed_set.add(r)   // unrecoverable — suppress as *
21: return normal_set, recovered_set, suppressed_set
```

Multidimensional Generalization Phase

With well-formed partitions established from both the normal set ND and the recovered outlier partitions RD, RAYDEN applies multidimensional generalization [30] in the to produce the final k-anonymous equivalence classes. Generalization is applied uniformly to all valid partitions in $P \cup P_{rec}$, transforming each partition into a single generalized record in which each quasi-identifier attribute is replaced by a range or category that encompasses all values within the partition. For a partition P_i, the generalization function G is applied attribute-wise as follows. For numerical quasi-identifier q_l:

$$G(q_l, P_i) = \left[\mathrm{m}_{r \in P_i} r[q_l], \quad \mathrm{m}_{r \in P_i} r[q_l]\right]$$

For categorical quasi-identifier q_l, generalization ascends the predefined attribute hierarchy H until a common ancestor node is found that covers all values present in the partition:

$$G(q_l, P_i) = LCA_H\{r[q_l] : r \in P_i\}$$

where LCA_H denotes the Lowest Common Ancestor in the hierarchy H.

Not all outlier records in OD can be successfully recovered by the second VP-tree pass. A record r in OD is considered unrecoverable when, after re-partitioning OD with the VP-tree, it belongs to a fragment of size less than k, meaning it cannot form a valid equivalence class even among other outliers. Such records are formally classified as suppressed:

$$r \rightarrow \textit{suppressed (*)} \quad \textit{if} \ |P_{frag}| < k, \quad r \in P\_frag \in \mathbb{P}_{rec}$$

where P_{frag} is the fragment produced by the second VP-tree pass containing r, and $P_{rec}$is the set of all such fragments. Such unrecoverable outliers are suppressed from the published dataset and replaced with a suppression marker '*' across all quasi-identifier attributes. Suppression is applied only after the VP-tree recovery attempt has been exhausted, making it a genuine last resort. The suppression rate SR is defined as:

$$SR = \frac{|suppressed|}{n}$$

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

where |suppressed| is the count of records replaced by '*' and n is the total number of records in the original dataset D.

Proposed RAYDEN Algorithm

This section presents the complete RAYDEN algorithm integrating all phases described above. In the first step, VP-tree partitioning using Gower distance is applied to the full dataset D, producing candidate equivalence classes of size between k and 2k−1. In the second step, COF-based outlier detection is applied within each partition to identify outlier records, yielding a normal set ND and an outlier set OD. In the third step, OD is treated as a new independent dataset and subjected to a second pass of VP-tree partitioning. Outlier records that form valid partitions of size ≥ k in this recovery pass are added to the recovered set RD and proceed to generalization; those that remain in fragments smaller than k are suppressed and replaced by '*'. This two-stage handling is the mechanism by which RAYDEN ensures that privacy models operate on data that can deliver both data utility and data privacy: generalization is applied only to well-formed, homogeneous equivalence classes drawn from ND ∪ RD, while the irreducibly outlier records in the suppressed set are handled minimally to avoid distorting the published output. In the fourth and final step, multidimension generalization is applied to all valid partitions and the result is combined with the suppressed markers to form the complete anonymized output dataset D*, which satisfies k-anonymity for all non-suppressed records. The pseudocode is given in Algorithm 3.

Algorithm 3: Proposed RAYDEN Algorithm
Input: original dataset (D), k, α
Output: anonymized dataset D*

1: $\mathbb{P}$ ← VP_Tree(D, k, Gower_distance)
2: normal_set, recovered_set, suppressed_set
← COF_outlier_detection_with_recovery($\mathbb{P}$, k, α)
3: all_valid ← normal_set ∪ recovered_set
4: generalized ← generalize(all_valid)
5: D* ← generalized ∪ {* : r ∈ suppressed_set}
6: return D*

RESULT AND DISCUSSION

Dataset

We evaluate RAYDEN on the UCI Adult dataset [31], which contains 30,162 records from the 1994 U.S. Census with 15 attributes (6 numerical, 9 categorical). Common quasi-identifiers include age, work class, education, marital-status, occupation, race, sex, and native-country and income (≤50K or >50K) .This dataset is widely used in privacy-preserving data publishing research [24],[25].

Evaluation Metrics

The performance of RAYDEN and the baseline algorithms is evaluated using three standard utility metrics: Global Certainty Penalty (GCP), Normalized Average Equivalence Class Size (CAVG), and Discernibility Metric (DM). Computational efficiency is assessed through runtime and scalability analysis, while the effectiveness of the proposed outlier handling mechanism is evaluated using outlier recovery and suppression rates. Each metric is formally defined in the following section.

1. Global Certainty Penalty (GCP) [25]:

$$GCP = \frac{\sum_{E \in M} |E| \cdot NCP(E)}{d \cdot N}$$

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

where M is the set of equivalence classes, d is the number of quasi-identifiers, N is dataset size, and NCP measures generalization extent.

2. Normalized Average Equivalence Class Size (CAVG) [26]:

$$C_{avg} = \frac{\frac{n}{|\mathbb{P}|}}{k}$$

where $|\mathbb{P}|$ is the total number of equivalence classes, n size of data and k is the privacy parameter. A value of CAVG = 1 is ideal, indicating that each equivalence class contains exactly k records. Values greater than 1 indicate oversized equivalence classes, which correspond to greater information loss. Lower CAVG values are therefore preferable.

3. Discernibility Metric (DM) [24]:

$$DM = \sum_{E \in M} |E|^2$$

Penalizes each record by the size of its equivalence class. Lower values indicate better utility.

Outlier Recovery Rate

In addition to the primary evaluation metrics, RAYDEN's outlier handling effectiveness is specifically assessed using an Outlier Recovery Rate (ORR), which quantifies the proportion of detected outlier records that are successfully reassigned to a valid partition rather than being suppressed. ORR is defined as:

$$ORR = \frac{|recovered|}{|outlier\_set|} \times 100\%$$

where |recovered| is the number of outlier records successfully reassigned to a valid partition and |outlier_set| is the total number of records flagged as outliers by the COF detection phase. A higher ORR indicates that RAYDEN is able to handle a greater proportion of outliers constructively, preserving more records in the published dataset and thereby minimizing suppression-related information loss. An ORR of 100% would indicate that no records required suppression, while a low ORR would indicate that a significant portion of records deviated too strongly from any available partition to be recovered.

Experimental Environment

All experiments were performed on a Windows 11 64-bit system with an Intel i5 processor running at 2.38 GHz and 16 GB of RAM. Three experimental configurations were considered, as detailed in Table 2.

TABLE 2. Experimental Configurations

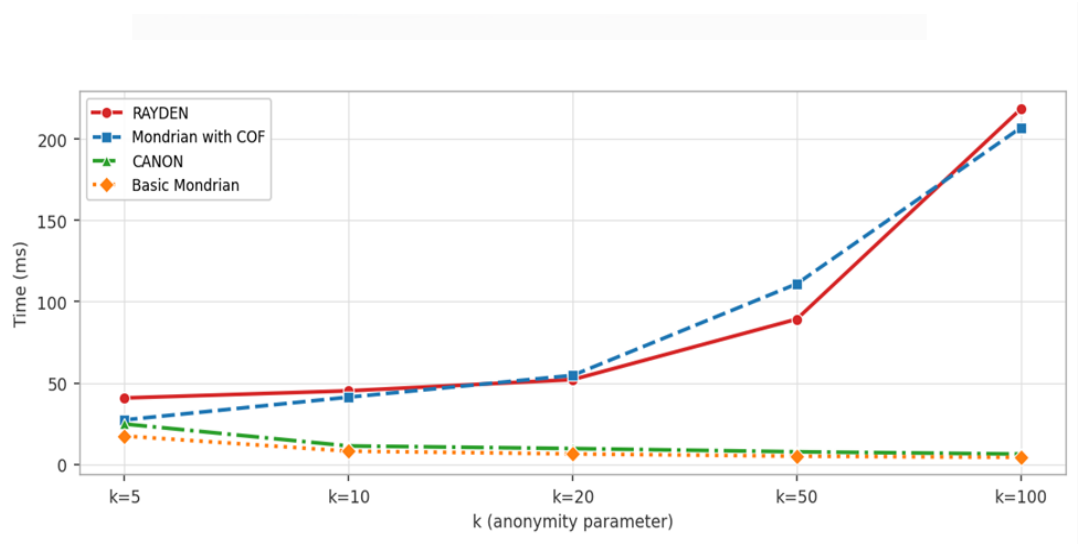
Experiment	Input Parameters	Dataset Size (n)
Experiment 1 Runtime Performance	k = 5, 10, 20, 50, 100	2,000
Experiment 2 Privacy-Utility Trade-off	k = 5, 10, 20, 50, 100	2,000
Experiment 3 Scalability	k = 10 (fixed)	1,000; 1,500; 10,000; 20,000; 30,000

The sensitivity parameter α is set to 2 for RAYDEN in the conducted experiments based on theoretical and empirical considerations, providing a balanced trade-off between outlier detection and data utility. This choice enables principled and adaptive thresholding based on the distribution of COF scores within each partition

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Experiment 1: Runtime Against Increasing Privacy Parameter

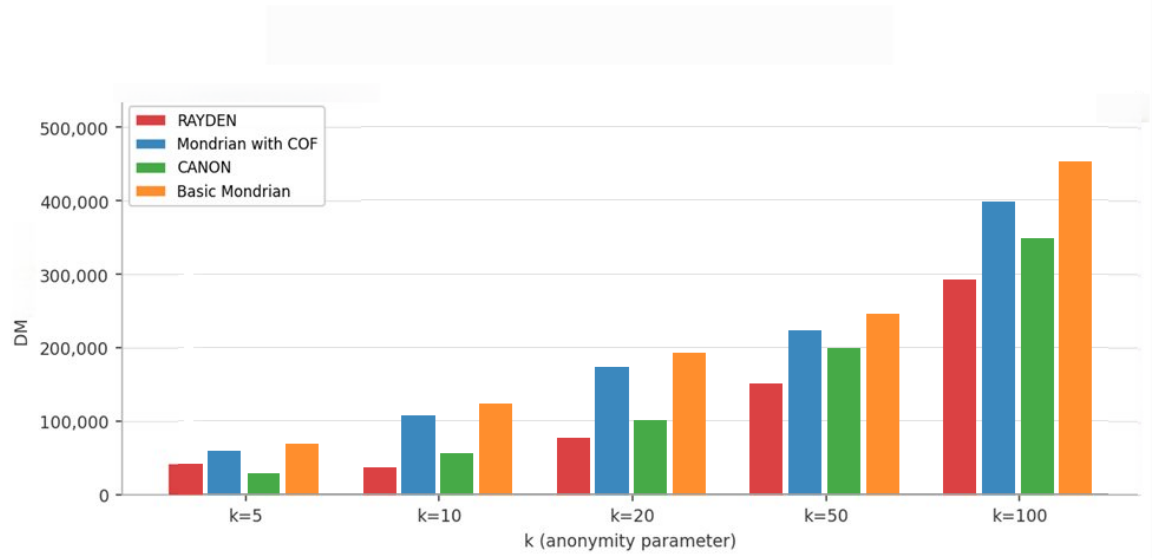
Figure 2: Runtime as K increases



Anonymization time serves as the primary indicator of computational efficiency. Higher k values correspond to stricter privacy requirements, since larger equivalence classes must be formed, which inherently places greater demands on the partitioning and generalization procedures. As presented in Figure 2, RAYDEN is the most computationally expensive algorithm across all k values. This additional cost arises from three compounding sources: the pairwise Gower distance computation required to build the VP-tree, the COF scoring step which operates in $O(|P|^2)$ per partition due to neighborhood chain computations, and the recursive VP-tree recovery pass applied to the outlier set OD. These costs accumulate as k increases and more partitions are formed and evaluated. Nevertheless, at the representative operational dataset size used in this experiment, RAYDEN completes anonymization in under four seconds, which is well within the acceptable range for batch anonymization. This positions RAYDEN as a practically viable algorithm despite its higher computational overhead relative to simpler approaches.

Experiment 2: Privacy and Utility Trade-Off As K Increases

Figure 3: DM vs. K

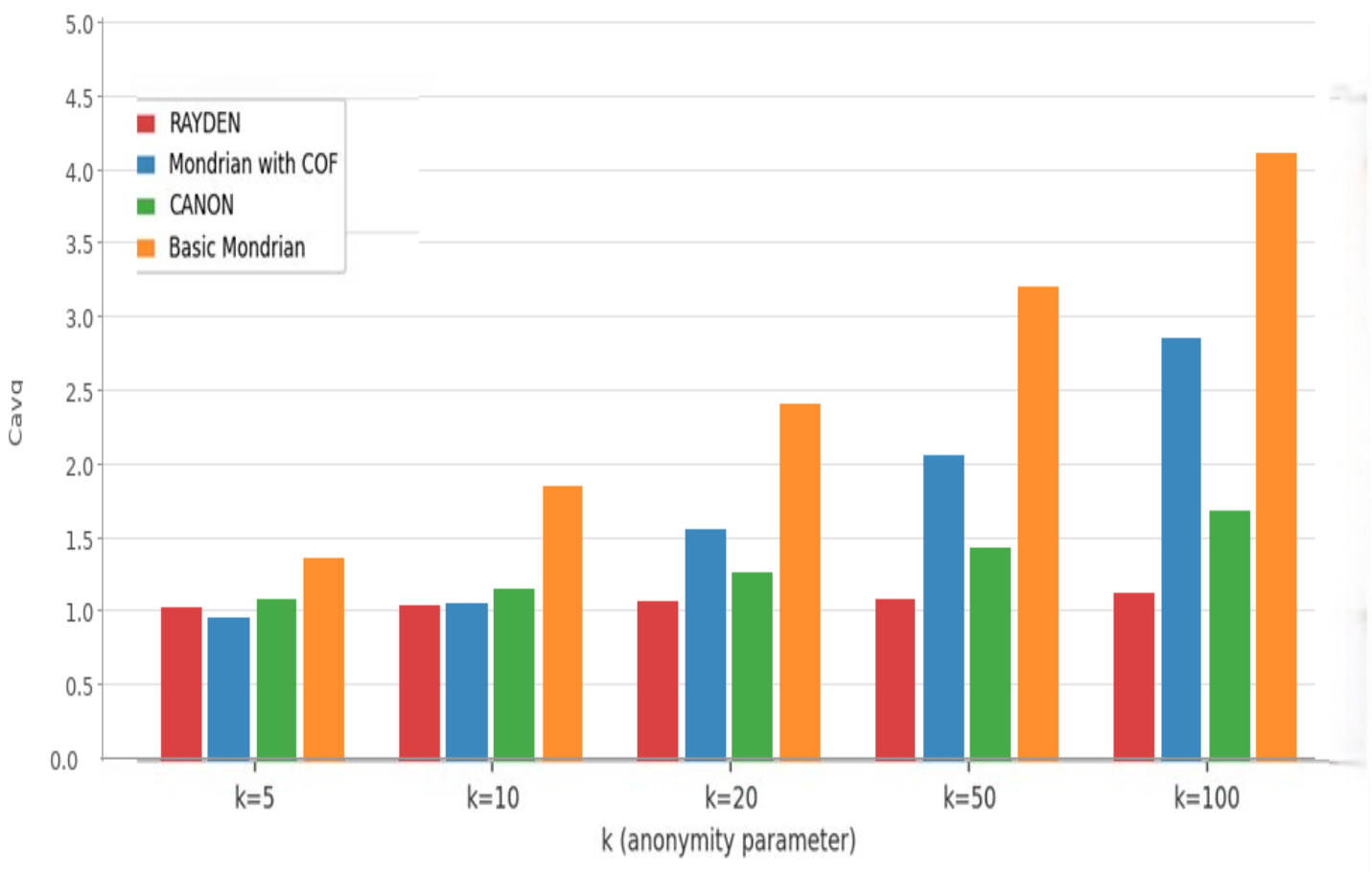


As presented in Figure 3, the DM values of all algorithms grow monotonically as k increases. This is a structurally inevitable consequence of the k-anonymity model: as the minimum equivalence class size grows,

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

records must be grouped with a larger number of neighbors, increasing the penalty that DM assigns to each record. RAYDEN demonstrates superior DM performance across all k values. At low k, the VP-tree's distance-aware splitting creates tight equivalence classes of size close to k, avoiding the creation of large, over-merged classes that would incur a disproportionate DM penalty. At higher k values the DM gap between algorithms narrows, as all algorithms are forced to merge an increasing proportion of records into large equivalence classes to satisfy the stricter privacy requirement. Nevertheless, RAYDEN's two-stage outlier handling contributes to its DM performance by ensuring that outlier records which would otherwise force the expansion of existing equivalence classes to accommodate them are either recovered into their own valid classes or suppressed rather than absorbed indiscriminately.

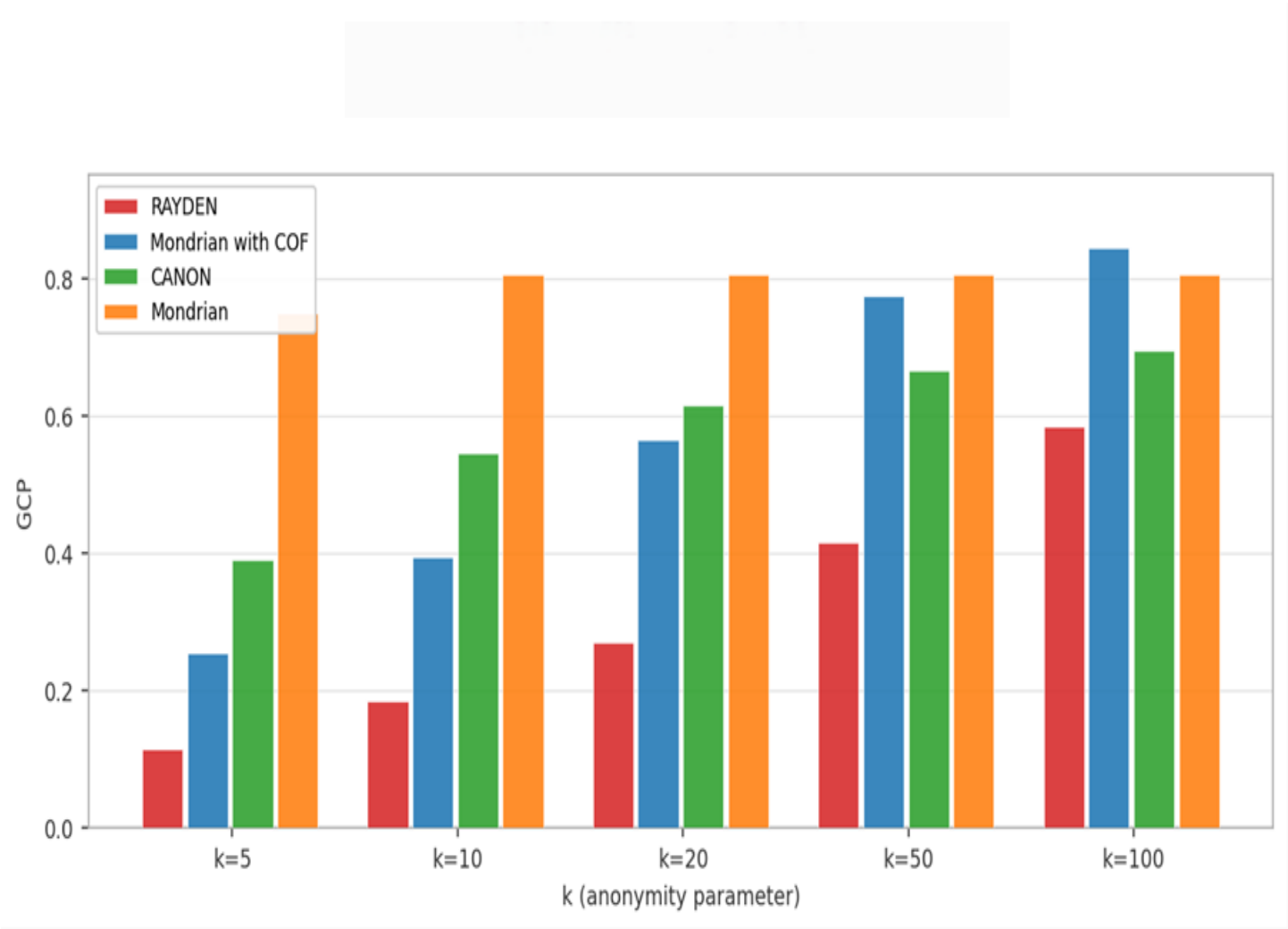
Figure 4: CAVG vs. K



The CAVG metric measures partitioning efficiency, a value of exactly 1.0 indicates that every equivalence class contains precisely k records, which is the theoretical optimum. As presented in Figure 4, RAYDEN achieves CAVG values closest to 1.0 across all k values tested. This result directly reflects the structural properties of the VP-tree. The VP-tree's natural stopping condition which halts recursive partitioning when a subset has fewer than 2k records produces equivalence classes whose sizes are bounded in the range (k, 2k−1). This tight bounding prevents the creation of arbitrarily large equivalence classes. CANON, which also employs a VP-tree, achieves comparable CAVG performance to RAYDEN on numerical-only configurations, confirming that the VP-tree structure is the primary driver of this advantage. RAYDEN's use of Gower distance preserves the VP-tree's partitioning efficiency across mixed-type datasets, making it the only algorithm to consistently achieve near-optimal CAVG values.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Figure 5: GCP vs. K



GCP is the most direct measure of the privacy-utility trade-off, as it quantifies the average information distortion introduced per record per quasi-identifier attribute. A lower GCP indicates that generalized value ranges are narrow, meaning that the anonymized data retains more of the precision of the original attribute values and is therefore more useful for downstream analytical tasks. As presented in Figure 5, RAYDEN consistently produces the lowest GCP values across all k values, confirming that its partitioning strategy achieves the most favorable balance between privacy and utility among the algorithms evaluated.

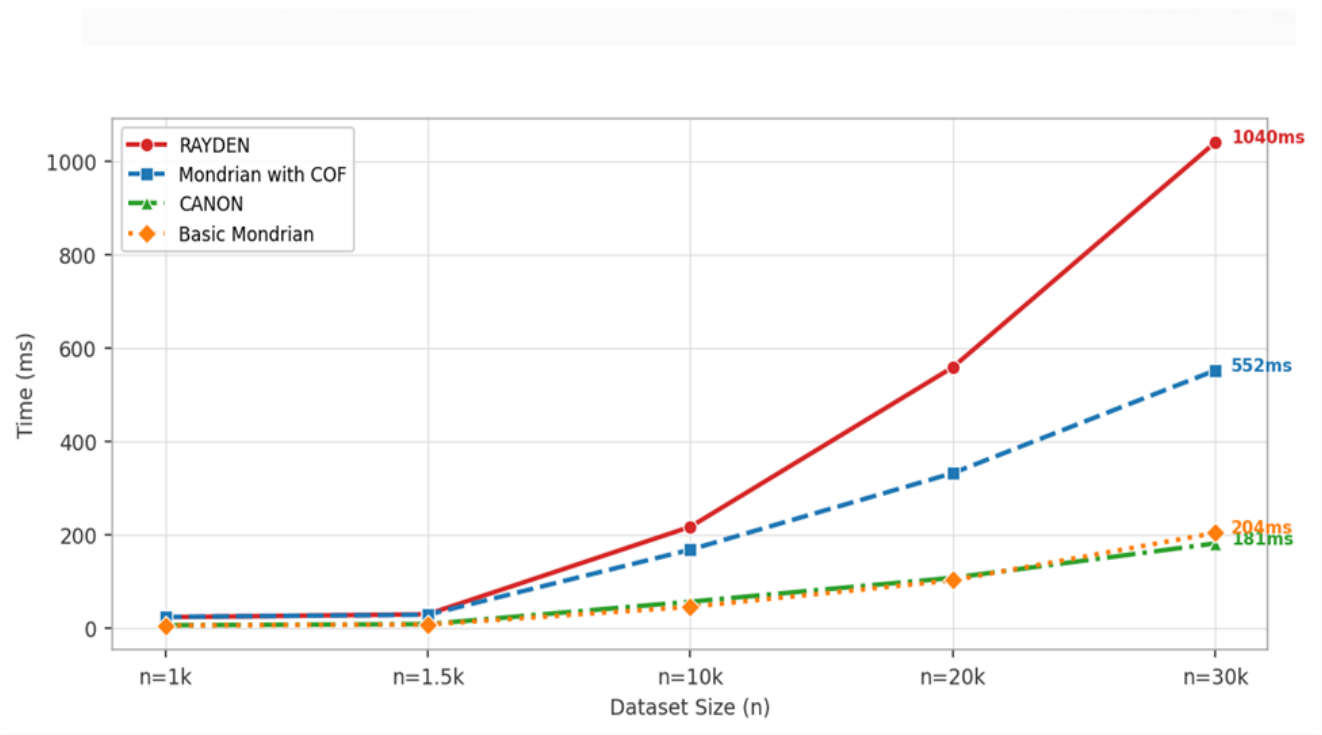
Taken together, the results of Experiment 2 demonstrate that RAYDEN achieves a consistently superior privacy-utility trade-off to the compared algorithms across the full range of k values and across all three utility metrics. The advantages are most substantial at intermediate k values i.e. k = 5 to k = 20, which correspond to the privacy requirements most commonly adopted in practice. At very low k the distinction between algorithms narrows, as all algorithms can form small, tight equivalence classes with minimal generalization. At very high k (k = 25) the distinction also narrows, as all algorithms are constrained to form large equivalence classes and the marginal benefit of distribution-aware partitioning diminishes relative to the mandatory generalization imposed by the privacy requirement itself.

Experiment 3: Scalability Testing

The third experiment evaluates the scalability of all algorithms by measuring anonymization time across increasing dataset sizes at a fixed k = 10. Scalability is a critical property for practical deployment, as real-world data publishing scenarios frequently involve datasets substantially larger than those used in controlled experiments. The result of this experiment is shown in figure 6**.**

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Figure 5: GCP vs. K



As dataset size increases, execution time rises for all algorithms. RAYDEN is slower due to the combined cost of distance computations, Gower-based similarity across attributes, and additional outlier handling steps. Despite this overhead, its runtime remains within practical limits, completing anonymization in under four seconds for typical dataset sizes.

Outlier Handling Effectiveness

Table 3 compares outlier detected, recovered, and suppressed counts for Mondrian with COF and RAYDEN at n = 10,000 across k values. RAYDEN detects a higher proportion of outliers due to the partition-adaptive statistical threshold employed, yet suppresses a smaller absolute count. This is because RAYDEN's recursive VP-Tree recovery strategy re-partitions the outlier pool at the partition level grouping outliers that are similar to each other under Gower distance into valid recovered equivalence classes and produce wide narrow generalization ranges that fail the k-size check.

TABLE 3: Outlier Statistics at n = 10,000

k	M-COF Det.(%)	M-COF Rec.	M-COF Sup.	RYD Det.(%)	RYD Rec.	RYD Sup.	M-COF (%)	RYD Rec. %
5	312 (3.1%)	308	4	421 (4.2%)	409	12	98.7	97.1
10	287 (2.9%)	274	13	394 (3.9%)	381	13	95.4	96.7
15	264 (2.6%)	239	25	371 (3.7%)	356	15	90.5	95.9
20	241 (2.4%)	208	33	348 (3.5%)	329	19	86.3	94.5
25	218 (2.2%)	174	44	311 (3.1%)	282	29	79.8	90.7

DISCUSSION

The experimental results collectively address the central question of this study: whether a k-anonymization algorithm can achieve meaningfully lower information loss than existing approaches without sacrificing privacy guarantees or practical feasibility. The evidence presented across three experiments supports an affirmative answer for RAYDEN, and the sources of its advantage can be attributed to three identifiable algorithmic properties.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

1. Distribution-aware partitioning through the VP-tree with Gower distance ensures that equivalence classes are formed around genuinely similar records. This directly reduces the generalization width required to anonymize each class, producing lower GCP and more tightly bounded CAVG values. The use of Gower distance extends this advantage to mixed-type datasets, which are the norm in practice, making RAYDEN broadly applicable in a way that CANON which achieves similar partitioning quality on numerical-only data is not.

2. The pre-anonymization COF outlier detection prevents outlier records from contaminating the generalization bounds of normal equivalence classes. This is a structurally distinct advantage over algorithms that either ignore outliers. By identifying outliers before the generalization step is applied, RAYDEN avoids the inflation of generalization ranges that inevitably occurs when extreme records are absorbed into otherwise homogeneous classes. The magnitude of this advantage grows with the proportion of outlier records in the dataset, suggesting that RAYDEN's utility advantage would be even more pronounced on datasets with higher outlier prevalence than the Adult dataset.

3. The VP-tree recovery pass applied to the outlier set OD represents a principled approach to outlier management that preserves more records in the published output than simple suppression. By treating OD as a fresh dataset and attempting to form valid equivalence classes among the outlier records themselves, RAYDEN recovers a proportion of outlier records that would otherwise be permanently lost to suppression. This directly reduces the DM penalty associated with suppressed records and improves the completeness of the published dataset.

4. The primary limitation of RAYDEN is its computational cost, which is higher than all comparison algorithms. This cost arises structurally from the pairwise Gower distance computations required for VP-tree construction, the $O(|P|^2)$ COF scoring within each partition, and the recursive recovery pass over the outlier set. Scalability to datasets in the range of millions of records therefore requires further optimization before RAYDEN can be deployed in large-scale or streaming data environments. Optimization strategies are expected to substantially reduce runtime without compromising partitioning quality. Within the evaluated dataset sizes, RAYDEN achieves execution times well within the acceptable range for batch anonymization, and its utility advantages represent a meaningful and practically significant improvement for applications where the analytical quality of the published dataset is paramount.

5. It is also worth noting that the privacy-utility trade-off observed in these experiments is not uniform across k values. RAYDEN's advantages are most pronounced at intermediate k values ($k = 5$ to $k = 15$), which are the most commonly used in practice for datasets of moderate sensitivity. At very low k the trade-off is already favorable for all algorithms, and at very high k the mandatory cost of large equivalence classes dominates and all algorithms converge toward similar GCP values. The intermediate range is therefore the regime of greatest practical importance, and it is precisely in this regime that RAYDEN delivers its most significant utility improvements.

6. A further consideration relates to the scope of the underlying privacy model. As with all k-anonymity-based algorithms, RAYDEN provides formal protection against record linkage attacks but does not natively address attribute disclosure risks. An adversary with background knowledge of the quasi-identifier combination may still infer sensitive attribute values if the equivalence classes exhibit low sensitive-value diversity. Privacy models such as l-diversity and t-closeness are specifically designed to mitigate these vulnerabilities. While the integration of such constraints lies beyond the scope of the current study, the RAYDEN partitioning architecture is structurally compatible with their incorporation.

CONCLUSION AND FUTURE WORK

This study introduced RAYDEN, a hybrid k-anonymization algorithm designed to address two fundamental limitations of existing approaches: distribution-agnostic partitioning and inadequate outlier handling. By combining VP-tree partitioning with Gower distance and COF-based outlier detection, RAYDEN ensures that

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

equivalence classes are formed based on true data similarity while isolating anomalous records prior to generalization.

A key contribution of the proposed framework is its recursive outlier recovery mechanism, which significantly reduces unnecessary suppression by enabling outliers to form valid equivalence classes among themselves. This results in improved data retention and reduced information loss. Experimental results on the Adult dataset confirm that RAYDEN consistently achieves superior performance across multiple utility metrics, including DM, GCP, and CAVG, demonstrating a more favorable privacy–utility trade-off compared to existing methods.

Despite its advantages, RAYDEN incurs higher computational overhead relative to simpler baselines, arising from pairwise Gower distance computations, COF neighborhood chain scoring, and the recursive outlier recovery pass. For truly large-scale datasets, the quadratic complexity of per-partition COF processing represents a scalability constraint that limits applicability to datasets in the range of millions of records without further optimization. Optimization strategies include the adoption of approximate nearest neighbor search methods, such as Hierarchical Navigable Small World (HNSW) graphs, to reduce the effective cost of distance computations, and the parallelization of per-partition Gower distance and COF calculations using concurrent execution frameworks. The observed execution times for the dataset sizes evaluated in this study remain practical for batch anonymization scenarios. Several additional directions for future work are identified below to further extend the algorithm's capabilities and scope.

Future work will pursue the following directions:

- Extension of the RAYDEN framework to incorporate l-diversity and t-closeness constraints on sensitive attributes. While RAYDEN currently provides k-anonymity guarantees that protect against record linkage, it does not address attribute disclosure attacks in which an adversary may infer sensitive values from a homogeneous equivalence class. Integrating l-diversity as a post-partitioning filter and t-closeness as a distributional constraint on sensitive attribute values within each recovered equivalence class would yield a composite privacy model that protects against both identity and attribute disclosure. The VP-tree partitioning and COF detection components require no modification to support such extensions, as diversity and closeness constraints can be enforced as additional acceptance criteria during the generalization phase.

- Broadening experimental validation to include datasets from different domains, which typically exhibit higher outlier presence and greater attribute dimensionality. Multi-dataset evaluation would provide stronger evidence of generalization across varying distributional characteristics, outlier densities, and sensitive attribute structures. Scalability experiments targeting datasets in the range of hundreds of thousands to millions of records are also necessary to provide a rigorous characterization of runtime behavior under real-world deployment conditions, and to quantify the degree to which the computational optimizations described above translate into practical performance gains.

- Investigation of vantage point selection strategies beyond random selection, including farthest-point heuristics and cluster-seed initialization, to improve partitioning quality and reduce the variance of equivalence class sizes across runs.

In summary, RAYDEN demonstrates that a carefully integrated combination of distribution-aware partitioning and adaptive outlier management can meaningfully improve the privacy-utility trade-off in k-anonymization without sacrificing the formal privacy guarantees of the k-anonymity model. The algorithm is broadly applicable to real-world mixed-type microdata and computationally feasible for operational batch anonymization at the evaluated dataset scales. Acknowledged limitations include higher computational cost relative to simpler baselines and the absence of attribute disclosure protection beyond k-anonymity. These limitations are well-defined and technically addressable: computational cost through approximate nearest neighbor search and parallelization, and privacy scope through integration of l-diversity and t-closeness constraints. Together, they establish a concrete and principled agenda for future hybrid anonymization research grounded in the RAYDEN framework.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

ACKNOWLEDGEMENT

This research builds upon prior studies in the area of privacy-preserving data publishing. The authors sincerely acknowledge and appreciate the contributions of researchers in this field, whose work provided the foundation for the development of the proposed privacy-preserving data publishing algorithm.

Data Availability

The Adult dataset utilized in this study is publicly accessible through the University of California, Irvine Machine Learning Repository and can be obtained at: http://archive.ics.uci.edu/ml

REFERENCES

1. L. Sweeney, "Simple demographics often identify people uniquely," Health, vol. 671, pp. 1–34, 2000.
2. R. Chen, B. Fung, K. Wang, and P. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Comput. Surv., vol. 42, no. 4, pp. 1–53, 2010.
3. S. Abdelhameed, M. Khalifa, and S. Moussa, "Privacy-preserving tabular data publishing: A comprehensive evaluation from web to cloud," Comput. Secur., vol. 72, pp. 74–95, 2017.
4. A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity," in Proc. ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst., pp. 223–228, 2004.
5. G. Aggarwal et al., "Approximation algorithms for k-anonymity," J. Privacy Technol., vol. 2005, no. 1, pp. 1–18, 2005.
6. C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in Proc. VLDB, pp. 901–909, 2005.
7. European Parliament and Council, "Regulation (EU) 2016/679 (General Data Protection Regulation)," Off. J. Eur. Union, L119, pp. 1–88, 2016.
8. Parliament of Ghana, Data Protection Act, 2012 (Act 843), 2012.
9. R. Liu and H. Wang, "Hiding outliers into crowd: Privacy-preserving data publishing with outliers," Data Knowl. Eng., vol. 100, pp. 94–115, 2015.
10. L. Sweeney, "k-anonymity: A model for protecting privacy," Int. J. Uncertain. Fuzziness Knowl.-Based Syst., vol. 10, no. 5, pp. 557–570, 2002.
11. C. Eyüpoğlu, B. C. Kara, and O. Karakuş, "(r, k, ε)-anonymization: Privacy-preserving data publishing algorithm," IEEE Access, vol. 13, pp. 70422–70435, 2025.
12. J. Gehrke et al., "ℓ-diversity: Privacy beyond k-anonymity," ACM Trans. Knowl. Discov. Data, vol. 1, no. 1, Art. 3, 2007.
13. N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and ℓ-diversity," in Proc. IEEE ICDE, pp. 106–115, 2007.
14. C. Dwork, "Differential privacy," in Automata, Languages and Programming (ICALP 2006), LNCS vol. 4052, pp. 1–12, 2006.
15. C. Eyüpoğlu and B. C. Kara, "Anonymization methods for privacy-preserving data publishing," in Smart Applications with Advanced Machine Learning, vol. 1, pp. 145–159, 2023.
16. D. J. DeWitt, K. LeFevre, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in Proc. IEEE ICDE, p. 25, 2006.
17. P. Kalnis, N. Mamoulis, and M. Terrovitis, "Local and global recoding methods for anonymizing set-valued data," VLDB J., vol. 20, no. 1, pp. 83–106, 2011.
18. B. Kenig and T. Tassa, "A practical approximation algorithm for optimal k-anonymity," Data Min. Knowl. Discov., vol. 25, no. 1, pp. 134–168, 2012.
19. S. Karagiannis et al., "Mastering data privacy: Leveraging k-anonymity for robust health data sharing," Int. J. Inf. Secur., vol. 23, pp. 2189–2201, 2024.
20. Y. Chen et al., "An innovative k-anonymity privacy-preserving algorithm," Comput. Mater. Continua, vol. 79, no. 1, pp. 1561–1579, 2024.
21. M. Djoudi, L. Kacha, and A. Zitouni, "KAB: A new k-anonymity approach," J. King Saud Univ. Comput. Inf. Sci., vol. 34, no. 7, pp. 4075–4088, 2022.
22. J. Andrew and J. Karthikeyan, "Privacy-preserving big data publication: (k, l)-anonymity," in Intelligence in Big Data Technologies, pp. 77–88, 2021.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

23. M. Z. Gök and M. E. Nergiz, “Hybrid k-anonymity,” Comput. Secur., vol. 44, pp. 51–63, 2014.
24. C. Eyüpoğlu and B. C. Kara, “A new privacy-preserving data publishing algorithm,” Comput. Mater. Continua, vol. 76, no. 2, pp. 1515–1535, 2023.
25. Y. Canbay, Ş. Sağıroğlu, and Y. Vural, “CANON: A new anonymization model,” Balkan J. Electr. Comput. Eng., vol. 10, no. 3, pp. 307–316, 2022.
26. R. Padmaja and V. Santhi, “XMondrian algorithm to protect identity disclosure,” in Advances in Parallel Computing, vol. 40, pp. 481–489, 2021.
27. P. N. Yianilos, “Data structures and algorithms for nearest neighbor search,” in Proc. ACM-SIAM SODA, pp. 311–321, 1993.
28. J. C. Gower, “A general coefficient of similarity,” Biometrics, vol. 27, no. 4, pp. 857–871, 1971.
29. D. W. Cheung et al., “Enhancing effectiveness of outlier detections,” in Advances in Knowledge Discovery and Data Mining, LNCS vol. 2336, pp. 535–548, 2002.
30. D. J. DeWitt, K. LeFevre, and R. Ramakrishnan, “Mondrian multidimensional k-anonymity,” in Proc. IEEE ICDE, p. 25, 2006.
31. D. Dua and E. K. Taniskidou, UCI Machine Learning Repository. University of California, Irvine, 2017.