Autoencoder-Enhanced Hierarchical Mondrian Anonymization via Latent Representations

Junpeng Hu , Tao Hu , Jinan Shen , Minghui Zheng *

Posted Date: 3 February 2026

doi: 10.20944/preprints202602.0124.v1

Keywords: data anonymization; k-anonymity; autoencoder; latent-space partitioning; microaggregation; linkage-attack risk

 doi:10.20944/preprints202602.0124.v1

Article

Autoencoder-Enhanced Hierarchical Mondrian Anonymization via Latent Representations

Junpeng Hu [1,2], Tao Hu [2], Jinan Shen [2] and Minghui Zheng [1,2,*]

[1] Sichuan University, Chengdu, China
[2] Hubei Minzu University, Enshi, China
* Correspondence: mhzheng3@163.com

Abstract

Releasing structured microdata requires balancing utility and privacy under group-based disclosure risks. We propose AE-LRHMA, a hybrid anonymization framework that performs Mondrian-style hierarchical partitioning in an autoencoder-learned latent space and integrates local (k,e) -microaggregation. To explicitly control sensitive-value concentration and diversity within each equivalence class, we introduce a tunable constraint set consisting of *k*, a maximum sensitive proportion threshold, and an optional sensitive-entropy threshold (used as a hard gate when enabled and otherwise as a soft term in split scoring). The anonymized output is generated via standard interval/set generalization in the original space. Experiments on Adult and Bank Marketing demonstrate that AE-LRHMA yields lower information loss and more stable group structures than representative baselines under comparable settings. We further report linkage-attack-oriented risk metrics to empirically characterize relative disclosure trends, without claiming formal guarantees such as differential privacy.

Keywords: data anonymization; k-anonymity; autoencoder; latent-space partitioning; microaggregation; linkage-attack risk

1. Introduction

With the increasing sharing and reuse of structured data in finance, healthcare, education, and other domains, achieving both regulatory compliance and analytical utility in data publication has become a central challenge in data governance and privacy protection. Structured microdata typically consist of quasi-identifiers (QIs) and sensitive attributes (SAs). When an adversary possesses auxiliary information and can match records on QIs, linkage attacks may re-identify individuals or infer their sensitive values, resulting in privacy leakage. Consequently, regulations such as the GDPR and China's Data Security Law and Personal Information Protection Law impose stricter requirements on the processing and release of personal data, making it necessary to provide explainable and adjustable trade-offs between privacy protection and data utility in practical data publishing scenarios [1,2].

In anonymizing structured tabular data, k-anonymity and its extensions have long served as mainstream frameworks [1]. Representative methods such as Mondrian and MDAV focus on constructing equivalence classes (ECs): by partitioning or grouping records and applying interval/set generalization to QIs, they ensure that each record is indistinguishable from at least k−1 others within its EC, thereby reducing re-identification risk under standard linkage assumptions [3–5]. Moreover, ℓ-diversity and t-closeness further incorporate constraints on sensitive-value distributions, complementing k-anonymity by mitigating attribute-inference risks that cannot be sufficiently controlled by EC size alone [6,7].

However, when data exhibit high dimensionality, strong correlations, and nonlinear relationships among attributes, existing approaches still face notable limitations.

 doi:10.20944/preprints202602.0124.v1

1. Partitioning or distance measurement performed directly in the original high-dimensional QI space often fails to capture nonlinear correlations, which can cause metric distortion, fragmented ECs, and unstable information loss [4,8].
2. Sensitive-distribution constraints are frequently enforced in a coarse or post-hoc manner, offering limited process-level controllability over worst-case situations (e.g., excessive concentration of a sensitive value or overly low within-class entropy) [9,10].
3. Although representation learning has been explored in privacy-related studies, systematic and workflow-level integration between learned representations and classic k-anonymity-family mechanisms remains insufficient—especially for simultaneously controlling EC geometry and sensitive distribution within a unified, interpretable framework [11–15].

To address these issues, we propose **AE-LRHMA**, an autoencoder-enhanced hierarchical Mondrian anonymization method based on latent representations. The method first normalizes/encodes QIs and trains an autoencoder to learn a low-dimensional latent space that regularizes nonlinear correlation structures. It then performs Mondrian-style hierarchical binary partitioning in the latent space and adopts a **constraint-aware splitting** strategy: at each split decision, EC size constraints and sensitive-distribution constraints are incorporated jointly, shifting the workflow from "split first, repair later" to a **constraint-driven** mechanism that more stably controls EC shapes and sensitive distributions. After that, a local *(k,e)*-microaggregation procedure is used as a stabilizer to improve within-group structure. Finally, the obtained partitions are mapped back to the original space to generate publishable data via standard interval/set generalization. For utility evaluation, we use NCP to quantify information loss, while for risk evaluation under the publishing–linkage-attack threat model we adopt linkage-attack-oriented indicators (e.g., ERR/UMR) to characterize empirical trends in re-identification and attribute-inference risks.

We consider a one-time release scenario and a publishing–linkage-attack threat model. As illustrated in Figure 1, the publisher anonymizes an original table containing QIs and SAs and releases the anonymized table. An attacker may also hold an auxiliary table with explicit identifiers and overlapping QIs and can perform record linkage by matching QIs, thereby re-identifying individuals and inferring their sensitive values. Our goal is to achieve a more stable and controllable privacy–utility trade-off under this threat model, rather than providing formal guarantees in the sense of differential privacy.

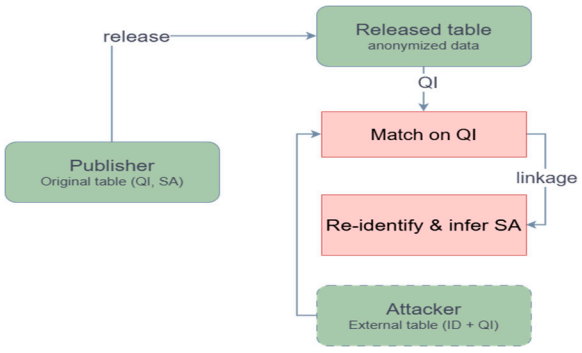

Figure 1. Publishing–linkage-attack threat model.

The main contributions of this paper are summarized as follows:

1. **AE-LRHMA anonymization framework.** We present an end-to-end pipeline for one-time release of static structured data, integrating autoencoder-based latent representations, hierarchical Mondrian partitioning, and local microaggregation to obtain an interpretable privacy–utility compromise.
2. **Latent-space hierarchical binary partitioning.** By learning compact latent representations for high-dimensional and strongly correlated QIs, we perform Mondrian-style recursive splitting in the latent space, alleviating metric distortion in the original space and stabilizing EC geometry.

 doi:10.20944/preprints202602.0124.v1

3. **Constraint-aware splitting mechanism.** We incorporate both EC size constraints and sensitive-distribution constraints directly into split feasibility checking and split scoring, reducing the instability and extra information loss caused by post-hoc "repair" operations.
4. **Tunable hybrid privacy constraints and coupling strategy.** We provide explicit "knobs" (e.g., maximum sensitive proportion, entropy-related control) during partitioning, and couple them with microaggregation-stage constraints to form an integrated risk-control chain from partitioning to final generalization output.
5. **Systematic evaluation under linkage attacks.** On the Adult and Bank Marketing datasets, we evaluate utility (NCP), efficiency, and attacker-oriented empirical risk indicators (ERR/UMR), demonstrating that the proposed method achieves a more reasonable privacy–utility balance than strong baselines.

2. Related Work and Background

2.1. Equivalence-Class Anonymization Models and Sensitive-Distribution Constraints

For one-time release of structured tabular data, equivalence-class (EC) anonymization remains one of the most widely adopted paradigms. k-anonymity partitions records into ECs and generalizes quasi-identifiers (QIs) so that each record becomes indistinguishable from at least k−1 others within its EC, thereby reducing re-identification risk under a typical publishing–linkage-attack assumption [1]. In practice, representative methods such as Mondrian and MDAV follow a relatively stable implementation workflow of "partitioning/aggregation + generalization output," which facilitates deployment in real-world anonymization pipelines [16,17].

However, k-anonymity mainly constrains EC size and does not directly regulate potential homogeneity of sensitive attributes within an EC; thus, attribute inference may still occur. To mitigate this limitation, ℓ-diversity and t-closeness incorporate constraints on sensitive-attribute (SA) distributions: ℓ-diversity emphasizes diversity of sensitive values within each EC, while t-closeness limits the deviation of an EC's SA distribution from the global distribution, reducing worst-case concentration of sensitive values[6,7].

Although these models bring SA distributions into the constraint set, many implementations still treat sensitive-distribution conditions as an end-stage check or a coarse-grained gate[18,19]. This makes it difficult to process-wise and finely control worst-case situations inside ECs (e.g., an excessively high proportion of a sensitive value or overly low sensitive entropy) [20,21]. Motivated by an "interpretable and tunable" risk-control objective under publishing–linkage attacks, we augment the classical k-anonymity framework with two indicators—maximum sensitive proportion and sensitive entropy—to quantify SA concentration and diversity, respectively. Importantly, we move these indicators forward into split feasibility checking and split scoring, so that sensitive-distribution control shifts from post-hoc inspection to a constraint-driven mechanism, which directly supports the tunability goal of our framework.

2.2. Partition-Based Anonymization for High-Dimensional Correlated Data: Mondrian Variants and Limitations

Among EC-based anonymization algorithms, Mondrian represents a canonical paradigm of "recursive binary partitioning + leaf generalization release." It repeatedly selects a split dimension and a cut point to divide a record set into subsets until size/stopping conditions are met, and then applies interval/set generalization to QIs at leaf nodes. This workflow is intuitive, interpretable, and easy to implement, and it can be extended to accommodate more complex privacy constraints; hence, Mondrian and its variants have been widely used and continuously improved in both research and practice.

Existing research around Mondrian has mainly progressed along three directions: (i) extending Mondrian to satisfy more sophisticated privacy models and constraints (e.g., ℓ-diversity) while preserving usability; (ii) developing distributed/parallel implementations for large-scale processing;

 doi:10.20944/preprints202602.0124.v1

and (iii) improving partition-structure quality by introducing structured mappings or stronger organization mechanisms to stabilize partition boundaries and EC shapes [4,8,9].

Nevertheless, when QIs are high-dimensional, strongly correlated, and exhibit nonlinear relationships, axis-aligned splitting in the original attribute space becomes prone to metric distortion and fragmentation. Heuristics based on spans or local distances may fail to capture nonlinear correlation structures, leading to irregular and overly fragmented ECs and making information loss harder to control in a stable manner [4,8]. Consequently, prior work has explored structured mapping mechanisms to improve Mondrian's partition structure and alleviate irregularity and fragmentation caused by original-space axis-aligned splitting.

In this context, our positioning is as follows: Mondrian's key advantage lies in its interpretable hierarchical grouping backbone, whereas its main weakness in high-dimensional correlated settings is that original-space splitting is difficult to stabilize while preserving structure. Therefore, we first learn low-dimensional latent representations of QIs via an autoencoder, and then perform Mondrian-style hierarchical partitioning in the latent space. This design aims to regularize correlations and scale differences geometrically, providing a more reliable basis for stable grouping and controllable information loss.

2.3. Microaggregation and Grouping Stability: MDAV/(k,e)-MDAV and Partition–Aggregation Synergy

Different from Mondrian's "partition-based generalization" route, MDAV exemplifies another important line—microaggregation—which groups similar records and applies representative/interval-based transformations to satisfy group-size constraints while reducing information loss. For large-scale data, MDAV and its variants (including accelerated versions for big data) have evolved toward better efficiency and scalability, forming a relatively mature aggregation-based anonymization family [3–5].

From a publishing–linkage-attack perspective, however, microaggregation may encounter practical tensions under complex distributions: when distributions are heterogeneous, constraints are numerous, or local density is uneven, aggregation can produce boundary residual samples and unstable small clusters, making the overall EC structure difficult to control[22,23]. Accordingly, a more robust trend in research and practice is to use hierarchical partitioning to provide an interpretable grouping framework, and then use local microaggregation to enhance within-group stability, thereby obtaining more controllable information loss and more stable EC structures in complex settings.

Moreover, as feature sources become more complex, multi-view collaborative microaggregation has been used to accommodate multi-source and multi-view feature organizations, better exploiting complementary information and maintaining grouping stability [24]. This further suggests that when the goal goes beyond merely meeting "group size" and also requires structure preservation and multi-constraint coordination, a single aggregation strategy may be insufficient, and partition–aggregation synergy better matches practical data-release needs.

Following this line of thinking, the (k,e)-MDAV component in our framework is best interpreted as a local stabilizer: after latent-space hierarchical partitioning outlines candidate EC contours, we perform local microaggregation within leaf nodes to handle boundary effects and residual-induced local instability, while using parameter e to constrain local sensitive dissimilarity. Together with the constraint-aware splitting mechanism, this forms an end-to-end coordinated control chain from partitioning to final EC stabilization.

2.4. Latent-Representation-Driven Anonymization: Representation Learning and Integration with the k-Anonymity Workflow

In recent years, the intersection of representation learning and privacy protection has attracted increasing attention. Particularly in high-sensitivity domains such as healthcare, surveys and empirical studies emphasize that real-world data sharing requires balancing utility and risk control, and they summarize the trade-offs of different anonymization strategies in practice [11–13]. In

 doi:10.20944/preprints202602.0124.v1

parallel, risk-evaluation studies under publishing–linkage attacks have gradually developed operational indicator systems, enabling comparison of relative risk trends across anonymization strategies under a unified threat assumption [25–27].

Along the direction of "representation learning + EC anonymization," prior work has attempted to learn structured mappings or low-dimensional organization mechanisms to improve the stability of original-space partitioning. Overall, however, workflow-level and interpretable integration of deep representation learning (e.g., autoencoders) with the hierarchical partitioning process of the k-anonymity family remains limited, and systematic solutions that can simultaneously control EC geometry and sensitive distributions within a unified framework are still lacking [14,15].

From a procedural viewpoint, traditional hierarchical partitioning anonymization often follows the pattern of "split by geometric heuristics (e.g., span/variance), check constraints, and patch if necessary." This paradigm has two typical drawbacks: (i) splitting directions are mainly geometry-driven and may yield ECs that satisfy size constraints but severely violate sensitive-distribution balance; and (ii) post-hoc patching introduces extra perturbations and can amplify information loss, making the privacy–utility trade-off less stable.

Based on this research landscape, our entry point can be summarized as process-level constraint-driven control: we perform Mondrian-style hierarchical splitting in the autoencoder-learned latent space, and embed both size constraints and sensitive-distribution constraints (maximum sensitive proportion and sensitive-entropy threshold) directly into feasibility checking and scoring for each split decision. This shifts the workflow from "split first, patch later" to a "constraint-driven" mechanism, enabling more stable control of EC shapes and SA distributions in high-dimensional correlated scenarios. Meanwhile, we do not claim formal privacy guarantees (e.g., differential privacy); instead, under a publishing–linkage-attack setting we employ empirical indicators (e.g., ERR/UMR) to characterize relative risk trends and support interpretable and tunable analysis and comparison.

3. Preliminaries

3.1. Problem Setting and Threat Model

We focus on a one-time release scenario of static structured (tabular) data. As illustrated in Figure 1, a data publisher anonymizes an original dataset containing quasi-identifiers (QIs) and sensitive attributes (SAs) and then releases the anonymized table. Meanwhile, an attacker may possess an external auxiliary table that includes explicit identifiers and overlaps with the released table on the QI fields. By matching records on QIs, the attacker can perform record linkage to re-identify individuals and infer their sensitive values. Our objective under this publishing–linkage-attack setting is to achieve a more stable and controllable privacy–utility trade-off and a clear risk reduction trend, rather than providing formal guarantees in the sense of differential privacy.

Under this setting, anonymization is typically implemented by interval/set generalization on QIs so that records sharing the same generalized QI representation form an equivalence class (EC), thereby reducing individual distinguishability in the QI space. For any given generalization scheme, the dataset can be partitioned into multiple ECs, where each EC consists of the records having identical generalized QI values. The classical k-anonymity model requires that the size of every EC be at least k, which can reduce re-identification risk under certain assumptions in linkage attacks. However, constraining EC size alone is insufficient to regulate attribute-inference risk caused by sensitive-value homogeneity within an EC. Therefore, it is necessary to additionally control whether the sensitive distribution inside an EC is overly concentrated or sufficiently diverse.

In summary, under the publishing–linkage-attack setting, EC size constraint k alone cannot effectively suppress sensitive-value homogeneity within ECs and may still lead to attribute inference. To make risk control operational, we adopt sensitive-distribution constraints (as introduced in Section 2.1) and explicitly treat them as *process-level* criteria and optimization objectives during

 doi:10.20944/preprints202602.0124.v1

subsequent splitting and within-group adjustment, enabling an interpretable and tunable risk reduction behavior.

3.2. Basic Workflow and Limitations of Mondrian Hierarchical Partitioning

Mondrian-style approaches represent a typical hierarchical partitioning–based generalization paradigm in EC anonymization. They recursively construct a binary partition tree: at each level, a split dimension and a cut point are selected to divide a record set into two subsets, until size and stopping conditions are satisfied; then interval/set generalization is applied to leaf nodes to produce the released data. This workflow is intuitive, interpretable, easy to implement, and can be extended to support more complex constraints; hence, it has been widely adopted in both research and engineering practice [4,8,9].

However, when QIs are high-dimensional, strongly correlated, and exhibit nonlinear dependence, directly performing axis-aligned splitting in the original QI space (or describing similarity using common distance metrics) often fails to reflect the latent structure accurately. On the one hand, metric distortion may occur; on the other hand, partition boundaries can become unstable and ECs may become fragmented, which in turn leads to fluctuating information loss and imbalanced sensitive distributions[4,8]. Therefore, while preserving the interpretability of hierarchical partitioning, improving the stability of partition structures and reducing fragmentation is a fundamental issue that subsequent method design must address.

3.3. Microaggregation and an Intuitive View of (k,e)-MDAV

Microaggregation methods group similar records and apply representative/interval-based transformations to satisfy group-size constraints while minimizing information loss. For large-scale data, MDAV and its variants have been extended toward improved efficiency and scalability [3,5]. Moreover, multi-view collaborative microaggregation can adapt to multi-source and multi-view feature organization, demonstrating the potential of microaggregation to handle complex data structures [24]. In scenarios with complex distributions and multiple constraints, relying solely on "partitioning" or solely on "aggregation" is often insufficient to simultaneously maintain interpretability and within-group stability. A more robust strategy is to combine hierarchical partitioning + local microaggregation: hierarchical partitioning provides an interpretable grouping backbone, while local microaggregation enhances within-group similarity and alleviates boundary effects as well as small-cluster/residual-sample issues caused by uneven density, thus producing more stable final EC structures.

In our framework, k corresponds to the minimum EC size requirement. The parameter e serves as a local constraint in the microaggregation stage of (k,e)-MDAV, restricting the spread of sensitive dissimilarity during local aggregation so that sensitive-risk control remains interpretable while grouping is stabilized. Importantly, e does not change the main mechanism of our approach (latent-space hierarchical partitioning and constraint-aware splitting); instead, it acts as a local refinement component that works collaboratively with them to improve the stability of the overall privacy–utility trade-off.

3.4. Evaluation Metrics

To compare the privacy–utility trade-offs of different anonymization strategies and parameter settings under a unified framework, we evaluate both utility and risk (Figure 2). On the utility side, we use NCP (Normalized Certainty Penalty) to quantify information loss caused by QI generalization: for numerical attributes, NCP measures the generalization strength via the EC interval length relative to the global range; for categorical attributes, it measures ambiguity via the EC value-set size relative to the total number of categories. The overall information loss is computed as a weighted aggregation of attribute-wise NCP, which directly reflects the degradation of analytical precision in the released data.

 doi:10.20944/preprints202602.0124.v1

On the risk side, we adopt an operational evaluation pipeline from the publishing–linkage-attack perspective. The attacker performs record linkage by matching QI fields between the external auxiliary table and the released table, and empirical risk indicators are computed to characterize trends in re-identification and attribute-inference risks [25–27]. Specifically, we use containment linkage to construct candidate EC sets: if the QI values of an external record are contained by the generalized interval/set of an EC in the released table for every QI, then this EC is considered a candidate. A smaller candidate set implies a smaller attacker search space; when the candidate set size equals 1, the external record can be uniquely localized to a single EC.

Based on the candidate sets, we use two complementary empirical risk metrics. UMR (Unique Match Rate) measures the proportion of external records that can be uniquely mapped to exactly one candidate EC. ERR (Expected Re-identification Risk) estimates an upper bound of identity-hit risk under a "most favorable candidate" strategy: the attacker prioritizes the smallest EC in the candidate set and uniformly guesses an individual within that EC, yielding the expected hit rate. We emphasize that ERR/UMR are empirical risk measures for comparative analysis under a specified threat model and are not equivalent to formal privacy guarantees such as differential privacy. In this paper, they are used together with NCP to form a two-sided utility–risk evidence chain, enabling clearer presentation of trade-offs across anonymization strategies and parameter settings (Figure 2).

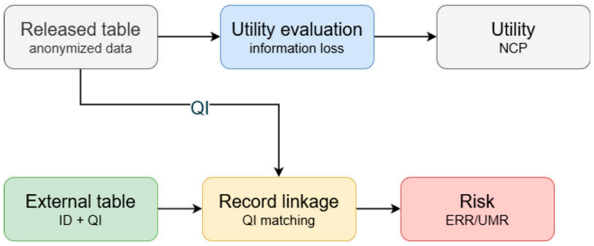

Figure 2. Utility–risk evaluation pipeline under linkage attacks.

4. AE-LRHMA Anonymization Algorithm

4.1. Design Rationale and Overall Framework

Under the publishing–linkage-attack threat model, we propose an autoencoder-enhanced hierarchical Mondrian anonymization method, termed AE-LRHMA. The key idea is to use latent representations as a bridge to unify the workflow of *data preprocessing* → *representation learning* → *hierarchical partitioning* → *generalization release*, and to introduce a constraint-aware splitting mechanism during partitioning to avoid the instability caused by the conventional "split first, repair later" paradigm. The overall framework and the data flow among modules are illustrated in Figure 3.

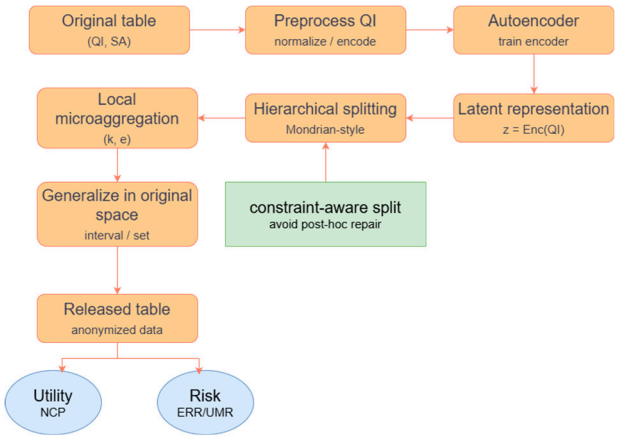

Figure 3. Overall workflow of AE-LRHMA.

 doi:10.20944/preprints202602.0124.v1

Within this framework, we further define a **hybrid privacy constraint** that jointly characterizes both equivalence-class (EC) size and sensitive-attribute (SA) distribution, providing explicit criteria for constraint-aware splitting and subsequent local microaggregation. Let the original dataset be $D = \{(q_i, s_i)|i = 1, \cdots, n\}$, where $q_i \in \mathbb{R}^d$ denotes the quasi-identifier (QI) vector of record i (e.g., age, occupation, education, and contact type), and s_i denotes the corresponding sensitive attribute (SA) value. Let Q be the set of QI attributes and S be the set of SA attributes. Given a generalization scheme, records with the same generalized QI values form an equivalence class (EC) $G \subseteq D$. Denote the EC partition of D as $\mathcal{G} = \{G_1, \cdots, G_m\}$. A consolidated list of symbols and notations is provided in Supplementary Table S3.

To jointly characterize EC size and SA distribution, we introduce two indicators on top of the classical k-anonymity model: the maximum sensitive proportion and the sensitive entropy. For a given EC G, the empirical distribution of SA values is defined as:

$$P_G(v) = \frac{|\{(q_i, s_i) \in G | s_i = v\}|}{|G|} \tag{1}$$

Based on the above distribution, the maximum sensitive proportion and the sensitive entropy are respectively defined as:

$$p_{\max(G)} = \max_v P_G(v) \tag{2}$$

$$H(G) = -\sum_v P_G(v) \log P_G(v) \tag{3}$$

Here, $p_{\max}(G)$ reflects the concentration of a certain sensitive value within an EC, while $H(G)$ reflects the diversity of sensitive values. A higher $p_{\max}(G)$ indicates a higher risk of attribute inference, and a lower entropy indicates weaker diversity. We therefore adopt the following hybrid privacy constraints to restrict EC size and sensitive distribution simultaneously:

$$|G| \geq k, p_{max}(G) \leq p_{max}, H(G) \geq h_{min} \tag{4}$$

In addition, a microaggregation is integrated into the framework to further stabilize utility and provide finer-grained control over EC characteristics. Specifically, if there exists an EC partition $\mathcal{G}$ such that every $G \in \mathcal{G}$ satisfies the constraints in (4), and the sensitive-attribute dissimilarity during local microaggregation does not exceed a threshold e, then the dataset D is said to satisfy $(k, e, p_{\max}, h_{\min})$-anonymity. Compared with traditional k-anonymity and ℓ-diversity models, this definition retains the EC-size constraint while explicitly controlling both concentration and diversity of sensitive information within each EC via p_{max} and an optional entropy hard-gate threshold $h_{\min}$, thereby offering a more fine-grained "knob" at the group-privacy level.

The core idea of AE-LRHMA is to use an autoencoder to perform nonlinear feature extraction and dimensionality reduction for QIs, conduct hierarchical Mondrian-style partitioning in the latent representation space, and locally integrate (k,e)-MDAV together with sensitive-distribution constraints, so as to balance privacy protection and data utility in high-dimensional settings[4]. Unlike partitioning directly in the original attribute space, the latent representations learned via reconstruction can better preserve global structural relationships among records in a low-dimensional space, making partition hyperplanes more regular and EC shapes closer to convex sets, which helps reduce information loss caused by generalization.

Overall, AE-LRHMA links data preprocessing, representation learning, hierarchical partitioning, and generalization output into four stages. First, in the preprocessing stage, the QI set Q and SA set S are selected according to the application scenario, and missing-value imputation, categorical encoding, and numerical normalization are applied to obtain a standardized QI matrix. Second, in the representation learning stage, an autoencoder is trained to extract low-dimensional latent representations from the QI matrix, capturing both global and local structural relationships among records. Third, in the latent space, Mondrian-style recursive binary partitioning is performed, and local (k,e)-MDAV microaggregation together with the $p_{\max}$ / $h_{\min}$ constraints are used to produce candidate ECs that satisfy the hybrid privacy model. Finally, the latent-space partitions are mapped back to the original QI space, where numerical attributes are generalized into intervals and

 doi:10.20944/preprints202602.0124.v1

categorical attributes into value sets, yielding an anonymized dataset that achieves a reasonable privacy–utility trade-off.

4.2. Constraint-Aware Splitting Mechanism

Traditional hierarchical partitioning anonymization often follows a "split by span, check constraints, and patch if needed" workflow, which leads to two issues. First, splitting directions are mainly driven by geometric structure and can easily produce ECs that meet size requirements but exhibit excessive concentration in sensitive distributions. Second, post-hoc patching introduces additional perturbations and amplifies information loss. To address this, AE-LRHMA moves constraint checking into every split decision, shifting the process from "post-hoc patching" to "constraint-driven" splitting.

Concretely, for a current node corresponding to a candidate EC, we generate multiple split candidates in the latent space (e.g., binary threshold splits along latent dimensions). For each candidate split producing left and right subsets, AE-LRHMA jointly evaluates:

1. Size constraint (k): if any subset has fewer than k records, the split is infeasible.
2. Upper bound on sensitive concentration ($p_{\max}$): if the maximum proportion of any sensitive value in a subset exceeds $p_{\max}$, the split is infeasible or receives a strong penalty in scoring.
3. Entropy-based diversity control ($h_{\min}$, optional hard gate): if a subset's sensitive entropy is below $h_{\min}$, the split is infeasible or receives a strong penalty; when $h_{\min}$=0, we disable the hard entropy gate but still encourage entropy-balanced splits through the entropy-improvement term in the composite split-scoring function (Eq. (7)).

Among feasible candidates, AE-LRHMA further leverages latent-space structural information (e.g., span/variance) to select a splitting direction that yields lower information loss and is more favorable for subsequent microaggregation. In this way, k, $p_{\max}$, and $h_{\min}$ are not merely "final constraints" but explicit control variables that directly participate in split decisions, explaining their controllable and tunable role in privacy-risk adjustment.

In addition, the parameter e constrains local sensitive dissimilarity in the (k,e)-MDAV microaggregation stage, limiting the spread of sensitive differences during local aggregation. Notably, e does not change the main contribution (latent-space hierarchical partitioning with constraint-aware splitting) but works collaboratively to improve the overall privacy–utility balance.

4.3. Algorithm Workflow and Key Modules

Let the input dataset be D, where q_i is the QI vector and s_i is the SA value. Key hyperparameters include: k-anonymity parameter k, local sensitive-dissimilarity threshold e, maximum sensitive proportion threshold $p_{\max}$, sensitive optional entropy hard-gate threshold $h_{\min}$, maximum group size max_group_size, latent dimension r, autoencoder training epochs, and weight coefficients in the composite split-scoring function.

(1) Data preprocessing.

First, determine the QI set and SA set and extract corresponding fields. For numerical QIs, perform missing-value imputation, outlier clipping, and z-score normalization. For categorical QIs, we map categories into numeric codes to obtain fixed-dimensional input vectors for the AE (integer encoding in our experiments). We note that integer encoding may introduce an artificial ordinal relationship; thus we discuss this choice as a limitation. After preprocessing, all QI features form a matrix X, where each row represents a record in a d-dimensional QI feature space.

(2) Autoencoder-based representation learning.

An autoencoder composed of an encoder and a decoder is trained in an unsupervised manner with X as input. The encoder maps the input from d dimensions to an r-dimensional latent space through stacked fully connected layers with nonlinear activations; the decoder uses a roughly

 doi:10.20944/preprints202602.0124.v1

symmetric structure to reconstruct inputs back to the original dimension. Parameters are optimized by minimizing the mean reconstruction loss:

$$\mathcal{L}(\theta, \phi) = \frac{1}{n}\sum_{n=1}^{n} \left\| x_i - g_\phi\left(f_\theta(x_i)\right)\right\|_2^2 \quad (5)$$

Regularization and dropout may be used to improve generalization. After training converges, only the encoder is kept to compute latent representations for all samples:

$$Z = f_\theta(X_Q) \in \mathbb{R}^{n\times r} \quad (6)$$

where the i-th row vector is the latent representation of record i.

(3) Latent-space Mondrian partitioning and composite scoring.

Using the latent matrix as input, the algorithm recursively performs Mondrian-style binary partitioning in the latent space. For a current node, if its size is below 2k or it reaches the maximum group size max_group_size, it is marked as a leaf and no further split is performed. Otherwise, the algorithm computes per-dimension statistics (min, max, span), selects the dimension with the largest span as the split axis, and splits the node at the median (or a quantile) into left and right subsets. It then computes sensitive-distribution statistics on both sides, including $p_{\max}$and sensitive entropy H. To preserve structural information while controlling sensitive distributions, a composite split-scoring function is constructed, e.g.,

$$Score(G \to G_L, G_R) = \alpha \cdot \Delta_{\text{span}} + (1-\alpha)\cdot \Delta_H \quad (7)$$

where one term measures the weighted change in latent-space span before/after splitting, another term reflects the entropy improvement on both sides relative to the parent, and the weights are tunable coefficients. A split is accepted only if both subsets satisfy feasibility checks (size constraint and threshold-based sensitive-distribution constraints), and the composite score exceeds a preset threshold; otherwise, the current node becomes a leaf. This yields candidate ECs that satisfy size constraints while accounting for sensitive distributions.

(4) Local (k,e)-MDAV microaggregation and generalization output.

Although latent-space partitioning outlines EC structures, boundary effects and non-uniform distributions may still produce small clusters or leftover records that violate local constraints. Therefore, within each candidate leaf node, AE-LRHMA invokes a Local_LR_KEMDAV subroutine to perform local microaggregation following the (k,e)-MDAV principle: it ensures each final EC has at least k records, and uses distance constraints to control within-group similarity and sensitive dissimilarity. Remaining records are assigned to the nearest valid EC based on a combined distance in latent space and sensitive space. Finally, ECs are mapped back to the original QI space: numerical attributes are generalized to intervals, categorical attributes to value sets, while sensitive attributes remain unchanged to support downstream analysis and modeling. Merging all ECs yields a published dataset satisfying $(k, e, p_{\max}, h_{\min})$-anonymity.

4.4. Risk Upper-Bound Estimation Under Linkage Attacks

To avoid restricting privacy claims to empirical metrics, we provide an interpretable upper-bound estimate of re-identification risk under a given external-knowledge setting and attack model. Suppose an attacker holds an external table T drawn from the same distribution and knows the quasi-identifier set Q used by the released data. The attacker performs containment linkage: for any external record , if for every , the value of x [q] is contained in the generalized interval/set of q in some equivalence class of the released table D, then is regarded as a candidate. This yields the candidate equivalence-class set:

$$C(x) = \{EC_j \,|\, x[Q] \text{ is } contained\ in\ EC_j[Q]\}. \quad (8)$$

The candidate set size reflects the attacker's localization difficulty: the smaller it is, the easier it is to narrow the search space; when it equals 1, the attacker can uniquely localize the record to a single EC. Note that localizing to an EC is not the same as identifying an individual. Without additional

 doi:10.20944/preprints202602.0124.v1

priors (e.g., using sensitive attributes), the attacker's ability to pinpoint an individual is primarily constrained by EC size. Consider the attacker's most favorable strategy: select the smallest EC among candidates. Define:

$$m(x) \triangleq \min_{EC \in C(x)} |EC| \tag{9}$$

When the candidate set is non-empty, under the assumption of "no extra priors and uniform guessing within an EC," admits a **proxy upper bound** under the stated assumption:

$$P_r(hit|x) \leq \frac{1}{m(x)} \tag{10}$$

If the candidate set is empty, the record cannot be matched to any EC and its contribution is treated as zero. Taking expectation over all external records yields an upper-bound estimate of the overall re-identification risk:

$$ERR \triangleq E_{x \sim T}\left[\frac{1}{m(x)}\right] \tag{11}$$

This indicator captures the expected conservative estimate of identity-hit risk under containment linkage and the most favorable candidate strategy; smaller values indicate lower overall re-identification risk. In addition to expected risk, we also consider the probability of being uniquely localized to an EC[25–27], defined as the Unique Match Rate (UMR):

$$UMR \triangleq P_r(|C(x)| = 1) \tag{12}$$

UMR measures how frequently external records are uniquely localized to an EC via containment linkage. Importantly, UMR and the expected risk bound focus on different aspects: UMR measures uniqueness of EC localization, while the expected risk bound measures the expected proxy upper bound of identity hits after localization under the most favorable candidate. They may not move in the same direction (e.g., UMR can be high while the expected risk bound remains low if the uniquely matched EC is still large). We emphasize that these are upper-bound estimates under stated assumptions rather than reproductions of the attacker's true success rate[25–27].

These bounds correspond directly to the publishing constraints. k-anonymity ensures each EC contains at least k records; even if an attacker achieves unique localization, identity uncertainty remains among at least k individuals, limiting success probability by roughly 1/k under uniform guessing. The sensitive-distribution constraints and control concentration and diversity of sensitive values within ECs: the best single-point sensitive guess is bounded by and increasing encourages more balanced sensitive distributions, suppressing worst-case inference risk. Note that the risk metrics mainly characterize localization risk via QIs; their variation is more directly influenced by generalization strength and EC structure, while / affect EC structures and attack behavior indirectly through the constraint-aware splitting and generalization procedures[25–27]. Section 5.7 further provides empirical validation of these trends.

4.5. Pseudocode and Complexity Analysis

The full pseudocode of AE-LRHMA is provided in Supplementary Table S6 (Algorithm 1). The algorithm takes as input the original dataset, QI set, SA set, privacy parameters k,e, , , autoencoder parameters, and related thresholds, and outputs an anonymized dataset. It first preprocesses QIs and trains the autoencoder to obtain low-dimensional latent representations; then performs Mondrian-style recursive partitioning in latent space, maintaining a queue of subsets to process; it uses sensitive dissimilarity and the composite constraint function Check_Constraints to select candidate subsets and a remainder set; next, it calls Local_LR_KEMDAV within each candidate subset to complete (k,e)-MDAV-based local grouping; finally, remaining records are assigned to the nearest EC, and interval/set generalization is performed in the original QI space to produce the final anonymized dataset[5].

From a complexity perspective, let the original QI dimension be d, latent dimension be r, number of records be n, and the number of training epochs be E. In the representation learning stage, the

 doi:10.20944/preprints202602.0124.v1

dominant cost per epoch is approximately linear in n and d, giving a time complexity on the order of:

$$O(E \cdot n \cdot d) \tag{13}$$

In the latent-space hierarchical partitioning stage, each split requires sorting and scanning samples along a chosen dimension. Assuming the partition tree height is on the order of $\log n$, the total time complexity for the entire partitioning tree is approximately:

$$O(r \cdot n \log n) \tag{14}$$

In the local (k,e)-MDAV microaggregation stage, if final EC sizes are roughly on the order of k, the total aggregation overhead can be estimated accordingly. Overall, the total time complexity of AE-LRHMA can be approximated as:

$$O(E \cdot n \cdot d + r \cdot n \log n + n \cdot k) \tag{15}$$

In practical settings, r is typically a small constant relative to d, so the algorithm scales well.

For space complexity, the algorithm needs to store the original QI matrix X, the latent representation matrix Z, the hierarchical partition tree, and EC partition results. The total space cost is approximately:

$$O\big(n \cdot (d + r)\big) \tag{16}$$

For the dataset scales used in our experiments, this memory footprint is easily manageable on standard server configurations and can support larger-scale structured data anonymization.

5. Algorithm Performance Comparison Experiments

5.1. Datasets and Preprocessing

In this study, we use two public datasets—UCI Bank Marketing and Adult Census Income—as experimental benchmarks. Both contain typical demographic and behavioral features, making them suitable for evaluating the privacy–utility trade-off of tabular data anonymization methods[28,29]. Supplementary Tables S1–S2 list the selected fields and their roles used in this paper.

Bank Marketing comes from a Portuguese bank's real telemarketing calls, with 45,211 records and 17 attributes. We select age, marital_status, education, contact, duration, and campaign as quasi-identifiers, set job_categorical as the sensitive attribute, and use client_id only as an index for reconstruction. During preprocessing, we first remove records with severe missingness or obvious anomalies; then apply integer encoding to categorical fields and z-score normalization to numeric QIs to remove scale effects and stabilize autoencoder training. Categorical QIs remain discrete and are generalized according to their semantics.

Adult is derived from the 1994 U.S. Census, with 32,561 records and 15 fields in the original training set. For comparability with Bank Marketing, we construct a field set isomorphic to Supplementary Table S1 (Supplementary Table S2): we keep key attributes such as age, education, marital-status, occupation, and hours-per-week, and form the aligned fields via renaming/merging (e.g., age, education, marital_status, job_categorical, duration), where job_categorical remains the sensitive attribute. We remove records containing "?" first, apply label encoding to all categorical fields, and perform type validation plus z-score normalization on continuous attributes such as age, duration, and campaign, ensuring Adult and Bank Marketing match in both field count and type.

5.2. Experimental Settings

All experiments are conducted on 10,000 randomly sampled records from each dataset to ensure comparability and control runtime cost. Experiments run on Windows 11; implementations are in Python. Hardware: AMD Ryzen 5 5600 (6 cores) and 32 GB RAM. Key hyperparameter settings for AE-LRHMA and baselines are listed in Table 1.

Autoencoder setup: numeric QIs are z-score normalized; categorical QIs are integer-encoded and fed into the AE. The AE is a fully-connected encoder–decoder with ReLU activations; latent

 doi:10.20944/preprints202602.0124.v1

dimension r follows Table 3. Training uses Adam (lr = 1e−3), epoch = 50, batch = 128, MSE loss, and a fixed random seed (seed = 42).

Table 1. Key hyperparameter settings of AE-LRHMA.

Category	Symbol / parameter	Meaning	Default
Privacy parameters	k	k-anonymity (minimum EC size)	8
	e	sensitive-difference threshold for local (k,e)-MDAV	3
	$p_{\max}$	upper bound of maximum sensitive proportion	0.8
	$h_{\min}$	optional hard-gate entropy threshold for sensitive distribution	0.0
Split stopping	variance_threshold	stop splitting if latent-space variance is below this value	0.05
Group size cap	max_group_size	maximum equivalence-class size	3k = 24
AE	r	latent dimension	3
	E	AE training epochs	50

5.3. Evaluation Metrics

To evaluate anonymization from multiple aspects—privacy strength, sensitive-distribution balance, and computational efficiency—we adopt the following metrics on both datasets:

(1) Equivalence-class (EC) size metrics. The number of equivalence classes, the average group size, and the maximum group size reflect the grouping granularity after anonymization. More ECs and smaller average group size usually imply finer partitioning and potentially stronger privacy, but often with higher information loss; fewer ECs and overly large ECs may cause over-generalization and obscure fine-grained structure. We report these three EC-size metrics for all methods to compare their privacy–utility trade-offs [29].

(2) Sensitive-distribution metrics. To characterize concentration and diversity of sensitive values within each EC, we use the maximum sensitive proportion $p_{max}(G)$ and sensitive entropy H(G) defined earlier (see Eqs. (2)–(3)). Here, $p_{\max}(G)$ reflects whether a sensitive value is overly concentrated, while H(G) measures diversity. Since we care about overall trends, we report the average $p_{\max}$ and average H across all ECs: smaller $p_{\max}$ indicates less concentration; larger H indicates more diverse and balanced sensitive distributions.

(3) Runtime. We record the total runtime of one anonymization run under the same hardware and parameter settings. Since Bank and Adult are similar in scale and dimensionality, runtime directly reflects overhead differences across hierarchical splitting and microaggregation.

(4) Information loss. We use the NCP (Normalized Certainty Penalty)-based normalized information loss to quantify the overall precision loss. For numeric attributes, NCP uses the ratio of the EC interval length to the global range; for categorical attributes, NCP uses the ratio of the number of categories in an EC to the total number of categories[9,30]. Overall information loss is a weighted average across attributes; larger values indicate more severe loss. The "Information loss" column in Table 2 corresponds to this metric.

Based on these metrics, Table 4 summarizes results of MDAV, APMCA, and AE-LRHMA on both datasets.

5.4. Visualization of AE-LRHMA Anonymization Results

To visually demonstrate grouping effects of AE-LRHMA in the latent space, we select the first three dimensions of the autoencoder latent representation and perform 3D visualization for both datasets before and after anonymization. Specifically, for each dataset we plot two subfigures: "latent distribution before anonymization" and "grouping results after AE-LRHMA," corresponding to

 doi:10.20944/preprints202602.0124.v1

Figure 4(a)(b) and Figure 5(a)(b). To enable fair visual comparison, the two subfigures for the same dataset use identical coordinate ranges and camera views; only point colors/labels differ to avoid visual bias caused by automatic axis scaling.

From Figure 4(a) and Figure 5(a), we observe that before anonymization the latent-space distributions span a wide range, with clear local density differences, and there exist isolated points and tiny clusters—indicating complex and irregular latent structures. After AE-LRHMA, we first obtain equivalence classes that satisfy k-anonymity and sensitive-attribute constraints, then apply KMeans on EC centroids to merge ECs into 10 coarse groups for coloring in Figure 4(b) and Figure 5(b). (KMeans is used only for visualization; it does not affect anonymization outputs or the statistics in Table 4 and Supplementary Tables S4–S5.) After processing, samples within the same coarse group form multiple compact sub-clusters, different groups are more clearly separated, and isolated points/tiny clusters are reduced. This suggests that under privacy constraints, AE-LRHMA can better preserve global latent-space structure, improving interpretability for downstream analysis and modeling.

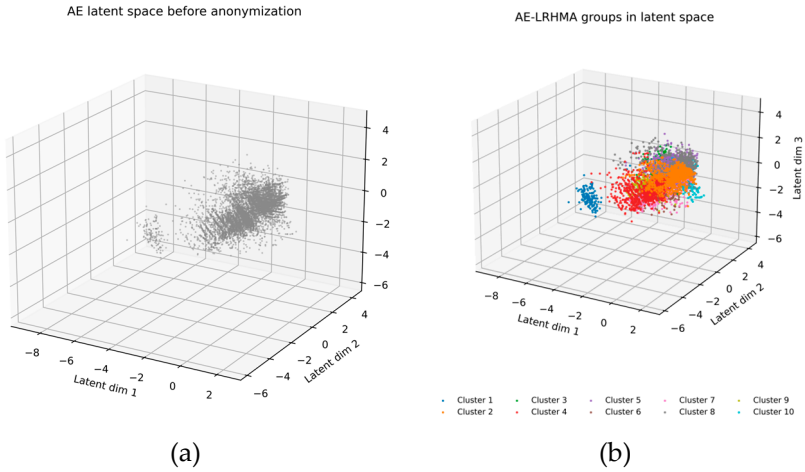

Figure 4. (a). Latent space before anonymization. (b). AE-LRHMA groups in latent space.

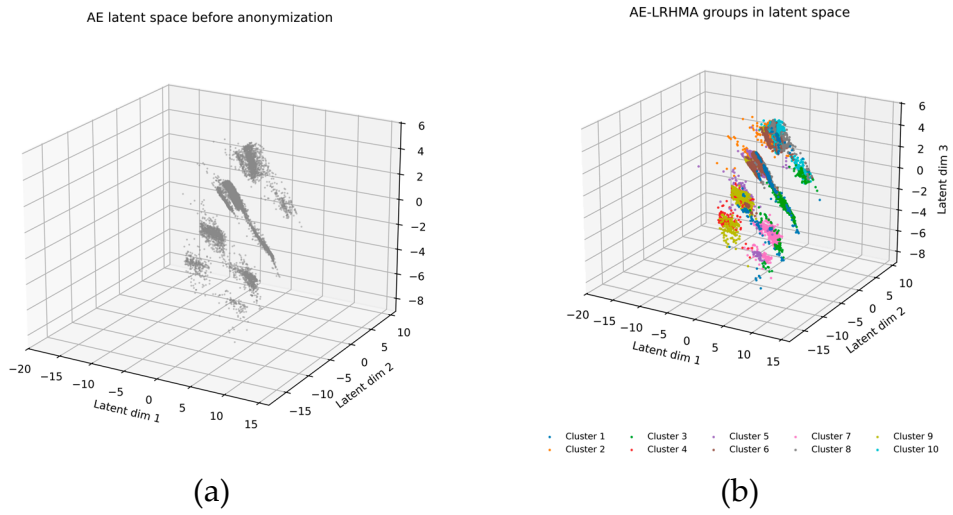

Figure 5. (a). Latent space before anonymization. (b). AE-LRHMA groups in latent space.

5.5. Comparison and Analysis of Three Methods

To quantitatively compare MDAV[3], APMCA[10], and AE-LRHMA, we report EC-size metrics, sensitive-distribution metrics (average , average entropy H), runtime, and NCP-based information loss on both datasets (Table 2). The three methods show clear differences in grouping granularity, sensitive-distribution balance, and computational cost.

 doi:10.20944/preprints202602.0124.v1

Table 2. Performance comparison of three algorithms on Bank and Adult.

Dataset	Method	#ECs	Avg group size	Max group size	Avg. p_{max}	Avg. entropy H	Runtime	NCP
Adult	MDAV	1250	8	8	0.3447	2.2333	194.63 s	0.4431
Adult	APMCA	411	25	147	0.3198	2.4973	7.87 s	0.3861
Adult	AE-LRHMA	506	20	24	0.3124	1.7690	7.92 s	0.2808
Bank	MDAV	1251	8	8	0.3919	2.0392	197 s	0.4516
Bank	APMCA	914	11	22	0.3792	2.1501	12.91 s	0.4087
Bank	AE-LRHMA	428	23	24	0.4303	1.5064	7.96 s	0.3373

Adult. MDAV yields 1250 equivalence classes (ECs), with both the average and maximum group sizes equal to 8, indicating that most groups collapse to the k lower bound and the partition becomes highly fragmented. Although k-anonymity is satisfied, the resulting fine-grained intervals/category sets incur larger information loss, and the runtime is close to 200 s (the highest among the three). APMCA forms 411 ECs (average size ≈ 25; maximum size = 147), reflecting a tendency toward over-generalization; it achieves $p_{\max}$=0.3198, H=2.4973, and NCP = 0.3861, improving over MDAV. AE-LRHMA produces 506 ECs (average size ≈ 20) with the maximum size capped within 24, leading to a more controlled group structure and the lowest information loss ($p_{\max}$=0.3124, H=1.7690, NCP = 0.2808). In terms of runtime, AE-LRHMA (≈ 7.92 s) is comparable to APMCA (≈ 7.87 s) and substantially faster than MDAV. The lower entropy H of AE-LRHMA relative to APMCA reflects a utility-oriented trade-off in sensitive-value dispersion. Here, we set $h_{\min} = 0$ to disable the hard entropy gate, while still encouraging entropy-balanced splits via the entropy-improvement term in the split-scoring function (Eq. (7)); the primary privacy control is enforced by $p_{\max}$ together with group-size constraints. The ERR/UMR linkage evaluation further suggests acceptable empirical disclosure risk under the configured thresholds.

Bank. Similar patterns are observed on Bank Marketing. MDAV produces a large number of small equivalence classes (1251 ECs; average/maximum size = 8) with NCP = 0.4516 and an approximate runtime of 197 s. APMCA forms 914 ECs (average size ≈ 11; maximum size = 22), resulting in lower NCP (0.4087) and improved sensitive-distribution indicators ($p_{\max}$=0.3792, H=2.1501) compared with MDAV. AE-LRHMA forms 428 ECs (average size ≈ 23; maximum size constrained within 24) and achieves the lowest NCP (0.3373). Although its sensitive-distribution indicators ($p_{\max}$=0.4303, H=1.5064) are higher than those of APMCA, they remain within the configured thresholds, providing an interpretable privacy–utility compromise. In terms of runtime, AE-LRHMA (≈ 7.96 s) is faster than APMCA (≈ 12.91 s) and substantially faster than MDAV.

Overall, MDAV tends to compress ECs to the minimum size, causing fragmentation, higher information loss, and the longest runtime—serving as a conservative "privacy-first" baseline. APMCA significantly reduces runtime and information loss via aggregation, but EC sizes can become large in some cases, risking over-generalization. AE-LRHMA performs hierarchical partitioning in the latent space and combines local (k,e)-MDAV with sensitive-distribution controls, achieving a more balanced trade-off across privacy, information loss, and efficiency, with the lowest NCP (0.2808/0.3373) and acceptable runtime on both datasets—making it more suitable for practical tabular data anonymization.

5.6. Ablation Study

To further analyze the role of each component in AE-LRHMA, we construct three ablation variants on both datasets while keeping other settings unchanged:

(1) **full-AE-LRHMA:** full model with autoencoder representation learning, latent-space hierarchical splitting, and sensitive-distribution constraints;
(2) **w/o-diversity:** keep AE and latent-space partitioning but remove maximum sensitive proportion and entropy constraints, leaving only k-anonymity and EC-size-related constraints;

 doi:10.20944/preprints202602.0124.v1

(3) **w/o-AE:** remove AE representation learning and perform hierarchical splitting plus local microaggregation directly in the original QI space. Results are summarized in Table 3.

Table 3. Ablation results on Adult and Bank.

Dataset	Method	#ECs	Avg. group size	Max group size	Avg. p_{max}	Avg. entropy H	Runtime	NCP
Adult	full-AE-LRHMA	506	20	24	0.3124	1.7690	7.92 s	0.2808
Adult	w/o-diversity	520	20	24	0.3213	1.7431	7.12 s	0.2778
Adult	w/o-AE	1132	9	24	0.3842	1.4485	4.73 s	0.1340
Bank	full-AE-LRHMA	428	23	24	0.4303	1.5064	7.96 s	0.3373
Bank	w/o-diversity	442	23	24	0.4326	1.4887	6.567 s	0.3293
Bank	w/o-AE	971	10	24	0.4784	1.2525	3.52 s	0.3023

Note: " ↓ " indicates lower information loss, but it does not necessarily imply stronger privacy.

First, comparing full-AE-LRHMA with w/o-AE shows that the autoencoder representation learning module is crucial for stabilizing EC structure and grouping granularity. Without AE, both datasets exhibit clear fragmentation: on Adult, #ECs increase from 506 to 1132 and average group size drops from 20 to 9; on Bank, #ECs rise from 428 to 971 and average size drops from 23 to 10, though max group size is still capped by max_group_size (=24). This indicates that performing Mondrian-style recursive splits directly in the original high-dimensional QI space is more likely to cause axis-aligned "hard cuts" due to inconsistent scales, complex correlations, and sparse outliers, repeatedly generating small ECs near the k lower bound and destabilizing the overall grouping structure.

Although w/o-AE yields a lower NCP on both datasets, this decrease should not be interpreted as improved anonymization quality. It is mainly a degeneracy effect: smaller equivalence classes reduce the numerical width of interval generalization, which can lower NCP while simultaneously fragmenting group structure and weakening risk controllability. In contrast, full AE-LRHMA reconstructs and regularizes high-dimensional quasi-identifier geometry in a compact latent space, so that split-dimension selection and median-based splitting better follow the underlying sample density. This reduces distance distortion and local noise, leading to more stable equivalence classes in both size and shape.

Second, with AE retained, comparing full-AE-LRHMA and w/o-diversity isolates the effect of sensitive-distribution constraints. The two are almost identical in structural metrics (#ECs, average size, max size), indicating that in the latent space, EC geometry is mainly determined by AE + hierarchical splitting. However, after removing $p_{\max}$ and entropy constraints, average $p_{\max}$ increases and average entropy decreases, suggesting that without explicit diversity control, local correlations between sensitive attributes and some QIs may still cause sensitive values to over-concentrate in certain ECs. With sensitive-distribution constraints enabled, the CheckConstraints module can reject or further split candidate partitions with "excessive concentration or insufficient diversity," suppressing extreme homogenization without significantly changing group-size distributions.

Finally, in terms of efficiency, w/o-AE is fastest because it avoids AE training and latent mapping; w/o-diversity incurs slightly more overhead than the full model due to additional sensitive-constraint checks and scoring computations. Overall, the added cost is acceptable and yields more stable EC structures and more controllable sensitive distributions. In summary, AE mainly stabilizes EC shapes and scales, while sensitive-distribution constraints suppress worst-case sensitive

 doi:10.20944/preprints202602.0124.v1

concentration; the two complement each other in AE-LRHMA, enabling a more stable and interpretable trade-off among privacy strength, utility, and efficiency.

5.7. Linkage-Attack Evaluation from the Attacker's Perspective

To validate the empirical trend and stability of the containment-linkage risk estimation (conservative/proxy upper-bound under stated assumptions) introduced earlier, we construct an external table T for both datasets and perform containment linkage under the setting that the QI set Q is known. For any external record $x \in T$, we compute its candidate EC set C(x) via Eq. (8). Assuming no sensitive-attribute prior knowledge, the attacker adopts the "most favorable candidate" strategy—guessing using the smallest EC in C(x) (Eq. (9)). We then use Eq. (11) as the expected proxy upper bound of identity-hit risk (ERR) and Eq. (12) as the probability of being uniquely located into an EC (UMR). Importantly, UMR measures uniqueness of EC localization, while ERR measures the expected proxy upper bound of further identity hit under the most favorable candidate; therefore, they need not change in the same direction.

We vary the external sampling ratio ext_frac $\in$ {0.05, 0.10, 0.20} and repeat experiments with different random seeds. Tables 4 and 5 report ERR/UMR under different ext_frac and seeds. To visualize the overall "utility loss–linkage risk" trade-off, we further plot NCP vs. ERR as a risk–utility scatter plot (Figure 6), where each point corresponds to one experimental configuration; detailed tables are provided in Supplementary Tables S4–S5.

Table 4. Linkage-attack risk on Adult dataset.

Method	ext_frac=0.05 (ERR / UMR)	ext_frac=0.10 (ERR / UMR)	ext_frac=0.20 (ERR / UMR)
APMCA	0.040748±0.000777/ 0.9188±0.0104	0.040948±0.001193/ 0.9262±0.0073	0.040968±0.000418/ 0.9306±0.0021
AE-LRHMA	0.082667±0.002392/ 0.0032±0.0023	0.081917±0.001319/ 0.0054±0.0011	0.082383±0.001169/ 0.0050±0.0019
MDAV	0.124950±0.000112/ 0.0000±0.0000	0.124950±0.000068/ 0.0000±0.0000	0.124913±0.000034/ 0.0003±0.0003

Note: ERR and UMR are computed by Eqs. (11) and (12), respectively; when C(x)=∅, the external record is considered unmatched and contributes 0 risk.

Table 5. Linkage-attack risk on Bank dataset.

Method	ext_frac=0.05 (ERR / UMR)	ext_frac=0.10 (ERR / UMR)	ext_frac=0.20 (ERR / UMR)
APMCA	0.090957±0.001188/ 0.9996±0.0009	0.091268±0.000715/ 0.9998±0.0004	0.091432±0.000320/ 0.9998±0.0003
AE-LRHMA	0.049080±0.000435/ 0.0080±0.0035	0.048895±0.000305/ 0.0076±0.0036	0.049078±0.000222/ 0.0082±0.0021
MDAV	0.125000±0.000000/ 0.0020±0.0035	0.124975±0.000056/ 0.0026±0.0009	0.124913±0.000056/ 0.0018±0.0004

From Tables 4–5, the relative ranking across methods is generally stable under different ext_frac and random seeds, indicating robustness of the attack evaluation. The overall trends are consistent with EC-structure and utility-loss analyses: MDAV often produces many small ECs near the k lower bound, making candidate sets easier to shrink and yielding higher ERR; AE-LRHMA strikes a balance between EC size and generalization strength, controlling risk at an acceptable level while achieving lower utility loss; APMCA's risk behavior depends on its aggregation pattern and can vary across datasets.

As shown in Figure 6, the risk–utility positions of different methods are clear: AE-LRHMA maintains lower NCP on both datasets with moderate risk; APMCA has lower risk on Adult but

 doi:10.20944/preprints202602.0124.v1

higher risk on Bank; MDAV lies in a region that is not advantageous in either risk or utility. Together, Tables 4–5 and Figure 6 indicate that reporting only information loss or runtime is insufficient to reflect real privacy risk; attacker-perspective linkage evaluation complements the evidence chain and makes conclusions about the privacy–utility trade-off more complete.

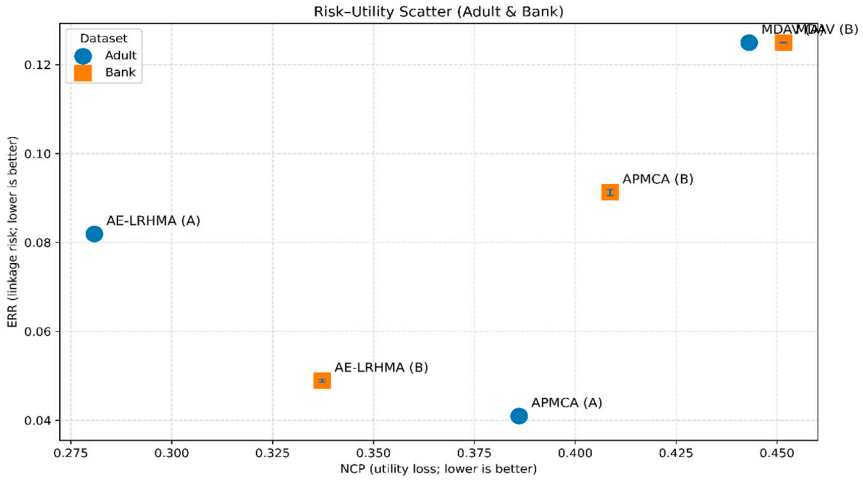

Figure 6. Risk–Utility Scatter (NCP–ERR) on Adult & Bank.

6. Conclusions

This paper addresses the privacy leakage risks that arise when high-dimensional structured data are published and shared. Building on the classical hierarchical Mondrian anonymization framework, we introduce low-dimensional latent representations learned by an autoencoder and propose an autoencoder-enhanced hierarchical anonymization method, AE-LRHMA. The method first performs unsupervised representation learning over the quasi-identifiers to compress the original high-dimensional attributes into a more compact latent space. It then conducts Mondrian-style hierarchical partitioning in the latent space, integrates local (k,e)-MDAV microaggregation, and incorporates sensitive-distribution constraints to explicitly control equivalence-class size as well as the concentration and entropy level of sensitive attributes within each equivalence class. Finally, it maps the partitioning results back to the original attribute space and applies interval/set generalization to produce released data that satisfy the anonymization constraints.

Experiments on two public datasets, Bank Marketing and Adult, show that AE-LRHMA achieves a more balanced performance than classic MDAV and APMCA in terms of equivalence-class granularity, sensitive-attribute distribution balance, and runtime. On the one hand, it avoids the severe information loss caused by MDAV's tendency to generate a large number of extremely small equivalence classes; on the other hand, it mitigates the over-generalization issue of APMCA, where equivalence classes can become overly large. Overall, AE-LRHMA can effectively strengthen group-level privacy protection while preserving data utility, demonstrating practical value and application potential for structured-data anonymization in real-world scenarios. This finding is consistent with recent research trends in medical data anonymization, health data sharing, and privacy-enhancing machine learning, highlighting the practical importance of jointly considering privacy strength and data utility when designing anonymization algorithms[11–15].

Nevertheless, this study has limitations. First, AE-LRHMA remains within the equivalence-class-based k-anonymity paradigm. The thresholds on maximum sensitive proportion, optional entropy hard-gate threshold, and local sensitive dissimilarity are heuristic, empirically motivated controls; they are not formal privacy guarantees (e.g., differential privacy) and do not provide strict theoretical upper bounds. Therefore, performance under stronger threat models (e.g., interactive queries or active injection) requires further investigation. Second, our evaluation focuses on Adult and Bank Marketing; settings with multiple sensitive attributes, highly imbalanced distributions, or

 doi:10.20944/preprints202602.0124.v1

temporal/tabular sequences have not been systematically validated. Third, the current implementation is single-machine and serial, leaving room for improvement on large-scale data. Future work includes integrating principled randomization mechanisms to better connect with formal privacy frameworks and extending the method to multi-sensitive and temporal scenarios with improved scalability (e.g., parallelization and incremental processing).

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org: Tables S1–S6. Tables S1–S2 report the field configurations for the Bank Marketing and Adult datasets; Table S3 lists the notations and symbols used in this manuscript; Tables S4–S5 provide complete linkage-attack metrics on both datasets (mean ± std over five random seeds); Table S6 contains the full algorithm listing and pseudocode of AE-LRHMA.

Author Contributions: Conceptualization, Minghui Zheng; Methodology, Junpeng Hu and Minghui Zheng; Software, Tao Hu; Validation, Junpeng Hu and Tao Hu; Formal analysis, Junpeng Hu; Investigation, Junpeng Hu; Data curation, Tao Hu and Jinan Shen; Writing—original draft, Tao Hu and Junpeng Hu; Writing—review and editing, Minghui Zheng, Tao Hu and Junpeng Hu; Visualization, Tao Hu; Supervision, Minghui Zheng; Project administration, Minghui Zheng. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 62262020).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The Adult Census Income and Bank Marketing datasets analyzed in this study are publicly available from the UCI Machine Learning Repository. The implementation code is available at: https://github.com/sakula328/AE-LRHMA.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aufschläger R, Folz J, März E, Guggumos J, Heigl M, Buchner B, et al. Anonymization Procedures for Tabular Data: An Explanatory Technical and Legal Synthesis. Information. 2023;14(9):487. doi:10.3390/info14090487.
2. Sanchez-Serrano P, Rios R, Agudo I. Privacy-Preserving Tabular Data Generation: Systematic Literature Review. In: Garcia-Alfaro J, et al., editors. Computer Security. ESORICS 2024 International Workshops. Lecture Notes in Computer Science. Cham: Springer; 2025. p.170–180. doi:10.1007/978-3-031-82349-7_12.
3. Rodríguez-Hoyos A, Estrada-Jiménez J, Rebollo-Monedero D, Mezher AM, Parra-Arnau J, Forné J. The Fast Maximum Distance to Average Vector (F-MDAV): An algorithm for k-anonymous microaggregation in big data. Eng Appl Artif Intell. 2020;90:103531. doi:10.1016/j.engappai.2020.103531.
4. Bloomston A, Burke E, Cacace M, Diaz A, Dougherty W, Gonzalez M, et al. Core Mondrian: Basic Mondrian beyond k-anonymity. arXiv [Preprint]. 2025 [cited 2025 Dec 29]. arXiv:2510.09661.
5. Thaeter F, Reischuk R. Improving Time Complexity and Utility of k-anonymous Microaggregation. In: Samarati P, et al., editors. E-Business and Telecommunications. ICETE 2021. Communications in Computer and Information Science, vol 1795. Cham: Springer; 2023. p.195–223. doi:10.1007/978-3-031-36840-0_10.
6. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. ℓ-diversity: Privacy beyond k-anonymity. ACM Trans Knowl Discov Data. 2007;1(1):Article 3. doi:10.1145/1217299.1217302.
7. Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity. In: Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE); 2007. p.106–115. doi:10.1109/ICDE.2007.367856.
8. Padmaja R, Santhi V. An Extended Mondrian Algorithm – XMondrian to Protect Identity Disclosure. In: Ambeth Kumar VD, et al., editors. Smart Intelligent Computing and Communication Technology. Advances in Parallel Computing. Amsterdam: IOS Press; 2021. p.480–490. doi:10.3233/APC210088.

 doi:10.20944/preprints202602.0124.v1

9. Ashkouti F, Khamforoosh K, Sheikhahmadi A. DI-Mondrian: Distributed Improved Mondrian for Satisfaction of the L-diversity Privacy Model Using Apache Spark. Inf Sci. 2021;546:1–24. doi:10.1016/j.ins.2020.07.066.
10. Xiong J, Chen X, Lan X. Adaptive perturbation anonymization algorithm based on multi-layer clustering and its application [in Chinese]. Modern Electronics Technique. 2025;48(14):85–89. doi:10.16652/j.issn.1004-373x.2025.14.014.
11. Olatunji IE, Rauch J, Katzensteiner M, Khosla M. A review of anonymization for healthcare data. Big Data. 2024;12(6):538–555. doi:10.1089/big.2021.0169.
12. Karagiannis S, Ntantogian C, Magkos E, Tsohou A, Landeiro Ribeiro L. Mastering data privacy: Leveraging k-anonymity for robust health data sharing. Int J Inf Secur. 2024;23:2189–2201. doi:10.1007/s10207-024-00838-8.
13. Onesimu JA, Karthikeyan J, Eunice J, Pomplun M, Dang H. Privacy preserving attribute-focused anonymization scheme for healthcare data publishing. IEEE Access. 2022;10:86979–86997. doi:10.1109/ACCESS.2022.3199433.
14. Languré ADL, Zareei M. Privacy-preserving emotion detection: Evaluating the trade-off between k-anonymity and model performance. IEEE Access. 2025;13:105901–105910. doi:10.1109/ACCESS.2025.3578958.
15. Bozorgpanah A, Torra V. Explainable machine learning models with privacy. Prog Artif Intell. 2024;13(1):31–50. doi:10.1007/s13748-024-00315-2.
16. Yan Y, Herman EA, Mahmood A, Feng T, Xie P. A weighted K-member clustering algorithm for k-anonymization. Computing. 2021;103:2251–2273. doi:10.1007/s00607-021-00922-0.
17. Coelho KK, Okuyama MM, Nogueira M, Vieira AB, Silva EF, Nacif JAM. A new k-anonymity method based on generalization-first k-member clustering for healthcare data. IEEE Trans Dependable Secure Comput. 2025;1–12. doi:10.1109/TDSC.2025.3615541.
18. Wang H, He J, Zhu N. Improving data utilization of k-anonymity through clustering optimization. Trans Data Priv. 2022;15(3):177–192.
19. Khatir RA, Izadkhah H, Razmara J. Designing a novel approach using a greedy and information-theoretic clustering-based algorithm for anonymizing microdata sets. Entropy. 2023;25:1613. doi:10.3390/e25121613.
20. Kacha L, Zitouni A, Djoudi M. KAB: A new k-anonymity approach based on black hole algorithm. J King Saud Univ Comput Inf Sci. 2022;34(7):4075–4088. doi:10.1016/j.jksuci.2021.04.014.
21. Kacha L. NKAB: An optimization approach for k-anonymity based on Black Hole Algorithm. Computers & Security, 2025, 157: 104612. doi:10.1016/j.cose.2025.104612.
22. Patil A, Wang B. Advancing data privacy: A novel k-anonymity algorithm with dissimilarity tree-based clustering and minimal information loss. Int J Recent Innov Trends Comput Commun. 2023;11(8):323–330. doi:10.17762/ijritcc.v11i8.8005.
23. Wang F, Chen H, Zhou Y. A privacy protection application of consumer personal information based on an improved k-anonymity algorithm. In: Proceedings of the 2024 5th International Conference for Emerging Technology (INCET); 2024 May 24–26; Karnataka, India. p. 1–7. doi:10.1109/INCET61516.2024.10593091.
24. Zouinina S, Bennani Y, Rogovschi N, Lyhyaoui A. Data Anonymization through Collaborative Multi-view Microaggregation. J Intell Syst. 2021;30:327–345. doi:10.1515/jisys-2020-0026.
25. Jiang Y, Mosquera L, Jiang B, Kong L, El Emam K. Measuring re-identification risk using a synthetic estimator to enable data sharing. PLoS One. 2022;17(6):e0269097. doi:10.1371/journal.pone.0269097.
26. Sondeck LP, Laurent M. Practical and ready-to-use methodology to assess the re-identification risk in anonymized datasets. Sci Rep. 2025;15:23223. doi:10.1038/s41598-025-04907-3.
27. Lee S, Kim Y, Kwon Y, Cho S. Secure privacy-preserving record linkage system from re-identification attack. PLoS One. 2025;20(1):e0314486. doi:10.1371/journal.pone.0314486.
28. Wu Q, Tang J, Dang S, Chen G. Data Privacy and Utility Trade-Off Based on Mutual Information Neural Estimator. arXiv [Preprint]. 2021 [cited 2025 Dec 29]. arXiv:2112.09651.
29. Mesana P, Vial G, Jutras P, Caporossi G, Crowe J, Gambs S. Measuring privacy/utility tradeoffs of format-preserving strategies for data release. J Bus Anal. 2025;8(3):147–169. doi:10.1080/2573234X.2025.2461507.

 doi:10.20944/preprints202602.0124.v1

30. Wang PS, Huang PY, Tsai YA, Tso R. An Enhanced Mondrian Anonymization Model based on Self-Organizing Map. In: Proceedings of the 15th Asia Joint Conference on Information Security (AsiaJCIS); 2020. p.97–100. doi:10.1109/AsiaJCIS50894.2020.00026.