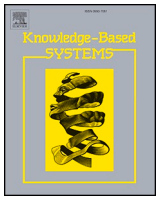
Knowledge-Based Systems 300 (2024) 112174

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



UMAP-SMOTENC: A simple, efficient, and consistent alternative for privacy-aware synthetic data generation

Goncalo Almeida *, Fernando Bacao

NOVA Information Management School, Universidade Nova de Lisboa, Campus de Campolide, Lisboa 1070-312, Portugal

ARTICLE INFO

Keywords:
Anonymization techniques
Machine learning
SMOTE
Synthetic data generation
UMAP

ABSTRACT

The intensification of governmental legislation and the social awareness around data privacy protection severely constrains organizations' data utilization capabilities. As a result, the interest in data anonymization techniques, which should preserve the patterns present in the original data but mitigate the risks of privacy leakage, has also increased. While conventional methods may compromise privacy, recently proposed deep learning generative approaches are computationally expensive and unreliable when used in tabular datasets, hindering the democratization and usability of data. In this paper, we explore this trade-off between privacy and the quality of the anonymized data, establishing a new equilibrium obtained using a synthetic oversampling technique, SMOTE-NC, on a non-linear compressed version of the input space, achieved with the application of UMAP. The introduced approach, UMAP-SMOTENC, constitutes an efficient and consistent solution that can be used without significant efforts on hyperparameter tuning or resourcing to massive computing infrastructures. An experiment was conducted to evaluate the robustness of the proposed solution, comparing several metrics and models across eight datasets with diverse characteristics. The results achieved suggest that the presented method can efficiently synthesize privacy-aware data while conserving the relevant patterns of the real dataset, particularly those required for classification tasks.

1. Introduction

As organizations progress towards evermore data-based decision-making processes, the ability to store, share, and use this resource has created a massive competitive advantage, while its combination with machine learning techniques has potentiated advances in crucial areas such as medical research [81], credit risk scoring [54] and fraud detection [65]. At the same time, new standards, both at the legal and social levels, are emerging, profoundly impacting how organizations interact with personal data [28]. It is under this context that the recent General Data Protection Regulation (GDPR) (Regulation [[72] 2016/679, Recital 26] was developed, imposing a clear distinction between the use of personal and anonymized information, defining the latter, in Recital 26, as "*information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable*". While anonymized data is free from the legal implications of the regulation, personal data is severely constrained by it. The confrontation of these opposite forces, represented by the data's value and the need for compliance, unsurprisingly exposes the urgency for research and adoption of anonymization techniques capable of exempting organizations from the strict guidelines of the GDPR and other similar acts. Examples of these directives include the Canadian Personal Information Protection and Electronic Documents Act [69] and the California Consumer Privacy Act [14].

Conventional anonymization methods, including the combination of data transformations such as generalization and suppression [38], noise addiction [99], or data swapping [18], do not break the one-to-one linkage between data points in the anonymized and original dataset, opening the door for diverse exploitative attacks. Additionally, these methods rely on limiting the patterns present in the original data as the tool for mitigating potential privacy leakage, which can severely hinder the usability of the obtained dataset [71]. As an alternative, recent research has focused on generative techniques that, by definition, do not suffer from re-identification issues, since any direct mapping between elements of the generated and the original datasets should not exist. In a simplified manner, these methods attempt to learn the patterns present in data, to then generate new entries by sampling according to a learned

* Corresponding author.
E-mail address: m20200594@novaims.unl.pt (G. Almeida).

https://doi.org/10.1016/j.knosys.2024.112174
Received 4 September 2023; Received in revised form 26 April 2024; Accepted 17 June 2024
Available online 21 June 2024

distribution. Several approaches have been proposed in the literature, such as adaptations of Variational Autoencoders (VAEs) [44] and Generative Adversarial Networks (GANs) [27] to the tabular data framework. However, despite the virtues of these techniques, most are computationally demanding and require a significant effort in hyperparameter tuning [49], which may hinder their mass adoption. Simultaneously, different methods suffer from specific pain points that limit their usefulness and reliability, for instance, the known issue with mode collapse and convergence stability when working with GANs [46]. As a result, there is still a gap in research waiting for a robust and efficient data anonymizing method capable of breaking the one-to-one relation between the original and the private datasets while preserving the core patterns of the original data.

Existing research has largely overlooked the potential use of Synthetic Minority Oversampling Techniques (SMOTE) [15] and its extensions in this context. While there are valid concerns about the plain use of this method due to the potential for generating entries prohibitively similar to the original ones, its success as a robust out-of-the-box solution for addressing imbalanced machine learning tasks should not be ignored. Therefore, in the present paper, we suggest two modifications of SMOTE's framework, intended to adapt it to the anonymization domain, mitigating potential privacy risks. More concretely, we introduce a new step, meant to compress the multivariate data space to a lossy but representative two-dimensional version using the Uniform Manifold Approximation and Projection technique (UMAP) [60], where the synthetic sampling will take place, reducing the risk of overexposure to the entries present on the original dataset. Then, instead of using this method for class balancing, we perform it across all the present classes, creating a fully synthetic dataset with the same class cardinality as the original one.

In addition, this paper also proposes a benchmark framework for evaluating the trade-off between privacy and utility of the synthetic data, with a particular focus on machine learning classification use cases. This framework allowed us to test the proposed approach, UMAP-SMOTENC, and compare its performance across several different datasets with a significant degree of variability in terms of dimensionality, complexity, and structure. In our evaluation, we not only compared the performance of UMAP-SMOTENC against the original SMOTE-NC method, which appears as a natural baseline for our proposed approach, but also against the Gaussian Copula [67], the Tabular VAE (TVAE) and the Conditional Tabular GAN (CTGAN) [98], three methods supported by a high-quality and efficient open-source initiative [67], being the reference framework in the field and widely popular among industry practitioners. The results obtained suggest that the proposed method surpasses the previously achieved equilibrium in the privacy and utility preservation trade-off, in a statistically significant way.

Therefore, the main contributions of this paper are three-fold. Firstly, it introduces a new and better method capable of generating fully synthetic data without requiring a rigorous hyperparameter tuning phase or a Graphical Processing Units (GPUs) backed infrastructure. Secondly, it defines guidelines for a comparative framework that can be used for future research on the trade-off between privacy and utility preservation for machine learning tasks. Finally, the proposed method's code implementation was open-sourced,[1] making UMAP-SMOTENC accessible to both researchers and practitioners.

This paper is structured as follows: Section 1 introduces the context around the need for privacy when handling data; Section 2 covers some of the classic and state-of-the-art (SOTA) approaches to data anonymization and debates potential data privacy and utility measures. Section 3 explains the suggested solution, succinctly presenting the algorithms on which it is based; Section 4 focuses on the methodology, experimental procedure, and datasets used; Section 5 summarizes and discusses the experimental results achieved; Section 6 presents the conclusions of the paper's findings.

[1] Available at https://github.com/gdalmeida99/UMAP-SMOTENC

2. Anonymization techniques

The definition of anonymized data introduced by the GDPR imposes two essential points of debate that we shall address before analyzing the anonymization techniques found in the literature.

Firstly, although the concept of information unrelated to an "identified or identifiable natural person" (Regulation [[72]2016/679, Recital 26] may seem clear, its practical implications are vague. While it is obvious that direct identifiers like names, email addresses, and social security numbers fall under the GDPR's definition of personal data, organizations used to assume that removing such variables would render the data free of personal information [85]. However, a study based on data from the 1990 U.S. Census [86] demonstrated that by using quasi-identifiers like zip codes, gender, and date of birth, it was possible to identify 87 % of the American population. The percentage decreased to 53 % and 18 % when broader location or county of residence were used instead of zip codes. Similarly, another study [79] showed how combining public patient clinical history (which includes quasi-identifiers) with a voter registry list allowed mapping back to the individuals, collecting confidential information in the process. Therefore, variables that indirectly identify an individual when combined with external or internal information should also be considered personal data. However, the line becomes even more blurred when scenarios like the ones presented by Malin and Sweeney [57] are considered, where individuals were successfully identified from a dataset primarily consisting of genomic information by using non-direct matching techniques against publicly available data.

The second component of the GDPR's definition ("rendered anonymous in such a manner that the data subject is not or no longer identifiable" (Regulation [[72]2016/679, Recital 26]) allows for data to be dissociated from personal identification. Nonetheless, it does not specify how to ensure true anonymity. We have established that removing direct identifiers is insufficient for anonymizing a dataset. Conversely, eliminating quasi-identifiers and other variables containing sensitive information (such as social network comments) would greatly diminish the value of most datasets, especially for machine learning applications. As suggested by Gulcher et al. [30], one possible approach is to apply data encryption techniques, where the original data is replaced with encoded values using a mapping function to protect confidential information while retaining the variables. However, under GDPR, this process would only achieve pseudonymization, which is still subject to regulation since the data "could be attributed to a natural person by the use of additional information" (Regulation [[72]2016/679, Recital 26]. This aspect sheds light on what constitutes a true anonymization mechanism. It must not only sever the linkage to the original person but also ensure irreversibility, even for the organization responsible for the process (e.g., through a mapping function).

Having clarified the concept of anonymization, we can now discuss the main techniques found in the literature. We will first examine conventional methods that involve data transformations to achieve specific privacy criteria, highlighting why they fail to provide a satisfactory solution. Then, we will explore generative techniques, which are also not free from pitfalls.

2.1. Transformative approaches

Even though it is possible to find older debates on data privacy protection in the literature, *k*-Anonymity, initially suggested by Samarati and Sweeney [79] and significantly amended by Sweeney [88], constituted a milestone in the research of anonymization techniques. While it does not directly constitute a method to anonymize datasets, it offers a criterion that transformative approaches should strive to achieve. According to this concept, a dataset can be subjected to growing privacy degrees, depending on the frequency of repeated

quasi-identifier patterns. The starting step to reach *k*-Anonymity involves identifying the quasi-identifiers and how frequently each combination, or key, occurs in the dataset. Then, through aggregation and suppression [78,87], a new dataset is obtained where each key appears at least a *k* number of times, meaning each entry becomes similar to *k*-1 others. By reducing each key's uniqueness, any linkage attempt will trace back to at least *k* entries. Despite *k*-Anonymity's popularity, several authors [20,55,90] expressed concerns about potential privacy gaps, particularly when this method is confronted with homogeneity and background knowledge attacks, in scenarios where a sensitive attribute is intended to be protected (e.g., disease of a patient. Homogeneity attacks occur when all elements of a key share the same sensitive attribute, thereby failing to protect the linkage between individuals and the sensitive attribute. Background knowledge attacks involve combining external information with a key's data to probabilistically disclose sensitive information about an individual.

To address some of the weaknesses of k-Anonymity, Machanavajjhala et al. [55] introduced l-Diversity. It adds the requirement for each group of keys to have at least *l* well-represented values per sensitive attribute, aiming to enhance robustness. However, l-Diversity has also been found to be susceptible to background knowledge attacks [58], and Domingo-Ferrer and Torra [20] and Li et al. [52] argue that it is not a necessary condition for achieving privacy. Additionally, two other types of attacks are used to evidence that this method offers insufficient protection of sensitive attributes. Although *l*-Diversity ensures protection against homogeneity attacks, it does not protect against similarity attacks, which can produce the same effect. These occur when different sensitive attributes offer the same semantic value (e.g., a key in which sensitive attributes are either stomach or colorectal cancer). Simultaneously, a key may have a skewed distribution of the sensitive attribute compared to the overall population, making any inference attempt about the individuals represented by that key significantly more conclusive than for the overall dataset.

More recently, *t*-Closeness [52] was proposed. To achieve this property, for each key present in the dataset, the distribution of the sensitive attributes should be at most *t* units of distance from the complete dataset distribution. Logically, by enforcing *t*-Closeness, the data becomes significantly less exposed to the attacks previously discussed, as no inference could be made for a restricted group that was not already present in the complete dataset. Nevertheless, this approach has other inherent issues. Firstly, while the paper suggests different methods to measure the value of *t* and consequently verify if a given t-Closeness criterion is met, it does not propose a process to enforce it, leaving options to conventional transformative methods that may involve high levels of suppression. Furthermore, even the authors admit that imposing *t*-Closeness is highly restrictive on data utility, as the correlation between keys and sensitive attributes is completely erased for low *t* values and severely impacted even for higher thresholds.

All methods presented this far suffer from additional common pain points. The parameter choice (*k, l,* and *t*) directly impacts the trade-off between anonymity and utility. Despite some existing research on optimizing their selection [19], the process is still domain and goal-specific, opening the door for human error in the anonymization process. On the same note, as identifying quasi-identifiers and sensitive variables is required, it is unclear how these methods would handle scenarios such as the ones described by Malin and Sweeney [57], where variables' assignment to those groups is not trivial, given that indirect matching can be leveraged against poorly anonymized public datasets. These techniques also fail when confronted with high-dimensional spaces [1], which have become increasingly frequent and where privacy is harder to preserve. Lastly, their applicability to numeric variables is non-trivial, requiring binarization and consequent loss of information.

2.2. Synthetic data generation

Mechanisms based on data transformations are vulnerable to diverse attacks while offering a prohibitive steep negative trade-off between data's utility and privacy. An inherently different approach is offered by generating synthetic datasets, which aim to preserve the original data properties while creating a disconnection between data points and real individuals. In this section, we will cover the research on synthetic data generation, significantly emphasizing some of the most prominent generative techniques: Copula Modelling, Variational Autoencoders, and Generative Adversarial Networks.

2.2.1. Early research and copula modelling

The research around the use of synthetic datasets was heralded by Little [53] and Rubin [75], with the first only focusing on the imputation of variables with a considerable risk of disclosure and the latter suggesting the theoretical ground for fully synthetic datasets. Since then, the field has seen interesting developments, as exemplified by Caiola and Reiter [13], Drechsler [23], and Reiter [73], where imputation models based on Random Forests [11], Support Vector Machines [10] and Classification and Regression Trees [12], have been introduced.

Copula Modelling represents a relevant family of data generation techniques, with several variants being addressed in the literature [43,67,84]. For the purpose of this paper, we will focus on the Gaussian Copula (GC) proposal [67] due to its high-quality implementation and consistent framework with the remaining techniques of interest [67]. GC synthesizes new data by relying on the original dataset's univariate probability distribution and the intra-variable relations of dependence. For the numeric variables, the aim is to find a good enough proxy (e.g., Gaussian distribution) for each variable's cumulative distribution function (CDF), given that computing the actual CDF may be prohibitively expensive, while for the categorical variables, the authors suggest the modelling of a Truncated Gaussian distribution based on each category's relative frequencies. Then, before calculating the covariance matrix, the multivariate Gaussian Copula is applied to transform all previously found distributions into the standard normal one, according to:

$$(\mu_0\ ,\ \mu_1,\ \ldots,\ \mu_n\ ) = \left[\Phi^{-1}\ (F_0(\mathrm{x}_0)),\ \Phi^{-1}\ (F_1\ (\mathrm{x}_1)),\ \ldots,\ \Phi^{-1}\ (F_n\ (\mathrm{x}_n))\right] \tag{1}$$

where for a dataset with n features, $(\mathrm{x}_0,\ \mathrm{x}_1,\ \ldots,\ \mathrm{x}_n)$ represents a row from the original dataset, $(F_0,\ F_1,\ \ldots,\ F_n)$ the CDFs found, and $\Phi^{-1}(F_i\ (\mathrm{x}_i))$ the application of the inverse Gaussian distribution's CDF. The sampling process can now be conducted with the covariance matrix and the variables' data distributions, whose combination can conceptually be understood as a highly compressed representation of the original multivariate data distribution.

Despite its efficiency and simplicity, the accuracy of the results obtained by this method relies heavily on how closely these CDFs' proxies match the original data distribution. Therefore, this method faces substantial challenges when dealing with complex distributions commonly found in the tabular domain. In the following sub-section, we will delve into a family of techniques specifically designed to address such scenarios.

2.2.2. Deep generative modelling

The success attained by GANs and VAEs in generating artificial data in the computer vision field has made a compelling case for their adaptation to tabular datasets [21]. Nonetheless, as argued initially by Xu et al. [98] and then by Zhao et al. [102], the synthesis of tabular data is far more complex than its counterpart in the image domain, offering challenges such as non-gaussian, multimodal, and highly skewed distributions.

Variational Autoencoders, introduced by Kingma and Welling [44], combine two convolutional neural networks to create an

encoder-decoder architecture, similar to what was originally proposed for deterministic Autoencoders (AE) [76] but introducing a probabilistic behaviour on the latent space, enabling the generation of new data points. In a deterministic AE, an encoder network maps the input vector, x, to a low-dimensional latent space's representation, z. Then, a decoder reconstructs the original input from the compressed space, obtaining x'. Relying on the backpropagation of the error, measured as the difference between x and x', which is known as the reconstruction loss, the AE learns how to effectively map $x \rightarrow z \rightarrow x'$.

In contrast, VAE's working principle is to learn the latent variables' distribution instead of a single value per latent space's feature. While this distribution is typically assumed to be Gaussian, more complex distributions can also be used [45], including distributions capable of properly handling categorical features [39]. Regardless of the distribution choice, learning the distribution makes it possible to sample according to it in the latent space. This leads to the synthesis of new vectors that the decoder returns to the original input space. To accommodate this concept, a new way to measure the network's error was proposed, as relying solely on the reconstruction loss could potentially lead to scenarios where the VAE reverts to the behaviour of an AE. Therefore, in addition to the reconstruction loss, a regularization factor was added to the loss function, which is calculated as the Kullback-Leibler divergence [42] between the learned latent variable distribution and the predefined prior distribution, aiming to enforce an approximation between the two.

Generative Adversarial Networks [27] also rely on a two neural networks' structure, but where both players, generator (G) and discriminator (D), perform competitively. The former aims to mislead the latter by generating synthetic data that closely resembles the original one. Conversely, the latter aims to distinguish between the real and synthesized data entries. To this, the generator starts with an input sampled from the noise space that typically follows a simple Gaussian distribution and maps it into the original data space. The discriminator is then presented with entries that can be fed by the generator or sampled from the original dataset. It returns a value in the range [0, 1] according to the estimated probability of its input belonging to the real data distribution. The generic value function of these two min-max players can be defined as:

$$\begin{aligned} \min_G \max_D V(D, G) &= \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})] \\ &\quad + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \end{aligned} \tag{2}$$

where x is sampled from the $p_{data}(\mathbf{x})$ and z from the $p_z(\mathbf{z})$, with $p_{data}(\mathbf{x})$ and $p_z(\mathbf{z})$ representing the original and noise distributions, respectively. This vanilla version faces severe limitations related to convergence and mode collapse, as discussed by Kodali et al. [46]. Convergence issues arise due to vanishing gradients during early learning stages, while mode collapse occurs when the generator and discriminator interact in a way that leads to the generator focusing excessively on specific regions of the space where the discriminator performs poorly. In an effort to address these challenges, Wasserstein GAN (WGAN) was introduced by Arjovsky et al. [4]. The main contribution of WGAN was replacing the value function used in the minimization of the Jensen–Shannon divergence (derived from Eq. (2), under an optimal discriminator) with the minimization of the Wasserstein distance. This change results in smoother gradients, even when dealing with significantly different distributions, thereby alleviating convergence issues. Furthermore, by potentiating a better training process, allowing the discriminator to achieve optimal performance, it becomes robust against mode collapse scenarios, forcing the generator to keep searching the space. The new value function of the min-max game can be represented as:

$$\begin{aligned} \min_G \max_{D\in\mathscr{D}} V(D, G) &= \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})} D(\mathbf{x}) \\ &\quad - \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})} D(G(\mathbf{z})) \end{aligned} \tag{3}$$

where $D \in \mathscr{D}$, constrains the critic to be a 1-Lipschitz function. Note that the notation was slightly adapted from the one presented by Arjovsky et al. [4] and Gulrajani et al. [31], to be consistent with the one employed in Eq. (2). Building on the WGAN architecture, the WGAN-GP (where GP stands for Gradient Penalty) [31] was introduced as an improved way to enforce the 1-Lipschitz's constraint, increasing the robustness of the architecture further.

In the literature, numerous approaches can be found attempting to transpose these deep generative models from unstructured data to tabular datasets. TableGAN [66] closely mimics the structure of a vanilla GAN, maintaining a convolutional architecture and requiring each entry to be reshaped to a square matrix (similar to the image format). TGAN [97] uses attention mechanisms [92] to sequentially model columns, relying on existing correlations between variables. GReaT [9], REaL-TabFormer [82], TabuLa [100], and TabMT [29] rely on the use of transformer architecture, from language models, in the generative process. MedGAN [17] was developed with a particular focus on the healthcare field, being designed to handle a high diversity of distributions between different columns. It combines an Autoencoder with a GAN, where the generator works on the Autoencoder's learned latent space, having its output inputted to the decoder before being presented to the discriminator. RTVAE [2] applies a VAE in the context of tabular data but uses the learned distribution for anomaly detection instead of data generation. TabDDPM [48] uses a different family of deep generative models based on stable diffusion. CTAB-GAN [102] proposes a new methodology to encode mixed datatypes, while its revised version, CTAB-GAN+ [101], improves upon the previously used loss function and training stability. Similarly to those, CTGAN and TVAE [98] rely on a fully connected layer-based architecture, while also introducing several specially tailored features to handle tabular data, aiming to make the generation process more effective. In the remainder of this section, we will delve deeper into these two models, not only because both have outperformed several other generative architectures, but also because, in alignment with the main objective of this paper, these methods offer the most complete and practical open-source implementation available, being widely accepted as the standard for practitioners.

CTGAN mixes a strong pre-processing component intended to solve the tabular domain's specific issues, with an architecture adapted from the WGAN-GP. It relies on the concept of mode-specific normalization to deal with non-gaussian and multimodal distributions, decomposing each continuous feature into a matrix constituted by one-hot vectors and a vector of scalars. The first step in this process is using the variational Gaussian mixture model [7] to infer the number of modes of the continuous column's distribution. From this, every column's value probability of belonging to each mode is calculated. Then, a mode is sampled for each value according to the previously calculated probability densities, and the result is one-hot encoded. Finally, the original value is standardized using the mode's mean and standard deviation. To address skewed categorical features, which may lead to improper learning of less-represented categories, CTGAN extends the constraint of Conditional Generative Adversarial Networks (CGAN) [62] to ensure even sampling of all categories.

The TVAE also employs mode-specific standardization for continuous features and one-hot encoding for categorical features. Each row in TVAE consists of concatenated one-hot vectors for categorical features and a combination of one-hot vectors and scalars for continuous features. In the latent space, scalars are assumed to follow a Gaussian distribution, while one-hot vectors are modelled using a categorical probability mass function.

Both CTGAN and TVAE were evaluated against various generative algorithms. Although they show a significant improvement in the utility of generated data for machine learning tasks, there remains a noticeable gap compared to the utility provided by the original data [98].

2.3. Measuring privacy and utility

In Section 2.1, we discussed how several concepts that are frequently considered privacy quantifiers fail to provide enough guarantees when

applied to transformative anonymization methods. The inevitable steep negative correlation between these privacy definitions and the datasets' utility offers a major reason for concern. As Domingo-Ferrer and Torra [20] pointed out, resourcing to high thresholds, as enforcing *t*-Closeness, may strip the dataset's utility completely. Conversely, looser definitions significantly open the door for disclosing sensitive information. Extending the criticisms further, these privacy concepts seem disconnected not only from the constant advances introduced in the literature but also from the challenging nature of real-world datasets [1]. For once, *k*-Anonymity and related improvements cannot be meaningfully applied to fully generated data. For example, suppose that some unique keys in the synthetic dataset do not represent any entry in the real dataset. While conceptually, this should not constitute a privacy risk as the unique quasi-identifiers cannot be linked to any real individual, the synthetic dataset does not respect *k*-Anonymity for $k \neq 1$.

Additionally, while useful for illustrating simple privacy attacks, the conventional segregation of data into quasi-identifiers and sensitive attributes may not address more complex scenarios. In a study presented by Narayanan and Shmatikov [63], an adversarial attack was proposed against a publicly available dataset of movie reviews. The dataset had undergone the removal of personal identifiers and only included a sample of ratings and corresponding dates per user, with the addition of noise. Surprisingly, the authors demonstrated that de-anonymization was possible even without relying on variables typically considered quasi-identifiers. While the dataset was not subjected to *k*-Anonymity or similar alternatives, it is defended that if any of those privacy definitions had been applied, the data's utility would have been completely removed before offering a guarantee of privacy that was enough to surpass the attacker model's ability to handle imperfect information.

The use of adversarial attacks to identify potential privacy gaps has led to the identification of two families of attacks that are suitable for synthetic data: attribute-based attacks [3,47] and membership inference attacks [37,80]. In attribute-based attacks, an attacker tries to disclose sensitive attributes of actual data based on patterns extracted from the synthetic dataset. However, in real-world scenarios, the most sensitive attribute is usually the "target" variable, making this privacy quantification method difficult to balance with the concept of machine learning utility preservation. Due to this, attribute-based attacks are progressively being disregarded as a potential indicator of privacy violation [40]. On the other hand, membership-based attacks aim to infer if a particular individual is present in the original population used to construct the synthetic dataset. There are diverse types of attacks that fall under this category, with some requiring knowledge about the anonymization method used to construct the synthetic dataset and access to a subset of the original dataset [83], while others requiring little to no knowledge about the synthetic data generation process, as methods based on the distance between the synthetic and the real entries. For a more practical outlook on the concepts involving adversarial attacks, the reader is directed to the TAPAS toolkit [36], which offers an interesting open-source implementation of several attacks discussed here.

As just mentioned, the proximity between synthetic points and the original entries can be thought of under the lenses of a membership inference attack. Nevertheless, the literature often presents it as a heuristic, being one of the most commonly used methods for privacy quantification in synthetic datasets [66,93,102]. The intuition for this is relatively straightforward: the use of synthetic data generation methods should break the linkage between the data and real individuals, and therefore, re-identification should not be possible. Nevertheless, this only holds if the synthetic data generation process is capable of preserving the original data patterns at the aggregate level, while ensuring that at the individual level, every synthetic entry is as dissimilar as possible to the real ones. Therefore, measuring the Euclidean distance between each synthetic data point and its closest neighbours in the original dataset, as illustrated in Fig. 1, tends to be a simple and effective practice to infer privacy preservation. Zhao et al. [102] further enhance

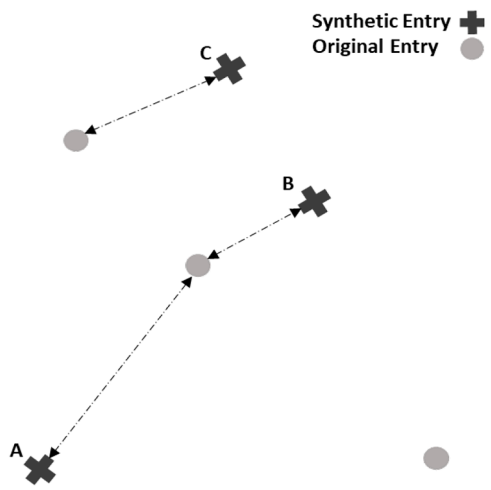


Fig. 1. Example of privacy quantification based on the nearest real neighbor: according to which point C is more exposed than point A.

this method by considering the distance to the two closest neighbours for each generated point. They then calculate the ratio of the distance to the closest neighbour over the distance to the second closest neighbour, resulting in a value between 0 and 1. Values closer to 1 indicate that synthetic entries are located in denser regions of the original data space, making linking them back to real individuals more challenging. This is demonstrated in Fig. 2, where synthetic point B is equidistant to two points from the original dataset. On the other hand, when the ratio is closer to 0, as shown with synthetic point A, a malicious attacker may attempt to establish a connection between the synthetic point and its closest neighbour in the original dataset. However, it is important to note that due to the synthetic nature of the data, such extrapolation may lead the attacker to incorrect inferences.

The last concept of privacy that will be addressed is Differential Privacy (DP) [24]. Proposed as a strong and mathematically quantifiable definition, it relies on noise addiction and was originally developed to face disclosure risk when querying a statistical database. The underlying idea can be intuitively understood as offering a guarantee to each individual that their presence or absence in a database should not lead to significantly different results. Despite DP not guaranteeing the assurance of absolute privacy, it presents a rigorous way to assess how high the impact of a single entry in a dataset can be, with this value represented as ε. Recently, some generative methods have been proposed that offer compliance with the DP concept, such as the DPGANs [96] and the PATEGANs [41]. While for ε values above 1, data generated by these

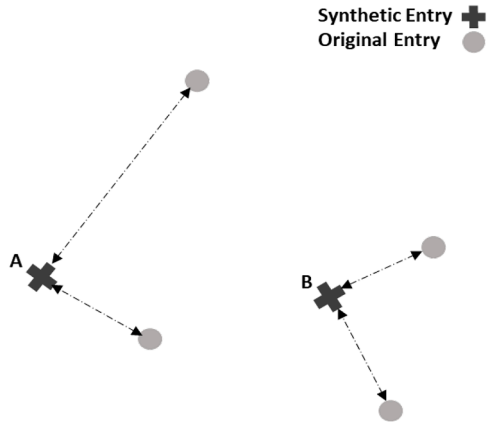


Fig. 2. Example of privacy quantification based on the two nearest real neighbors' ratio: according to which point A is more exposed than point B.

GANs' architecture tends to converge to the quality achieved by non-differential private GANs, enforcing stricter privacy levels leads to severe utility degradation. Building on this and considering that the levels of data utility, even when generated by SOTA methods, lag considerably behind the ones achieved with the original data, it is unclear if a new equilibrium point that further prioritizes privacy at the cost of utility is desirable. Finally, DP has recently been criticized for introducing bias in datasets, creating a significantly less accurate depiction of certain data groups with lower representation [70,89]. In real-world applications, this can be particularly dangerous, allowing for unfair data-based policies from companies toward less-represented subsets of the population.

Anonymized data's utility can be measured from different perspectives depending on the use case in which it is intended. In the context of this paper, the focus is on evaluating the quality of anonymized data for machine learning tasks. This evaluation is often conducted by training predictive models using the generated data and assessing their performance on an unseen portion of the original dataset, as done by Xu et al. [98], Kamthe et al. [43], and Zhao et al. [102]. This strategy is closely aligned with the industry's needs. For example, a credit institution may have anonymized historical client data that it desires to use to make informed decisions about potential new clients' propensity to default on loans. While the machine learning system will be built using synthetic data as input, the goal is for the model to extrapolate well to real individuals.

3. Proposed method

Despite the major improvements offered by state-of-the-art generative models over conventional approaches, the gap between the utility of the anonymized and original data is still an unsolved problem, particularly for machine learning tasks. In this section, we will present our suggested approach, which combines UMAP [60] and an extension of the SMOTE framework [15], that can handle mixed datatypes, the SMOTE-NC (where NC stands for Nominal and Continuous).

3.1. Synthetic minority oversampling techniques

Inspired by data augmentation techniques used in computer vision tasks [32], SMOTE was proposed aiming to solve imbalanced learning scenarios (which occur when, in a classification task, the target class presents a significantly skewed distribution) in tabular datasets, by oversampling the minority classes, until all classes have the cardinality of the majority one. This technique, when applied to a binary dataset, can be decomposed in the following steps:

1. For each point to be synthesized, sample an entry, n_i, from the minority class.
2. Using the Euclidean distance, obtain the k nearest neighbors of n_i, that belong to the minority class. Any k value in the range [1, $n_{minority}$-1] is valid, corresponding $n_{minority}$ to the number of entries from the minority class.
3. Randomly select one of the k neighbors, k_i, and obtain the synthetic point, s_i, following:

$$s_i = n_i + (n_i - k_i) \times u_i \tag{4}$$

where u_i represents a vector with the same shape as n_i and k_i, populated with randomly selected values between 0 and 1. Following this equation assures that the synthetic point belongs to the line segment connecting n_i and k_i, as represented in Fig. 3. Expanding this framework for multi-class datasets relies on repeating the presented process for all the desired minority classes.

The convenience and effectiveness achieved by SMOTE led to strong research interest, with the development of several enhancements. As a non-exhaustive review, some of the proposed methods were: Borderline SMOTE [33], which focuses on oversampling points near the data decision boundary, intending to expand the minority classes in problematic frontiers; ADASYN (Adaptative Synthetic Sampling Method) [34] adaptatively synthesizes points according to the space density of the minority class, generating more entries near the instances that are harder to learn; K-Means SMOTE [50] applies the K-Means algorithm to the full dataset and generates minority points only in clusters where the minority class is the most represented one, aiming to lower the introduction of noise in the oversampling process; Geometric SMOTE [22] generalizes the original SMOTE by allowing the synthesis of points in a hyper-sphere/spheroid instead of solely on the line segment linking two points.

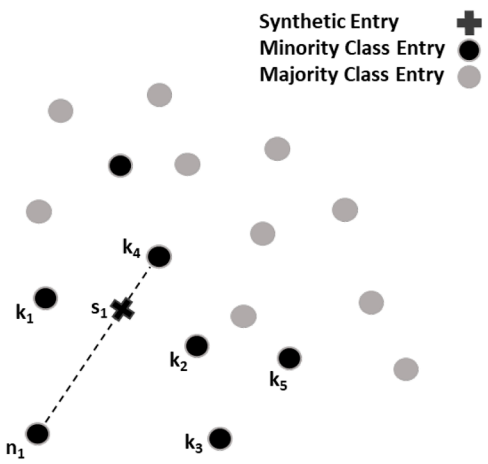


Fig. 3. Illustration of the SMOTE generative process: in this case between n_1 and its 4th nearest neighbor.

Although the aforementioned methods are only able to handle numeric variables, in the original SMOTE paper [15], the authors proposed SMOTE-NC as an extension to the original framework, capable of facing datasets with mixed datatypes. Firstly, the median standard deviation across all numeric features from minority class instances is computed. Then, when finding the k nearest neighbors to n_i, for every candidate being evaluated, the computed median value will be included as a parameter for the Euclidean distance for each categorical feature that does not match the one present in n_i. This can be considered an attempt to penalize different categories in a way comparable with the distance calculation for numeric features. Lastly, while the numeric features of the synthetic point will be filled following Eq. (4), each categorical feature will be populated with the mode category across all the k neighbours.

3.2. Uniform manifold approximation and projection

UMAP is a recently introduced method for dimensionality reduction, which, similarly to other dimensionality reduction techniques, aims to uncover a low dimensional latent space where it is possible to effectively perform data analysis by visual inspection and be free from the implications of the curse of dimensionality [6] while preserving a large portion of the patterns present in the original space. Unlike traditional approaches, such as Principal Component Analysis (PCA) [56], where each component results from a linear combination of the original variables, UMAP constructs a non-linear low-dimensional representation of the input space based on the concept of graph neighborhood. Furthermore, in comparison with the state-of-the-art dimensionality reduction technique for visualization, t-distributed Stochastic Neighbor Embedding (t-SNE) [91], UMAP provides a framework that is significantly more scalable and better at preserving the global topology of the input

space, while being as effective in preserving the local one [5,60].

The UMAP algorithm can be decomposed into two distinct steps: the creation of a graph in the high-dimensional space and its effective projection to a low-dimensional one. For the mathematical foundation and assumption required for the UMAP application, the reader is referred to McInnes et al. [60].

A radius is defined for each point to construct the graph in high-dimensional space. This radius equals the distance between a point and its k^{th} most distant neighbor, where k is a hyperparameter that influences the graph's structure. A larger value of k leads to more global structures, while a smaller value results in more compact structures. If the radii of two points overlap, a potential connection is formed, and a weighted edge is created between them. The weight of the edge represents the likelihood of the points being connected, which is inversely related to the distance between them, adjusted by the distances to their closest neighbors. As a result, every pair of connected points will be linked by two directed edges that must be combined to form a single entity, whose weight will represent the probability that at least one of the links actually exists. A complete weighted graph is achieved once all points in the original space have undergone this process. In topological data analysis terms, it constitutes a fuzzy simplicial complex, approximating the assumed locally connected and uniformly distributed original manifold.

The second part of the process involves the progressive approximation of the obtained graph to a low-dimensional representation. This is done using a force-directed graph layout algorithm. In this algorithm, attractive forces are applied to the edges, pulling connected vertices closer together, while repulsive forces are applied between a sample of vertices, pushing them apart. Initially, a graph is initialized in the low-dimensional space, which can be random or follow a more sophisticated initialization method. Then, an interactive process takes place to approximate the two topological representations, by optimizing the cross-entropy between the weighted edges of the original graph and the low-dimensional graph.

Despite being originally proposed as an unsupervised dimensionality reduction tool, the official algorithm's implementation [61] has since then evolved to allow for two important properties: the use of the framework in a supervised manner and the projection of points from the latent space to the original one, according to the originally learned mapping between high dimensional and latent space. For supervised use, the working principle involves the construction of an additional graph based on the class of belonging, which is used to prune the connections of the original one. Regarding the projection of points to the original input space, it relies on the construction of an adjacency matrix to identify the neighbourhood relations between the original points projected to the latent space and the new points in the latent space that are intended to be projected to the original input space. Then, through a process similar to the original progressive approximation of the low-dimensional and high-dimensional graph, an optimization process is performed aiming to preserve: (i) the relationship between the original points in the high-dimensional space and their position on the latent space; (ii) the relative position between the original points in the latent space and the new points to be projected to the original input space. This reprojection component was empirically validated, and concluded to work well when the new points are positioned inside the convex hull of the latent space. Otherwise, an alternative can be presented using Parametric UMAP [77], which through an autoencoder, learns a parametric mapping between the original input space and the latent space, and its reverse.

3.3. UMAP-SMOTENC: anonymization using UMAP and SMOTE-NC

In the existing literature, Walia et al. [93], Endres et al. [25], and Kotelnikov et al. [48] have employed the SMOTE framework as a means of generating synthetic data within the context of anonymization. While Endres et al. [25] present a minor modification to the SMOTE algorithm aimed at enhancing the generative process, it does not specifically address the associated privacy concerns. On the other hand, both Walia et al. [93] and Kotelnikov et al. [48] adopt plain SMOTE as a straightforward baseline, yielding unsatisfactory outcomes in terms of privacy protection. These outcomes are quantified by evaluating the Euclidean distance between each synthetic data point and its closest neighbor in the original dataset, as discussed in Section 2.3. The observed results are in line with expectations and have also been acknowledged by Douzas and Bacao [22], where the tendency of plain SMOTE to overfit the original dataset is briefly mentioned. This overfitting arises due to the limited diversity introduced when synthesizing data along the line segment connecting two given points. Moreover, another notable critique of the SMOTE framework, specifically its inability to effectively handle high-dimensional data, is examined by Blagus and Lusa [8].

To address both these challenges, this paper introduces a novel approach called UMAP-SMOTENC, which combines the UMAP technique with the SMOTE framework. Prior research has demonstrated that incorporating dimensionality reduction techniques before applying SMOTE can effectively mitigate the curse of dimensionality [64,94]. This behavior is expected to be further enhanced by leveraging the intriguing topological properties offered by UMAP, which align closely with the preservation of local structure required by SMOTE. By allowing oversampling to occur in a non-linear compressed representation of the input space, the potential for overfitting the original dataset is significantly reduced. While in the two-dimensional space, a synthetic entry will still be generated on the line segment connecting two given points, in the original input space, this line segment is potentially represented by a significantly more complex structure, allowing for higher diversity in the generative process. Moreover, both the transformation from the original space to the two-dimensional representation and its reverse involve stochastic processes, thereby introducing a larger disconnection between the synthetic and the real datasets. Furthermore, UMAP-SMOTENC leverages the supervised UMAP framework to improve the preservation of decision boundaries between different classes.

The proposed methodology, depicted in Fig. 4, involves a sequential application of UMAP and SMOTE-NC, followed by the projection of synthetic data points back to the original feature space. The process begins by projecting the numerical features onto a two-dimensional space using the supervised UMAP algorithm. If the original dataset includes categorical features, they are appended to the two-dimensional representation of x_c. A new dataset will be generated using the SMOTE-NC method (which reverts to the plain SMOTE behaviour if no categorical features are present). Contrary to the conventional use case, instead of synthesizing for each non-majority class $n_m - n_i$ instances (where n_m represents the number of instances from the majority class and n_i the number of instances from a given minority class), according to the proposed method, for each class, SMOTE-NC generates as many points as each class' cardinality, resulting in a fully synthetic dataset. Using the UMAP technique, the synthetic numeric features are then projected back to the original input space and the synthetic categorical variables are concatenated to them. Finally, if any originally discrete feature exists, it is rounded to the nearest unit to ensure consistency, as the proposed method cannot guarantee the preservation of this data type. Note that using UMAP to reproject features to the original input space can only be safely performed, as mentioned in Section 3.2., because SMOTE's working principle ensures that all synthetic points fall inside the original latent space convex hull.

4. Methodology

This section describes the steps followed to evaluate the suggested approach. Particularly, we present the datasets, algorithms, and metrics used in the experiment, as well as the steps followed to ensure their validity.

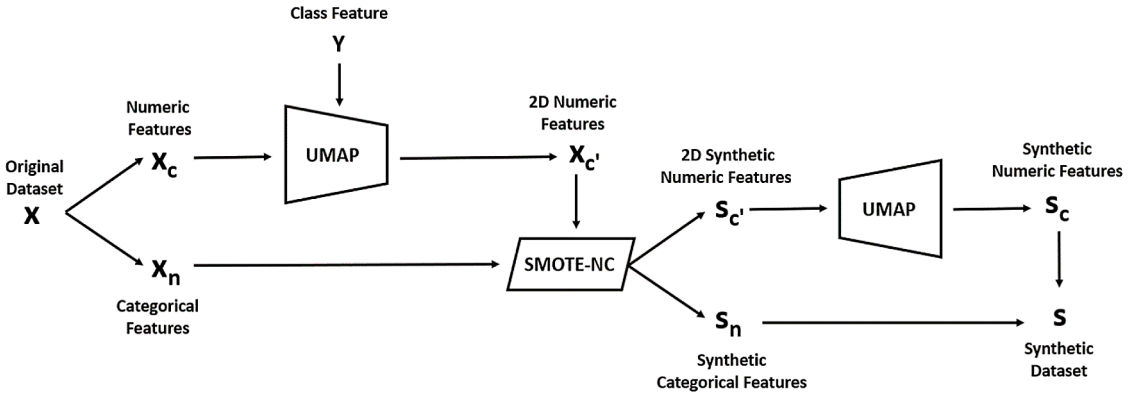


Fig. 4. Schematization of the UMAP-SMOTENC process.

4.1. Datasets

We recurred to eight different classification datasets to generate synthetic data, which reference can be found in the appendix section of this document. The selection intention was to show the robustness of the method proposed to diverse degrees of data complexity, structure, and cardinality at the row, feature, and class levels. From the UCI machine learning repository, we selected the Adult, Dry Beans, Magic Gamma, Optical Recognition of Handwritten Digits, Occupancy and Cover Type datasets, and from Kaggle, the Credit Card and Loan datasets. Some high-level statistics can be found in Table 1.

Adult and Loan represent typical datasets containing individual personal information, with both categorical and numeric features. Credit Card and Occupancy offer less conventional anonymization problems where the data refers to individuals' activity and not directly to their characteristics. Digits Recognition is a multiclass dataset composed exclusively of discrete features, intended to test how effective the proposed method is when dealing with this type of feature, given the above-mentioned incapacity to preserve this datatype during the process. Finally, Dry Beans, Magic Gamma, and Cover Type were selected to enrich the diversity of the experiment further.

4.2. Evaluation metrics

The experiment evaluation closely followed the key concepts discussed in Section 2.3. To quantify the level of privacy provided, we decided to use distance-based methods, partially due to their frequent use in the literature and partially because they offer one of the most agnostic privacy quantification methods available, not assuming any pre-existent knowledge from the attacker about the anonymization process or the real dataset (except for the information about the target individuals). In this regard, we decided to use the Euclidean distance between each synthetic data point and its nearest neighbour in the original dataset as the main privacy indicator, following standard practice found in the literature [66,93,102]. Additionally, inspired by the proposal of Zhao et al. [102], we calculated the ratio between the distances to the two nearest neighbours and the ratio between the nearest neighbor and the tenth neighbour of each synthetic point. This approach aimed to better assess the isolation of synthetic points within the original feature space. Mean values were computed for each privacy metric across the datasets.

To evaluate the utility of the generated synthetic data for machine learning tasks, the "train on synthetic and evaluate on real data" methodology was employed. Considering the target data distribution, the F1 score was chosen as the evaluation metric. The F1 score combines precision (the ratio between true positives (TP) and the sum of TP and false positives (FP)) and recall (the ratio between TP and the sum of TP and false negatives (FN)) using a harmonic mean. This metric provides equal weighting to both FP and FN cases, which is particularly relevant in imbalanced learning scenarios. The macro-averaged F1 score was used for multiclass datasets, representing the unweighted mean of the F1 scores achieved for each class.

Recognizing the importance of balancing privacy and utility without compromising either, the aforementioned metrics were combined to form a single indicator, aiming to quantify this trade-off. The mean F1 score across all classifiers used was chosen as the overall utility metric, while the mean of the three privacy scores discussed served as the overall privacy metric. The privacy scores were individually scaled to the range [0,1] to ensure consistency, where 0 represents the absence of privacy protection provided by the original training set, and 1 represents the highest privacy score obtained across the diverse synthetic training sets. Similarly, the utility score was also scaled to the same range, with 0 representing the utility obtained by training on the lowest-performing synthetic training set and 1 by training on the real training set. Balancing the aggregated privacy and utility scores using the harmonic mean gave both aspects equal importance, resulting in a final trade-off score. This trade-off score serves as the primary performance indicator for the anonymization mechanism:

$$\textit{Trade Off Score} = 2 \times \frac{(\mathrm{OP} \times \mathrm{OU})}{(\mathrm{OP} + \mathrm{OU})} \tag{5}$$

where OP refers to the overall privacy score and the OU to the overall utility score.

Table 1
Datasets description after an initial brief pre-processing step.

Datasets	# Rows	# Features	# Numeric Features	# Categorical Features	# Classes	Largest Imbalance Ratio
Adult	48,842	14	5	9	2	3.2
Loan	5000	12	6	6	2	9.4
Cover Type	25,000	12	10	2	7	103.3
Credit Card	284,807	30	30	0	2	577.9
Occupancy	20,560	5	5	0	2	3.3
Dry Beans	13,611	16	16	0	7	6.8
Magic Gamma	19,020	10	10	0	2	1.8
Digit Recognition	20,000	16	16	0	26	1.1

4.3. Implementation and algorithms

All the code used in this experiment was written in the Python programming language and can be consulted on the GitHub associated with this project. To construct the experiment, we recurred to several open-source libraries, namely Umap-Learn [61], Skicit-Learn [68], XGBoost [16], Synthetic Data Vault [67], and Imbalanced-Learn [51].

The proposed method was tested in comparison with four different algorithms: Plain SMOTE-NC was used as a baseline that offers limited privacy guarantees; GC as an intuitive and easy-to-use non-deep generative approach; and both TVAE and CTGAN were selected as the SOTA methods for anonymization. In alignment with the central aim of the current paper (the introduction of a method that can be leveraged by researchers and practitioners alike), all the benchmarking algorithms used were selected not only based on their academic relevance but also based on their impact on industry settings, through high-quality and well-maintained open-source algorithm implementations. Particularly, SMOTE-NC is supported by the Imbalanced-Learn project [51], and the GC, TVAE, and CTGAN are supported by the Synthetic Data Vault project [67].

All algorithms followed the official implementation and were trained with the suggested hyperparameters. The same is valid for the proposed algorithm (both for the UMAP and SMOTE-NC components). Note that for the deep generative models, this implies relying on the best hyperparameters found according to the original paper, whose evaluation partially overlaps with the datasets and metrics used in the current paper.

For the machine learning task, and considering we are interested in showing its usability in a model-agnostic way, we selected several different classifiers, where a standard parameter tuning phase was applied: Logistic Regression (LR) [59], Naïve Bayes (NB), Classification Decision Tree (DT) [12], Random Forest (RF) [11], Extreme Gradient Boosting (XGB) [16] and Multi-Layer Perceptron (MLP) [74]. This selection encompasses a range of commonly used techniques in both academic and industrial settings.

From all the algorithms discussed in this section, only the deep generative models (CTGAN and TVAE) required non-standard computational resources. Despite the existence of the option to run on a standard machine, in practice, its application demands access to Graphical Processing Units (GPUs).

4.4. Experimental procedure

The experimental procedure begins with a brief manual preprocessing step for the datasets, which involves identifying categorical and numeric features and removing direct identifiers. The original dataset is then divided into five non-overlapping chunks using 5-fold stratified cross-validation, with each portion maintaining the target's data distribution of the overall dataset. Successively, each chunk serves as the test set, with the remaining four serving as the train set. Five different synthetic train datasets are generated from each original train dataset, using the generative models under study. Unlike CTGAN, TVAE, and GC implementations that handle feature standardization internally, SMOTE-NC and UMAP-SMOTENC require a scaling step before the generative phase. This step aims to normalize the numeric features to a common range of [0, 1], being this procedure reverted afterwards.

Having the previously obtained synthetic datasets, for the Machine Learning evaluation, each combination of classifier and set of hyperparameters is, sequentially, fitted to every train dataset and evaluated in the test dataset. Previous to this step, the categorical features need to be one-hot encoded, and the numeric ones preferentially scaled, depending on the classifier's selection. Lastly, as the privacy quantification relies on the Euclidean distance, all datasets' numeric features are scaled to the same range, using the original train set as a reference, whose features' values are compressed to the range [0, 1]. Using a common referential ensures that equal entries from different datasets are represented by the same values, allowing for the correct risk assessment.

Additionally, we computed the privacy risk for the test set. To exemplify the reasoning behind this addiction, consider a dataset containing individuals' data that was split into a non-overlapping train and test set. Under this scenario, and given that no individual can be simultaneously on both sets, the test set represents precisely what is desirable to be achieved with generative anonymization tools: a group of "individuals", that do not correspond to a real person (at least in the sense of not being on an organization's database, in this case, represented by the train set), but that present the same overall characteristics of the real individuals. While acknowledging that future research may be needed to validate this threshold (quantified by the privacy in the test set) sufficiency, we will recur to it as the floor privacy assurance we expect to see enforced by any anonymization mechanism.

The process discussed above, and schematized in Fig. 5, is repeated three times per original dataset, using different initialization seeds. The final results represent the average performance achieved over the 15 different evaluation runs.

5. Results and discussion

In this section, the experimental results are reported and discussed. Firstly, the performance of the different anonymization mechanisms over the trade-off metric addressed in Section 4.2 is presented. Then, using statistical tests, we confirm that the proposed method offers a new equilibrium in this trade-off between utility and privacy: offering adequate privacy assurances, contrary to the plain SMOTE-NC while generating more useful data for classification tasks than the current SOTA techniques. Lastly, in Section 5.2, we discuss the main inferences that can be made from the results achieved, being this discussion complemented, in Section 5.3, with a reflection on potential future directions to be followed.

5.1. Results

Following the proposed methodology, the performance of the different generators was evaluated. For the most relevant metric, the trade-off score between privacy and utility, an analysis both at the rank and absolute level was conducted, with the results being displayed in Tables 2 and 3, respectively. Additionally, to better understand the equilibrium point achieved in the trade-off score, we also present the individual metrics that were used for the utility and privacy quantification in Tables 4 and 6, respectively. All the results were validated with the relevant statistical tests.

As abovementioned, Table 2 reports the mean rank achieved across the three repetitions of the five stratified folds for all datasets and evaluation procedures. The ranks were obtained by assigning a relative score for each anonymizer's performance (with 1 representing the best and 5 the worst) for each combination under analysis. On the found ranks, we applied the Friedman test [26], with the null hypothesis being the non-existence of a statistically significant difference across the performance levels achieved by the different anonymization tools. As the null hypothesis is rejected, at the significance level of $\alpha = 0.05$, it is suggested that a statistically significant difference in the utility and privacy trade-off is offered by the different mechanisms.

In addition to the Friedman test, we employed the Wilcoxon signed-rank test [95], adjusting the p-values for multiple pairwise comparisons using the Holm method [35]. In this sense, Table 3 follows the same logic as Table 2, but presents the mean absolute trade-off scores across datasets, and the corresponding relevant statistical validation. As displayed in the table, by rejecting the null hypothesis, that there is no statistically significant difference between the trade-off offered by the UMAP-SMOTENC and the remaining generative methods, at the significance level of $\alpha = 0.05$, it is shown that the proposed method introduces a new, and superior equilibrium point in this trade-off.

Regarding the machine learning utility, in Table 4, we summarize the

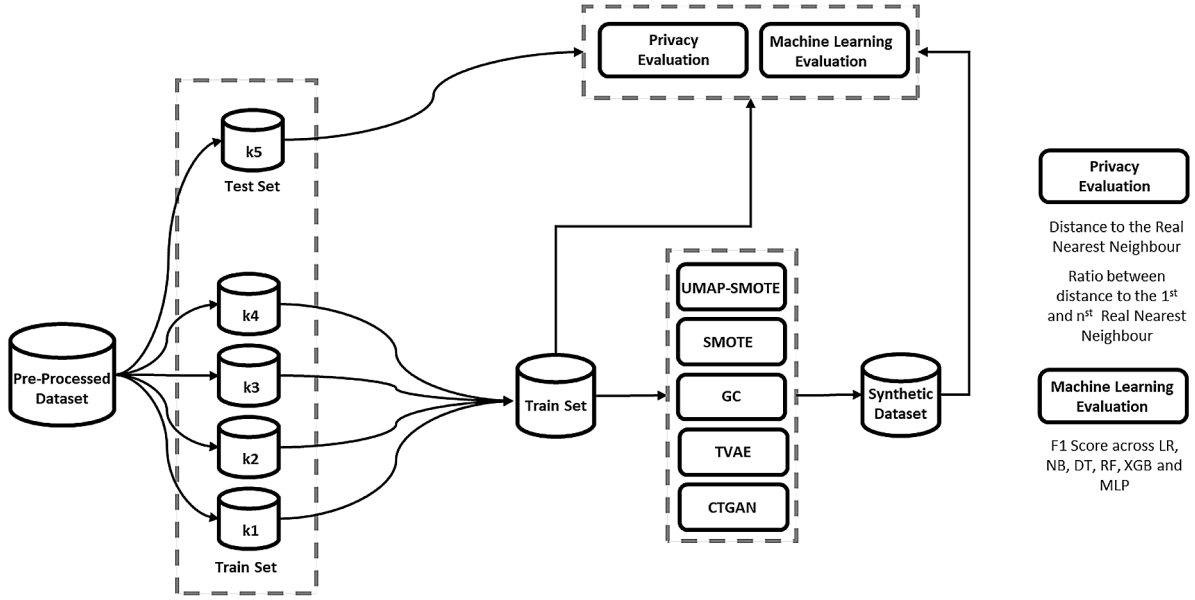


Fig. 5. Schematization of the experimental procedure followed.

Table 2
Mean Rank Scores for the privacy and utility trade-off score and results of the corresponding Friedman test.

Generator	Trade Off Score
GC	4.45±0.46
CTGAN	2.69±0.60
TVAE	2.76±0.54
SMOTE-NC	3.86±0.44
UMAP-SMOTENC	**1.23±0.24**
p-Value / Significance	4.6e-04 / **True**

Table 3
Mean Absolute for the privacy and utility trade-off score and results from the Wilcoxon signed-rank test with Holm's adjusted p-values for the pairwise comparation with UMAP-SMOTENC.

Generator	Trade Off Score	p-Value / Significance
GC	0.24±0.11	3.1e-02 / **True**
CTGAN	0.63±0.09	3.1e-02 / **True**
TVAE	0.63±0.11	3.1e-02 / **True**
SMOTE-NC	0.47±0.00	3.1e-02 / **True**
UMAP-SMOTENC	**0.83±0.02**	N/A

mean absolute scores obtained per classifier across the diverse synthetic train sets and the original one, and in Table 5, the relevant statistical validation is performed comparing the performance of UMAP-SMOTENC and the remaining methods. The null hypothesis assumes that there is no significant pairwise difference between the mean results achieved using the UMAP-SMOTENC generated dataset and the other synthetic datasets, across the various classifiers analyzed. By rejecting the null hypothesis at a significance level of α=0.05 for each classifier (except for Naïve Bayes), we conclude that there is a significant improvement in machine learning performance when utilizing UMAP-SMOTENC as the generator, in comparison with CG, TVAE, and CTGAN. On the other hand, plain SMOTE-NC demonstrated significantly better classification performance than the proposed method, except for Naïve Bayes and Logistic Regression classifiers. This outcome was expected since plain SMOTE-NC has the potential to generate entries that are extremely close to the original ones.

Lastly, Table 6 shows the mean absolute privacy scores obtained by the different generative techniques compared to the privacy baseline offered by the test set, and discussed in Section 4.4. Simultaneously, the statistical significance of the results is validated by testing the null hypothesis that there is no significant pairwise difference between the test set baseline and the anonymization methods. By rejecting the null hypothesis, at the significance level of $\alpha = 0.05$, for every method, we conclude that all methods provide statistically significant differences in privacy protection compared to the proposed baseline. However, while GC, CTGAN, TVAE, and UMAP-SMOTENC provide better privacy protection than the baseline, SMOTE-NC offers lower privacy protection. This reinforces the conclusion that SMOTE-NC cannot be safely used as an anonymization technique, as it fails to offer the necessary privacy assurances.

5.2. Discussion

The results presented in the previous section demonstrate that UMAP-SMOTENC consistently outperforms CTGAN, TVAE and GC in what concerns the quality of the anonymized, regardless of the classifier used for the downstream machine learning task. In the opposite direction, SMOTE-NC offers a better preservation of utility than the proposed method. Nevertheless, this was something to be expected. As discussed

Table 4
Mean Absolute for the machine learning utility across classifiers.

Generator	LR	NB	DT	RF	XGB	MLP
GC	0.39±0.04	0.29±0.05	0.37±0.06	0.33±0.02	0.37±0.04	0.35±0.05
CTGAN	0.61±0.04	0.51±0.02	0.60±0.05	0.63±0.04	0.63±0.05	0.64±0.04
TVAE	0.63±0.09	0.49±0.03	0.61±0.08	0.64±0.09	0.66±0.11	0.66±0.12
SMOTE-NC	**0.76±0.03**	0.53±0.01	**0.80±0.02**	**0.85±0.02**	**0.87±0.01**	**0.83±0.02**
UMAP-SMOTENC	0.75±0.02	**0.57±0.02**	0.77+0.05	0.82±0.02	0.83±0.02	0.81±0.02
Original Train Set	0.75±0.02	0.53±0.01	0.83±0.02	0.87±0.02	0.89±0.01	0.85±0.02

Table 5
Results from the Wilcoxon signed-rank test with Holm's adjusted p-values for the pairwise comparation with UMAP-SMOTENC, regarding the machine learning utility across classifiers.

Generator	LR	NB	DT	RF	XGB	MLP
GC	3.1e-02 / **True**	1.6e-01 / False	3.1e-02 / **True**	3.1e-02 / **True**	3.1e-02 / **True**	3.1e-02 / **True**
CTGAN	3.1e-02 / **True**	4.5e-01 / False	3.1e-02 / **True**	3.1e-02 / **True**	3.1e-02 / **True**	3.1e-02 / **True**
TVAE	3.1e-02 / **True**	4.5e-01 / False	3.1e-02 / **True**	3.1e-02 / **True**	3.1e-02 / **True**	3.1e-02 / **True**
SMOTE-NC	4.6e-01 / False	9.5e-01 / False	3.1e-02 / **True**	3.1e-02 / **True**	3.1e-02 / **True**	3.1e-02 / **True**

in Section 3.1., SMOTE is often criticized in the literature for generating samples that are near duplicates of the original ones, which in the context of data anonymization raises severe concerns. This is further validated by the results presented in Table 6, according to which SMOTE-NC is the only method unable to surpass the minimum privacy threshold defined by the test set. Contrarily, the proposed method not only surpassed this threshold, in a statistically significant way, but it also performed at par or better than the TVAE.

Although Tables 4 and 6 provide evidence that SMOTE-NC is not a secure technique in the context of anonymization, and GC often provides an utility level that is less than half of the original one, evaluating the quality of synthetic data based solely on privacy or utility offers a limited view on the effectiveness of the anonymization process. In this regard, by proposing a trade-off metric between privacy and utility, displayed in Tables 2 and 3, a more complete analysis can be performed. Incidentally, it becomes evident that the proposed method largely surpasses the other generative models in analysis. In this sense, this trade-off metric is partially vindicated by highlighting the poor performance obtained by both SMOTE-NC and GC. On the other hand, with a more interesting performance, both TVAE and CTGAN present the same level on the equilibrium between privacy and utility, showcasing that the selection between these two methods is case-specific: TVAE tends to offer higher levels of utility, while CTGAN offers better privacy preservation. Finally, UMAP-SMOTENC outperforms even these methods significantly, with an average improvement of around 30 % on the trade-off score, suggesting a new and better equilibrium point.

The interesting results obtained by the proposed method can be decomposed in its performance on both sides of the problem: utility and privacy. On the utility side, and similarly to other SMOTE frameworks, the proposed method does not attempt to learn a global data distribution as most families of generative methods. Instead, it focuses on local relations between data points, which removes the need to overcome the challenging nature associated with the data distribution on the tabular domains. This resulted in the generation of synthetic points more aligned with the original dataset, leading to better utility preservation. While this was to be expected due to the nature of the method and the success SMOTE-based methods have encountered on tabular settings, it is on the privacy side that the proposed method surpasses expectations. In this regard, there are several reasons why we hypothesize that UMAP-SMOTENC can provide a safe generation of points, while SMOTE-NC fails to do so. Firstly, by introducing UMAP into the process, we are forcing the projection of the original dataset to a low-dimensional space, which strongly constrains the exposition of the SMOTE framework to the original data. Additionally, as the synthetic oversampling occurs in a non-linear projection of the original input space, the constraint to generate points across the line segment connecting two given points does not hold, introducing more diversity on the generative process, and therefore less tendency for overfitting the original data. Lastly, the stochastic nature of both the original projection to the low dimensional space and its reverse is also responsible for introducing a desirable degree of noise in the generative procedure.

5.3. Future directions

The results obtained in this work highlight the importance of properly quantifying the quality of the synthetic data generated and which equilibrium is achieved in the trade-off between privacy and utility. In this sense, future researchers can leverage our proposal of a unified metric for measuring the trade-off between privacy and utility, exploring how different balances between these two factors are appropriate in different domains and when faced with a differential level of cost between privacy and utility preservation. Additionally, under more restricted settings, where potential attackers may be aware of certain properties of the original dataset, different privacy quantification measures may be added to the trade-off metric, particularly, by quantifying some of the attacks identified in Section 2.3. Similarly, while here the utility was considered under the light of a classification task, depending on the desired downstream task, different metrics may be applied to the trade-off score.

Regarding the proposed method itself, it would be important to analyse UMAP-SMOTENC as a new step towards a more applied-centric approach to this particular field, and not as an end framework in itself, despite the promising results achieved. To this, future researchers may consider the expansion of the proposed framework to more complex SMOTE variants, which may further improve the results achieved. This process should, nonetheless, be carefully defined and focus on SMOTE variants that do not change the original data distribution. Simultaneously, future researchers may be interested in understanding the impact of including categorical features when constructing the UMAP graph. A simple proposal for this may rely on the construction of an additional graph, similar to what was done for the target variable. In addition, while for the scope of the current paper we intentionally kept the SMOTE-NC and UMAP default hyperparameters, the impact of a carefully designed tuning phase on the trade-off between data's utility and privacy constitutes an important research avenue. Finally, the current method can be generalized to work with datasets that do not pose a classification challenge. Although this can be achieved in a trivial manner, by considering the full dataset as an unique class, more robust

Table 6
Mean Absolute Scores for the privacy scores and results from the Wilcoxon signed-rank test with Holm's adjusted p-values for the pairwise comparation with baseline introduced by the Test Set.

Generator	Mean 1st NN	p-Value / Significance	Mean Ratio 1st and 2nd NN's	p-Value / Significance	Mean Ratio 1st and 10th NN's	p-Value / Significance
GC	0.18±0.00	3.9e-02 / **True**	**0.94±0.00**	3.9e-02 / **True**	**0.83±0.00**	3.9e-02 / **True**
CTGAN	**0.18±0.00**	3.9e-02 / **True**	0.92±0.01	3.9e-02 / **True**	0.79±0.01	3.9e-02 / **True**
TVAE	0.11±0.00	3.9e-02 / **True**	0.89±0.00	3.9e-02 / **True**	0.72±0.01	3.9e-02 / **True**
SMOTE-NC	0.03±0.00	3.9e-02 / **True**	0.52±0.00	3.9e-02 / **True**	0.24±0.00	3.9e-02 / **True**
UMAP-SMOTENC	0.11±0.00	3.9e-02 / **True**	0.92±0.00	3.9e-02 / **True**	0.75±0.00	3.9e-02 / **True**
Test Set Baseline	0.06±0.00	N/A	0.75±0.00	N/A	0.47±0.00	N/A

approaches could leverage unsupervised techniques, such as clustering, to infer some pre-existing structures in data and use those as the dataset's classes.

6. Conclusion

Despite the debate around the need for privacy protection mechanisms being far from recent, the field is still in the early stages of development, with several untapped gaps. Particularly, most literature seems disconnected from the real-world application, where methods based on transformative approaches are still frequently used, due to their simplicity, despite their fragilities being widely acknowledged.

Aware of this gap, this paper introduced UMAP-SMOTENC, a new method to generate fully-synthetic anonymized datasets. This approach relies on UMAP to project the original data to a compressed non-linear representation of the input space, where the SMOTE process occurs. The rigorous evaluation procedure conducted in this paper confirms that not only does UMAP-SMOTENC offer a significantly better balance between privacy and utility preservation than the benchmarked approaches, but it also builds on two methods that are widely used by practitioners (UMAP and SMOTE), easing the path towards its adoption. On this same note, the proposed method is also simple to apply and robust across domains, not requiring a rigorous hyperparameter tuning phase or resource to domain expertise. Therefore, we envision this work as one of the first steps to close the significant disconnection between the industry and academia regarding the use of synthetic data for anonymisation purposes.

Finally, outlining the commitment to establish bridges between industry and academia, we make available an implementation of the proposed method in the GitHub associated with the project,[2] that both researchers and practitioners can leverage to either apply directly or build upon.

Funding sources

This work was supported by a grant of the Portuguese Foundation for Science and Technology ("Fundação para a Ciência e a Tecnologia"), DSAIPA/DS/0116/2019, and project UIDB/04152/2020—Centro de Investigação em Gestão de Informação (MagIC).

CRediT authorship contribution statement

Goncalo Almeida: Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Fernando Bacao:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The Code is available at https://github.com/gdalmeida99/UMAP-SMOTENC and the Data pointer is provided in the Appendix.

Appendix

The following datasets were used in the evaluation procedure:

Adult: https://archive.ics.uci.edu/ml/datasets/adult

Magic: https://archive.ics.uci.edu/ml/datasets/magic+gamma+telescope

Digits: https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits

Beans: https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset

Occupancy: https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection

Cover Type: https://archive.ics.uci.edu/ml/datasets/covertype

Credit Card: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Loan: https://www.kaggle.com/datasets/teertha/personal-loan-modeling

[2] Available at https://github.com/gdalmeida99/UMAP-SMOTENC

References

[1] C.C. Aggarwal, On k-anonymity and the curse of dimensionality, in: Proceedings of the 31stInternational Conference on Very Large Data Bases, Norway, 2005, pp. 901–909, https://doi.org/10.5555/1083592.1083696.
[2] Akrami, H., Aydore, S., Leahy, R.M., & Joshi, A.A. (2020). Robust variational autoencoder for tabular data with beta divergence. 10.48550/arxiv.2006.08204.
[3] F. Al Zamal, W. Liu, D. Ruths, Homophily and latent attribute inference: inferring latent attributes of twitter users from neighbors, in: Proceedings of the International AAAI Conference on Web and Social Media 6, 2021, pp. 387–390, https://doi.org/10.1609/icwsm.v6i1.1434.
[4] Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. 10.48550/arxiv.1701.07875.
[5] E. Becht, L. McInnes, J. Healy, C.A. Dutertre, I.W.H. Kwok, L.G. Ng, F. Ginhoux, E.W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP, Nat. Biotechnol. 37 (2018) 38–44, https://doi.org/10.1038/nbt.4314.
[6] R.E. Bellman, The theory of dynamic programming, Bull. Am. Math. Soc. 60 (6) (1954) 503–515.
[7] Bishop, C.M. (2006). In M. Jordan, J. Kleinberg, & B. Schölkopf (eds.), Pattern Recognition and Machine Learning. Springer Science+Business Media, LLC.
[8] R. Blagus, L. Lusa, SMOTE for high-dimensional class-imbalanced data, BMC. Bioinformatics. 14 (2013) 106, https://doi.org/10.1186/1471-2105-14-106.
[9] V. Borisov, K. Sessler, T. Leemann, M. Pawelczyk, G. Kasneci, Language models are realistic tabular data generators, in: International Conference on Learning Representations (ICLR, 2023, https://doi.org/10.48550/arXiv.2210.06280.
[10] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992, pp. 144–152, https://doi.org/10.1145/130385.130401.
[11] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[12] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, 1st Ed., Routledge, 1984 https://doi.org/10.1201/9781315139470.
[13] G. Caiola, J.P. Reiter, Random forests for generating partially synthetic, categorical data, Trans. Data Priv. 3 (2010) 27–42, https://doi.org/10.5555/1747335.1747337.
[14] California Consumer Privacy Act. (2018). Cal. Civ. Code § 1798.100 (2018).
[15] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2011) 321–357, https://doi.org/10.1613/jair.953.
[16] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.
[17] E. Choi, S. Biswal, B. Malin, J. Duke, W.F. Stewart, J. Sun, Generating multi-label discrete patient records using generative adversarial networks, in: Proceedings of Machine Learning for Healthcare 2017, 2017, p. 68, https://doi.org/10.48550/arXiv.1703.06490.
[18] T. Dalenius, S.P. Reiss, Data-swapping: a technique for disclosure control, J. Stat. Plan. Inference 6 (1) (1982) 73–85, https://doi.org/10.1016/0378-3758(82)90058-1.
[19] R. Dewri, I. Ray, I. Ray, D. Whitley, On the optimal selection of k in the k-anonymity problem, in: IEEE 24thInternational Conference on Data Engineering, Mexico, 2008, pp. 1364–1366, https://doi.org/10.1109/ICDE.2008.4497557.
[20] J. Domingo-Ferrer, V. Torra, A critique of k-anonymity and some of its enhancements, in: Third International Conference on Availability, Reliability and Security, Spain, 2008, pp. 990–993, https://doi.org/10.1109/ARES.2008.97.
[21] G. Douzas, F. Bacao, Effective data generation for imbalanced learning using conditional generative adversarial networks, Expert. Syst. Appl. 91 (2018) 464–471, https://doi.org/10.1016/j.eswa.2017.09.030.
[22] G. Douzas, F. Bacao, Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE, Inf. Sci. 501 (2019) 118–135, https://doi.org/10.1016/j.ins.2019.06.007.
[23] Drechsler, J. (2010). Using support vector machines for generating synthetic datasets. In Domingo-Ferrer, J., Magkos, E. (eds.), Privacy in Statistical Databases. Lecture Notes in Computer Science, 6344, 148–161. Springer, Berlin, Heidelberg. 10.1007/978-3-642-15838-4_14.
[24] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In S. Halevi, T. Rabin (eds.), Theory of Cryptography. Lecture Notes in Computer Science, 3876, 256–284. Springer, Berlin, Heidelberg. 10.1007/11681878_14.

[25] M. Endres, A. Mannarapotta Venugopal, T.S. Tran, Synthetic data generation: a comparative study, in: Proceedings of the 26th International Database Engineered Applications Symposium, 2022, pp. 94–102, https://doi.org/10.1145/3548785.3548793.
[26] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Am. Stat. Assoc. 32 (200) (1937) 675–701, https://doi.org/10.1080/01621459.1937.10503522.
[27] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Adv. Neural Inf. Process. Syst. 21 63 (11) (2014) 139–144, https://doi.org/10.1145/3422622.
[28] N. Gruschka, V. Mavroeidis, K. Vishi, M. Jensen, Privacy issues and data protection in big data: a case study analysis under GDPR, in: 2018 IEEE International Conference on Big Data (Big Data), USA, 2019, pp. 5027–5033, https://doi.org/10.1109/BIGDATA.2018.8622621.
[29] M.S. Gulati, P.F. Roysdon, TabMT: generating tabular data with masked transformers, in: Conference on Neural Information Processing Systems (NeurIPS), 2023, https://doi.org/10.48550/arXiv.2312.06089.
[30] J.R. Gulcher, K. Kristjánsson, H. Gudbjartsson, K. Stefánsson, Protection of privacy by third-party encryption in genetic research in Iceland, Eur. J. Hum. Genet. 8 (10) (2000) 739–742, https://doi.org/10.1038/SJ.EJHG.5200530.
[31] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of Wasserstein GANs, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 5769–5779, https://doi.org/10.5555/3295222.3295327.
[32] T.M. Ha, H. Bunke, Off-line, handwritten numeral recognition by perturbation method, IEEe Trans. Pattern. Anal. Mach. Intell. 19 (5) (1997) 535–539, https://doi.org/10.1109/34.589216.
[33] Han, H., Wang, W.Y., & Mao, B.H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In D. Huang, X. Zhang, G. Huang (eds.), Advances in Intelligent Computing. Lecture Notes in Computer Science, 3644, 878–887. Springer, Berlin, Heidelberg. 10.1007/11538059_91.
[34] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 1322–1328, https://doi.org/10.1109/IJCNN.2008.4633969.
[35] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 (2) (1979) 65–70.
[36] F. Houssiau, J. Jordon, S.N. Cohen, O. Daniel, A. Elliott, J. Geddes, C. Mole, C. Rangel-Smith, L. Szpruch, TAPAS: a toolbox for adversarial privacy auditing of synthetic data, in: SyntheticData4ML NeurIPS Workshop, 2022, https://doi.org/10.48550/arXiv.2211.06550.
[37] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P.S. Yu, X. Zhang, Membership inference attacks on machine learning: a survey, ACM Comp. Surv. 54 (11s) (2022) 1–37, https://doi.org/10.1145/3523273.
[38] V.S. Iyengar, Transforming data to satisfy privacy constraints, in: Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, 2002, pp. 279–288, https://doi.org/10.1145/775047.775089.
[39] E. Jang, S. Gu, B. Poole, Categorical Reparameterization with Gumbel-Softmax, in: International Conference on Learning Representations (ICLR), 2017, https://doi.org/10.48550/arXiv.1611.01144.
[40] Jordan, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S.N., & Weller, A. (2022). Synthetic data - what, why and how?. 10.48550/arXiv.2205.03257.
[41] J. Jordon, J. Yoon, M. Van Der Schaar, PATE-GAN: generating synthetic data with differential privacy guarantees, in: International Conference on Learning Representations, USA, 2019.
[42] J.M. Joyce, Kullback-Leibler divergence, in: M. Lovric (Ed.), International Encyclopedia of Statistical Science, Springer, Berlin, Heidelberg, 2011, pp. 720–722, https://doi.org/10.1007/978-3-642-04898-2_327.
[43] Kamthe, S., Assefa, S., & Deisenroth, M. (2021). Copula flows for synthetic data generation. 10.48550/arxiv.2101.00598.
[44] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, in: International Conference on Learning Representations, Canada, 2013, https://doi.org/10.48550/arxiv.1312.6114.
[45] D.P. Kingma, M. Welling, An Introduction to Variational Autoencoders, Now Foundations and Trends, 2019. https://ieeexplore.ieee.org/document/9051780.
[46] Kodali, N., Abernethy, J., Hays, J., & Kira, Z. (2017). On convergence and stability of GANs. 10.48550/arxiv.1705.07215.
[47] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, in: Proceedings of the National Academy of Sciences (PNAS) 110, 2013, pp. 5802–5805, https://doi.org/10.1073/pnas.1218772110.
[48] Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2022). TabDDPM: modelling tabular data with diffusion models. 10.48550/arXiv.2209.15421.
[49] K. Kurach, M. Lučić, X. Zhai, M. Michalski, S. Gelly, A large-scale study on regularization and normalization in GANs, in: International Conference on Machine Learning, 2019, pp. 3581–3590, https://doi.org/10.48550/arXiv.1807.04720.
[50] F. Last, G. Douzas, F. Bacao, Oversampling for imbalanced learning based on K-means and SMOTE, Inf. Sci. 465 (2017) 1–20, https://doi.org/10.1016/j.ins.2018.06.056.
[51] G. Lemaître, F. Nogueira, C.K. Aridas, Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 7 (2016) 1–5, https://doi.org/10.48550/arXiv.1609.06570.
[52] N. Li, T. Li, S. Venkatasubramanian, t-Closeness: privacy beyond k-anonymity and ℓ-diversity, in: 2007 IEEE 23rdInternational Conference on Data Engineering, Turkey, 2007, pp. 106–115, https://doi.org/10.1109/ICDE.2007.367856.
[53] R.J.A. Little, Statistical analysis of masked data, J. Off. Stat. 9 (2) (1993) 407–426.
[54] C. Luo, D. Wu, D. Wu, A deep learning approach for credit scoring using credit default swaps, Eng. Appl. Artif. Intell. 65 (2017) 465–470, https://doi.org/10.1016/J.ENGAPPAI.2016.12.002.
[55] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, ℓ-Diversity: privacy beyond k-anonymity, in: 22ndInternational Conference on Data Engineering, USA, 2006, p. 24, https://doi.org/10.1109/ICDE.2006.1.
[56] A. Maćkiewicz, W. Ratajczak, Principal components analysis (PCA), Comput. Geosci. 19 (3) (1993) 303–342, https://doi.org/10.1016/0098-3004(93)90090-R.
[57] B. Malin, L. Sweeney, Re-identification of DNA through an automated linkage process, in: Proceedings of the AMIA Symposium, 2001, pp. 423–427.
[58] D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, J.Y. Halpern, Worst-case background knowledge for privacy-preserving data publishing, in: 2007 IEEE International Conference on Data Engineering, 2007, pp. 126–135, https://doi.org/10.48550/arxiv.0705.2787.
[59] P. McCullagh, J.A. Nelder, Generalized Linear Models, Chapman And Hall, 1989.
[60] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. 10.48550/arxiv.1802.03426.
[61] L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: uniform manifold approximation and projection, J. Open Source Softw. 3 (29) (2018) 861, https://doi.org/10.21105/joss.00861.
[62] Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. 10.48550/arxiv.1411.1784.
[63] A. Narayanan, V. Shmatikov, Robust de-anonymization of large sparse datasets, in: 2008 IEEE Symposium on Security and Privacy, USA, 2008, pp. 111–125, https://doi.org/10.1109/SP.2008.33.
[64] M. Naseriparsa, M.M.R. Kashani, Combination of PCA with SMOTE resampling to boost the prediction rate in lung cancer dataset, Int. J. Comput. Appl. 77 (3) (2013) 33–38, https://doi.org/10.5120/13376-0987.
[65] P. Nerurkar, S. Bhirud, D. Patel, R. Ludinard, Y. Busnel, S. Kumari, Supervised learning model for identifying illegal activities in Bitcoin, Appl. Intell. 51 (2020) 3824–3843, https://doi.org/10.1007/S10489-020-02048-W.
[66] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, Y. Kim, Data synthesis based on generative adversarial networks, in: Proceedings of the VLDB Endowment 11, 2018, pp. 1071–1083, https://doi.org/10.14778/3231751.3231757.
[67] N. Patki, R. Wedge, K. Veeramachaneni, The synthetic data vault, in: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Canada, 2016, pp. 399–410, https://doi.org/10.1109/DSAA.2016.49.
[68] F. Pedregosa, V. Michel, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Cournapeau, G. Varoquaux, A. Gramfort, B. Thirion, V. Dubourg, A. Passos, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830, https://doi.org/10.48550/arXiv.1201.0490.
[69] Personal Information Protection and Electronic Documents Act, S.C. c.5 (2000).
[70] D. Pujol, R. McKenna, S. Kuppam, M. Hay, A. Machanavajjhala, G. Miklau, Fair decision making using privacy-protected data, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*'20), Association for Computing Machinery, USA, 2019, pp. 189–199, https://doi.org/10.48550/arxiv.1905.12744.
[71] K. Purdam, M. Elliot, A case study of the impact of statistical disclosure control on data quality in the individual UK samples of anonymised records, Environ. Plan. A 39 (5) (2007) 1101–1118, https://doi.org/10.1068/A38335.
[72] Regulation (EU), On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Eur. Parliam. Council (2016).
[73] J.P. Reiter, Using CART to generate partially synthetic public use microdata, J. Off. Stat. 21 (3) (2005) 441–462.
[74] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (6) (1958) 386–408, https://doi.org/10.1037/H0042519.
[75] D.B. Rubin, Discussion: statistical disclosure limitation, J. Off. Stat. 9 (2) (1993) 461–468.
[76] D.E. Rumelhart, J.L. McClelland, Learning internal representations by error propagation. Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations, MIT Press, 1987, pp. 318–362. https://ieeexplore.ieee.org/document/6302929.
[77] T. Sainburg, L. McInnes, T.Q. Gentner, Parametric UMAP embeddings for representation and semi-supervised learning, Neural Comput. (2021) 33 (11) (2021) 2881–2907, https://doi.org/10.1162/neco_a_01434.
[78] P. Samarati, Protecting respondents' identities in microdata release, IEEe Trans. Knowl. Data Eng. 13 (6) (2001) 1010–1027, https://doi.org/10.1109/69.971193.
[79] Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. https://doi.org/10.1184/R1/6625469.v1.
[80] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: Proceedings of the IEEE Symposium on Security and Privacy (SP), 2017, pp. 3–18, https://doi.org/10.48550/arXiv.1610.05820.

[81] J.A.M. Sidey-Gibbons, C.J. Sidey-Gibbons, Machine learning in medicine: a practical introduction, BMC. Med. Res. Methodol. 19 (64) (2019) 1–18, https://doi.org/10.1186/S12874-019-0681-4.
[82] Solatorio, A.V., & Dupriez, O. (2023). REaLTabFormer: generating realistic relational and tabular data using transformers. 10.48550/arXiv.2302.02041.
[83] Stadler, T., Oprisanu, B., & Troncoso, C. (2020). Synthetic data–anonymisation groundhog day. 10.48550/arXiv.2011.07018.
[84] Y. Sun, A. Cuesta-Infante, K. Veeramachaneni, Learning vine copula models for synthetic data generation, in: Proceedings of the 33rd AAAI Conference on Artificial Intelligence 30, 2019, pp. 5049–5057, https://doi.org/10.48550/arxiv.1812.01226.
[85] L. Sweeney, Guaranteeing anonymity when sharing medical data, the Datafly system, in: Proceedings of the AMIA Annual Fall Symposium, 1997, pp. 51–55.
[86] L. Sweeney, Simple Demographics Often Identify People Uniquely, Carnegie Mellon University, Data Privacy, 2000, https://doi.org/10.1184/R1/6625769.v1.
[87] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, Int. J. Uncertain., Fuzziness Knowl.-Based Syst. 10 (5) (2002) 571–588, https://doi.org/10.1142/S021848850200165X.
[88] L. Sweeney, k-ANONYMITY: a model for protecting privacy, Int. J. Uncertain., Fuzziness Knowl.-Based Syst. 10 (5) (2002) 557–570, https://doi.org/10.1142/S0218488502001648.
[89] Tran, C., Dinh, M.H., Beiter, K., & Fioretto, F. (2021). A fairness analysis on private aggregation of teacher ensembles. 10.48550/arxiv.2109.08630.
[90] T.M. Truta, B. Vinay, Privacy protection: P-Sensitive k-Anonymity property, in: 22nd International Conference on Data Engineering Workshops (ICDEW'06), USA, 2006, p. 94, https://doi.org/10.1109/ICDEW.2006.116.
[91] L. Van Der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[92] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31stInternational Conference on Neural Information Processing Systems, USA, 2017, pp. 6000–6010, https://doi.org/10.48550/arxiv.1706.03762.
[93] M. Walia, B. Tierney, S. Mckeever, Synthesising tabular data using Wasserstein conditional GANs with gradient penalty (WCGAN-GP), in: AICS 2020: 28th Irish Conference on Artificial Intelligence and Cognitive Science, Ireland, 2020, https://doi.org/10.21427/E6WA-SZ92.
[94] J. Wang, M. Xu, H. Wangf, J. Zhang, Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding, in: 2006 8th International Conference on Signal Processing Proceedings, China 3, 2006, https://doi.org/10.1109/ICOSP.2006.345752.
[95] Wilcoxon, F. (1992). Individual comparisons by ranking methods. In S. Kotz, N.L. Johnson (eds.), *Breakthroughs in Statistics*. Springer Series in Statistics: Vol. II, 196–202. Springer, New York. 10.1007/978-1-4612-4380-9_16.
[96] Xie, L., Lin, K., Wang, S., Wang, F., & Zhou, J. (2018). Differentially private generative adversarial network. 10.48550/arxiv.1802.06739.
[97] Xu, L., & Veeramachaneni, K. (2018). Synthesizing tabular data using generative adversarial networks. 10.48550/arxiv.1811.11264.
[98] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling tabular data using conditional GAN, in: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Canada 32, 2019, https://doi.org/10.48550/arxiv.1907.00503.
[99] Yancey, W.E., Winkler, W.E., & Creecy, R.H. (2002). Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer (eds), *Inference Control in Statistical Databases*. Lecture Notes in Computer Science, 2316, 135–152. Springer, Berlin, Heidelberg. 10.1007/3-540-47804-3_11.
[100] Zhao, Z., Birke, R., & Chen, L.Y. (2023). TabuLa: harnessing language models for tabular data synthesis. 10.48550/arXiv.2310.12746.
[101] Zhao, Z., Kunar, A., Birke, R., & Chen, L.Y. (2022). CTAB-GAN+: enhancing tabular data synthesis. 10.48550/arXiv.2204.00401.
[102] Z. Zhao, A. Kunar, H. Van der Scheer, R. Birke, L.Y. Chen, CTAB-GAN: effective table data synthesizing, in: Proceedings of the Asian Conference on Machine Learning, 2021, pp. 97–112, https://doi.org/10.48550/arxiv.2102.08369.