IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.1120000

Privacy mechanisms and evaluation metrics for Synthetic Data Generation: A systematic review

PABLO A. OSORIO-MARULANDA[1 2], GORKA EPELDE[2 3], MIKEL HERNANDEZ[2 4], IMANOL ISASA[2], NICOLAS MORENO REYES[1], and ANDONI BERISTAIN IRAOLA[2 3 4]

[1]School of Applied Sciences and Engineering, Universidad EAFIT, Medellín, Colombia
[2]Digital and Biomedical Technologies, Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Donostia - San Sebastián, Spain
[3]eHealth Group, Biogipuzkoa Health Research Institute, Donostia - San Sebastian, Spain
[4]Computer Science and Artificial Intelligence Department, University of the Basque Country (UPV/EHU), Computer Science Faculty, Donostia - San Sebastian, Spain

Corresponding author: Gorka Epelde (e-mail: gepelde@vicomtech.org)

ABSTRACT The growth of data publishing, sharing, and mining mechanisms in various fields of industry and science has led to an increase in the flow of data, making it an important asset that needs to be protected and managed effectively. To this end, different mechanisms have been used across different domains, including Privacy Enhancing Technologies like Synthetic Data Generation, which aim to protect user-sensitive data and prevent misuse among different domains. Then, Synthetic data has been used not only to augment datasets and balance classes but also in applications of data analysis paradigms that aim to provide useful insights in terms of utility while preserving the privacy of sensitive data. Still, there is a gap in the conceptual and state-of-the-art understanding of the level of privacy synthetic data generators can provide and how they affect various industries and fields. This systematic review attempts to address how privacy has been assessed and measured in the framework of synthetic data generation, and getting to know which metrics have been used to evaluate those mechanisms. We provide an overview a total of 105 recent studies in this field after a screening process and identify future open research directions. The main findings include a high prevalence of differential privacy as a privacy-preserving technique and privacy budget cost as a trade-off metric, with a high percentage of GAN-based model implementations, and mainly healthcare applications. Our systematic review covers multiple privacy domains and can be understood as a general framework for privacy measurement applied in Synthetic Data Generation.

INDEX TERMS Anonymization, Confidentiality, Privacy, Privacy metrics, Privacy-preserving big data analytics, Synthetic Data, Synthetic Data Generation

I. INTRODUCTION

THE collection and use of data has significantly driven advances in different domains, such as science, healthcare, and industry. However, this data-driven progress has also raised significant concerns, particularly in the context of privacy and security. Since data collection involves the creation of a data-sharing network to facilitate discovery and ensure transparency, personal data, which often contains sensitive information, requires privacy preservation during its processing and dissemination [1].

Different types of sensitive information, such as where users go, what health problems they face, what are their likes, dislikes, and habits, among others, are susceptible to exposure and use by different types of attackers [2]. This leads us to consider how necessary it is to preserve data privacy and how important it is to maintain the privacy of the results. However, it is crucial to be aware not only of data privacy but also of the usefulness that data may have. The latter characteristic is seriously affected by the number of attributes or general information lost from a source dataset after applying privacy-preserving mechanisms. The goal is to strike a balance between data utility and privacy, i.e. to maintain the privacy of individuals' information while preserving the use of such information in data analysis paradigms for secondary purposes [2].

According to Matwin [3], it is not uncommon to confuse and even underestimate data privacy when comparing it to data security. Although both terms are related, since they seek to guard two constant parts of the data flow, they represent different concepts. Data security seeks to protect the flow of information when it has not reached its final destination, preventing intrusion by intruders, unauthorized access, and data theft. On the other hand, data privacy considers the moment when the data reaches its final destination and tries not to reveal sensitive or personal information that has been collected in the data now in the possession of the recipient. Therefore, this approach should ideally have a treatment from the data itself, and not in the way it is shared.

With the above, considering data privacy nowadays becomes even more important, due to the constant increase in the flow of information through sharing facilities, the amount of data that can be shared is immense, and very rich in terms of information, making threats such as membership inference attacks, or model inversion attacks have the ability to reveal information about the specific contents of a database. Even so, there are current regulations that are rightly committed to guaranteeing data privacy, such as the European Union's General Data Protection Regulation.

Among the many risks that exist considering data loss or disclosure of information from a threat, identity theft and impersonation, financial risks, reputational damage, phishing, cyber-attacks, and unethical information sharing are some of them.

As proposed by Majeed [1], there are three challenges while handling sensitive information among individuals: (i) to prevent the misuse of personal data, to enable fair and unbiased decision-making concerning real-world entities, (ii) to restrict target profiling, and (iii) to improve the quality of personal data for the welfare of societies, enhancing data quality, when available data is limited. These challenges can be addressed by resolving the dilemma between privacy and utility using different approaches complemented with Privacy Enhancing Techniques, making available larger privacy-preserving datasets with higher quality and utility.

A. MOTIVATION

Data privacy and confidentiality play a critical role, particularly in healthcare contexts. The sharing of data raises significant privacy concerns, especially during data-sharing processes. Recent research has introduced Synthetic Data Generation (SDG) techniques as a Privacy Enhancing Technology to address these concerns.

In this context, Rankin et al. [4] propose the utilization of synthetic data for developing machine learning models, which can be outsourced from healthcare departments. These models can then be retrained and adopted for clinical practice or to inform policy decisions. Additionally, Hernandez et al. [5] demonstrate the integration and automation of SDG within a control data processing workflow, emphasizing the preservation of privacy in the health and wellness domain.

However, several recent works have established that SDG models are insufficient to guarantee certain privacy standards. According to Stadler et al. [6], synthetic data is insufficient to preserve privacy. The study shows that synthetic data do not prevent inference attacks and preserve the utility of the data. In addition, studies tend to overestimate the anonymization capacity resulting from the use of these technologies. This fact is also supported by Zhang et al. [7], who show that partially synthetic data is highly vulnerable to membership inference attacks, while fully synthetic data is substantially more resistant to such attacks, even marginally susceptible. Also, the study developed by Torfi et al. [8] mentions that although the process of generating synthetic data from a Generative Adversarial Network (GAN) is not reversible from the data, the naive use of GANs for SDG does not guarantee that the system preserves privacy based solely on the fact that GANs are not reversible.

To overcome those weaknesses, different privacy techniques have been developed to reinforce privacy in synthetically generated data. The implementation of those, although widely deployed, still needs to overcome a significant number of challenges, such as the recognition of the appropriate privacy techniques for a given data generation context.

The motivation behind this work is to analyze current research at the intersection of SDG and privacy techniques. Our goal is to offer a comprehensive overview that considers the diverse aspects of research, including the field of application (with a special focus on health data), generation models, and the nature of the data. We intend for this resource to guide future research adopters.

In addition, there is a lack of metrics that can correctly capture the level of privacy and usability of the privately generated data. Therefore, developing rigorously validated evaluation criteria for future efforts is crucial, considering the discussion about the inverse relationship between those terms.

B. RESEARCH QUESTIONS

The main objective of this review is to analyze the privacy mechanisms used together with SDG, and the metrics used to evaluate the privacy level of those combinations with a special focus on privacy-preserving data publishing and sharing. With this in mind, the research questions that directed this review were:

- **RQ1** Which privacy techniques can be integrated with synthetic data generation models?
- **RQ2** To what extent and criteria different metrics have been applied to evaluate the privacy of generated data by different SDG methods combined with privacy mechanisms? How are those methods applied in a privacy utility trade-off evaluation?
- **RQ3** How have privacy techniques in the framework of synthetically generated data been used and applied in health data?

C. CONTRIBUTIONS

This work presents a comprehensive and understandable study on the implementation of privacy techniques and metrics in the framework of SDG. To achieve this goal, we performed a systematic review of the topic. The review aimed to start from the advances made in the area, resolve doubts regarding the implementation of privacy techniques in certain fields, and verify the most used methods in recent years with a certain set of models. Although work has been done in this area, to the best of our knowledge, this is the work that covers the most recent literature and is framed in the general application of synthetic data generation, as well as in the segmented implementation of privacy techniques in different application domains, moreover, similar works are discussed in Section II-G. Likewise, the study aims to analyze approaches regarding the consideration of the privacy-utility trade-off in recent studies, to establish a general summary from which to start future investigations.

D. ARTICLE OUTLINE

Section II contains background information describing the privacy risks associated with tabular data, the definition of important terms such as Privacy Preserving Data Models and Synthetic Data, and how they are covered in this line; a description of SDG models in the context of privacy-preserving guarantees, along with the most commonly used performance metrics, privacy metrics, and finally an overview of other works that have been done on the subject and have similar objectives to those of this article. Subsequently, Section III describes the process of collecting the articles, including the search methodology, selection criteria, and data extraction. Section IV provides an overview of selected publications. Section V describes data privacy techniques in synthetic data generation. Section VI describes data privacy metrics in the SDG, including an overview of metrics and categories of metrics. Section VII gives an evaluation of the fields of implementation of the mentioned works, Section VIII provides a discussion which includes the main findings, limitations of current research, and future directions, and Section IX elaborates on the conclusions of the research.

A graphical summary of the systematic review is presented in Figure 1.

II. BACKGROUND

To understand the use cases of synthetic data generation, and how it creates a general framework to be analyzed in terms of privacy, it is important to take into account some general concepts in which this topic becomes relevant. The first of these is the notion of Privacy Preserving Data Publishing (PPDP), a process that provides methods and tools to publish useful information while preserving privacy. If a dataset is released for data analysis, some techniques are needed to reduce the risk of identifying sensitive information about individuals by linking data. PPDP performs a data transformation that guarantees its usefulness, while preserving privacy, in the publication of large data collections. While Privacy-Preserving Data Mining (PPDM) performs data mining tasks in private databases, PPDP's main goal is how to *publish* the data for data mining tasks [9].

The second concept is the privacy risk, which is associated with how an individual or user may have their data in danger from different types of threat. Finally, it is necessary to conceptualize a privacy metric and a privacy mechanism, where the first concept encapsulates how privacy can be measured, and the second, how privacy can be achieved under the transformation of a dataset, all of this focused on the general structure that establishes the generation of synthetic data.

A. PRIVACY RISKS

For a structured data set, it is possible to establish a configuration on the sensitivity of the attributes associated with each of the individuals, being able to categorize them as follows:

- *Explicit identifier (ID)*: A set of attributes that uniquely identifies a record owner.
- *Quasi-identifier (QID)*: A set of attributes that cannot uniquely identify a record owner but potentially identify the target if combined with some auxiliary information.
- *Sensitive attribute (SA)*: Sensitive information that the record owner intends to keep private from authorized parties.
- *Non-sensitive attribute (NSA)*: An attribute that does not violate the record owner's privacy if disclosed.

Additionally, the disclosure types of instances where the release of personal information with the potential to affect individuals can be classified as follows:

- *Identity disclosure*: When an adversary reveals the identity of a victim.
- *Attribute disclosure*: When an adversary successfully links a victim to their SA information with a high probability.
- *Membership disclosure*: It occurs when an adversary successfully infers the existence of a targeted victim in the published dataset with high probability.
- *Linkage attack*: The adversary may re-identify the identity and discover the SA values of a targeted record owner by matching the auxiliary QID values with the published table T'.
- *Homogeneity attack*: This attack discloses the SA values of a target when there is insufficient homogeneity in the SA. That is, the combination of QID is mapped to one SA value only.
- *Background knowledge attack*: This attack utilizes logical reasoning and additional knowledge about a target to breach the SA values.
- *Skewness attack*: When the overall distribution of SA in the original data is skewed, SA values can be inferred. The SA values have different degrees of sensitivity.
- *Similarity attack*: This attack discloses SA values when the semantic relationship of distinct SA values in an equivalence class is close.

B. PRIVACY PRESERVING DATA MODELS

For a proper description of what is meant by privacy-preserving data models, it is also necessary to consider the previous definition of the term Privacy Enhancing Technologies (PETs).

To ensure that privacy and usability criteria are met, techniques framed in PETs are used, which offer a wide variety

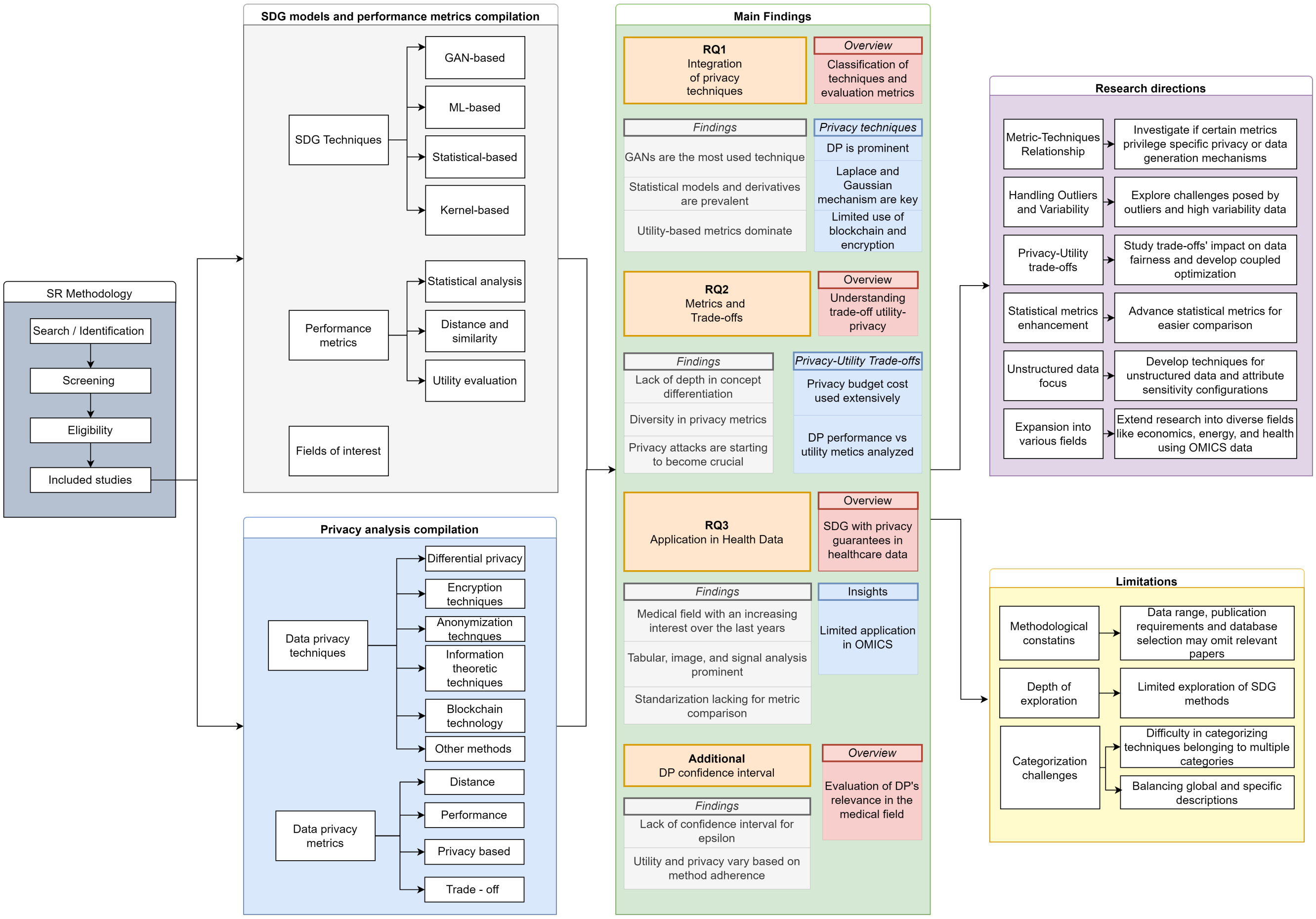
SR Methodology
Search / Identification
Screening
Eligibility
Included studies
SDG models and performance metrics compilation
SDG Techniques
GAN-based
ML-based
Statistical-based
Kernel-based
Performance metrics
Statistical analysis
Distance and similarity
Utility evaluation
Fields of interest
Privacy analysis compilation
Data privacy techniques
Differential privacy
Encryption technques
Anonymization technqus
Information theoretic techniques
Blockchain technology
Other methods
Data privacy metrics
Distance
Performance
Privacy based
Trade - off
Main Findings
RQ1
Integration of privacy techniques
Overview
Classification of techniques and evaluation metrics
Findings
GANs are the most used technique
Statistical models and derivatives are prevalent
Utility-based metrics dominate
Privacy techniques
DP is prominent
Laplace and Gaussian mechanism are key
Limited use of blockchain and encryption
RQ2
Metrics and Trade-offs
Overview
Understanding trade-off utility-privacy
Findings
Lack of depth in concept differentiation
Diversity in privacy metrics
Privacy attacks are starting to become crucial
Privacy-Utility Trade-offs
Privacy budget cost used extensively
DP performance vs utility metics analyzed
RQ3
Application in Health Data
Overview
SDG with privacy guarantees in healthcare data
Findings
Medical field with an increasing interest over the last years
Tabular, image, and signal analysis prominent
Standarization lacking for metric comparison
Insights
Limited application in OMICS
Additional
DP confidence interval
Overview
Evaluation of DP's relevance in the medical field
Findings
Lack of confidence interval for epsilon
Utility and privacy vary based on method adherence
Research directions
Metric-Techniques Relationship
Investigate if certain metrics privilege specific privacy or data generation mechanisms
Handling Outliers and Variability
Explore challenges posed by outliers and high variability data
Privacy-Utility trade-offs
Study trade-offs' impact on data fairness and develop coupled optimization
Statistical metrics enhancement
Advance statistical metrics for easier comparison
Unstructured data focus
Develop techniques for unstructured data and attribute sensitivity configurations
Expansion into various fields
Extend research into diverse fields like economics, energy, and health using OMICS data
Limitations
Methodological constains
Data range, publication requirements and database selection may omit relevant papers
Depth of exploration
Limited exploration of SDG methods
Categorization challenges
Difficulty in categorizing techniques belonging to multiple categories
Balancing global and specific descriptions

FIGURE 1. A summary of the systematic review with the main findings of the research

of solutions adapted to specific scenarios. PETs or Privacy-Preserving data models can be described as an algorithm that, starting from some input data, generates a more private version of it according to the criteria in which this feature is measured. For a correct understanding of these techniques, Carvalho et al. [10] complete a taxonomy for their classification, contemplating three essential categories: perturbative, non-perturbative, and de-associative. The first ones refer to techniques that distort the data before they are released, even taking into account the non-alteration in the statistics or usefulness that they may have, reducing the level of detail or partially suppressing information from the original data while preserving the veracity. On the other hand, models in the non-perturbative category, sort the data before release in such a way that the statistics to be measured do not differ significantly from the data obtained by statisticians in the original database. Finally, the de-associative models' main objective is to create buckets to curb the correlation between QID and sensitive attributes.

Among the techniques considered as PETs are secure multiparty computation, federated learning, differential privacy, and anonymization. These techniques alter, remove, or encrypt identifiers that can directly reveal a person's information or allow it to be linked to an external database. SDG is also included in the PETs, running a mathematical model for creating data that is not directly related to the registers taken directly from different users, getting the behavioral patterns, and preserving the statistical properties of the original database.

On the other hand, Majeed [1] briefly proposes a taxonomy based on a scheme describing SOTA privacy tracks. Within this classification, we have the following categories: Syntactic privacy methods, Clustering privacy methods, Semantic privacy methods, AI-based privacy methods, Attack-specific privacy methods, Application-specific privacy methods, Data-specific privacy methods, Domain-specific privacy methods, Hybrid privacy methods, Attribute-centric privacy methods, and Synthetic data-based privacy methods.

C. SYNTHETIC DATA

According to Jordon et al. [11], synthetic data can be defined as *data that has been generated using a purpose-built mathematical model or algorithm, to solve a (set of) data science task(s)*. This model can take many forms, from a Bayesian Network to a GAN or a Variational Auto-encoder. The idea around the existence of this model, which we can denote as generator $\mathbf{G}$, is that considering a real database $\mathbf{D}$, the generator is capable of producing a $\hat{D} \sim \mathbf{G}(\mathbf{D})$, which denotes the synthetic data generated by $\mathbf{G}$. Although they proposed a definition that covers sufficient categories in this area, this concept does not have a widely accepted meaning.

Synthetic data has several uses in this field and has regained relevance in recent years due to the growing need to extract information from large databases for the development of data-driven models, where the quality of the model output depends directly on the amount of data used for training. Jordon et al. [11] give a detailed description of the different implementations result of this trend, among which they consider: Machine Learning tool development, Software testing, Bias reduction/elimination, What-if scenario generation, Data augmentation-data labeling, and finally the use of synthetic data as a private data release.

Although the generation of synthetic data is part of the techniques framed under PETs, which may lead one to think that synthetically generated data is inherently private mistakenly, it has been shown that various generators may have the ability to leak information about real data. Such is the case with [12] and [13].

D. SDG TECHNIQUES

In this section, SDG models are listed in 4 categories that will be used for classifying the literature models. These categories were created to segment the model according to its internal functionality and the general structure used for generating synthetic data.

1) GAN-based Synthetic Generation Models

GAN-based models have been regaining great relevance since their introduction in 2014, by Goodfellow et al. [14]. The original architecture is based on two different artificial neural networks trained simultaneously in a competitive manner. One of them, the generator, represented by G_θ, has the objective of generating the most realistic possible data, where the parameter θ represents the weights of the neural networks, taking an input a Gaussian random variable, with an output $G_\theta(Z)$, where the distribution of $G_\theta(Z)$ is denoted by P_θ. The goal of this network is to choose a θ such that the output $G_\theta(Z)$ has a distribution close to the real data. The second network is called the discriminator, which has the opposite aim of the generator, trying to distinguish the realistic data from the synthetic data the best it can. It is represented by G_ω, parameterized by weights ω. The goal of the discriminator is to assign 1 to the samples from the real distribution P_X, and 0 to the generated samples (P_θ). GANs are represented by the next optimization function:

$$\underset{G}{min}\,\underset{D}{max}\,E_X[\log(D_\omega(X))] + E_Z[\log(1 - D_\omega(G_\theta(Z)))]$$

where E_X is the expected value over all real data instances, and E_Z is the expected value over all random inputs to the generator.

Posterior to the classification of D, the G is trained again with the error signal from D using backpropagation. This equation is the log of the probability of D predicting that the real data is genuine and the log probability of D classifying synthetic data as not genuine. The mathematical representation of the optimal discriminator is:

$$\underset{G}{min} JS(P_x||P_\theta)$$

where JS represents the Jensen-Shannon divergence between the probability of the real data and the probability of the generated data.

Maintaining the basic structure of a GAN, different approaches to this concept have been implemented, to correct different shortcomings of the model, the practical interest, or the data used. Among the different sub-models, we have the WGAN [15] [16]. It was first introduced by Arjovsky et al. [17], to insert different amounts of noise in the training process in the gradients of the discriminator, which improves the stability of learning and provides meaningful learning curves useful for debugging and hyperparameter searches.

Another sub-structure is the DPGAN model, which follows the conventional differentially private mechanisms by introducing deliberate noise to gradients of the Wasserstein distance during the learning process, intending to do gradient clipping and clip solely on weights [18].

DCGAN [19] is another type of GAN, introduced as a type of CNN especially good at learning the hierarchy of representations from objects parts to scenes in both, the generator and the discriminator, where the first becomes fractional-strided convolutions, and the second with strided-convolutions, using a batch norm in both structures.

WaveGAN was based on the previous architecture [20]. It was proposed by Donahue et al. [21], which focuses on generating synthetic audio waveforms in an unsupervised manner. It is trained by minimizing the Wasserstein-1 distance between the distribution of the real and generated data. AC-GAN [22], formulated by Odena et al. [23], adds more structure to the GAN latent space along with specialized cost functions. So, every generated sample has a corresponding class label $c \sim p_c$. The generator uses this label and a z noise vector as an entry, and then the discriminator gives a probability distribution over sources and a probability distribution over the class labels: $P(S|X)$, and $P(C|X) = D(X)$. The objective function has two different parts: the log-likelihood of the correct source L_S and the log-likelihood of the correct class, L_C. D is trained to maximize $L_S + L_C$ while G is trained to maximize $L_C - L_S$.

CTGAN [24] [25], another GAN derivation that stands for Conditional Tabular GAN, addresses some of the key problems for tabular data synthesis, simulating records in the training process one by one, selecting one of the variables, then randomly selecting a value for that variable, and according to that value the algorithm finds a matching row from the training data, also generating the rest of the variables conditioning on the selected one. Finally, PATEGAN [26], which stands for Private Aggregation of Teacher Ensembles GAN.

This mechanism replaces the GAN discriminator with a PATE mechanism so that the discriminator is differentially private, but requires the student version to allow back-propagation to the generator [27]. Since the discriminator is replaced with the PATE mechanism, a set of k teacher-discriminators and a student-discriminator is built. So, teachers are now being trained to improve their loss respect to the generator, the generator is being trained to improve its loss respect to the student, and the student is trained to improve its loss of respect to the teachers.

2) Machine Learning based Synthetic Generation models

The models compiled in this category address those that use a structure based on some machine learning to generate synthetic data. Here, we group some based on Neural Networks (NN), Deep Learning (DL), and linear regression modeling

One of the methods used is K Nearest-Neighbor (KNN). Beigi et al. [28] use this mechanism under the assumption that with tabular data, any strong conditional relationships for a given measurement are only detectable among highly similar individuals, ignoring strong conditions that are only detectable in a few individuals that will have an impact on the overall analysis. Additionally, most of the critical global conditional relationships can be detected with a lower dimension represented by the dimension d. With this in mind, the algorithm takes data from a space embedded by Principal Components Analysis (PCA), and new records are generated from a seed, with their nearest neighbors, and then, the number of attributes is randomly selected from the nearest neighbors to generate a synthetic data point.

Sequential Encoder-Decoder is a structure proposed to learn the mapping between raw real data and their representations in the latent space. The model used by Yoon et al. [29] is combined with a GAN for the generation of synthetic data. Since the data consists of 5 different categories: measurement time, static numerical, static categorical, temporal numerical, and temporal categorical, the encoder-decoder model is trained with a weighted sum of reconstruction losses for each feature category. The reconstruction losses include mean squared error for numerical features, softmax cross-entropy for categorical features, and binary cross-entropy for mask features.

The CART model, which stands for Classification and Regression Trees, is a non-parametric method for generating partially synthetic data. It has been used in [30]–[32]. The CART models, which were formulated by Breiman et al. [33], are highly flexible for estimating the conditional distribution of a univariate outcome given multivariate predictors. The model partitions the space into subsets of units with relatively homogeneous outcomes. The partitions are found by recursively splitting the predictors into binary subsets. A tree structure can effectively represent the series of splits, with the leaves corresponding to the subsets of units. This approach is useful for analyzing complex data sets and identifying the most important predictors for a given outcome [34]. Since this method is non-parametric, it can be used to impute missing data, using leaves of trees as imputation classes, assuming the data are missing at random. To generate partially synthetic data, the imputer selects the values from the observed data that will be replaced with imputations, and then imputes new values, considering the conditional distributions of the variables to be replaced. The imputer controls the disclosure risk and data utility by choosing values to replace, pruning the CART trees, and sampling from the leaves of the trees.

The DP-CSM model proposed by Yao et al. [35] for generating a synthetic set of trajectories uses the original trajectories in a location generalization module. In this module, some coresets are constructed to implement a k-means algorithm and obtain k centers. Then, the model generates a set of candidate trajectories according to the generalized location sets via the set Cartesian product alike operation. The DP-CSM model also employs a trajectory selection module, where a scoring function is defined to measure the similarity between the original and candidate trajectories based on the spatio-temporal features. The model then selects the top-k candidate trajectories with the highest scores as the synthetic trajectories, which preserve the privacy and utility of the original data.

A Variational Autoencoder (VAE) is a type of neural network architecture that can be used to generate synthetic data. It is a variant of the autoencoder, where the central hidden layer is replaced with latent dimensions instead of ordinary nodes. This forces the network to map the input data into one or more normal distributions. The result is that new data can be sampled from the latent dimensions, and will have a similar distribution as the training data, eliminating the need to access the original data in the data reconstruction process [36].

Another DL structure is used by Benarous et al. [37] and Sasada et al. [38], where an extended Long Short-Term Memory (LSTM) is proposed. It is based on RNN while solving the vanishing gradient problem. An LSTM unit is made of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell. The forget gates decide what information to discard from a previous state by assigning a value between 0 and 1 compared to a current input. Input gates decide which pieces of new information to store in the current state, using the same system as forget gates. Output gates control which pieces of information in the current state to output by assigning a value from 0 to 1 to the information, considering the previous and current states. Selectively outputting relevant information from the current state allows the LSTM network to maintain useful, long-term dependencies to make predictions, both in current and future time steps.

A framework based on called Siamese Neural Network (SNN) can be found in [20]. An SNN has the goal of finding similarities between input samples, using at least two identical networks to train the model to identify similar and dissimilar inputs. One common approach is to train the model with triple-loss, using three identical networks. The triple loss is a distance-based function and operates with three inputs:

a as the reference sample (anchor), p as a similar sample relative to a, and a dissimilar one as negative called n. The loss function is defined as:

$$L = \max(d(a,p) - d(a,n) + a, 0)$$

To achieve a clear separation between the two classes, the margin a is introduced. The function is minimized to reduce the distance between the anchor and the positive sample while increasing the distance between the anchor and the negative sample. Additionally, three identical models are used for the three inputs, which share the same architecture and weights

Dandekar et al. [39] used the ridge regression model, a variant of linear regression, to train models that predict the response attribute given the predictor attributes on a private dataset. To generate synthetic predictor attributes for a set of records, they sampled from histograms constructed over predictor attributes in the private dataset. These values were used as inputs for the regression models to generate a synthetic attribute.

3) Statistical-based Synthetic Generation Models

The models compiled in this category are defined by an internal structure based on statistics (Bayesian or frequentist), which learns the underlying patterns and distributions of the data and generates new data samples that meet specific needs or conditions. Here, we compile marginal, stochastic, histogram, and density-based algorithms.

One of the most common methods in this category is Markov chains, which are used by Vie et al. [40] and Benarous et al. [37]. This probabilistic graphical model relies on a probability transition for jumping from one action to another: $P_{su} = P(j_{t+1} = u | j_t = s)$ is the probability of jumping from action s to action u. The Markov Chain is trained on the existing corpus of actions, and once the matrix P is estimated, it can be used to sample random walks from action to action. This model is memoryless since the next actions depend only on the current action: $P(j_{t+1}|j_t, \cdots, j_1) = P(j_{j+1}|j_t)$.

An extension of the previous method is the Variable-order Markov Model (VMM), which defines the states similarly to the standard definitions. In contrast to fixed-order Markov models, where the orders are the same for all positions and contexts, VMMs allow the order to vary for each position based on its context. Thus, VMMs provide the means for capturing both large and small orders, reducing the memory needed to store the model, but requiring an increase in computation time.

Chen et al. [41] used a method for generating synthetic data that preserves the correlation between attributes in a dataset while ensuring privacy. The method involves two modules: a marginal sampling module and a data generation module. The marginal sampling module is used to sample from the original data to obtain two-way marginals. The sampling process is based on mutual information, which is updated iteratively to retain, as much as possible, the correlation between attributes. The data generation module is used to extract the synthetic data from the sampled two-way marginals. This method can be used to generate synthetic data for high-dimensional datasets while preserving privacy and correlation information between attributes.

Also, some works have been made using Bayesian Networks [42], [43]. A Bayesian network over a set of random variables is a way to compactly describe their joint distribution, by specifying conditional independence among certain random variables. Considering a set of attributes $\mathcal{A}$ a fully connected set of attributes, and a set of *attribute-parent* (AP) *pairs*, $\{(A_{k+1}, \Pi_{k+1}), \cdots, (A_d, \Pi_d)\}$, for a certain $k \geq 1$. the fully connected set of attributes and the $(d-k)AP$ pairs in the Bayesian Network $\mathcal{N}$, define a way to approximate $P(\mathcal{A})$ using a joint distribution $P(A_1, \cdots, A_k)$, and $(d-k)$ conditional distributions $P(A_1|\Pi_1), P(A_2|\Pi_2), \cdots, P(A_k|\Pi_k)$. Finally, $P_{\mathcal{N}}(\mathcal{A})$, is expressed as follows:

$$P_{\mathcal{N}}(\mathcal{A}) = P(A_1, \cdots, A_k) \prod_{i=k+1}^{d} P(A_i|\Pi_i)$$

4) Kernel-based Synthetic Generation Models

There are very few works listed in this category. Although it could be considered a subcategory of the previous one, it was decided to elaborate it based on the particularities of these methods. The most common SDG model is based on Kernel Density Estimation (KDE), which is a way of estimating the probability density function of a random variable based on a sample of data points. It works by placing a kernel function, such as a Gaussian, on each data point and summing them up to obtain a smooth curve that approximates the true density function. The kernel density estimate model is defined as:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Where h represents a bandwidth parameter that controls the kernel's width and affects the estimated density's smoothness. A larger bandwidth leads to a smoother density but may oversmooth the data and lose some details. A smaller bandwidth leads to a more detailed density, but may be too noisy and sensitive to outliers; K is the kernel function, and x is a single data point. The KDE method can be applied to univariate or multivariate data and can be used for generating synthetic data by sampling from the estimated density function. Particularly, the works done by Harder et al. [44], Cunningham et al. [45], and Pozi and Omar [46] use this approach. Harder et al. [44]uses random feature representations of kernel mean embeddings. They introduced a model called DP-MERF, which learns the distribution of an unlabeled dataset, by minimizing the random feature representation of the Maximum Mean Discrepancy (MMD), a metric that compares two probability functions in terms of all possible moments. Cunningham et al. [45] adapt the KDE method with a Laplacian Kernel for generating Road Network - and Geography Aware data, and Pozi and Omar [46] used a Gaussian kernel KDE modeling, combined with a feature selection mechanism to determine the dependent and independent variables, and model them according to this nature in two separate cases.

E. PERFORMANCE METRICS

Compiling the resemblance and utility techniques implemented in the different research papers for synthetically generated data, we also carried out a classification based on the general concept, structure, and internal functioning of the metric used, contemplating 3 different categories

1) Statistical analysis

Statistical analysis techniques are employed to explore and interpret inherent patterns, variations, and relationships within datasets. This category includes classical statistical methods such as hypothesis testing (e.g., Welch t-tests [30], Mann-Whitney U [22]), measures of central tendency and dispersion (mean, standard deviation), tests for distributional differences (Kolmogorov-Smirnov [28], [31], [47]–[49]), and correlation measures like Pearson Correlation [18], [28], [50], [51]. Other techniques, such as Kaplan-Meier curves [28], entropy-based techniques [52], or Kullback-Leibler divergence-based techniques [25], aim to reveal the statistical significance of observed phenomena, providing insights into the nature of the data and potential differences between real and synthetic data.

2) Distance and similarity measures

Distance and similarity measures quantify the relationships between data points or distributions. These metrics help assess the dissimilarity or similarity of data instances. Common measures include Euclidean distance [25], [53], cosine similarity [51], cosine distance [38], [54], and statistical distance metrics (e.g., Hellinger distance [16]). These techniques are crucial for understanding the data structure, identifying patterns, and clustering similar data points to detect how far or close the synthetic data is to the real data.

3) Utility evaluation and Machine Learning Models

Utility evaluation metrics encompass techniques that evaluate the general performance of models, algorithms, or data. It involves evaluating how well the utility of data is preserved after applying privacy mechanisms. Metrics such as accuracy [6], [20], [26], [40], [41], [46], [55]–[64], precision [65]–[67], recall [60], [61], [65], F1 score [8], [60], [61], [68]–[71], and AUC [29], [32], [60], [72]–[75] are used to quantify the performance of models or the impact of privacy-preserving techniques. This category is essential for balancing privacy and utility trade-offs in various applications. Some of the models included are neural networks (CNN) [16], [72], Multi-Layer Perceptron (MLP) [16], regression [36], decision trees [36], and KNN [36], [76].

F. PRIVACY PRESERVING DATA METRICS

From a mathematical perspective, a metric is a function that measures the distance between two points in a metric space. A metric satisfies the following conditions: the distance between two points is always non-negative, the distance between a point and itself is zero, the distance between two points is the same regardless of the order in which they are considered, and the distance between three points satisfies the triangle inequality. However, when we speak in the context of privacy metrics we are not referring to a metric in the mathematical sense. It is necessary to keep in mind that there is no consensus on the conditions that a privacy metric should meet, but we will consider for the benefits of generalization the comment proposed by Wagner et al. [77], who in his review establishes a privacy metric as a measure that somehow describes the level of privacy. Even with the level of generalization that is handled, the review proposed by Wagner et al. [77] describes a compilation using authors about the conditions to define a privacy metric, among which are: that the metric be mathematically understandable, that it be orthogonal to the level of cost and utility, that it proposes bounds on how the adversary can effectively succeed in identifying individuals, that it be probability-based, that it returns the number of individuals an adversary cannot distinguish and how variant the adversary's predictions are, that it reflect how difficult it is for an adversary to succeed, and even that it be monotonic with increasing adversary strength. As can be seen, many of these conditions cannot be met for a single metric (e.g. that a metric returns a probability and at the same time returns the number of individuals that the adversary cannot distinguish).

G. PREVIOUS REVIEWS AND WORKS

Some reviews and compilation research have been conducted in the field of SDG models, privacy models, and privacy metrics, which build the taxonomy to identify the categories in these fields and provide a baseline for the construction of articles on the topics in this paper, at the technical and conceptual level. One of the main reviews to consider is by Majeed A. [1]. He described different privacy-preserving data publishing and major tracks of research, proposing a brief taxonomy and explanation of the difference between attribute-centric methods and synthetic-data-based privacy methods. In section 4 of his paper entitled "Discussion on Synthetic Data-Based Privacy Methods," he explored how synthetic data generation and data-based methods have been exploited in recent years. He showed that the mechanisms commonly used in this field are framed in differential privacy, probabilistic modeling, variational autoencoders, generative adversarial networks, neural networks, generative models, pipelines including differential privacy, clustering methods, and transformer models. Privacy is commonly achieved using DP or any other model with synthetic data generation and anonymization.

Another important review to consider is the one presented by Carvalho et al. [10]. They described privacy-preserving techniques for data publishing, describing those used in microdata de-identification, privacy measures suitable for several disclosure types, information loss, and predictive performance measures. They defined the de-identification process as starting from deleting the identifiers to measuring the utility of the data and implementing privacy-preserving techniques. Also, they ensured that the results complied with risk and utility thresholds before releasing the data. Carvalho et al. [10] compile a taxonomy of privacy-preserving techniques, data utility measures, available software, and disclosure attack types.

Boudewijn A. et al. [78] recently published a study that compiles the state-of-the-art and future research directions regarding privacy measurement in Tabular Synthetic Data. They discussed different synthetic privacy risks, properties of statistical privacy indicators, attack mechanisms, and other relevant terms in this field. Their work is relevant in terms of the discussion of the privacy definition properties found, showing that there is no consensus in the choice of the parameters, being one of the causes of the difficulty of choosing appropriate values in practice. Additionally, they proposed several avenues for future research, such as standardizing privacy metrics assessment, outlier protection, and incorporating privacy into generators.

Wagner and Eckhoff [77] presented a systematic review of technical privacy metrics. They compiled the various privacy metrics, taxonomizing them according to their output category, and presented a method for choosing privacy metrics based on 9 questions to help identify the right privacy metric for a particular scenario. Countinho-Almeida et al. [79] conducted a study on current GAN implementations adapted to tabular healthcare data. They focused mainly on the models employed, the datasets used, and the metrics reported on the quality of the data generated in terms of usability, privacy, and how they compare to each other. Monreale et al. [80] discussed advances in privacy-preserving mobility data publishing. They described adversarial attacks and privacy models typically considered for mobility data, as well as frameworks for privacy risk assessment. Ghatak et al. [81] performed a conceptual analysis of different privacy methods, privacy-utility trade-off, and different real applications, focusing on microdata applications. Ficek et al. [82] reviewed the use of differential privacy in the field of medicine. They addressed the concept of privacy utility trade-off, but without focusing primarily on the generation of synthetic data. Tran et al. [2] conducted a study covering both the systematic and multidimensional view of the preservation of privacy of big data analytics in an integrated framework with consideration of different typical practical scenarios. Fung et al. [9] conducted a survey conceptually covering recent developments in privacy-preserving data publishing. They evaluated different approaches to the term, and different study challenges considered the importance of identifying the non-technical difficulties faced by decision-makers in applying such privacy techniques, such as degradation of data/service quality, loss of valuable information, increased costs, and increased complexity. Additionally, various studies, such as those by Endres et al. [83] and Dankar & Ibrahim [84], conduct experiments to draw general conclusions about the advantages and disadvantages of using SDG methods. These experiments are carried out under different contexts, such as data preprocessing techniques and the use of various datasets.

The systematic review conducted in this work aims to provide a current temporal horizon by performing a detailed comparison of the methods used to preserve privacy when synthetic data is generated, categorizing them, and returning a general overview of the state of art applied to different areas of development. To address research gaps in the literature on various SDG techniques and their connection with privacy mechanisms and metrics, we identified and established these relationships. This provides a general overview of the topic and lays the groundwork for future developments in the field.

III. SYSTEMATIC REVIEW PROCESS

In this section, we describe in detail the process of compiling the articles, in order to ensure the reproducibility of the review, and to justify the set of articles selected for subsequent inspection. To carry out this systematic review, the methodology designed by Uman [85] was used as a starting point. First, a search route is established through a strategy that constructs the way to find the publications of interest. Subsequently, a bibliographic search is performed in different databases with the established criteria, to extract the information from the results through a process of reading and synthesis. A summarized graph can be seen in Figure 2.

A. SEARCH STRATEGY

1) Search string and engines

To identify relevant literature, we used a query that included the following keywords: *"((Privacy) OR (Privacy-utility) OR (Privacy utility trade-of)) AND ((Synthetic Data Generation) OR (SDG))"*. We searched for these terms in the title and abstract fields of articles published between January 1, 2018, and October 1, 2023, in the following databases: Web of Science, PubMed, Google Scholar, and Scopus. For the Google Scholar engine, using the equivalent search string, we conducted a pre-screening criterion by selecting only the texts related to the research topic.

2) Selection criteria

The selection criteria for this study are divided into two categories: inclusion and exclusion criteria. The inclusion criteria consist of the following: (1) papers must directly address privacy and utility, exploring concepts, models, or methods related to PET or PPDP in the context of SDG; (2) only peer-reviewed research articles, conference papers, and proceedings will be considered to ensure the quality and credibility of the sources; and (3) papers published in English or other relevant languages that the research team can effectively review will be included. In the exclusion criteria phase, papers with irrelevant topics, no peer review, insufficient information, or not available full text will be eliminated. By using these criteria to select the appropriate studies, we ensure that the sources chosen to contribute to the central theme of the survey, uphold the quality and credibility of the sources included in the analysis, prevent redundancy, and

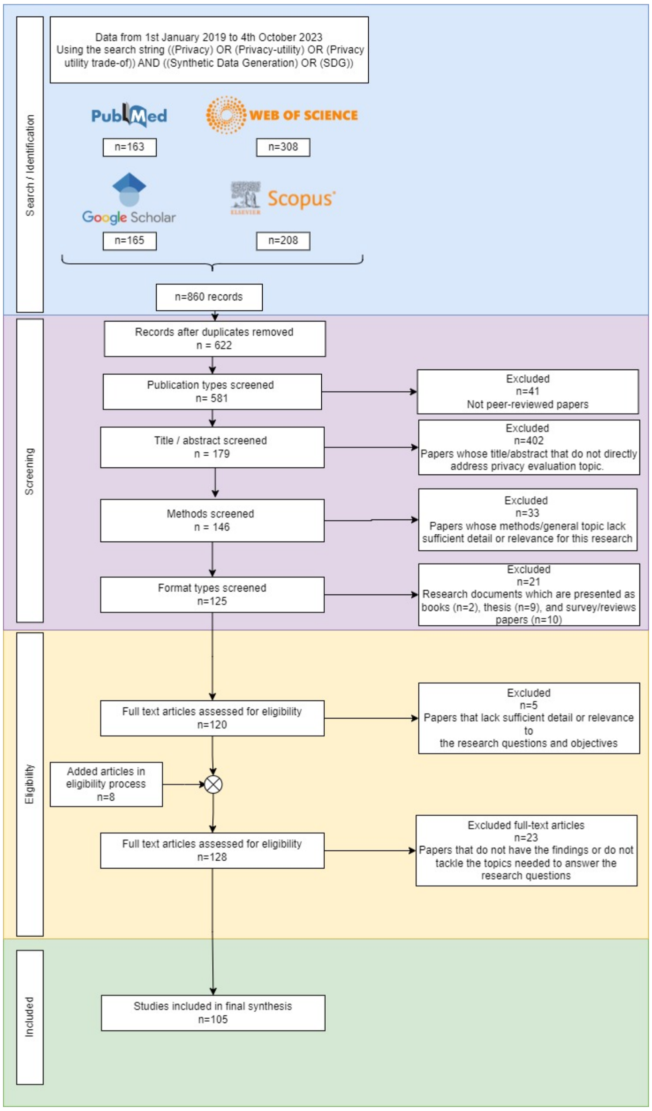
Search / Identification
Data from 1st January 2019 to 4th October 2023
Using the search string ((Privacy) OR (Privacy-utility) OR (Privacy utility trade-of)) AND ((Synthetic Data Generation) OR (SDG))
PubMed
n=163
WEB OF SCIENCE
n=308
Google Scholar
n=165
Scopus
n=208
n=860 records
Screening
Records after duplicates removed
n = 622
Publication types screened
n= 581
Excluded
n=41
Not peer-reviewed papers
Title / abstract screened
n = 179
Excluded
n=402
Papers whose title/abstract that do not directly address privacy evaluation topic.
Methods screened
n = 146
Excluded
n=33
Papers whose methods/general topic lack sufficient detail or relevance for this research
Format types screened
n=125
Excluded
n=21
Research documents which are presented as books (n=2), thesis (n=9), and survey/reviews papers (n=10)
Eligibility
Full text articles assessed for eligibility
n=120
Excluded
n=5
Papers that lack sufficient detail or relevance
to
the research questions and objectives
Added articles in eligibility process
n=8
Full text articles assessed for eligibility
n=128
Excluded full-text articles
n=23
Papers that do not have the findings or do not tackle the topics needed to answer the research questions
Included
Studies included in final synthesis
n=105

FIGURE 2. Flow diagram of the systematic review process

exclude sources with unavailable full text helps ensure that all selected papers can be thoroughly examined and analyzed. A detailed list of the criteria is listed below.

- **Inclusion criteria**
  1) Relevance to Privacy: Papers must directly address privacy and utility, exploring concepts, models, or methods related to PET or PPDP in the context of SDG.
  2) Peer-Reviewed Journals and Conference Proceedings: Consider only peer-reviewed research articles, conference papers, and proceedings to ensure the quality and credibility of the sources.
  3) Language: Include papers published in English or other relevant languages that the research team can effectively review.
- **Exclusion criteria**
  1) Irrelevant Topics: Exclude papers that do not directly address or are not completely related to the privacy evaluation in the context of SDG.
  2) Duplicate Publications: Exclude duplicate papers, conference abstracts, or papers that do not provide substantial additional information compared to other included sources.
  3) Publication Types: Exclude sources that are not peer-reviewed, such as blog posts, magazine articles, non-academic websites, or non-scientific publications.
  4) Insufficient Information: Exclude papers that lack sufficient detail or relevance to the research questions and objectives.
  5) Unavailable Full Text: Exclude sources for which the full text is unavailable or not accessible for review.

B. DATA EXTRACTION

The phase of extracting data to answer the research questions involves the following steps: (1) identifying the relevant studies that meet the inclusion criteria; (2) extracting the necessary data from each study, including publication year, study design, privacy mechanisms used, utility measures, trade-off assessments, and key findings; and (3) summarizing the key findings from each study, focusing on how they address the trade-off between privacy and utility. By following these steps, we can ensure that the data extracted is relevant to the research questions and can be used to answer them effectively.

C. SYNTHESIS OF RESULTS

After completing the data extraction phase, we summarized the key findings according to the comments for solving each research question. Then, we categorized the studies based on their different attributes, considering similarities in privacy mechanisms, utility metrics, or other relevant characteristics. Using these groups, we concluded the studies and discussed the implications for privacy-preserving data generation methods. Finally, we highlighted some limitations and how they affect the generalizability of our findings.

IV. OVERVIEW OF THE SELECTED PUBLICATIONS

In this section we present a general review of the publications selected through the selection process described in the previous section, categorizing the papers in their most global aspects regarding SDG models and performance metrics, as well as the various fields of application in which they are used. As can be seen in Figure 2, a total of 860 publications were collected through search engines, with the later addition of 8 publications added manually, as they had highly relevant content but were outside the search period. A total of 238 duplicates were eliminated to begin the screening process. Then, a total of 41 papers were excluded for not meeting the pre-reviewed publication requirements, to perform a title/abstract screening of the remaining 581 publications. Those publications whose titles and abstracts did not have direct relevance to the topic of privacy were excluded. At the end of this process, a total of 179 remaining works were thoroughly evaluated to determine the quality of their methods and the level of detail that the paper proposed, excluding 33 papers in the process. Finally, the publication format and its presentation were evaluated in the last step of the screening, discarding those publications that were contemplated as books, theses, and surveys/reviews, the latter being evaluated for comparative analysis.

Moving on to the eligibility step, a total of 120 papers were selected for the final stage, where the amount of detail and relevance to answering the research questions is evaluated first, followed by relevant conclusions for the same cause.

In the end, 105 research papers were collected for review, to conduct a thorough analysis and the potential to answer the research questions with them.

Table 1 provides a brief description of the selected papers. Considering that the central objective of this Systematic Review is to evaluate privacy techniques in the context of synthetic data generation, the mentioned table contains a categorized description of the methods used for synthetic data generation, as well as their performance metrics, which include both data utility and resemblance methods. For synthetic data generation methods, a classification was made considering the following categories: GAN-based, ML-based, Statistical-based, and Kernel-based, taking as a reference the review by Hernandez et al. [86]. Performance metrics were classified into statistical analysis, distance and similarity measures, utility evaluation, and machine learning models or techniques. Also, we classified the papers according to their field of development or application, as well as a subcategory based on the method, technique, or sub-area to have a better understanding of the former applications of the research papers. The classification of the papers in table 1 allows a more refined inspection of the papers investigated, and the essential characteristics for the SDG methods and the performance metrics used.

The classification process involved a thorough analysis of the papers, with a focus on the relevance of the content to the research question.

Finally, Figure 3 shows the increase in interest and the number of articles published on the topic presented, noting an evident jump from 2020 to 2021. Nevertheless, there is a decrease in the trend for the period 2022-2023, probably because the compilation of papers was made before the 2023 year ended.

V. DATA PRIVACY TECHNIQUES IN SYNTHETIC DATA GENERATION

The objective of this section is to respond to **RQ1**, which raises the need to know what privacy mechanisms are used in SDG. In Table 2, it is possible to observe in detail which techniques are used in the collection of carried-out works, along with a compilation of the SDG mechanisms that are used together with these techniques. Considering how diverse the techniques can be, they will be stratified and explained according to different categories.

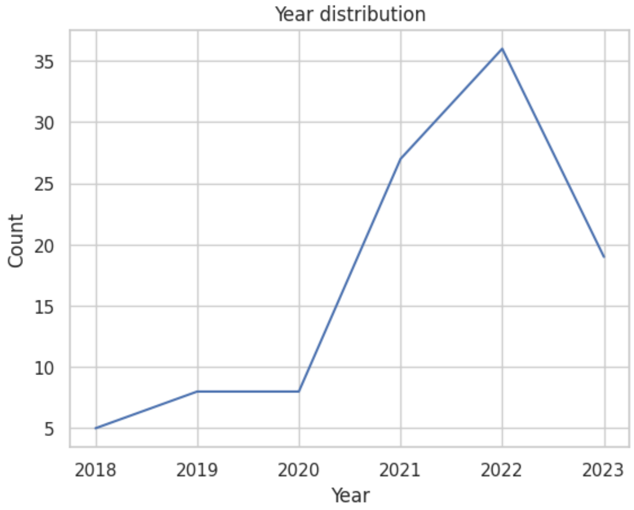

FIGURE 3. Year distribution of the selected publication in the date interval of research

A. DIFFERENTIAL PRIVACY (DP)

Differential Privacy is by far the most widely used privacy mechanism, as its properties place it as a state-of-the-art method. This model is based on a perturbation approach. The objective is to mask the differences in computation results of a function f on neighboring datasets, which differ on at most one data item. DP acquires the intuition that releasing aggregate results should not reveal too much information about any individual data item that contributes to these results [2]. The formal definition considers a randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{\mathcal{X}}$, where $\mathcal{X}$ represent the universe of databases X (collection of records). $\mathcal{M}$ is (ϵ, δ)-differentially private if for all $S \subseteq Range(\mathcal{M})$ and for all $X, Y \in \mathbb{N}^{\mathcal{X}}$ such that $\|X - Y\|_1 \leq 1$:

$$P[\mathcal{M}(X) \in S] \leq exp(\epsilon)P[\mathcal{M}(Y) \in S] + \delta$$

If $\delta = 0$, we say that $\mathcal{M}$ is ϵ-differentially private [134]. Another derivation of DP is Local Differential Privacy (LDP). A privacy algorithm satisfies ϵ-Local Differential Privacy if and only if any input β and β', $\forall s \in Range(\mathcal{M})$ we have

$$\frac{P[\mathcal{M}(\beta) = s]}{P[\mathcal{M}(\beta') = s]} \leq e^{\epsilon}$$

where $Range(\mathcal{M})$ is the value of the algorithm $\mathcal{M}$ area. LDP aims to protect the privacy of individuals' data by controlling the similarity of the results of any two inputs. This ensures that attackers cannot deduce the input data through the result. Unlike classic differential privacy, local differential privacy is less constrained by neighboring datasets. This is because local differential privacy does not require a trusted third party to complete the privacy processing of the data [108].

Renyi Differential Privacy (RDP) is another derivation of the differential private algorithm. Consider the randomized algorithm $\mathcal{M}$ previously defined as $(\alpha, \epsilon) - RDP$ for all the neighbor datasets X and Y if:

$$D_{\infty}(\mathcal{M}(x)||\mathcal{M}(y)) \leq \epsilon$$

TABLE 1. Overview of included publications

Reference	Field	SDG Techniques				Performance metrics		
		GAN based	ML based	Statistical based	Kernel based	Statistical analysis	Distance and similarity measures	Utility evaluation
[54]	Facial Recognition	x				x	x	
[30]	Sports		x					x
[55]	Health-Image analysis	x						x
[68]	Health-Image analysis	x						x
[29]	Health-Tabular records	x	x			x		x
[72]	Health-Image analysis	x						x
[73]	Health-Image analysis	x						x
[87]	Health-Image analysis							
[47]	Health-Tabular records	x				x		
[88]	Health-Tabular records	x				x	x	x
[89]	Health-hybrid							
[28]	Health-Tabular records	x	x	x		x		
[90]	Theoretical development-Method	x	x					
[91]	Health-Tabular records	x					x	
[92]	Health-Tabular records	x						x
[53]	Health-Image analysis	x					x	x
[93]	Compilation							
[15]	Health-Image analysis	x				x	x	x
[94]	Energy	x						
[95]	Health-Signals	x						
[96]	Health-Tabular records							x
[97]	Health-Signals	x					x	
[98]	Health-Tabular records							
[99]	Health-Demographic							
[100]	Health-Image analysis							
[101]	Image Processing	x				x	x	
[102]	Health-Tabular records							x
[22]	Health-Tabular records	x				x		
[103]	Health-Tabular records							
[65]	Recommendation systems		x					x
[35]	Mobility-Trajectory		x			x	x	
[74]	Health-Tabular records	x						x
[36]	Demographic-Method		x					x
[44]	Theoretical development-Method				x		x	
[48]	Social graphs					x	x	
[16]	Health-Image analysis	x						x
[40]	Education		x	x				x
[104]	Theoretical development-Method		x					x
[56]	Mobility-Trajectory	x						
[37]	Geoprivacy-Geospatial location		x	x		x	x	x
[105]	Image processing-Method	x	x			x	x	
[8]	Health-hybrid(Signals, Tabular Data)	x				x		x
[106]	Energy							
[107]	Theoretical development-Method	x				x		
[108]	Graphs-Method			x		x		x
[57]	Health-Tabular records	x						x
[109]	Trajectory		x			x		x
[58]	Theoretical development-Method			x		x		x
[49]	Health-Tabular records			x		x	x	x
[38]	Generative text model-Method		x				x	
[50]	Mobility-Trajectory	x				x		
[110]	Theoretical development-Method							x
[59]	Image Processing		x					x
[111]	Graphs -Method					x		
[75]	Health-Tabular records			x			x	x
[112]	Health-Signals	x		x				x
[113]	Theoretical development-Method	x						
[18]	Theoretical development-Method	x				x		x
[114]	Theoretical development-Method			x				x
[76]	Theoretical development-Method			x				x

Reference	**Field**	**SDG Techniques**				**Performance metrics**		
		GAN based	**ML based**	**Statistical based**	**Kernel based**	**Statistical analysis**	**Distance and similarity measures**	**Utility evaluation**
[115]	Image Processing		x					x
[116]	Theoretical development-Method			x		x		x
[41]	Theoretical development-Method			x				x
[25]	Theoretical development-Method	x				x	x	
[66]	Theoretical development-Method							x
[117]	Theoretical development-Method	x				x	x	x
[60]	Theoretical development-Method	x						x
[118]	Image Processing-Forensic	x						x
[67]	Health-Sharing Data System	x				x		
[51]	Trading-Financial Market		x			x	x	
[119]	Theoretical development-Method			x		x		x
[120]	Trajectory-Vertical partitioning method			x			x	x
[69]	Mobile computing-Location			x		x		x
[121]	Theoretical development-Method			x				x
[122]	Image Processing	x					x	
[123]	Energy							
[70]	Theoretical development-Method			x				x
[42]	Theoretical development-Method			x				x
[43]	Theoretical development-Method			x			x	x
[45]	Trajectory-Location			x	x		x	x
[61]	Economy-Method		x	x				x
[62]	Theoretical development-Method			x				x
[71]	Theoretical development-Method			x				x
[124]	Theoretical development-Method			x				
[31]	Taxes-Policies		x			x		
[63]	Theoretical development-Method			x				x
[125]	Theoretical development-Method							
[126]	Theoretical development-Method			x				
[127]	Framework -Method		x				x	x
[20]	Health-Signals	x	x					x
[24]	Theoretical development-Method	x	x					
[128]	Theoretical development-Method			x		x		x
[129]	Theoretical development-Method			x				
[32]	Microbiology		x	x				x
[64]	Image Processing	x				x		x
[52]	Theoretical development-Method			x		x		x
[130]	Theoretical development-Method			x				
[131]	Theoretical development-Method			x				
[132]	Health-Signals	x				x		
[46]	Theoretical development-Method				x	x		x
[26]	Browser Fingerprinting	x	x					x
[39]	Theoretical development-Method		x					x
[133]	Theoretical development-Method							
[6]	Scenario evaluation -Method							x
[7]	Health-Attacks	x						

where D_∞ refers to the Rényi divergence.

DP-SGD is a common technique that stands for differentially private stochastic gradient descent. Since DL models are trained by minimizing some loss function $f(X;\theta) := \frac{1}{m}\sum_{i=1}^{m} f(x_i;\theta)$ on a dataset $X = \{x_i \in \mathbb{R}^n\}_{i=1}^{m}$, a usual method to find the optimal value is to perform stochastic gradient descent (SGD) on a batch B of sampled data points iteratively:

$$B \leftarrow BATCHSAMPLE(X)$$

$$\theta \leftarrow -\eta \frac{1}{|B|} \sum_{i \in B} \nabla_\theta f(x_i, \theta)$$

To make SGD private Abadi et al. [135] proposed to first clip the gradient of each sample to ensure the l_2-norm at most C:

$$CLIP(x, C) := x \min\left(1, C/\left\|x\right\|_2\right)$$

To make a private quantity g that can be used for the descent step $\theta \leftarrow \theta - \eta \cdot g$ we estimate using a multivariate Gaussian noise parametrized by ψ:

$$g \leftarrow \frac{1}{|B|}\left(\sum_{i \in B} CLIP(\nabla_\theta f(c_i, \theta), C) + \mathcal{N}(0, C^2\psi^2 I)\right)$$

Jorgenser et al. [136] introduce the term Personalized differential Privacy (PDP), considering that not all users require the same level of privacy. Considering a privacy budget $\mathcal{S} = \{\epsilon_1, \cdots, \epsilon_n\}$, a set of users $\mathcal{U} = \{u_1, \cdots, u_n\}$. A randomized mechanism $\mathcal{M}$ which satisfies $\mathcal{S}$-PDP if for any pair $X, Y \in \mathcal{D}$ which differs in one arbitrary user u_i, and for all sets $O \in R$ of possible outputs

$$P[\mathcal{M}(X) \in O] \leq e_i^{\epsilon} P[\mathcal{M}(Y) \in O]$$

There are other combined mechanisms, or applied to specific fields, that have been implemented based on the concept of DP. Such is the case ϵ-geo-indistinguishability [56], which introduces the use of distance between two objects to make it applicable to protect users' location, or $(k-\epsilon-\delta)$ anonymization, which improves the privacy of datasets by implementing DP and k-anonymization using KD tree and random sampling mechanisms.

In Table 4, it is possible to observe a subcategorization of the citations of the review that use DP as a model to ensure privacy. As given by definition, different mechanisms are used to ensure that the differential privacy rate is met. Those mechanisms work by adding random noise to the data, which masks the identity of individuals while still providing useful information.

1) DP mechanisms

A randomized query mechanism $\mathcal{M}$ for a query function will randomly output a number with a probability distribution. A necessary definition of these types of mechanisms is the concept of sensitivity. l_1-sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$ is:

$$\Delta f = \max_{\substack{x,y \in \mathbb{N}^{|\mathcal{X}|} \\ \|x-y\|_1 = 1}} |f(x) - f(y)|_1$$

This function captures the magnitude by which a single individual's data can change the function f in the worst case [134]. among the different DP mechanisms, the most used are the Laplace mechanism, the Gaussian Mechanism, the exponential mechanism, the geometric mechanism, and the staircase mechanism.

- **Laplace mechanism [134]**
  Consider the Laplace distribution centered at 0, with scale b, defined by:
  $$Lap(x|b) = \frac{1}{2b} exp\left(-\frac{|x|}{b}\right)$$
  with a variance $\delta^2 = 2b^2$. Given a function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$, the Lapalce mechanism is defined as
  $$\mathcal{M}_L(x, f(\cdot), \epsilon) = f(X) + (Y_1, \cdots, Y_k)$$
  where Y_i are i.i.d random variables dram from $Lap(\Delta f/\epsilon)$.
- **Staircase mechanism [35]**
  For a given function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$, the Staircase mechanism adds noise to the query result:
  $$\mathcal{M}_s(x, f(\cdot), \epsilon) = f(x) + g(\Delta f, \epsilon)$$
  where $g(\Delta f, \epsilon)$ is the noise generator sampling noise from a staircase-shaped probability distribution, which is defined as:
  $$h_\gamma(X, \Delta, \epsilon) = \begin{cases} e^{-j\epsilon}\alpha(\gamma) & \text{if } \|X\|_1 \in [j\Delta f, (j+\gamma)\Delta f] \\ e^{-(j+1)\epsilon}\alpha(\gamma) & \text{if } \|X\|_1 \in [(j+\gamma)\Delta f, (j+1)\Delta f] \end{cases} \tag{1}$$
  where $j \in \mathbb{N}$, and $\gamma \in [0, 1]$, $\alpha(\gamma)$ is a normalization factor to make
  $$\int\int \cdots \int_{\mathbb{R}^k} h_\gamma(x) dx_1 dx_2 \cdots dx_k = 1$$
  defined as
  $$\alpha(\gamma) = \frac{k!}{2^k (\Delta f)^k \sum_{j=1}^{k} \frac{k!}{j!(k-j)!} c_{k-j}(b + (1-b)\gamma^j)}$$
  where $b = e^\epsilon$, $c_j = \sum_{i=0}^{+\infty} i^j b^i$ and $\gamma = \frac{1}{1 + e^{\epsilon/2}}$.
- **Gaussian mechanism [104]**
  For a given function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$, the Gaussian mechanism is to add noise to the query result:
  $$\mathcal{M}_{Ga}(x, f(\cdot), \epsilon) = f(x) + z$$
  where z is the noise generator sampling noise from a Normal probability distribution $\mathcal{N}(0, \sigma^2 \Delta f^2)$, which is defined as:
  $$\mathcal{N}(0, \sigma^2 \Delta f^2) = \frac{1}{\sigma \Delta f \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma \Delta f}\right)^2}$$
  This mechanism holds the condition
  $$\sigma^2 \epsilon^2 \geq 2 \ln(1.25/\delta) \Delta f^2$$
- **Geometric mechanism [137]**
  The geometric mechanism for a given function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$ defined as
  $$\mathcal{M}_{Ge}(x, f(\cdot), \epsilon) = f(x) + m$$
  where m is a random variable distributed by a two-sided geometric distribution $P(M = m) = \frac{1-\alpha}{1+\alpha}\alpha^{|m|}$.
- **Exponential mechanism [66]**
  This algorithm uses a score function to determine the quality of a query result. It then generates a query result with a probability that is proportional to an exponential function. Specifically, given the output range $\mathbb{R}^k$ and a query score function $f(x, r)$, where r is the output of a query function on a dataset x, this algorithm outputs $r \in \mathbb{R}$ with a probability proportional to $exp\left(\frac{\epsilon f(x, r)}{2\Delta f}\right)$.

B. ENCRYPTION TECHNIQUES

These techniques use mathematical algorithms to transform data into an unreadable format unless decrypted. Among the techniques used in this category, we have encryption, Homorphic encryption, and pailler cryptosystem. Encryption is used by Khowaja et al. [68] with a spike learning technique for model weights by representing the spatial domain data into the temporal axis, making it resilient against leakage attacks. Homomorphic encryption is also used for securely generating synthetic V-histograms over distributed datasets in [103]. Pailler cryptosystem is also used in [123], with two Horner parameters by leveraging homomorphism, for line-loss calculation in Residential Areas.

C. ANONYMIZATION TECHNIQUES

This category involves techniques that modify data to prevent the identification of individuals whose information is included in the dataset.

k-anonymity model is used against record linkage in which the adversary can uniquely identify the victim's record with the help of some additional information. This measure states that any released data item must be linked to at least k individuals with equal probability [138]. Let $T\{A_1, \cdots A_n\}$ be a table which represents a structured dataset. A QID of T is a set of attributes $\{A_i, \cdots, A_j\} \subseteq \{A_1, \cdots, A_n\}$ whose release must be controlled. Consider the tuple $t \in T$ where $t|A_i, \cdots, A_j|$ means the sequence of values $A_i, \cdots, A_j$ in t, and $T(A_i, \cdots .T)$ is a projection, maintaining the duplicate tuples of attributes $A_i, \cdots, A_j$ in T. Finally, $Q|_T$ denotes the set of quasi-identifiers associated with T, and $|T|$ denotes the number of tuples in T [139]. Each data release must be such that at least k individuals can indistinctly match every combination of quasi-identifier values. So, with a table $T(A_1, \cdots, A_n)$ and the quasi-identifiers associated $Q|_T$, T is said to satisfy k-anonymity if and only if for each quasi-identifier $QI \in Q|_T$ each sequence of value in $T[QI]$ appears at least with k occurrences in $T[QI]$.

l-diversity can deal with the main problem of k-anonymity, which considers that if all k values with the same quasi-identifiers also have the same value for the sensitive attribute, k-anonymity can be met while still revealing the sensitive value for an attribute to an adversary [138]. Let us define a q*-block to be the set of tuples in T* whose nonsensitive attribute values generalize to q*. A q*-block is l-diverse if contains at least l *well represented* values for the sensitive attribute S. A table is l-diverse if every q*-block is l-diverse [140]. Since *"well represented"* values can be a non-objective term, Machanavajjhala et al. [140] define it in three possible ways:

- **Distinct l-diversity**: At least l distinct values for the sensitive field in each equivalence class exist.
- **Entropy l-diversity**: A table is entropy l-diverse if for every block q*-block
$$-\sum s \in Sp_{(q*,s)} log(p_{(q*,s')}) \geq log(l)$$
where
$$p_{(q*,s)} = \frac{n_{(q*,s)}}{\sum_{s' \in S} n_{(q*,s')}}$$
represents the fraction of tuples in the q*-block with sensitive attribute value equals to s, and $p_{(q*,s)}$ is the fraction of records in the q*-block which has the sensitive value s.
- **Recursive $(c - l)$-diversity**: This definition ensures that the most common value does not appear too often while less common values are ensured to not appear too infrequently. Basically, in a given q*-block, let r_i denote the number of times the i^{th} most frequent sensitive value appears in that q*-block. Given a constant c, the q*-block satisfies recursive (c, l)-diversity if $r_1 < c(r_l + r_{l+1} + \cdots + r_m)$. A table T* satisfies recursive $(c - l)$-diversity if every q*-block satisfies recursive l-diversity.

A final approach from l-diversity is t-closeness. An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the attribute distribution in the whole table is no more than a threshold t. Table T* is said to have t-closeness if all equivalence classes have t-closeness. There are some metrics to measure the distance between two probabilistic distributions. Given two distributions $P = (p_1, p_2, \cdots, p_3)$ and $Q = (q_1, q_2, \cdots, q_3)$, the *variational distance* is defined as:

$$D[P, Q] = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i|$$

and the *Kullback-Leibler* (KL) distance is defined as

$$D[P, Q] = \sum_{i=1}^{m} p_i log \frac{p_i}{q_i} = H(P) - H(P, Q)$$

where $H(P) = \sum_{i=1}^{m} p_i \log p_i$ is the entropy of P and $H(P, Q) = \sum_{i=1}^{m} p_i \log q_i$. However, Ningui et al. [141] say those distances cannot reflect the semantic distance among values. They proposed using Earth Mover's distance, which is based on the minimal amount of work needed to transform one distribution to another by moving distributions mass. This can be formally defined as a function to minimize:

$$WORK(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}$$

where d_{ij} is the ground distance between element i of P and element j of Q. We want to find a flow $F = [f_{ij}]$ where f_{ij} is the flow of mass from element i of P to element j of Q. The optimization function is subject to:

$$f_{ij} \geq 0 \quad 1 \leq i \leq m, 1 \leq j \leq m$$

$$p_i - \sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m} f_{ji} = q_i \quad 1 \leq i \leq m$$

$$\sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1$$

D. INFORMATION THEORETIC TECHNIQUES

These techniques are based on quantifying and managing information gain and loss. One of them is Mutual information, used in [51] as a part of the generator, which in the training process attempts to privatize the input data, and the discriminator attacks the generator by maximizing the mutual information between the actual sensitive attributes and those retrieved from released data. Sensitivity is another technique used in [65], defined to describe the privacy guarantee for the original item at the item level, by calculating the relative similarity between the original data and the original data set, and it is used as part of a privacy regularizer. Given an original data $i \in \mathcal{I}_u$ and the synthetic data $v \in \mathcal{V}_u$, and q a feature vector of a particular user, it is defined as:

$$f_{sim}(q_i, q_v) = \frac{q_i^T q_v - \min(q_i)}{q_i^T q_i - \min(q_i)} \tag{2}$$

where $\min(\cdot)$ is the cosine similarity between item i and the item which is the most insensitive one in the item set $\mathcal{I}$. For a synthetic item to meet the user's privacy in terms of sensitivity, sensitivity must have a boundary $f_{sim}(q_i, q_v) \leq \gamma$.

E. BLOCKCHAIN TECHNOLOGY

The techniques in this category utilize decentralized and distributed digital ledgers to ensure transparency and security in transactions. Chen and Huang [51] use the blockchain mechanism to ensure access control and security, with blockchain smart contracts to store and verify user permissions that regulate access to smart patient records and also guarantee that only authorized agents can access the data, and depending on permissions, possibly only part of the data. It provides a programmable, distributed ledger with an immutable history of transactions, which means that the data stored in the blockchain cannot be altered or modified by a single organization or node. The blockchain also requires consensus among the participating organizations before any transaction can be written on the ledger, which prevents unauthorized or malicious actions.

F. OTHER SPECIFIC OR COMBINED METHODS

These are either specific methods for certain types of data or combinations of various techniques for enhanced privacy preservation.

Multi-Aspect Trajectory Classification (MARC) [50] is a method that can be used as a machine-learning model for user identification based on trajectory data, aiming to minimize the categorical cross-entropy loss provided by:

$$-\frac{1}{\mathcal{N}_{train}} \sum_{T \in \mathcal{T}_{train}} \sum_{L \in \mathcal{L}} 1_{T \in \mathcal{T}_{train}} \cdot \log p[T \in L]$$

where $\mathcal{T}_{train}$ is the set of trajectories used to train the model, $\mathcal{N}_{train}$ is the number of training instances, and $\mathcal{L}$ is the set of labels used to classify the trajectories.

Kuppa et al. [24] propose an Instance Level Privacy Coefficient for generated synthetic data to avoid MIA-type attacks. Given X_r and X_s, real and synthetic data set generated by some algorithm G, latent vectors in embedding space $Em(X_r), Em(X_s)$, the spherelet distance between $s(X_i)$ and $s(X_j)$ is measured:

$$d_S = d_R(X_i, X_j) + d_{Eu}(X_i, X_j)$$

where d_{Eu} is the euclidean distance, and d_R is Riemanian Geometric Divergence. With this, the distance between points by projecting the samples onto a sphere centered at c with radius r is measured. The Spherelet of X is denoted by $s(X) = S(V, c, r)$ where V determines an affine subspace the sphere lies in. The spherical error is defined by [24]:

$$\epsilon(X) = \frac{1}{n} \sum_{i=1}^{n} (||x_i - c|| - r)^2$$

If x_i lie on the sphere then $\epsilon(X) = 0$. It is possible to see which points are found in the data manifold with low probability, and how close the synthetic and real in nearest neighbor proximity, which can give us a proxy measure for over-fitting and data memorization [24], since a problem with synthetic data generators is data memorization from the model itself. With this consideration, since this model works on the data level can be used to filter synthetic data which does not accomplish privacy.

Subsample-and-aggregate is also used. It randomly splits a dataset X into k blocks of equal size, and then, a query function f is applied to each block to obtain a query result, which is an estimation of the query function over the entire dataset. A total of k estimations are produced and afterward aggregated using a deferentially-private aggregation function [66]. The aggregated result is used as the perturbed result for the query function.

Lyu et al. [54] implements a Makeup generation algorithm that proposes the perturbation of face images from transfer to render makeup in source faces according to the style of reference images while keeping the identity information intact. This is done by transferring the adversarial makeup in the UV space, using an encoder-decoder architecture as its backbone.

Vietri et al. [71] proposed a method to iteratively construct synthetic data by repeatedly selecting queries on which the SD currently represents poorly using DP with a Gaussian Mechanism and a private selection mechanism, called One-shot Report Noisy Top-K With Gumbel Noise. It takes as input a dataset $X \in \mathcal{X}^n$ with n rows, and a synthetic dataset $\hat{X} \in \mathcal{X}*$, a set of m statistical queries $Q = \{q_1, \cdots, q_m\}$ and a parameter ρ. Adding Gumbel noise to the queries:

$$\hat{y}_i = \left| q_i(X) - q_i(\hat{X}) \right| + Z_i$$

where $Z_i \sim$ Gumbel $(K/\sqrt{2}\rho n)$

Let $i_{(1)}, \cdots, i_{(i)}$ be an ordered set of indices such that $\hat{y}_{i_{(1)}}, \geq, \cdots, \geq \hat{y}_{i_{(i)}}$. The algorithm outputs the top-K indices $\{i_{(1)}, \cdots, i_{(k)}\}$ corresponding to the K queries where the answer to the synthetic and real data differs the most.

VI. DATA PRIVACY METRICS IN SYNTHETIC DATA GENERATION

The objective of this section is to answer **RQ2**, which raises the need to explore the extent to which different privacy metrics have been used in a framework of synthetic data generation to evaluate their effectiveness and quality. The metrics, listed in the Table 3 can then be stratified as follows:

TABLE 2. Privacy mechanisms in literature and its(their) associated SDG model(s).

Privacy Mechanism	Associated SDG technique	References
Encryption	GASCNN	[53], [68]
DP	2-way margins, 3-way margins, AC-GAN, AsgLDP*, Autoencoder, Bayesian networks, BGAN, CART, Copula Based, CopulaGan, Copula-Shirley, CTGAN, Density-Aware Grid, Discriminative-Generative Distillation, DP latent tree (DPLT), DP-auto-GAN, DP-BLSGD, DP-CGAN, DPCopula, DP-CSM, DP-CTGAN, DP-GAN, DP-Hflow, DP-MERF, DPView, DP-WGAN, DTG (RNN gradient descent algorithm with differential privacy), Encoder-decoder + Long Short-Term Memory, Extended short-term memory networks (LSTMs), FL-GAN, Fully Bayes Model, GAN, Gaussian Copula, GEDDP, GEM, GEP, G-PATE, GS-WGAN, Iterative proportional fitting with target of marginal distributions, Kernel Density Estimation (KDE), KNN, low-dimensional marginals, Markov chains (MC), Masked Autoregressive Flow, Mobility model, MST (Private-PGM :Query selection with marginals), MWEM, Normalizing flows, PART-GAN, PATE-CTGAN, PATE-GAN, PEP, Posterior inferences using k-way margins, Posterior Predictive Distribution (PPD), PPDU*, PPGAN, Private sampling, PrivSyn(Marginal selection + Noise addition + postprocessing + Graduate update method for synthesis), Pseudo posterior mechanism, QUAIL Method, r-anonymous microaggregation, RDP, Relaxed Adaptive Projection, Ridge regression, Route length distribution, Simulation based, Synthetic Data Vault, Trip Distribution, TSADP, TVAE, Variable-order Markov models (VMMs), Variational Autoencoder, Weighted Uniform Distribution (WUD), WGAN, Wind Power Obfuscation* + Transmission Capacity Obfuscation*, zCDP	[15], [16], [18], [22], [25], [26], [28], [35]–[39], [42]–[48], [57]–[59], [61], [63], [64], [66], [67], [69]–[71], [76], [87], [89], [90], [96], [99]–[107], [109], [111]–[117], [119]–[121], [124]–[126], [128]–[131]
epsilon-Geo-indistinguishability (DP)	GAN	[56]
Multi-Aspect Trajectory Classification (MARC)	GAN	[50]
(k- $\epsilon - \delta$)- Anonymization	—	[110]
Hashing for SI	CTGAN	[60]
Mutual Information	PPGAIN	[51]
l-diversity	CART	[31]
k-anonimity	WaveGAN, Siamese Neural Networks	[20]
Instance Level Privacy Coefficient	CTGAN, WGAN-GP, PATEGAN, DPGAN, RVAE	[24]
Makeup attack mechanism	3DAM-GAN	[54]
Homomorphic encryption	—	[103]
Sensitivity	UPC-SGD	[65]
kRR	2-ways marginal	[41]
t-closeness	CTGAN	[60]
Pailler cryptosystem	—	[123]
Statistical Disclosure	Ridge regression	[39]
Blockchain	—	[67]
One-shot Report Noisy Top-K With Gumbel Noise	Relaxed Adaptive Projection	[71]
Subsample-and-aggregate algorithm	—	[66]

A. DISTANCE METRICS

These metrics are used to measure the similarity between two data points. In the context of privacy-preserving, these metrics allow us to calculate how close is the original dataset compared to the synthetic dataset. According to the data structure, the data usage of the different techniques may vary according to their dimensionality, and also to their nature. Metrics found in this review include Hamming distance, Euclidean distance, Wasserstein distance, Hausdorff distance, Closet distance, and distance.

B. PERFORMANCE METRICS

Generally, these metrics are focused on calculating the performance of an attack model on a dataset. These metrics include accuracy, which identifies when an attacker has a successful prediction, misclassification cost, a ROC curve of private vs non-private implementation, and correct attribution probability.

These attacks aim to extract sensitive information from a dataset without authorization and measure the efficiency of the algorithm according to its capacity to keep the data private. Among the different privacy attacks, we can identify Minimum difference attacks, Membership inference attacks, Re-Identification attacks, Attribute inference attacks, outlier leakage attacks, and also pairwise attacks based on distance. In this category, we can also enclose the adversarial attacks, which aim to fool a model by manipulating the input data, such as Nearest Neighbour Adversarial Accuracy, Dominant set clusters, and Model inversion attacks.

C. PRIVACY BASED METRICS

These metrics are metrics that come from privacy models and measure their particular performance. Here we can include ϵ-identifiability, ϵ-privacy, Privacy Gain, Differential budget, and Privacy loss.

D. OTHER METRICS

These metrics do not fit into any of the above categories. Examples include Linear regression, @K score, Cumulative privacy loss, Outliers defense, RNNR, GCAP, K-NN, Memorization coefficient, Replicated uniques, G_ϵ, Distance evaluation, Shadow-modeling, RTS Similarity, Privacy budget, GroundHog, Attack models, Reverting SNN to ResNet attack, Synthetic Predictor, Generating class representation attack, Kernel density, Replacement ratio, Likelihood estimation, and Generative representation learning

E. TRADE OFF METRICS

Various metrics have been used not only to measure the privacy of data but also to see how privacy models alter the quality and performance of synthetic data when used to simulate, model, predict, forecast, or classify. Many of the models that use DP in their base use differential budget to analyze the trade-off, putting the ϵ rate used to generate differentially private data on one axis and a quality performance measure, usually accuracy, F1-score, or AUC, on the other axis. Fung [9], while considering anonymization for classification tasks, proposes a trade-off metric between privacy and accuracy (utility). The score proposed by him is defined as:

$$Score(v) = \frac{InfoGain(v)}{PrivLoss(v) + 1}$$

which works as a selection criterion for guiding their top-down refinement process to heuristically maximize the classification goal, with a *refinement v*.

To provide a privacy-utility trade-off strategy, Liu et al. [65] employs a privacy regularizer to constrain the differences of relative similarity between the original and synthetically generated data. From Eq. 2, the follow privacy regularizer is defined:

$$\mathcal{L}_s = \sum_{u,i,v} [f_{sim}(q_i, q_v) - \gamma_u]_+$$

where $[z]_+ = max(z, 0)$ refers to the standard hinge loss. and γ_u indicates the sensitivity for an user u, adopted as a safety margin. They also define the following utility regularizer:

$$\mathcal{L}_g = \sum_{(u,v)} - \ln \delta(p_u^T q_v)$$

Since their purpose is to create a recommendation system, the utility is maximized with the effectiveness of a recommendation of a n item V to the user u, being p the feature vector of the user, and q the feature vector of the item v. Finally, the loss function is formulated as:

$$\mathcal{L}_I = \lambda_s \mathcal{L}_S + \lambda_g \mathcal{L}_g$$

where λ_S and λ_g are hyperparameters that control the weights.

Galloni and Lendák [131] introduce an evaluation framework for evaluating not only a privacy guarantee (ϵ) but also macro-statistics and data utility. It is defined as:

$$G_\epsilon = \alpha\mu(X_s, X_p) + \beta\delta(X_s, X_p)$$

where α and β are weights, X_s is a synthetic data set, and X_p is a private dataset. The macro statistics measure is defined as follows:

$$\mu(X_s, X_p) = \frac{\|\phi_k(X_s) - \phi_k(X_p)\|_2}{m(m-1)/2}$$

where m is the number of features, and ϕ_k is a correlation coefficient defined in [142]. On the other hand, the utility measure δ is defined as

$$\delta(D_s, D_p) = \frac{1}{mKL} \textstyle\sum_{i=1}^{m} \sum_{k=1}^{K} \sum_{l=1}^{L} ||\mathrm{acc}^l(M_{X_{i,s}}^k) - \mathrm{acc}^l(M_{X_{i,p}}^k)||_2$$

where m is the number of machine learning tasks, K is the number of different Machine Learning Models, and L is the total number of different Accuracy Scores (or any other metric). So that, $M_{X_{i,s}}^k$ refers to the k-th ML model optimized using the data X_s and the attributes $X_{1,s}, X_{2,s}, \cdots, X_{i-1,s}, X_{i+1,s}, \cdots, X_{m,s}$ to predict the attribute $X_{i,s}$

VII. FIELDS OF INTEREST

The objective of this section is to answer **RQ3**. Figure 4 shows the distribution of the different applications found in this study, defined in the synthesis of results phase in the systematic review process, according to the main field of application, keywords, and the focus of the methodology in the paper. A clear result in terms of the number of articles in the health and theoretical development category can be observed. The vast majority of the works in the theoretical development category propose different techniques or methodologies as the main findings in their articles. In these, even presenting examples, they are generalizable to any other field of interest that adapts to the data structure and the models they propose. On the other hand, it is possible to observe in figure 5 the subfields in which the health field is segmented. Other fields of interest are energy and image processing applications.

VIII. DISCUSSION

The systematic review presented in this document analyzes and compiles advances in privacy techniques and metrics in the context of synthetic data generation. The review's findings provide evidence of the efforts made by different

TABLE 3. Privacy metrics in literature and its(their) associated SDG model(s).

Unique values metrics	Associated SDG technique	Reference
Minimum difference attack	GASCNN	[68]
Membership inference attack	GAN, Gaussian Multivariate, Synthetic Data, CTGAN, WGAN, WGANGP, Markov chains, IRT, Variational autoencoder, WGAN-GP, PATEGAN, DPGAN, RVAE, Sequential encoder-decoder networks	[7], [24], [29], [40], [59], [87], [88]
Pairwise attacks	ac-GAN, pGAN	[73]
Hamming distance	DP-CGANS, MST, Bayesian Networks	[47], [125]
Euclidean distance	DP-CGANS, Gaussian Multivariate, Synthetic Data Vault, CTGAN, WGANGP	[47], [88]
Linear regression	DP-CGANS, KNN, Gaussian Copula, CopulaGAN, TVAE, CTGAN	[28], [47]
ϵ-identifiability	ADS-GAN	[91]
Nearest Neighbour Adversarial Accuracy	PGAN, GAN	[74], [92]
Disaggregation using NILM	GAN, hybrid-GAN	[94]
ϵ-privacy	-	[99]
Re-Identification attack	Sequential encoder-decoder networks, GAN, Markov chains, IRT	[29], [40], [97]
Wasserstein distance	FL-GAN, Federated Learning	[105]
Similarity	Gaussian Multivariate, Synthetic Data Vault, CTGAN, WGANGP, Bayesian networks	[49], [88]
@K score	GAN	[50]
Tachycardia Privacy	MWEM, DP-CTGAN, PATE-CTGAN	[112]
Attack: Dominant set clusters	Variational Autoencoder	[115]
Cumulative privacy loss	Masked Autoregressive Flow	[116]
Earth Mover's Distance (EMD)	CTGAN	[60]
Bozorth3 (similarity score)	StyleGAN, ProgressiveGAN	[118]
Outliers defense	(Density-Aware Grid + Trip Distribution + Mobility model + Route length distribution)	[69]
RNNR	Bayesian Networks	[43]
GCAP (Generalized correct attribution probability)	Bayesian Networks	[62]
Closet-distance (Hamming distance, LP-distance)	MST, Bayesian Networks	[125]
K-NN	DP-CGANS, WaveGAN, Siamese Neural Network, CART, Synthetic Data Vault	[6], [20], [32], [47]
Memorization coefficient	CTGAN, WGAN-GP, PATEGAN, DPGAN, RVAE	[24]
Replicated uniques (ru) (Disclosure risk)	2-way margins, 3-way margins	[130]
G_ϵ	Bayesian Networks, Copula-Shirley, DPCopula	[131]
Utility : Accuracy	Variational Autoencoder	[59], [132]
Leading bit attack	GASCNN	[68]
Distribution attack	ac-GAN, pGAN	[73]
Hausdorff distance	Gaussian Multivariate, Synthetic Data Vault, CTGAN, WGANGP	[88]
Missclasification cost	-	[99]
Differential budget	FL-GAN, Federated Learning	[105]
Distance evaluation	Bayesian Networks	[49]
Privacy loss	Encoder-decoder + Long Short-Term Memory	[38]
ROC (private vs non-private)	Masked Autoregressive Flow	[116]
Correct attribution probability (CAP)	Bayesian Networks	[43]
Shadow-modeling	MST, Bayesian Networks	[125]
Model inversion attack	GASCNN	[68]
Attribute inference attack	Gaussian Multivariate, Synthetic Data Vault, CTGAN, WGANGP, Sequential encoder-decoder networks,GAN	[29], [88]
RTS Similarity	Gaussian Multivariate, Synthetic Data Vault, CTGAN, WGANGP	[88]
Privacy budget	Encoder-decoder + Long Short-Term Memory	[38]
GroundHog	MST, Bayesian Networks	[125]
Attack models: Logistic regression Random Forest	-	[6]
Reverting SNN to ResNet attack	GASCNN	[68]
Synthetic Predictor	MST, Bayesian Networks	[125]
Generating class representation attack	GASCNN	[68]
Kernel density	MST, Bayesian Networks	[125]
LP-Distance	MST, Bayesian Networks	[125]
Replacement ratio	UPC-SGD	[65]
Likelihood estimation	WGAN	[7]
Generative representation learning (GRL)	WGAN	[7]
Contrastive representation learning - local augmentation	WGAN	[7]

TABLE 4. Differential privacy mechanisms and variations

DP variation	Associated SDG Technique	Reference
Laplace mechanism	MST (Private-PGM : Query selection with marginals), Bayesian Networks, Simulation-based, GAN, (Wind Power Obfuscation*, Transmission Capacity Obfuscation*), BGAN, WGAN, DTG (RNN gradient descent algorithm with differential privacy), DPView, Pseudo posterior mechanism, Fully Bayes Model, Copula Based, Differentially private latent tree (DPLT), Weighted Uniform Distribution (WUD), Kernel Density Estimation (KDE), Bounded Laplace Method (BLM)	[45], [52], [57], [58], [66], [76], [89], [96], [101], [106], [109], [111], [114], [120], [125]
Staircase mechanism	DP-CSM	[35]
Gaussian mechanism	DTG (RNN gradient descent algorithm with differential privacy), Masked Autoregressive Flow, Normalizing flows, DP-Hflow, Autoencoder, Encoder-decoder + Long Short-Term Memory,low-dimensional marginals, PEP, GEM, Bayesian networks, Relaxed Adaptive Projection, MST (Private-PGM: Query selection with marginals), PrivSyn (Marginal selection + Noise addition+postprocessing + Graduate update method for synthesis), DP-WGAN	[38], [42], [52], [64], [67], [70], [71], [104], [109], [116], [119], [121], [125]
Geometric mechanism	k-way margins	[129]
Exponential mechanism	(Wind Power Obfuscation*,Transmission Capacity Obfuscation*)	[66], [106]
Personalized Differential Privacy (PDP)	-	[66]
DP-SGD	DP-CGANS, DP-auto-GAN, Masked autoregressive Flow	[47], [107], [116]
Renyi Differential Privacy	Normalizing flows, RDP-CGAN	[8], [70]
LDP	LDPGM-ORR, 2-ways marginal, DP-Federated-Generative-Autoencoder (Wasserstein Autoencoder (WAE)), Bounded Laplace Method (BLM)	[41], [52], [108], [127]

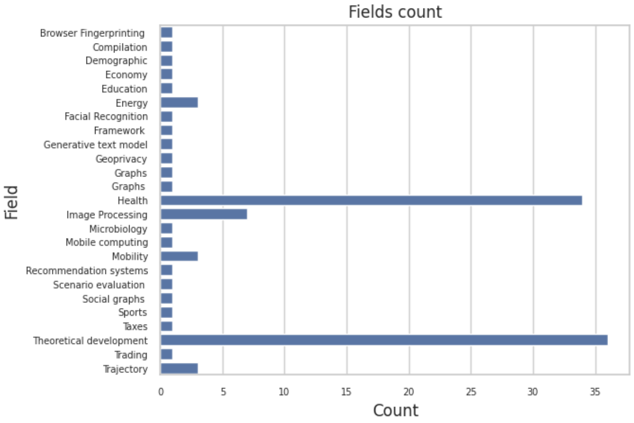

FIGURE 4. Distribution of the different fields of application

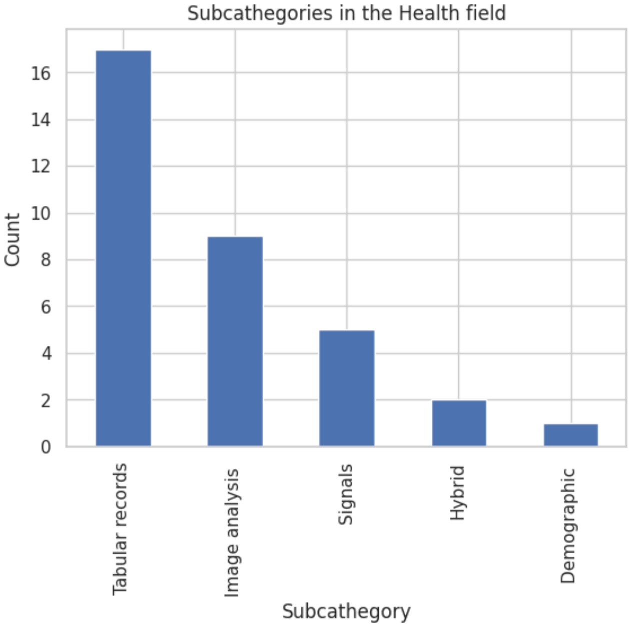

FIGURE 5. Distribution of the sub-categories of health field

fields to preserve privacy, under the premise that the use of unique synthetic data generation techniques poses a risk to the privacy of different users. SDG techniques are known for representing considerable potential in the progress of models executed in different fields of knowledge and in the enhancement of methods that can perform their functions while ensuring privacy guarantees are a clear demonstration of how implementations can be carried out responsibly. This work can be used to gain insight into current privacy metrics and techniques in the context of synthetic data generation, to serve as the root of eventual investigations that aim to search for synthetic data in various areas, as well as a compilation of which techniques are most used in different contexts to facilitate scientific access and development.

A. MAIN FINDINGS

To answer each of the research questions presented in the methodology section, the following compilation of findings is presented.

1) RQ1: Which privacy techniques can be integrated to evaluate the privacy of synthetically generated data?

To answer this question, it was first necessary to classify the techniques of synthetic data generation and their evaluation metrics, to recognize the existence of these techniques and the importance of a certain set of them in the general development framework of synthetic data. The first finding is represented by the common use of techniques based on GANs, used by 42% of the compiled works, and secondly, techniques based on statistical models and the variety derived from them stand out. On the other hand, the most common metrics for measuring performance are utility-based metrics, a finding that highlights the interest in achieving utility standards in synthetic data generation, rather than resemblance.

The most commonly used privacy technique throughout the compiled studies is Differential Privacy, confirming its position as a state-of-the-art technique, either with its general use or with uses derived from it (Section V-A). This technique has been integrated into SDG models of all categories. Within its use, as can be seen in Section V-A1, the Laplace mechanism and the Gaussian mechanism stand out as mechanisms for preserving DP, followed by the use of LDP and DP-SGD. The implementation of techniques based on Blockchain or Encryption stands out for their lack of use of methods for synthetic data generation. Likewise, the utilization of anonymization techniques such as l-diversity or k-anonymity are still present in a few works. Finally, it is possible to look through the table 1 a dominance of GAN-based methods, in addition to the fact that utility-based metrics are the most widely used when making implementations with privacy guarantees.

2) RQ2: To what extent and criteria different metrics have been applied to evaluate the privacy of generated data by different SDG methods combined with privacy mechanisms? How are those methods applied in a privacy utility trade-off evaluation?

To answer these questions, we compiled various privacy metrics used in the articles. Initially, there is a lack of depth in the concept and differentiation of privacy techniques, privacy metrics, and synthetic data. In this research, we try to differentiate the concepts from definitions used by Boudewijn et al. [78], Wagner et al. [77], and Carvalho et al. [10]. A data transformation to make synthetic data from original data is not a methodology for generating synthetic data, but the implementation of a mechanism. Thus, a metric is, as the name describes, a measure for data privacy, which can vary in different categories, as defined in this paper: distance metrics, performance metrics, privacy-based metrics, and other types of metrics. However, it should be noted that many privacy metrics are used within synthetic data generators to produce data with privacy guarantees, a situation in which such a metric would behave as a mechanism within that framework; or a situation in which a parameter of a privacy mechanism can be interpreted as a privacy metric.

There is a high diversity in the privacy metrics used, and the choice of their use in the different fields of development has no clear criteria. This means that in the methodology of elaboration of some papers, the privacy metric is more related to the objective of the paper than to the synthetic data generation technique used there. The latter is facilitated when the privacy mechanism used is differential privacy since this parameter can be useful when performing parallel analysis. Additionally, it is possible to observe how the models based on attacks are well used in the literature, which may be a sign of the interest of developing works in this field: avoiding privacy attacks.

The evaluation of the second part of this question is described in Section VI-E, where we describe works that implement privacy-utility trade-off mechanisms within their synthetic data generation models, seeking to solve an optimization problem, or globally evaluating the behavior of a synthetic data generation model with the use of a privacy mechanism. As described, most of the articles make a final implementation based on a privacy metric vs. a utility metric to perform an analysis in this sense, in particular, Privacy Budget has a large number of implementations, where the performance of the DP is compared with utility metrics such as ROC, AUC or accuracy.

3) RQ3: How have privacy techniques in the framework of synthetically generated data been used and applied in health data?

To answer this question, a classification by fields of application documented in the Section VII was carried out. It is possible to observe that the medical field is one of the most interested in producing synthetic data generation methodologies that guarantee the privacy of users or patients. Likewise, there is a considerable diversity in the subcategories of application, where applications in tabular data, image analysis, and signal analysis stand out. It is necessary to point out that although another of the most developed categories are theoretical developments, many of these works have direct applicability in the field of health. However, relevant implementations in the OMICS field were not found in this review. At last, there is still a lack of standardization to compare properly different metrics, as it was mentioned by Hernandez et al. [86].

Although the use of DP is becoming very relevant, even in the medical field, there is still no confidence interval that verifies that a certain value of *epsilon* guarantees privacy, since the utility and privacy gained lies in the specific method of fulfilling the conditions that ensure DP.

B. LIMITATIONS

The limitations of this study begin with the methodology designed for the search of papers, as the date range, publication requirements, the databases used for the search, and the screening performed may have discarded papers and documents relevant to the topic of interest. In particular, the selection used to extract the articles from Google Scholar may have increased the bias from earlier stages of the systematic review. Due to the number of articles explored and the objective of the review, many of the synthetic data generation methods were not explored in depth, which can lead to inaccuracies when describing them. In the same vein, important information may not have been considered in the descriptions of the privacy mechanisms and metrics implemented.

Finally, there may be inaccuracies at the time of categorization since many techniques could belong to more than one category. Therefore, the naming and description of the techniques were done seeking a balance between the global and the specific, that is, categories that better segmented the data but that internally could describe general characteristics of the methods that belong to them.

Future extensions of the systematic review could involve a detailed investigation into the specific nature of health data utilized in synthetic data generation studies, understanding the differences of data types, such as cancer or diabetes data, for providing valuable context for evaluating the effectiveness of privacy-preserving techniques in different healthcare research domains. Also, exploring the establishment of practical privacy thresholds for each type of study could offer some insights about the use of DP, determining the threshold at which mechanisms perform optimally would enable the identification of thresholds that balance privacy protection with data utility. Finally, an analysis of the relationship between the choice of model and the characteristics of the data across various could identify correlations between specific modeling approaches and the nature of the data, helping to know which models are most suitable for different types of health data.

C. RESEARCH DIRECTIONS

This work has found several possible avenues for future research with high relevance in the field of privacy mechanisms and their evaluation metrics in the context of synthetic data generation. One possible avenue for development involves researching whether the use of specific metrics is biased toward a particular set of privacy mechanisms or synthetic data generation methods. This exploration should consider the nature, structure, and volume of the data that the model uses for generating and measuring privacy.

Similarly, there is a lack of publications on how different privacy techniques are challenged when dealing with outliers or data with high variability, and how different synthetic data generators are involved in data with these characteristics. In this line, there is a lack of research on privacy-utility-bias trade-off issues, as well as on the effect of the use of this framework on the fairness of the data. On the other hand, more research is needed on metrics to determine how statistics such as covariance are preserved when different privacy mechanisms are implemented in the data.

Regarding the findings on the privacy-utility trade-off, an important line of research is the implementation of mechanisms that are internally coupled in the synthetic data generation models, and that in the best case, optimize these two characteristics of the data in each iteration. Also, advancements in statistical metrics to more easily compare studies are a possible avenue for future developments.

This research shows an advance in the development of privacy techniques for the generation of synthetic tabular data. However, the development of metrics and techniques in this framework for unstructured data is scarce, and no configuration on the sensitivity of the attributes associated with this type of data (e.g., images, time series) is described.

Finally, there is a lack of development of privacy techniques in the framework of synthetic data generation in highly relevant fields, such as economics, energy, LLM and text generation models, mobility, recommender systems, and health implementations using OMICS data.

IX. CONCLUSION

The research carried out in this systematic review compiles works that address privacy metrics and techniques implemented in the context of synthetic data generation, intending to provide an overview that serves as a state-of-the-art and current diagnosis of the topic. It highlights the applications in the field of health, the use of GAN-based generating models, the use of Differential Privacy as a method of privacy preservation, and the use of metrics such as the Privacy Budget or attack mechanisms for measuring privacy in synthetically generated data. The growing popularity of privacy techniques for synthetic data has been emphasized due to recent concerns about how SDG techniques alone can fail to maintain the privacy of the original data. The wide mixture of elements in theoretical developments is also considered, as well as the diversity of applications in different fields. This review lists solutions from a development perspective to implement privacy techniques among the different SDG algorithms and encourage the development of diversification and standardization of metrics, and sensitivity analysis of the different metrics to properly select them according to the structure of the data, and the SDG model used. It also provides an applicability perspective in fields of knowledge where synthetic data are relevant, and the sensitivity that they inherently contain makes them candidates for the implementation of privacy measures. In practice, this review recommends and justifies the use of Differential Privacy as a suitable technique for privacy preservation in different fields of knowledge, the use of GAN-based techniques, with their derivations, for the generation of synthetic data, and a diverse set of evaluation metrics. Unfortunately, the study may have suffered from the lack of an existing general framework for categorizing privacy techniques and metrics, as well as the limited use of metrics that guarantee covariance, bias, and fairness. Therefore, future research should focus on developing a general framework for categorizing privacy techniques and metrics, as well as on the use of metrics that guarantee covariance, bias, and fairness.

ACKNOWLEDGMENT

This work has been partially funded by the SYNTHEMA project, which is an initiative funded by the European Union's Horizon Europe Research and Innovation programme under grant agreement N° 101095530.

Funded by
the European Union

X. AUTHORSHIP CONTRIBUTION STATEMENT

Pablo A. Osorio Marulanda: Conceptualization, Methodology, Investigation, Writing - Original draft. **Gorka Epelde**: Conceptualization, Methodology, Review & Editing, Supervision. **Mikel Hernandez**: Conceptualization, Methodology, Review & Editing.**Imanol Isasa**: Conceptualization, Methodology, Review. **Nicolás Moreno**: Conceptualization, Methodology, Review. **Andoni Beristain Iraola**: Project Administration.

XI. ABBREVIATIONS AND ACRONYMS

The abbreviations used in this paper can be seen in Table XI.

Abreviation	Meaning
SDG	Synthetic Data Generation
PPDP	Privacy Preserving Data Publishing
PPDS	Privacy Preserving Data Sharing
PPDM	Privacy Preserving Data Minning
GAN	Generative Adversarial Networks
PETs	Privacy Enhancing Technologies
SDG	Synthetic Data Generation
ID	Explicit Identifier
QID	Quasi-Identifier
SA	Sensitive Identifier
NSA	Non-sensitive attribute
ML	Machine Learning
VAE	Variational Autoencoder
VMM	Variable Order Markov Model
KDE	Kernel Density Estimation
CNN	Convolutional Neural Networks
KNN	K Nearest Neighbor
MLP	Multilayer Perceptron
DP	Differential Privacy
RDP	Renyi Differential Privacy
DP-SDG	Differentially Private Stochastic Gradient Descent
PDP	Personalized Differential Privacy
LDP	Local Differential Privacy
MARC	Multi-Aspect Trajectory Classification
GASCNN	Generative adversarial networks and spike learning-based convolutional neural network
DP-CGAN	Differentially private Conditional Generative Adversarial Networks
TVAE	Tabular Variational Autoencoder
CTGAN	Conditional Tabular Generative Adversarial Networks
WGAN	Wasserstein Generative Adversarial Networks
GS-WGAN	Gradient-sanitized Wasserstein Generative Adversarial Networks
LSTM	Long Short-Term Memory Networks
MC	Markov Chains
BGAN	Boundary-Seeking Generative Adversarial Networks
PPGAN	Privacy-preserving Generative Adversarial Network
PPGAIN	Privacy-preserving Generative Adversarial Imitation Network
WGANGP	Wasserstein Generative Adversarial Network with Gradient Penalty
PATE-GAN	Private Aggregation of Teacher Ensembles Generative Adversarial Network
zCDP	Zero-Concentrated Differential Privacy
GEDDP	Gradient Embedding Perturbation with Differential Privacy
DP-BLSGD	Rényi Differentially Private Backtracking Line Search Based Sub-sampled Gradient Descent
RDP	Reparametrized gradient perturbation
TSADP	Tempered Sigmoid Activations with Differential Privacy
GEP	Gradient Embedding Perturbation
AC-GAN	Auxiliary Classifier Generative Adversarial Network
FL	Federated Learning
MEWM	Multiplicative Weights Exponential Mechanism
PART-GAN	Privacy-preserving Augmentation and Releasing scheme for Time series data via GAN
DPLT	Differentially Private Latent Tree
PEP	Private Entropy Projection
GEM	Generative networks with the exponential mechanism
WUD	Weighted Uniform Distribution

REFERENCES

[1] Abdul Majeed. Attribute-centric and synthetic data based privacy preserving methods: A systematic review. *Journal of Cybersecurity and Privacy*, 3(3):638–661, 2023.

[2] Hong-Yen Tran and Jiankun Hu. Privacy-preserving big data analytics a comprehensive survey. *Journal of Parallel and Distributed Computing*, 134:207–218, 2019.

[3] Stan Matwin. Privacy-preserving data mining techniques: survey and challenges. In *Discrimination and privacy in the information society: Data mining and profiling in large databases*, pages 209–221. Springer, 2013.

[4] Debbie Rankin, Michaela Black, Raymond Bond, Jonathan Wallace, Maurice Mulvenna, Gorka Epelde, et al. Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR medical informatics*, 8(7):e18910, 2020.

[5] Mikel Hernandez, Gorka Epelde, Andoni Beristain, Roberto Álvarez, Cristina Molina, Xabat Larrea, Ane Alberdi, Michalis Timoleon, Panagiotis Bamidis, and Evdokimos Konstantinidis. Incorporation of synthetic data generation techniques within a controlled data processing workflow in the health and wellbeing domain. *Electronics*, 11(5):812, 2022.

[6] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1451–1468, 2022.

[7] Ziqi Zhang, Chao Yan, and Bradley A Malin. Membership inference attacks against synthetic health data. *Journal of biomedical informatics*, 125:103977, 2022.

[8] Amirsina Torfi, Edward A Fox, and Chandan K Reddy. Differentially private synthetic medical data generation using convolutional gans. *Information Sciences*, 586:485–500, 2022.

[9] Benjamin CM Fung, Ke Wang, Rui Chen, and Philip S Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (Csur)*, 42(4):1–53, 2010.

[10] Tânia Carvalho, Nuno Moniz, Pedro Faria, and Luís Antunes. Survey on privacy-preserving techniques for data publishing. *arXiv preprint arXiv:2201.08120*, 2022.

[11] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

[12] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.

[13] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, pages 587–601, 2017.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[15] Tabea Kossen, Manuel A Hirzel, Vince I Madai, Franziska Boenisch, Anja Hennemuth, Kristian Hildebrand, Sebastian Pokutta, Kartikey Sharma,

Adam Hilbert, Jan Sobesky, et al. Toward sharing brain images: Differentially private tof-mra images with segmentation labels using generative adversarial networks. *Frontiers in artificial intelligence*, 5:85, 2022.

[16] Fahim Faisal, Noman Mohammed, Carson K Leung, and Yang Wang. Generating privacy preserving synthetic medical data. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE, 2022.

[17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[18] Stella Ho, Youyang Qu, Bruce Gu, Longxiang Gao, Jianxin Li, and Yong Xiang. Dp-gan: Differentially private consecutive data publishing using generative adversarial nets. *Journal of Network and Computer Applications*, 185:103066, 2021.

[19] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[20] Anja Shevchyk, Rui Hu, Kevin Thandiackal, Michael Heizmann, and Thomas Brunschwiler. Privacy preserving synthetic respiratory sounds for class incremental learning. *Smart Health*, 23:100232, 2022.

[21] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.

[22] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.

[23] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

[24] Aditya Kuppa, Lamine Aouad, and Nhien-An Le-Khac. Towards improving privacy of synthetic datasets. In *Annual Privacy Forum*, pages 106–119. Springer, 2021.

[25] Shuo Wang, Carsten Rudolph, Surya Nepal, Marthie Grobler, and Shangyu Chen. Part-gan: Privacy-preserving time-series sharing. In *Artificial Neural Networks and Machine Learning–ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part I 29*, pages 578–593. Springer, 2020.

[26] Katharina Dietz, Michael Mühlhauser, Michael Seufert, Nicholas Gray, Tobias Hoßfeld, and Dominik Herrmann. Browser fingerprinting: How to protect machine learning models and data with differential privacy? *Electronic Communications of the EASST*, 80, 2021.

[27] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[28] Mandis Beigi, Afrah Shafquat, Jason Mezey, and Jacob Aptekar. Simulants: Synthetic clinical trial data via subject-level privacy-preserving synthesis. In *AMIA Annual Symposium Proceedings*, volume 2022, page 231. American Medical Informatics Association, 2022.

[29] Jinsung Yoon, Michel Mizrahi, Nahid Farhady Ghalaty, Thomas Jarvinen, Ashwin S Ravi, Peter Brune, Fanyu Kong, Dave Anderson, George Lee, Arie Meir, et al. Ehr-safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *NPJ Digital Medicine*, 6(1):141, 2023.

[30] Mitchell Naughton, Dan Weaving, Tannath Scott, and Heidi Compton. Synthetic data as a strategy to resolve data privacy and confidentiality concerns in the sport sciences: Practical examples and an r shiny application. *International Journal of Sports Physiology and Performance*, 18(10):1213–1218, 2023.

[31] Claire McKay Bowen, Victoria Bryant, Leonard Burman, John Czajka, Surachai Khitatrakun, Graham MacDonald, Robert McClelland, Livia Mucciolo, Madeline Pickens, Kyle Ueyama, et al. Synthetic individual income tax data: Methodology, utility, and privacy implications. In *International Conference on Privacy in Statistical Databases*, pages 191–204. Springer, 2022.

[32] Markus Hittmeir, Rudolf Mayer, and Andreas Ekelhart. Utility and privacy assessment of synthetic microbiome data. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 15–27. Springer, 2022.

[33] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. CRC press, 1984.

[34] Jerome P Reiter. Using cart to generate partially synthetic public use microdata. *Journal of official statistics*, 21(3):441, 2005.

[35] Xin Yao, Juan Yu, Jianmin Han, Jianfeng Lu, Hao Peng, Yijia Wu, and Xiaoqian Cao. Dp-csm: Efficient differentially private synthesis for human mobility trajectory with coresets and staircase mechanism. *ISPRS International Journal of Geo-Information*, 11(12):607, 2022.

[36] Bo-Chen Tai, Szu-Chuang Li, Yennun Huang, and Pang-Chieh Wang. Examining the utility of differentially private synthetic data generated using variational autoencoder with tensorflow privacy. In *2022 IEEE 27th Pacific Rim International Symposium on Dependable Computing (PRDC)*, pages 236–241. IEEE, 2022.

[37] Maya Benarous, Eran Toch, and Irad Ben-Gal. Synthesis of longitudinal human location sequences: Balancing utility and privacy. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–27, 2022.

[38] Taisho Sasada, Masataka Kawai, Yuzo Taenaka, Doudou Fall, and Youki Kadobayashi. Differentially-private text generation via text preprocessing to reduce utility loss. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 042–047. IEEE, 2021.

[39] Ashish Dandekar and Stéphane Bressan. Who is alice? privacy risk, the case of regression. In *2018 International Workshop on Big Data and Information Security (IWBIS)*, pages i–vi. IEEE, 2018.

[40] Jill-Jênn Vie, Tomas Rigaux, and Sein Minn. Privacy-preserving synthetic educational data generation. In *European Conference on Technology Enhanced Learning*, pages 393–406. Springer, 2022.

[41] Xue Chen, Cheng Wang, Qing Yang, Teng Hu, and Changjun Jiang. Locally differentially private high-dimensional data synthesis. *Science China Information Sciences*, 66(1):112101, 2023.

[42] Ergute Bao, Xiaokui Xiao, Jun Zhao, Dongping Zhang, and Bolin Ding. Synthetic data generation with differential privacy via bayesian networks. *Journal of Privacy and Confidentiality*, 11(3), 2021.

[43] Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. Utility and privacy assessments of synthetic data for regression tasks. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 5763–5772. IEEE, 2019.

[44] Frederik Harder, Kamil Adamczewski, and Mijung Park. Dp-merf: Differentially private mean embeddings with randomfeatures for practical privacy-preserving data generation. In *International conference on artificial intelligence and statistics*, pages 1819–1827. PMLR, 2021.

[45] Teddy Cunningham, Graham Cormode, and Hakan Ferhatosmanoglu. Privacy-preserving synthetic location data in the real world. In *17th International Symposium on Spatial and Temporal Databases*, pages 23–33, 2021.

[46] Muhammad Syafiq Mohd Pozi and Mohd. Hasbullah Omar. A kernel density estimation method to generate synthetic shifted datasets in privacy-preserving task. *J. Internet Serv. Inf. Secur.*, 10(4):70–89, 2020.

[47] Chang Sun, Johan van Soest, and Michel Dumontier. Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *Journal of Biomedical Informatics*, page 104404, 2023.

[48] Chengkun Wei, Shouling Ji, Changchang Liu, Wenzhi Chen, and Ting Wang. Asgldp: collecting and generating decentralized attributed graphs with local differential privacy. *IEEE Transactions on Information Forensics and Security*, 15:3239–3254, 2020.

[49] Zhenchen Wang, Puja Myles, and Allan Tucker. Generating and evaluating synthetic uk primary care data: preserving data utility & patient privacy. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 126–131. IEEE, 2019.

[50] Ivan Fontana, Marc Langheinrich, and Martin Gjoreski. Gans for privacy-aware mobility modeling. *IEEE Access*, 11:29250–29262, 2023.

[51] Hsin-Yi Chen and Szu-Hao Huang. Generating a trading strategy in the financial market from sensitive expert data based on the privacy-preserving generative adversarial imitation network. *Neurocomputing*, 500:616–631, 2022.

[52] Mengmeng Yang, Longxia Huang, and Chenghua Tang. K-means clustering with local distance privacy. *Big Data Mining and Analytics*, 2023.

[53] Aaron S Coyner, Jimmy S Chen, Ken Chang, Praveer Singh, Susan Ostmo, RV Paul Chan, Michael F Chiang, Jayashree Kalpathy-Cramer, J Peter Campbell, Imaging, Informatics in Retinopathy of Prematurity Consortium, et al. Synthetic medical images for robust, privacy-preserving training of artificial intelligence: application to retinopathy of prematurity diagnosis. *Ophthalmology Science*, 2(2):100126, 2022.

[54] Yueming Lyu, Yue Jiang, Ziwen He, Bo Peng, Yunfan Liu, and Jing Dong. 3d-aware adversarial makeup generation for facial privacy protection. *arXiv preprint arXiv:2306.14640*, 2023.

[55] Bardia Khosravi, Pouria Rouzrokh, John P Mickley, Shahriar Faghani, A Noelle Larson, Hillary W Garner, Benjamin M Howe, Bradley J Erickson, Michael J Taunton, and Cody C Wyles. Creating high fidelity synthetic pelvis radiographs using generative adversarial networks: unlocking the potential of deep learning models without patient privacy concerns. *The Journal of Arthroplasty*, 38(10):2037–2043, 2023.

[56] Jong Wook Kim and Beakcheol Jang. Deep learning-based privacy-preserving framework for synthetic trajectory generation. *Journal of Network and Computer Applications*, 206:103459, 2022.

[57] Sana Imtiaz, Muhammad Arsalan, Vladimir Vlassov, and Ramin Sadre. Synthetic and private smart health care data generation using gans. In *2021 International Conference on Computer Communications and Networks (ICCCN)*, pages 1–7. IEEE, 2021.

[58] Chih-Hsun Lin, Chia-Mu Yu, and Chun-Ying Huang. Dpview: Differentially private data synthesis through domain size information. *IEEE Internet of Things Journal*, 9(17):15886–15900, 2022.

[59] Ruikang Yang, Jianfeng Ma, Yinbin Miao, and Xindi Ma. Privacy-preserving generative framework for images against membership inference attacks. *IET Communications*, 17(1):45–62, 2023.

[60] Anantaa Kotal, Aritran Piplai, Sai Sree Laya Chukkapalli, and Anupam Joshi. Privetab: Secure and privacy-preserving sharing of tabular data. In *Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics*, pages 35–45, 2022.

[61] Rudolf Mayer, Markus Hittmeir, and Andreas Ekelhart. Privacy-preserving anomaly detection using synthetic data. In *Data and Applications Security and Privacy XXXIV: 34th Annual IFIP WG 11.3 Conference, DBSec 2020, Regensburg, Germany, June 25–26, 2020, Proceedings 34*, pages 195–207. Springer, 2020.

[62] Markus Hittmeir, Rudolf Mayer, and Andreas Ekelhart. Efficient bayesian network construction for increased privacy on synthetic data. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5721–5730. IEEE, 2022.

[63] Joonas Jälkö, Eemil Lagerspetz, Jari Haukka, Sasu Tarkoma, Antti Honkela, and Samuel Kaski. Privacy-preserving data sharing via probabilistic modeling. *Patterns*, 2(7), 2021.

[64] Ren Yang, Xuebin Ma, Xiangyu Bai, and Xiangdong Su. Differential privacy images protection based on generative adversarial network. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1688–1695. IEEE, 2020.

[65] Fan Liu, Zhiyong Cheng, Huilin Chen, Yinwei Wei, Liqiang Nie, and Mohan Kankanhalli. Privacy-preserving synthetic data generation for recommendation systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1379–1389, 2022.

[66] Ben Niu, Yahong Chen, Boyang Wang, Zhibo Wang, Fenghua Li, and Jin Cao. Adapdp: Adaptive personalized differential privacy. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021.

[67] Vladimir Janjic, JKF Bowles, Andreas Francois Vermeulen, Agastya Silvina, Marios Belk, Christos Fidas, Andreas Pitsillides, Mohit Kumar, Michael Rossbory, Michael Vinov, et al. The serums tool-chain: ensuring security and privacy of medical data in smart patient-centric healthcare systems. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2726–2735. IEEE, 2019.

[68] Sunder Ali Khowaja, Kamran Dahri, Muhammad Aslam Jarwar, and Ik Hyun Lee. Spike learning based privacy preservation of internet of medical things in metaverse. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[69] Mehmet Emre Gursoy, Ling Liu, Stacey Truex, Lei Yu, and Wenqi Wei. Utility-aware synthesis of differentially private and attack-resilient location traces. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 196–211, 2018.

[70] Jaewoo Lee, Minjung Kim, Yonghyun Jeong, and Youngmin Ro. Differentially private normalizing flows for synthetic tabular data generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7345–7353, 2022.

[71] Giuseppe Vietri, Cedric Archambeau, Sergul Aydore, William Brown, Michael Kearns, Aaron Roth, Ankit Siva, Shuai Tang, and Steven Z Wu. Private synthetic data for multitask learning and marginal queries. *Advances in Neural Information Processing Systems*, 35:18282–18295, 2022.

[72] Ethan Schonfeld and Anand Veeravagu. Demonstrating the successful application of synthetic learning in spine surgery for training multi–center models with increased patient privacy. *Scientific Reports*, 13(1):12481, 2023.

[73] Hanxi Sun, Jason Plawinski, Sajanth Subramaniam, Amir Jamaludin, Timor Kadir, Aimee Readie, Gregory Ligozio, David Ohlssen, Mark Baillie, and Thibaud Coroller. A deep learning approach to private data sharing of medical images using conditional generative adversarial networks (gans). *Plos one*, 18(7):e0280316, 2023.

[74] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416:244–255, 2020.

[75] Zhenchen Wang, Puja Myles, and Allan Tucker. Generating and evaluating cross-sectional synthetic electronic healthcare data: preserving data utility and patient privacy. *Computational Intelligence*, 37(2):819–851, 2021.

[76] Yuichi Sei, J Andrew Onesimu, and Akihiko Ohsuga. Machine learning model generation with copula-based synthetic dataset for local differentially private numerical data. *IEEE Access*, 10:101656–101671, 2022.

[77] Isabel Wagner and David Eckhoff. Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)*, 51(3):1–38, 2018.

[78] Alexander T.P. Boudewijn, Andrea Filippo Ferraris, Daniele Panfilo, Vanessa Cocca, Sabrina Zinutti, Karel De Schepper, and Carlo Rossi Chauvenet. Privacy measurement in tabular synthetic data: State of the art and future research directions. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*, 2023.

[79] João Coutinho-Almeida, Pedro Pereira Rodrigues, and Ricardo João Cruz-Correia. Gans for tabular healthcare data generation: a review on utility and privacy. In *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24*, pages 282–291. Springer, 2021.

[80] Anna Monreale, Roberto Pellungrini, et al. A survey on privacy in human mobility. *TRANSACTIONS ON DATA PRIVACY*, 16(1):51–82, 2023.

[81] Debolina Ghatak and Kouichi Sakurai. A survey on privacy preserving synthetic data generation and a discussion on a privacy-utility trade-off problem. In *International Conference on Science of Cyber Security*, pages 167–180. Springer, 2022.

[82] Joseph Ficek, Wei Wang, Henian Chen, Getachew Dagne, and Ellen Daley. Differential privacy in health research: A scoping review. *Journal of the American Medical Informatics Association*, 28(10):2269–2276, 2021.

[83] Markus Endres, Asha Mannarapotta Venugopal, and Tung Son Tran. Synthetic data generation: a comparative study. In *Proceedings of the 26th International Database Engineered Applications Symposium*, pages 94–102, 2022.

[84] Fida K Dankar and Mahmoud Ibrahim. Fake it till you make it: Guidelines for effective synthetic data generation. *Applied Sciences*, 11(5):2158, 2021.

[85] Lindsay S Uman. Systematic reviews and meta-analyses. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 20(1):57, 2011.

[86] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45, 2022.

[87] Adriano Lucieri, Andreas Dengel, and Sheraz Ahmed. Translating theory into practice: assessing the privacy implications of concept-based explanations for biomedical ai. *Frontiers in Bioinformatics*, 3, 2023.

[88] Mikel Hernadez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. *Methods of Information in Medicine*, 2023.

[89] Fang Liu, Dong Wang, and Tian Yan. Some examples of privacy-preserving sharing of covid-19 pandemic data with statistical utility evaluation. *BMC Medical Research Methodology*, 23(1):1–18, 2023.

[90] Shiming Ge, Bochao Liu, Pengju Wang, Yong Li, and Dan Zeng. Learning privacy-preserving student networks via discriminative-generative distillation. *IEEE Transactions on Image Processing*, 32:116–127, 2022.

[91] Jingpu Shi, Dong Wang, Gino Tesei, and Beau Norgeot. Generating high-fidelity privacy-conscious synthetic patient data for causal effect estimation with multiple treatments. *Frontiers in Artificial Intelligence*, 5:918813, 2022.

[92] Rohit Venugopal, Noman Shafqat, Ishwar Venugopal, Benjamin Mark John Tillbury, Harry Demetrios Stafford, and Aikaterini Bourazeri. Privacy preserving generative adversarial networks to model electronic health records. *Neural Networks*, 153:339–348, 2022.

[93] João Coutinho-Almeida, Ricardo João Cruz-Correia, and Pedro Pereira Rodrigues. Dataset comparison tool: Utility and privacy. *Challenges of Trustable AI and Added-Value on Health*, page 23, 2022.

[94] Sanket Desai, Nasser Sabar, Rabei Alhadad, Abdun Mahmood, and Naveen Chilamkurti. Mitigating consumer privacy breach in smart grid using obfuscation-based generative adversarial network, 2022.

[95] Vajira Thambawita, Jonas L Isaksen, Steven A Hicks, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Christina Ellervik, Morten Salling Olesen, Torben Hansen, et al. Deepfake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. *Scientific reports*, 11(1):21896, 2021.

[96] MinDong Sung, Dongchul Cha, and Yu Rang Park. Local differential privacy in the medical domain to protect sensitive information: algorithm development and real-world validation. *JMIR Medical Informatics*, 9(11):e26914, 2021.

[97] Damian Pascual, Alireza Amirshahi, Amir Aminifar, David Atienza, Philippe Ryvlin, and Roger Wattenhofer. Epilepsygan: Synthetic epileptic brain activities with privacy preservation. *IEEE Transactions on Biomedical Engineering*, 68(8):2435–2446, 2020.

[98] SeHee Oh, MinDong Sung, Yumie Rhee, Namki Hong, and Yu Rang Park. Evaluation of the privacy risks of personal health identifiers and quasi-identifiers in a distributed research network: Development and validation study. *JMIR Medical Informatics*, 9(5):e24940, 2021.

[99] Yi Yang, Shuai Huang, Wei Huang, and Xiangyu Chang. Privacy-preserving cost-sensitive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2105–2116, 2020.

[100] Hafiz Imtiaz, Jafar Mohammadi, Rogers Silva, Bradley Baker, Sergey M Plis, Anand D Sarwate, and Vince D Calhoun. A correlated noise-assisted decentralized differentially private estimation protocol, and its application to fmri source separation. *IEEE Transactions on Signal Processing*, 69:6355–6370, 2021.

[101] Jinao Yu, Hanyu Xue, Bo Liu, Yu Wang, Shibing Zhu, and Ming Ding. Gan-based differential private image privacy protection framework for the internet of multimedia things. *Sensors*, 21(1):58, 2020.

[102] Jing Ma, Qiuchen Zhang, Jian Lou, Joyce C Ho, Li Xiong, and Xiaoqian Jiang. Privacy-preserving tensor factorization for collaborative health data analysis. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1291–1300, 2019.

[103] Yichen Jiang, Chenghong Wang, Zhixuan Wu, Xin Du, and Shuang Wang. Privacy-preserving biomedical data dissemination via a hybrid approach. In *AMIA Annual Symposium Proceedings*, volume 2018, page 1176. American Medical Informatics Association, 2018.

[104] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 510–526. Springer, 2019.

[105] Bangzhou Xin, Wei Yang, Yangyang Geng, Sheng Chen, Shaowei Wang, and Liusheng Huang. Private fl-gan: Differential privacy synthetic data generation based on federated learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2927–2931. IEEE, 2020.

[106] Vladimir Dvorkin and Audun Botterud. Differentially private algorithms for synthetic power system datasets. *IEEE Control Systems Letters*, 2023.

[107] Uthaipon Tao Tantipongpipat, Chris Waites, Digvijay Boob, Amaresh Ankit Siva, and Rachel Cummings. Differentially private synthetic mixed-type data generation for unsupervised learning. *Intelligent Decision Technologies*, 15(4):779–807, 2021.

[108] Yuxuan Zhang, Jianghong Wei, Xiaojian Zhang, Xuexian Hu, and Wenfen Liu. A two-phase algorithm for generating synthetic graph under local differential privacy. In *Proceedings of the 8th International Conference on Communication and Network Security*, pages 84–89, 2018.

[109] Xinyao Liu, Baojiang Cui, Junsong Fu, Zishuai Cheng, and Xuyan Song. Secure data publishing of private trajectory in edge computing of iot. *Security and Communication Networks*, 2022, 2022.

[110] Yao-Tung Tsou, Mansour Naser Alraja, Li-Sheng Chen, Yu-Hsiang Chang, Yung-Li Hu, Yennun Huang, Chia-Mu Yu, and Pei-Yuan Tsai. (k, ε, δ)-anonymization: privacy-preserving data release based on k-anonymity and differential privacy. *Service Oriented Computing and Applications*, 15(3):175–185, 2021.

[111] Lihe Hou, Weiwei Ni, Sen Zhang, Nan Fu, and Dongyue Zhang. Ppdu: dynamic graph publication with local differential privacy. *Knowledge and Information Systems*, 65(7):2965–2989, 2023.

[112] Arno Appenzeller, Moritz Leitner, Patrick Philipp, Erik Krempel, and Jürgen Beyerer. Privacy and utility of private synthetic data for medical data analyses. *Applied Sciences*, 12(23):12320, 2022.

[113] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[114] Jingchen Hu, Terrance D Savitsky, and Matthew R Williams. Private tabular survey data products through synthetic microdata generation. *Journal of Survey Statistics and Methodology*, 10(3):720–752, 2022.

[115] Olayinka Adeboye, Tooska Dargahi, Meisam Babaie, Mohamad Saraee, and Chia-Mu Yu. Deepclean: a robust deep learning technique for autonomous vehicle camera data privacy. *IEEE Access*, 10:124534–124544, 2022.

[116] Chris Waites and Rachel Cummings. Differentially private normalizing flows for privacy-preserving density estimation. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 1000–1009, 2021.

[117] Yi Liu, Jialiang Peng, JQ James, and Yi Wu. Ppgan: Privacy-preserving generative adversarial network. In *2019 IEEE 25Th international conference on parallel and distributed systems (ICPADS)*, pages 985–989. IEEE, 2019.

[118] Stefan Seidlitz, Kris Jürgens, Andrey Makrushin, Christian Kraetzer, and Jana Dittmann. Generation of privacy-friendly datasets of latent fingerprint images using generative adversarial networks. In *VISIGRAPP (4: VISAPP)*, pages 345–352, 2021.

[119] Kuntai Cai, Xiaoyu Lei, Jianxin Wei, and Xiaokui Xiao. Data synthesis via differentially private markov random fields. *Proceedings of the VLDB Endowment*, 14(11):2190–2202, 2021.

[120] Peng Tang, Xiang Cheng, Sen Su, Rui Chen, and Huaxi Shao. Differentially private publication of vertically partitioned data. *IEEE transactions on dependable and secure computing*, 18(2):780–795, 2019.

[121] Terrance Liu, Giuseppe Vietri, and Steven Z Wu. Iterative methods for private synthetic data: Unifying framework and new methods. *Advances in Neural Information Processing Systems*, 34:690–702, 2021.

[122] Kyungjune Baek and Hyunjung Shim. Commonality in natural images rescues gans: Pretraining gans with generic and privacy-free synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7854–7864, 2022.

[123] Yinqiao Xiong, Peidong Zhu, Zhizhu Liu, Hui Yin, Tiantian Deng, et al. Eplc: An efficient privacy-preserving line-loss calculation scheme for residential areas of smart grid. *Security and Communication Networks*, 2019, 2019.

[124] March Boedihardjo, Thomas Strohmer, and Roman Vershynin. Covariance's loss is privacy's gain: Computationally efficient, private and accurate synthetic data. *Foundations of Computational Mathematics*, pages 1–48, 2022.

[125] Ferdoos Hossein Nezhad, Ylenia Rotalinti, Puja Myles, and Allan Tucker. Privacy assessment of synthetic patient data. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–6. IEEE, 2023.

[126] Clément Pierquin, Bastien Zimmermann, and Matthieu Boussard. Practical considerations on using private sampling for synthetic data. *arXiv preprint arXiv:2312.07139*, 2023.

[127] Xue Jiang, Xuebing Zhou, and Jens Grossklags. Privacy-preserving high-dimensional data collection with federated generative autoencoder. *Proc. Priv. Enhancing Technol.*, 2022(1):481–500, 2022.

[128] Zhikun Zhang, Tianhao Wang, Ninghui Li, Jean Honorio, Michael Backes, Shibo He, Jiming Chen, and Yang Zhang. {PrivSyn}: Differentially private data synthesis. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 929–946, 2021.

[129] Michelle Nixon, Andres Barrientos, Jerome Reiter, and Aleksandra Slavkovic. A latent class modeling approach for differentially private synthetic data for contingency tables. *Journal of Privacy and Confidentiality*, 12(1), 2022.

[130] Gillian M Raab. Utility and disclosure risk for differentially private synthetic categorical data. In *International Conference on Privacy in Statistical Databases*, pages 250–265. Springer, 2022.

[131] Andrea Galloni and Imre Lendák. Differentially private copulas, dag and hybrid methods: A comprehensive data utility study. In *International Conference on Computational Collective Intelligence*, pages 270–281. Springer, 2023.

[132] Swarajya Madhuri Rayavarapu, Tammineni Shanmukha Prasanthi, Gottapu Santosh Kumar, Gottapu Sasibhushana Rao, and Aruna Singham.

Employing generative networks for synthetic phonocardiogram and electrocardiogram signal creation: A privacy-ensured approach to data augmentation in heart diagnostics. *Ingénierie des Systèmes d'Information*, 28(4), 2023.

[133] Abdul Majeed and Seong Oun Hwang. Quantifying the vulnerability of attributes for effective privacy preservation using machine learning. *IEEE Access*, 11:4400–4411, 2023.

[134] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[135] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[136] Zach Jorgensen, Ting Yu, and Graham Cormode. Conservative or liberal? personalized differential privacy. In *2015 IEEE 31St international conference on data engineering*, pages 1023–1034. IEEE, 2015.

[137] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 351–360, 2009.

[138] Chris Clifton. Privacy metrics., 2009.

[139] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, 1998.

[140] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. L-diversity: privacy beyond k-anonymity. *22nd International Conference on Data Engineering (ICDE'06)*, pages 24–24, 2006.

[141] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115, 2007.

[142] Max Baak, Rose Koopman, Hella Snoek, and Sander Klous. A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. *Computational Statistics & Data Analysis*, 152:107043, 2020.

PABLO A. OSORIO-MARULANDA received his bachelor's degree in Mathematical Engineering from the Faculty of Science and Engineering at Universidad EAFIT in 2022. Currently, he is advancing his academic journey by pursuing a master's degree in Applied Mathematics at the same institution. In addition to his studies, he is a research assistant in the Big Data Group within the Digital Health and Biomedical Technologies department at the Vicomtech Foundation. His current focus lies in the development of his Master's Thesis, where he delves into the critical area of privacy within the realm of synthetic data generation.

GORKA EPELDE is a Project Leader and Principal Researcher in Vicomtech. In 2014, Gorka obtained his Computer Science PhD from the University of the Basque Country. From 2000 until 2007 Gorka held the position of Assistant Researcher at Ikerlan. From 2007 onwards, Gorka has been a Staff Researcher at Vicomtech's eHealth and Biomedical Applications department. Since 2009, he has been part of the eHealth group under the Bioengineering Area of the BioGipuzkoa Health Research Institute. His fields of interest include interoperability architectures, data preparation and integration, synthetic data generation, as well as human-computer interaction.

MIKEL HERNANDEZ holds a bachelor's degree in Telecommunication Systems Engineering (2019) and a master's degree in Biomedical Technologies (2021) from the Faculty of Engineering of Mondragon University. Currently, he is working as a Researcher in the Big Data Health group of the Digital Health and Biomedical Technologies department from the Vicomtech Foundation, on which he is also developing his PhD Thesis in collaboration with the Unversity of the Basque Country. His PhD thesis is related to the design and implementation of workflows and tools for the synthetic generation of data and his fields of interest include data analysis, synthetic data generation and evaluation, and big data architectures in the health and wellbeing domain.

IMANOL ISASA earned his Bachelor's degree in Biomedical Engineering from Mondragon Unibertsitatea in 2021. Following an internship in the Department of Mechanical and Industrial Production, Imanol traveled to Germany (Hochschule Furtwangen University, HFU) to undertake his Bachelor's Thesis related to the analysis of Electrical Impedance Tomography (EIT) images in COVID-19 patients in collaboration with the Institute of Technical Medicine (ITeM). In September 2021, he started his Master's studies in Biomedical Technologies at Mondragon Unibertsitatea, and by the end of 2022, he began developing his Master's Thesis at Vicomtech, entitled "Healthcare-oriented generation of synthetic time series and associated subjects". Upon completing his master's studies, Imanol joined the Digital Health and Biomedical Technologies Department at the same institution, aiming to continue working on generative models, synthetic data, and data privacy.

NICOLÁS MORENO holds a bachelor's degree in mathematics with a doctorate in statistics conferred by the State University of Campinas. Currently, he serves as a professor at EAFIT University in Medellín, Colombia. His research focus lies within the field of probability and statistics, with a current emphasis on Bayesian inference, stochastic processes, and statistical applications.

ANDONI BERISTAIN IRAOLA obtained his degree in Computer Science at UPV/EHU (2005) and his PhD. in Computer Science at UPV/EHU (2009). He works as principal researcher and project manager in projects at Fundación Centro de tecnologías de Interacción Visual y Comunicaciones Vicomtech since 2023,and as senior researcher and project manager since 2010, which is part of the Basque Research & Technology Alliance (BRTA), in the Digital health and Biomedical Technologies department. He specializes in machine vision, data analysis and natural user interfaces. He is part of the eHealth research group at the Biodonostia Research Institute and the Computational Intelligence Group of the UPV / EHU. He has worked as a researcher on multiple FP6, FP7 and H2020 projects and has coordinated various state and regional projects, including direct industrial contracts. He has also collaborated in the coordination of European projects (e.g. ITEA 2 MEDIATE, CAPTAIN and SYNTHEMA) and as work package leader (CAPTAIN, grant agreement 769830). He is reviewer of scientific journals of international impact.