

Tabular synthetic data generation: A literature review

Joao Fonseca^{1*}, Fernando Bacao¹

¹NOVA Information Management School, Universidade Nova de Lisboa

*Corresponding Author

Postal Address: NOVA Information Management School, Campus de Campolide, 1070-312 Lisboa, Portugal

Telephone: +351 21 382 8610

The generation of synthetic data can be used for anonymization, regularization, oversampling, semi-supervised learning, self-supervised learning and various other tasks. The wide range of applications of these mechanisms motivated the development of new algorithms specialized in generating data for specific types of data and Machine Learning (ML) tasks. As a result, the analysis of the different types of generative models

1 Introduction

Synthetic data is obtained from a generative process based on properties of real data [1]. The generation of synthetic data is essential for various domains and tasks. For example, synthetic data is used as a form of regularizing neural networks (*i.e.*, data augmentation) [CITATION]. One form of anonymizing datasets is via the production of synthetic observations (*i.e.*, synthetic data generation) [CITATION]. In settings where only a small portion of training data is labeled, some techniques generate artificial data using both labeled and unlabeled data with a modified loss function to train neural networks (*i.e.*, semi-supervised learning) [2]. In imbalanced learning contexts, synthetic data can be used to balance the target classes' frequencies and reinforce the learning of minority classes (*i.e.*, oversampling) [3]. Some active learning frameworks use data generation to improve the quality of data selection and classifier training [4]. Other techniques employ data generation to produce deep neural networks without labeled data (*i.e.*, self-supervised learning) [5].

The breadth of these techniques span multiple domains, such as facial recognition [6], Land Use/Land Cover mapping [CITATION], medical image processing [CITATION], Natural Language Processing (NLP) [7] or credit card default prediction [8]. According to the domain and data type, the data generation techniques used may vary significantly. Generally speaking, some data generation mechanisms are specific to some domains, data types or tasks. For example, ... Most, if not all, of these techniques are applied on the input or output space.

However, there are various data generation techniques that are invariant to the task or data types used. These techniques can be either applied in the feature space [9] or in tabular datasets. On the one hand, data generation in the feature space uses a generative model to learn a manifold, lower-dimensional abstraction over the input space [10], defined here as the feature space. At this level, any tabular data generation mechanism can be applied and reconstructed into the input space if necessary. On the other hand, synthetic data generation on tabular data can be applied to most problems. Although, the choice of generation mechanism is still dependant on (1) the importance of the relationships found between the different features, (2) the ML task developed and (3) the motivation for the generation of synthetic data. For example, when generating data to address an imbalanced learning problem (*i.e.*, oversampling), the relationships between the different features are not necessarily kept since the goal is to reinforce the learning of the minority class by redefining an ML classifier’s decision boundaries. If the goal is to anonymize a dataset, perform some type of descriptive task, or ensure a consistent model interpretability, these relationships need to be kept.

Depending on the context, evaluating the quality of the generated data is a complex task. For example, for image and time series data, perceptually small changes in the original data can lead to large changes in the euclidean distance [1, 11]. The evaluation of generative models typically account primarily for the performance in a specific task, since good performance in one criterion does not imply good performance on another [11]. However, in computationally intensive tasks it is often impracticable to search for the optimal configurations of generative models. To address this limitation, other evaluation methods have been proposed to assist in this evaluation, which can be distinguished into statistical divergence metrics and precision/recall metrics [12]. The relevant performance metrics found in the literature are discussed in Section 5.

1.1 Motivation, Scope and Contributions

This literature review focuses on generation mechanisms applied to tabular data and the different ML techniques where tabular synthetic data is used. In addition, we focus on the ML perspective of synthetic data, as opposed to the practical perspective. From a practical sense, synthetic data is used as a proxy of real data. It is assumed to be inaccessible, essential and a secondary asset for tasks like education, software development, or systems demonstrations [13].

We focus on data generation techniques in the tabular and feature space (*i.e.*, embedded inputs), given its breadth in scope. Related literature reviews are mostly focused on specific algorithmic or domain applications, with little to no emphasis on the core generative process. For this reason, these techniques often appear “sandboxed”, even though there is a significant overlap between them. There are some related reviews published since 2019. Assefa et al. [1] provides a general overview of synthetic data generation for time series data anonymization in the finance sector. Hernandez et al. [14] reviews data generation techniques for tabular health records anonymization. Raghunathan [15] reviews synthetic data anonymization techniques that preserve the statistical properties of a dataset. Nalepa et al. [16] reviews data augmentation techniques for brain-tumor segmentation. Bayer et al. [17] distinguishes augmentation techniques for text classification into feature and data space, while providing an extensive overview of augmentation methods within this domain. However, the taxonomy proposed and feature space augmentation methods are not necessarily specific to the domain. Shorten et al. [18], Chen et al. [19], Feng et al. [7] and Liu et al. [20] also review data augmentation techniques for text data. Yi et al. [21] review Generative Adversarial Network architectures for medical imaging. Wang et al. [22] reviews face data augmentation techniques. Shorten et al. [23] and Khosla et al. [24] discuss techniques for image data augmentation. Iwana et al. [25] and Wen et al. [26] also review time series data augmentation techniques. Zhao et al. [27] review data augmentation techniques for graph data. The analysis of related literature

Table 1: Related literature reviews published since 2019.

Reference	Data type	ML problem	Domain	Observations
Assefa et al. [1]	—	Differential privacy	Finance	Analysis of applications, motivation and properties of synthetic data for anonymization.
Hernandez et al. [14]	Tabular	Differential privacy	Healthcare	Focus on GANs.
Raghunathan [15]	Tabular	Differential privacy	Statistics	Focus on general definitions such as differential privacy and statistical disclosure control.
Nalepa et al. [16]	Image	Segmentation	Medicine	Analysis of algorithmic applications on a 2018 brain-tumor segmentation challenge.
Bayer et al. [17]	Text	Classification	—	Distinguish 100 methods into 12 groups.
Shorten et al. [18]	Text	Deep Learning	—	General overview of text data augmentation.
Chen et al. [19]	Text	Few-shot Learning	—	Augmentation techniques for machine learning with limited data
Feng et al. [7]	Text	—	—	Overview of augmentation techniques and applications on NLP tasks.
Liu et al. [20]	Text	—	Various	Analysis of industry use cases of data augmentation in NLP. Emphasis on input level data augmentation.
Yi et al. [21]	Image	—	Medicine	Emphasis on GANs.
Wang et al. [22]	Image	Deep Learning	—	Regularization techniques using facial image data. Emphasis on Deep Learning generative models.
Shorten et al. [23]	Image	Deep Learning	—	Emphasis on data augmentation as a regularization technique.
Khosla et al. [24]	Image	—	—	Broad overview of image data augmentation. Emphasis on traditional approaches.
Iwana et al. [25]	Time series	Classification	—	Defined a taxonomy for time series data augmentation.
Wen et al. [26]	Time series	Various	—	Analysis of data augmentation methods for classification, anomaly detection and forecasting.
Zhao et al. [27]	Graph	Various	—	Graph data augmentation for supervised and self-supervised learning.
Khalifa et al. [28]	Image	—	Various	General overview of image data augmentation and relevant domains of application.

70 reviews ¹ is shown in Table 1.

71 The different taxonomies established in the literature follow a similar philosophy, but vary in terminology
72 and are often specific to the technique discussed. Regardless, it is possible to establish a broader taxonomy
73 without giving up on specificity. This study provides a joint overview of the different data generation
74 approaches, domains and ML techniques where data generation is being used, as well as a common
75 taxonomy across domains. It extends the analyses found in these articles and uses the compiled knowledge
76 to identify research gaps. We compare the strengths and weaknesses of the models developed within each

¹Results obtained using Google Scholar, limited to articles published since 2019, using the search query ("synthetic data generation" OR "oversampling" OR "imbalanced learning" OR "data augmentation") AND ("literature review" OR "survey"). Retrieved on August 11th, 2022. More articles were added later whenever found relevant.

77 of these fields. Finally, we identify possible future research directions to address some of the limitations
78 found. The contributions of this paper are summarized below:

- 79 • Bridge different ML concepts using synthetic data generation in its core (Algorithmic applications +
80 Review of the State-of-the-art).
- 81 • Propose a synthetic data generation/data augmentation taxonomy to resolve the ambiguity in the
82 literature (Data augmentation taxonomy).
- 83 • Characterize all relevant data generation methods using the proposed taxonomy.
- 84 • Discuss the ML techniques in which synthetic data generation/data augmentation is used, beyond
85 regularization and consolidate the current data generation mechanisms across the different techniques
86 (Algorithmic Applications).
- 87 • Bring to light the key challenges of synthetic data generation and put forward possible research
88 directions in the future.

89 1.2 Paper Organization

90 This paper is organized as follows: Section 2 defines and formalizes the different concepts, goals, trade-offs
91 and motivations related to synthetic data generation. Section 3 establishes the taxonomy used to categorize
92 all the methods described in the paper. Section ?? reviews synthetic data generation mechanisms in the
93 feature space. Section ?? reviews synthetic data generation mechanisms in the input space. Section 4
94 describes the applications of synthetic data in ML methods. Section 5 reviews performance evaluation
95 methods of synthetic data generation mechanisms. Section 6 summarizes the main findings and discusses
96 limitations and possible research directions in the state-of-the-art. Section 7 presents the main conclusions
97 drawn from this study.

98 2 Background

99

100 In this section we define basics concepts, common goals, trade-offs and motivations regarding the generation
101 of synthetic data in ML. We define synthetic data generation as the production of observations using
102 a generative model (regardless of its nature) that resemble naturally occurring observations within
103 a certain domain. It requires access to either a training dataset, a generative process, or a data
104 stream. However, additional requirements might be imposed depending on the ML task being developed.
105 For example, to generate artificial data for regularization purposes in supervised learning (*i.e.*, data
106 augmentation) the training dataset must be annotated [CITATION]. The generation of synthetic data for
107 anonymization purposes assumes synthetic datasets to be different from the original data, while following
108 the same statistical properties [CITATION]. Domain knowledge may also be necessary to encode specific
109 relationships among features into the generative process.

2.1 Use Cases

The breach of sensitive information is an important barrier to the sharing of datasets, especially when it concerns personal information [29]. A common solution for this problem is the generation of synthetic data without identifiable information. Generally speaking, ML tasks that require data with sensitive information are not compromised when using synthetic data. The experiment conducted by Patki et al. [30] using relational datasets showed that in 11 out of 15 comparisons ($\approx 73\%$), practitioners performing predictive modelling tasks using fully synthetic datasets performed the same or better than those using the original dataset. This topic is discussed in Section 4.1.

A common problem in the training of deep neural networks are their capacity to generalize [31] (*i.e.*, reduce the difference in classification performance between known and unseen observations). Data augmentation is a common method to address this problem. The generation of synthetic observations increases the range of the possible input space used in the training phase, which reduces the performance difference between known and unseen observations. Although other regularization methods exist, data augmentation is a useful method since it does not affect the choice in the architecture of the ML classifier and does not exclude the usage of other regularization methods. In domains such as computer vision and NLP, data augmentation is also used to improve the robustness of models against adversarial attacks [32, 33]. These topics are discussed into higher detail in Section 4.2.

In supervised learning, synthetic data generation is often motivated by the need to balance target class distributions (*i.e.*, oversampling). Since most ML classifiers are designed to perform best with balanced datasets, defining an appropriate decision boundary to distinguish rare classes becomes difficult [34]. Although there are other approaches to address imbalanced learning, oversampling techniques are generally easier to implement since they do not involve modifications to the classifier. This topic is discussed into higher detail in Section 4.4.

In supervised learning projects where labeled data is not readily available, but can be labeled, an Active Learning (AL) method may be used to improve the labelling process. AL aims to reduce the cost of producing training datasets by finding the most informative observations to label and feed into the classifier [35]. In this case, the generation of synthetic data is particularly useful to reduce the amount of labelled data required for a successful ML project and its costs. A similar motivation applies to the case of few-shot learning: small datasets may be expanded with synthetic data [36]. These topics are discussed in Sections 4.5 and 4.6.

The two other techniques reliant on synthetic data generation is Semi-supervised and Self-supervised learning. The former leverages both labeled and unlabeled data in the training phase, simultaneously. Most of the methods in the literature apply perturbations on the training data as part of the training procedure [37]. Self-supervised learning is a technique used to train neural networks in the absence of labeled data. Both techniques use synthetic data generation as an internal procedure for most of these methods. These techniques are discussed in Sections 4.7 and 4.8.

2.2 Problem Formulation

The original dataset, $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$, is a collection of real observations and is distinguished according to whether a target feature exists, $\mathcal{D}_L = ((x_i, y_i))_{i=1}^l$, or not, $\mathcal{D}_U = (x_i)_{i=1}^u$. All three datasets, \mathcal{D} ,

150 \mathcal{D}_L and \mathcal{D}_U consist of ordered collections with lengths $l + u$, l and u , respectively. Synthetic data
 151 generation is performed using a generator, $f_{gen}(x; \tau) = \tilde{x}$, where τ defines the generation policy (*i.e.*, its
 152 hyperparameters), $x \in \mathcal{D}$ is an observation and $\tilde{x} \in \mathcal{D}^s$ is a synthetic observation. Analogous to \mathcal{D} , the
 153 synthetic dataset, \mathcal{D}^s , is also distinguished according to whether there is an assignment of a target feature,
 154 $\mathcal{D}_L^s = ((\tilde{x}_j, \tilde{y}_j))_{j=1}^{l'}$, or not, $\mathcal{D}_U^s = (\tilde{x}_j)_{j=1}^{u'}$.

155 Depending on the ML task, it may be relevant to establish metrics to measure the quality of \mathcal{D}^s . In this
 156 case, a metric $f_{qual}(\mathcal{D}^s, \mathcal{D})$ is used to determine the level of similarity/dissimilarity between \mathcal{D} and \mathcal{D}^s . In
 157 addition, a performance metric to estimate the performance of a model on the objective task, f_{per} , may be
 158 used to determine the appropriateness of a model with parameters θ , *i.e.*, f_θ . The generator’s goal is to
 159 generate \mathcal{D}^s with arbitrary length, given $\mathcal{D} \sim \mathbb{P}$ and $\mathcal{D}^s \sim \mathbb{P}^s$, such that $\mathbb{P}^s \approx \mathbb{P}$, $x_i \neq x_j \forall x_i \in \mathcal{D} \wedge x_j \in \mathcal{D}^s$.
 160 $f_{gen}(x; \tau)$ attempts to generate a \mathcal{D}^s that maximizes either f_{per} , f_{qual} , or a combination of both.

161 3 Data Generation Taxonomy

162

163 The taxonomy proposed in this paper is a compilation of different definitions found in the literature,
 164 along with other traits that vary among domains and generation techniques. Within image data studies,
 165 Shorten et al. [23] and Khalifa et al. [28] divide data augmentation techniques into “basic” or “classical”
 166 approaches and deep learning approaches. In both cases, the former refers to domain-specific generation
 167 techniques, while the latter may be applied to any type of data. Iwana et al. [25] proposes a time-series data
 168 augmentation taxonomy divided in four families: (1) Decomposition, (2) Pattern mixing, (3) Generative
 169 models and (4) Decomposition. With exception to generative models, the majority of the methods
 170 presented in the remaining families are well established and domain specific. Hernandez et al. [14] defines
 171 a taxonomy for synthetic tabular data generation approaches divided in three types of approaches: (1)
 172 Classical, (2) Deep learning and (3) Others. Most taxonomies found followed similar definitions with
 173 variations in terminology or distinction criteria. In addition, all taxonomies with categories defined as
 174 “basic”, “traditional” or “classical” use these to characterize domain-specific transformations.

175 Within the taxonomies considered, none of them consider how a generation mechanism employs \mathcal{D} into the
 176 generation process or, if applicable, the training phase. However, it is important to understand whether a
 177 generation mechanism randomly selects x and a set of close neighbors, thus considering local information
 178 only, or considers the overall dataset or data distribution for the selection of x and/or generation of \tilde{x} .
 179 Our proposed taxonomy is depicted in Figure 1. It characterizes data generation mechanisms using four
 180 properties:

- 181 1. Architecture. Defines the broader type of data augmentation. It is based on domain specificity, archi-
 182 tecture type or data transformations using a heuristic or random perturbation process. Generation
 183 techniques that apply a form of random perturbation, interpolation or geometric transformation to
 184 the data with some degree of randomness are considered randomized approaches. Typical, domain-
 185 specific data generation techniques are considered traditional architectures. These techniques apply
 186 transformations to a data point using *a priori* domain knowledge. Generative models based on
 187 neural network architectures are defined as network-based. These architectures attempt to either
 188 generate observations in the feature space and/or by producing observations that are difficult to
 189 distinguish from the original dataset.
- 190 2. Application level. Refers to the phase of the ML pipeline where the generative process is included.

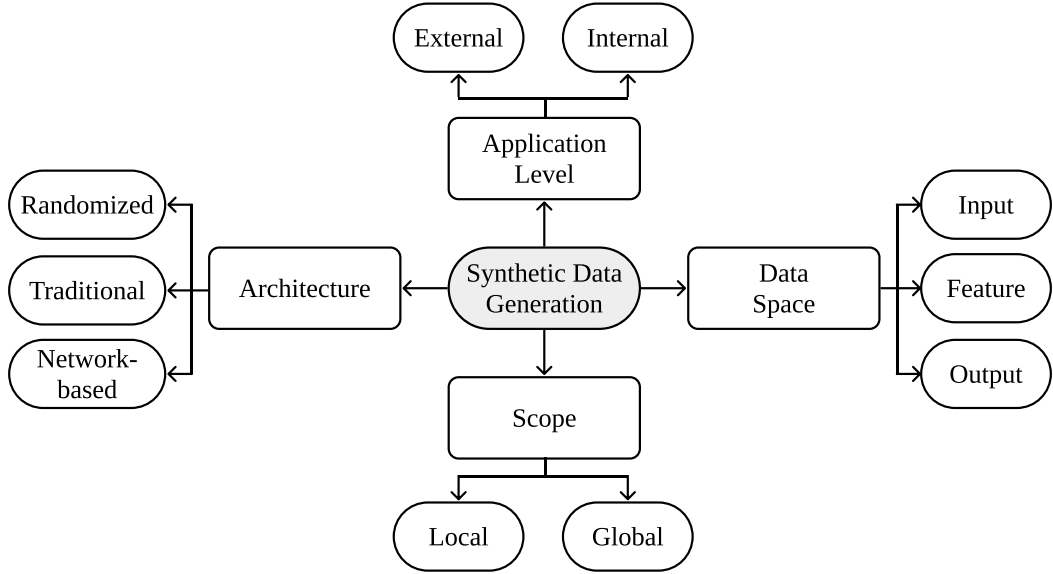


Figure 1: General taxonomy of data generation mechanisms proposed in this paper.

Generative models are considered internal if they are used alongside the primary ML task, whereas models used prior to the development of the primary ML task are considered external.

3. **Scope.** Considers the usage of the original dataset’s properties. Generative models that consider the density of the data space, statistical properties of \mathcal{D} , or attempt to replicate specific relationships found in \mathcal{D} are considered to have a global scope, whereas generative models that consider a single observation and/or a set of close neighbors are considered to have a local scope. On the one hand, generative models with a local scope do not account for \mathbb{P}^s but allow for a larger diversity of candidate x^s and higher variance within \mathcal{D}^s . On the other hand, generative models with a global scope have a higher capacity to model \mathbb{P}^s but produce candidate x^s with lower diversity and lower variance within \mathcal{D}^s .
4. **Data space.** Refers to the type data representation used to apply the generative model. Generation mechanisms can be applied using the raw dataset (*i.e.*, on the input space), an embedded representation of the data (*i.e.*, on the feature space) or based on the target feature (*i.e.*, on the output space).

Throughout the analysis of the different types of generation mechanisms, all relevant methods were characterized using this taxonomy and listed in Table 2.

Table 2: Summary of synthetic data generation mechanisms discussed in this work.

Algorithm	ML problem	Domain	Architecture	Level	Data Space	Scope
SynSys [38]	Reg.	Healthcare	[Statistical]	External	Input	Global
CTGAN [39]		—	Net.	External	Feature	Global
SenseGen [40]	Anon. + Reg.	Sensor	Net. + [Stat.]	External	Input	Global
SDV [30]	Anon.	—	[Copula]	External	Input	Global
MST [41]	DP	—	[Statistical]	External	Input	Global
QUAIL [42]	DP	—	—	External	—	—
SuperQUAIL [43]	DP					

Continued on next page

Table 2: Summary of synthetic data generation mechanisms discussed in this work.

Algorithm	ML problem	Domain	Architecture	Level	Data Space	Scope
MWEM [44]	DP	Healthcare				
MWEM-PGM [45]	DP					
PrivBayes [46]	DP					
DPGAN [47]	DP					
DPCTGAN [42]	DP					
DP-CTGAN [48]	DP					
PATE-GAN [49]	DP					
PATECTGAN [42]	DP					
FEM [50]	DP					
RAP [51]	DP					
Kamino [52]	DP					
RON-GAUSS [53]	DP					
SMOTE [54]	Ovs					
SMOTENC [54]	Ovs					
SMOTEN [54]	Ovs					
Borderline-SMOTE [55]	Ovs					
G-SMOTE [56]	Ovs					
ADASYN [57]	Ovs					
KernelADASYN [58]	Ovs					
Safe-level SMOTE [59]	Ovs					
LR-SMOTE [60]	Ovs					
K-means SMOTE [61]	Ovs					
CGAN [62]	Ovs					
K-means CTGAN [63]	Ovs					
G-SMOTER [64]	Ovs + Reg					
SMOTER [65]	Ovs + Reg					

4 Algorithmic applications

In this section we discuss the data generation mechanisms for the different contexts where they are applied. We emphasize the constraints in each problem that condition the way generation mechanisms are used.

CTGAN [39]

4.1 Dataset Anonymization

Synthetic data generation is a technique used to produce synthetic, anonymized versions of datasets [29]. It is considered a good approach to share sensitive data without compromising significantly a given data mining task [66, 67]. Traditional data anonymization techniques, as well as federated learning are two other viable solutions for privacy-preserving data publishing tasks, but contain drawbacks [14]. On the one hand, traditional data anonymization requires domain knowledge, is labor intensive and remains susceptible to disclosure [68]. On the other hand, federated learning is a technically complex task that

consists on training ML classifiers on edge devices and aggregating temporarily updated parameters on a centralized server, instead of aggregating the training data [69]. Although it prevents sharing sensitive data, its applicability is dependent on the task. Dataset anonymization via synthetic data generation attempts to balance disclosure risk and data utility in the final synthetic dataset. The goal is to ensure observations are not identifiable and the relevant data mining tasks are not compromised [70, 71].

The generation of synthetic datasets allow a more flexible approach to the successful implementation of ML tasks. To do this, it is important to guarantee that sensitive information in \mathcal{D} is not leaked into \mathcal{D}^s . Differential privacy (DP), a formalization of privacy, offers strict theoretical privacy guarantees [42]. A differentially private generation mechanism produces a synthetic dataset, regulated by the privacy parameter ϵ , with statistically indistinguishable results when using either \mathcal{D} or neighboring datasets $\mathcal{D}' = \mathcal{D} \setminus \{x\}$, for any $x \in \mathcal{D}$. A synthetic data generation model (f_{gen}) guarantees (ϵ, δ) -differential privacy if $\forall S \subseteq \text{Range}(f_{gen})$ all $\mathcal{D}, \mathcal{D}'$ differing on a single entry [44]:

$$Pr[f_{gen}(\mathcal{D}) \in S] \leq e^\epsilon \cdot Pr[f_{gen}(\mathcal{D}') \in S] + \delta \quad (1)$$

In this case, ϵ is a non-negative number defined as the privacy budget. A lower ϵ guarantees a higher level of privacy, but reduces the quality of the produced synthetic data. The generation of DP synthetic data is especially appealing since DP is not affected by post-processing; any ML pipeline may be applied using \mathcal{D}^s without a loss of differential privacy.

Despite the formalization and the ability to quantify differential privacy, there are popular synthetic data-based anonymization approaches that perform this task without DP guarantees. Specifically, the Synthetic Data Vault (SDV) [30] is a method for database anonymization that uses Gaussian Copula models for generating data. However, this method allows the usage of other generation mechanisms. A posterior extension of SDV was proposed to generate data using a CTGAN [39] and to handle sequential tabular data using a conditional probabilistic auto-regressive neural network [72].

The MST mechanism [41] is a general approach that produces differentially private data. It uses the Private-PGM mechanism [45] that relies on Probabilistic Graphical Models (PGM) to generate synthetic data. PGM models are most commonly used when it is important to maintain the pre-existing statistical properties and relationships between features [73].

The Quail-ified Architecture to Improve Learning (QUAIL) is a DP method that produces differentially private data by distributing the privacy budget between a DP classifier to attribute the target labels onto \mathcal{D}^s and the data generator. QUAIL works as a framework that involves the adoption of both a DP classifier and generator. Originally, it was experimented using DPGAN [47], DPCTGAN, MWEM [44], PATE-GAN [49] and PATE-CTGAN.

Anonymizing data using synthetic data generation in the financial sector [1].

A benchmark of various differentially private synthetic data generation mechanisms [74].

Guidelines for effective synthetic data generation [29]

4.2 Regularization in Supervised Learning

256 The performance of Machine Learning models is highly dependent on the quality of the training dataset
257 used [75, 76]. The presence of imbalanced and/or small datasets, target labels incorrectly assigned, outliers
258 and high dimensional input spaces reduce the prospects of a successful machine learning (ML) model
259 implementation [76, 77, 78]. In the case of deep learning, for example, these models are often limited by a
260 natural inclination to overfitting, label noise memorization and catastrophic forgetting [79]. Regularization
261 methods are the typical approach to address these problems, but producing robust ML solutions is still a
262 challenge [31].

263 It is frequently assumed that the training data is sampled from a fixed data source, it is balanced and does
264 not contain label noise. Under these conditions, the resulting ML classifier is expected to achieve good
265 generalization performance [80]. Although, in practical applications, this is rarely the case. When the
266 training data is not representative of the true population, or the model is over-parametrized, it becomes
267 particularly prone to overfitting [81]. Regularization methods attempt to address these limitations. They
268 can be divided into three categories [82]:

- 269 1. Output level modifications. Transforms the labels in the training data.
- 270 2. Algorithmic level modifications. Modifies the classifier’s architecture, loss function or other compo-
271 nents in the training procedure.
- 272 3. Input level modifications. Modifies the training dataset by expanding it with synthetic data.

273 The last approach, input level modifications, is known as data augmentation. Data augmentation is used to
274 increase the size and data variability of data in a training dataset, by producing synthetic observations [83,
275 84]. Since it is applied at the data level, it can be used for various types of problems and classifiers [85].

276 Problems such as fraud detection and healthcare are frequently tackled via synthetic data generation [86].

277 “Su et al. [78] show that 70.97% of images can be misclassified by changing just one pixel” Shorten et al.
278 [23]

279 “Moreover, the current research about so called adversarial attacks on CNNs showed that deep neural
280 networks can be easily fooled into misclassification of images just by partial rotations and image translation
281 [1], adding the noise to images [5] and even changing one, skillfully selected pixel in the image [6].”
282 Mikołajczyk et al. [87]

283 Data augmentation can also be used to improve a model’s robustness against adversarial attacks.

284 4.3 Time Series

285 Synsys [38] approaches time-series using both Hidden Markov and regression models. They show the
286 method’s effectiveness in the Healthcare domain with limited ground truth data by comparing it to models
287 trained using only real data. A related model, Sensegen [40], uses an adversarial training approach to
288 train an LSTM that predicts the parameters of Gaussian Mixture Models (GMM) at each time stamp,
289 using real data as an input. Finally, the GMM estimations are used to sample synthetic data.

290 Generative adversarial networks in time series

291 4.4 Oversampling

292

293 KernelADASYN [58]

294 The original author of SMOTE recently published the paper “Efficient Augmentation for Imbalanced Deep
295 Learning” [88]

296 4.5 Active Learning

297

298 4.6 Few-shot Learning

299

300 Analysis of six feature space data augmentation techniques for few-shot learning [36]

301 FlipDA [89]

302 Data generation can be used to address Few-shot learning in three ways [90]: (1) transforming samples
303 from the dataset, (2) transforming samples from a weakly labeled or unlabeled dataset, or (3) transforming
304 samples from similar datasets.

305 4.7 Semi-supervised Learning

306

307 Synthetic data generation for semi-supervised learning given limited labeled data regarding the COVID-19
308 pandemic [91].

309 Extensive literature review on semi-supervised learning [37]

310 4.8 Self-supervised Learning

311

312 5 Evaluating the Quality of Synthetic Data

313

314 The log-likelihood (and equivalently the Kullback-Leibler Divergence) is a de-facto standard to train and
315 evaluate generative models [11]. Other common metrics include Parzen window estimates, which Theis
316 et al. [11] show that these metrics behave independently and should generally be avoided. Therefore, it is
317 necessary to evaluate generative models with respect to the application these models are being developed
318 for.

319 The evaluation of generative models should quantify three key aspects of synthetic data [12]:

- 320 1. Fidelity
- 321 2. Diversity
- 322 3. Generalization

323 The 3-dimensional metric proposed by Alaa et al. [12] quantifies these aspects via the combination of
324 three metrics (α -Precision, β -Recall and Authenticity) for various application domains.

325 5.1 Statistical Divergence Metrics

326 5.2 Precision/Recall Metrics

327 6 Discussion

328

329 6.1 Main Findings

330 The combination of data generation strategies is an approach commonly found in different problems, such
331 as self-supervised learning [5]. It can be more frequently found in text data applications [17] and image
332 data [CITATION].

333 6.1.1 RQ1: bla bla bla

334 6.1.2 RQ2: bla bla bla

335 6.1.3 RQ3: bla bla bla

336 6.2 Limitations

337 Research across the different applications appears to be sandboxed even though all techniques integrate
338 synthetic data in its core.

339 Given the breadth and complexity of input-level and feature-level data generation mechanisms, it is
340 increasingly important to find a method to efficiently determine the most appropriate data generation
341 policies. However, the complexity of this task is determined by various factors: different data types, ML
342 problems, model architectures, computational resources, performance metrics and contextual constraints.
343 Auto-augmentation and meta learning aim to address this challenge and are still subject to active
344 research.

345 The evaluation of anonymization techniques lack standardized, objective and reliable performance metrics
346 and benchmark datasets to allow an easier comparison across classifiers to evaluate key aspects of data
347 anonymization (resemblance, utility, privacy and performance). These datasets should contain mixed data
348 types (*i.e.*, a combination of categorical, ordinal, continuous and discrete features) and the metrics should
349 evaluate the performance of different data mining tasks along with the anonymization reliability. This
350 problem appears to be universal across domains. For example, Hernandez et al. [14] observed the lack of
351 a universal method or metric to report the performance synthetic data generation algorithms for tabular
352 health records. Therefore, in order to facilitate the usage of these techniques in industry domains, these
353 benchmarks must also be realistic. Rosenblatt et al. [42] attempts to address this problem by proposing a
354 standardized evaluation methodology using standard datasets and real-world industry applications.

355 Computational cost and inconsistent quality of synthetic data generated with GANs (*e.g.*, mode collapse).

356 Unlike with data privacy solutions, data augmentation techniques generally do not consider the simi-
357 larity/dissimilarity of synthetic data. The study of quality metrics for supervised learning may reduce
358 computational overhead and experimentation time. No studies related to the relationship of quality
359 metrics and performance in the primary ML task were found [CONFIRM!!!].

360 There is not a clear understanding of what types of data augmentation methods are more appropriate
361 according to different model architectures, ML tasks or domains and the reason why they work better or
362 worse depending on the task. In addition, it is still unclear *why* data augmentation works. Research on
363 this topic lacks depth and fails to address the theoretical underpinnings [7].

364 “Dao et al. (2019) note that “data augmentation is typically performed in an ad-hoc manner with little
365 understanding of the underlying theoretical principles”, and claim the typical explanation of DA as
366 regularization to be insufficient.” [7]

367 There is a lack of research on oversampling solutions to generate synthetic data with mixed data types
368 and datasets with exclusively non metric features.

369 There is a lack of methods adapted to use categorical features for tabular data.

370 There is a paucity of research on the usage of probabilistic-based generation mechanisms in oversampling.

371 There is no clear understanding of the most appropriate data augmentation techniques used to train
372 self-supervised models and how their behavior and performance varies according to the data generation
373 method used.

374 oversampling does not seem to be a relevant source of bias in behavioral research and does not appear to
375 have an appreciably different effect on results for directly versus indirectly oversampled variables [92]

6.3 Research directions

Quantifying the quality of the generated data:

1. Realistic
2. Similarity
3. Usefulness (determine purpose and relevant performance metric)
4. Understand the relationship between the 3 factors

7 Conclusions

References

- [1] Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. “Generating synthetic data in finance: opportunities, challenges and pitfalls”. In: *Proceedings of the First ACM International Conference on AI in Finance*. 2020, pp. 1–8.
- [2] Samuli Laine and Timo Aila. “Temporal ensembling for semi-supervised learning”. In: *International Conference on Learning Representations (ICLR)*. Vol. 4. 5. 2017, p. 6.
- [3] Joao Fonseca, Georgios Douzas, and Fernando Bacao. “Improving imbalanced land cover classification with K-Means SMOTE: Detecting and oversampling distinctive minority spectral signatures”. In: *Information* 12.7 (2021), p. 266.
- [4] Yoon-Yeong Kim, Kyungwoo Song, JoonHo Jang, and Il-Chul Moon. “LADA: Look-Ahead Data Acquisition via Augmentation for Deep Active Learning”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 22919–22930.
- [5] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. “Bootstrap your own latent-a new approach to self-supervised learning”. In: *Advances in neural information processing systems* 33 (2020), pp. 21271–21284.
- [6] Jiang-Jing Lv, Xiao-Hu Shao, Jia-Shui Huang, Xiang-Dong Zhou, and Xi Zhou. “Data augmentation for face recognition”. In: *Neurocomputing* 230 (2017), pp. 184–196.
- [7] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. “A Survey of Data Augmentation Approaches for NLP”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 968–988. DOI: [10.18653/v1/2021.findings-acl.84](https://doi.org/10.18653/v1/2021.findings-acl.84). URL: <https://aclanthology.org/2021.findings-acl.84>.
- [8] Talha Mahboob Alam, Kamran Shaukat, Ibrahim A Hameed, Suhuai Luo, Muhammad Umer Sarwar, Shakir Shabbir, Jiaming Li, and Matloob Khushi. “An investigation of credit card default prediction in the imbalanced datasets”. In: *IEEE Access* 8 (2020), pp. 201173–201198.

- [9] Terrance DeVries and Graham W Taylor. “Dataset augmentation in feature space”. In: *arXiv preprint arXiv:1702.05538* (2017).
- [10] Diederik P Kingma, Max Welling, et al. “An introduction to variational autoencoders”. In: *Foundations and Trends® in Machine Learning* 12.4 (2019), pp. 307–392.
- [11] L Theis, A van den Oord, and M Bethge. “A note on the evaluation of generative models”. In: *International Conference on Learning Representations (ICLR 2016)*. 2016, pp. 1–10.
- [12] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. “How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 290–306.
- [13] Miro Mannino and Azza Abouzied. “Is this real? Generating synthetic data that looks real”. In: *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 2019, pp. 549–561.
- [14] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. “Synthetic Data Generation for Tabular Health Records: A Systematic Review”. In: *Neurocomputing* (2022).
- [15] Trivellore E Raghunathan. “Synthetic data”. In: *Annual Review of Statistics and Its Application* 8 (2021), pp. 129–140.
- [16] Jakub Nalepa, Michal Marcinkiewicz, and Michal Kawulok. “Data augmentation for brain-tumor segmentation: a review”. In: *Frontiers in computational neuroscience* 13 (2019), p. 83.
- [17] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. “A survey on data augmentation for text classification”. In: *ACM Computing Surveys* (2021).
- [18] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. “Text data augmentation for deep learning”. In: *Journal of big Data* 8.1 (2021), pp. 1–34.
- [19] Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. “An empirical survey of data augmentation for limited data learning in NLP”. In: *arXiv preprint arXiv:2106.07499* (2021).
- [20] Pei Liu, Xuemin Wang, Chao Xiang, and Weiye Meng. “A survey of text data augmentation”. In: *2020 International Conference on Computer Communication and Network Security (CCNS)*. IEEE. 2020, pp. 191–195.
- [21] Xin Yi, Ekta Walia, and Paul Babyn. “Generative adversarial network in medical imaging: A review”. In: *Medical image analysis* 58 (2019), p. 101552.
- [22] Xiang Wang, Kai Wang, and Shiguo Lian. “A survey on face data augmentation for the training of deep neural networks”. In: *Neural computing and applications* 32.19 (2020), pp. 15503–15531.
- [23] Connor Shorten and Taghi M Khoshgoftaar. “A survey on image data augmentation for deep learning”. In: *Journal of big data* 6.1 (2019), pp. 1–48.
- [24] Cherry Khosla and Baljit Singh Saini. “Enhancing performance of deep learning models with different data augmentation techniques: A survey”. In: *2020 International Conference on Intelligent Engineering and Management (ICIEM)*. IEEE. 2020, pp. 79–85.
- [25] Brian Kenji Iwana and Seiichi Uchida. “An empirical survey of data augmentation for time series classification with neural networks”. In: *Plos one* 16.7 (2021), e0254841.
- [26] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. “Time series data augmentation for deep learning: a survey”. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Zhi-Hua Zhou. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 4653–4660.
- [27] Tong Zhao, Gang Liu, Stephan Günnemann, and Meng Jiang. “Graph Data Augmentation for Graph Machine Learning: A Survey”. In: *arXiv preprint arXiv:2202.08871* (2022).

- [28] Nour Eldeen Khalifa, Mohamed Loey, and Seyedali Mirjalili. “A comprehensive survey of recent trends in deep learning for digital images augmentation”. In: *Artificial Intelligence Review* (2021), pp. 1–27.
- [29] Fida K Dankar and Mahmoud Ibrahim. “Fake it till you make it: Guidelines for effective synthetic data generation”. In: *Applied Sciences* 11.5 (2021), p. 2158.
- [30] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. “The synthetic data vault”. In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE. 2016, pp. 399–410.
- [31] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. “Understanding deep learning (still) requires rethinking generalization”. In: *Communications of the ACM* 64.3 (2021), pp. 107–115.
- [32] Yi Zeng, Han Qiu, Gerard Memmi, and Meikang Qiu. “A data augmentation-based defense method against adversarial attacks in neural networks”. In: *International Conference on Algorithms and Architectures for Parallel Processing*. Springer. 2020, pp. 274–289.
- [33] John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. “Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp”. In: *arXiv preprint arXiv:2005.05909* (2020).
- [34] José A Sáez, Bartosz Krawczyk, and Michał Woźniak. “Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets”. In: *Pattern Recognition* 57 (2016), pp. 164–178.
- [35] Joao Fonseca, Georgios Douzas, and Fernando Bacao. “Increasing the Effectiveness of Active Learning: Introducing Artificial Data Generation in Active Learning for Land Use/Land Cover Classification”. In: *Remote Sensing* 13.13 (2021), p. 2619.
- [36] Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and William Campbell. “A Closer Look At Feature Space Data Augmentation For Few-Shot Intent Classification”. In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. 2019, pp. 1–10.
- [37] Jesper E Van Engelen and Holger H Hoos. “A survey on semi-supervised learning”. In: *Machine Learning* 109.2 (2020), pp. 373–440.
- [38] Jessamyn Dahmen and Diane Cook. “SynSys: A synthetic data generation system for healthcare applications”. In: *Sensors* 19.5 (2019), p. 1181.
- [39] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. “Modeling tabular data using conditional gan”. In: *Advances in Neural Information Processing Systems* 32 (2019).
- [40] Moustafa Alzantot, Supriyo Chakraborty, and Mani Srivastava. “Sensegen: A deep learning architecture for synthetic sensor data generation”. In: *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE. 2017, pp. 188–193.
- [41] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. “Winning the NIST Contest: A scalable and general approach to differentially private synthetic data”. In: *Journal of Privacy and Confidentiality* 11.3 (2021).
- [42] Lucas Rosenblatt, Xiaoyan Liu, Samira Pouyanfar, Eduardo de Leon, Anuj Desai, and Joshua Allen. “Differentially private synthetic data: Applied evaluations and enhancements”. In: *arXiv preprint arXiv:2011.05537* (2020).
- [43] Lucas Rosenblatt, Joshua Allen, and Julia Stoyanovich. “Spending Privacy Budget Fairly and Wisely”. In: *arXiv preprint arXiv:2204.12903* (2022).
- [44] Moritz Hardt, Katrina Ligett, and Frank McSherry. “A simple and practical algorithm for differentially private data release”. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 2*. 2012, pp. 2339–2347.

- [45] Ryan McKenna, Daniel Sheldon, and Gerome Miklau. “Graphical-model based estimation and inference for differential privacy”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4435–4444.
- [46] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. “Privbayes: Private data release via bayesian networks”. In: *ACM Transactions on Database Systems (TODS)* 42.4 (2017), pp. 1–41.
- [47] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. “Differentially private generative adversarial network”. In: *arXiv preprint arXiv:1802.06739* (2018).
- [48] Mei Ling Fang, Devendra Singh Dhami, and Kristian Kersting. “Dp-CTGAN: Differentially private medical data generation using CTGANs”. In: *International Conference on Artificial Intelligence in Medicine*. Springer. 2022, pp. 178–188.
- [49] 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.
- [50] Giuseppe Vietri, Grace Tian, Mark Bun, Thomas Steinke, and Steven Wu. “New oracle-efficient algorithms for private synthetic data release”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9765–9774.
- [51] Sergul Aydore, William Brown, Michael Kearns, Krishnaram Kenthapadi, Luca Melis, Aaron Roth, and Ankit A Siva. “Differentially private query release through adaptive projection”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 457–467.
- [52] Chang Ge, Shubhankar Mohapatra, Xi He, and Ihab F Ilyas. “Kamino: Constraint-aware differentially private data synthesis”. In: *arXiv preprint arXiv:2012.15713* (2020).
- [53] Thee Chanyaswad, Changchang Liu, and Prateek Mittal. “Ron-gauss: Enhancing utility in non-interactive private data release”. In: *Proceedings on Privacy Enhancing Technologies* 2019.1 (2019), pp. 26–46.
- [54] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.
- [55] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. “Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning”. In: *International conference on intelligent computing*. Springer. 2005, pp. 878–887.
- [56] Georgios Douzas and Fernando Bacao. “Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE”. In: *Information Sciences* 501 (2019), pp. 118–135.
- [57] Haibo He, Yang Bai, Eduardo A Garcia, and Shutao Li. “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE. 2008, pp. 1322–1328.
- [58] Bo Tang and Haibo He. “KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning”. In: *2015 IEEE congress on evolutionary computation (CEC)*. IEEE. 2015, pp. 664–671.
- [59] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. “Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem”. In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2009, pp. 475–482.
- [60] XW Liang, AP Jiang, T Li, YY Xue, and GT Wang. “LR-SMOTE—An improved unbalanced data set oversampling based on K-means and SVM”. In: *Knowledge-Based Systems* 196 (2020), p. 105845.
- [61] Georgios Douzas, Fernando Bacao, and Felix Last. “Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE”. In: *Information Sciences* 465 (2018), pp. 1–20.

- [62] Georgios Douzas and Fernando Bacao. “Effective data generation for imbalanced learning using conditional generative adversarial networks”. In: *Expert Systems with applications* 91 (2018), pp. 464–471.
- [63] Chunsheng An, Jingtong Sun, Yifeng Wang, and Qingjie Wei. “A K-means Improved CTGAN Oversampling Method for Data Imbalance Problem”. In: *2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS)*. IEEE. 2021, pp. 883–887.
- [64] Luís Camacho, Georgios Douzas, and Fernando Bacao. “Geometric SMOTE for regression”. In: *Expert Systems with Applications* (2022), p. 116387.
- [65] Luís Torgo, Rita P Ribeiro, Bernhard Pfahringer, and Paula Branco. “Smote for regression”. In: *Portuguese conference on artificial intelligence*. Springer. 2013, pp. 378–389.
- [66] Jennifer Taub, Mark Elliot, Maria Pampaka, and Duncan Smith. “Differential correct attribution probability for synthetic data: an exploration”. In: *International Conference on Privacy in Statistical Databases*. Springer. 2018, pp. 122–137.
- [67] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. “Data Synthesis based on Generative Adversarial Networks”. In: *Proceedings of the VLDB Endowment* 11.10 (2018).
- [68] Jerome P Reiter. “New approaches to data dissemination: A glimpse into the future (?)” In: *Chance* 17.3 (2004), pp. 11–15.
- [69] Bin Yu, Wenjie Mao, Yihan Lv, Chen Zhang, and Yu Xie. “A survey on federated learning in data mining”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12.1 (2022), e1443.
- [70] Kalpana Singh and Lynn Batten. “Aggregating privatized medical data for secure querying applications”. In: *Future Generation Computer Systems* 72 (2017), pp. 250–263.
- [71] Ping Li, Tong Li, Heng Ye, Jin Li, Xiaofeng Chen, and Yang Xiang. “Privacy-preserving machine learning with multiple data providers”. In: *Future Generation Computer Systems* 87 (2018), pp. 341–350.
- [72] Kevin Zhang, Neha Patki, and Kalyan Veeramachaneni. “Sequential Models in the Synthetic Data Vault”. In: *arXiv preprint arXiv:2207.14406* (2022).
- [73] Jim Young, Patrick Graham, and Richard Penny. “Using Bayesian networks to create synthetic data”. In: *Journal of Official Statistics* 25.4 (2009), p. 549.
- [74] Yuchao Tao, Ryan McKenna, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. “Benchmarking differentially private synthetic data generation algorithms”. In: *arXiv e-prints* (2021), arXiv–2112.
- [75] Giuseppe Fenza, Mariacristina Gallo, Vincenzo Loia, Francesco Orciuoli, and Enrique Herrera-Viedma. “Data set quality in Machine Learning: Consistency measure based on Group Decision Making”. In: *Applied Soft Computing* 106 (2021), p. 107366.
- [76] Alon Halevy, Peter Norvig, and Fernando Pereira. “The unreasonable effectiveness of data”. In: *IEEE Intelligent Systems* 24.2 (2009), pp. 8–12.
- [77] Pedro Domingos. “A few useful things to know about machine learning”. In: *Communications of the ACM* 55.10 (2012), pp. 78–87.
- [78] Shaeke Salman and Xiuwen Liu. “Overfitting mechanism and avoidance in deep neural networks”. In: *arXiv preprint arXiv:1901.06566* (2019).
- [79] Zeke Xie, Fengxiang He, Shaopeng Fu, Issei Sato, Dacheng Tao, and Masashi Sugiyama. “Artificial neural variability for deep learning: On overfitting, noise memorization, and catastrophic forgetting”. In: *Neural computation* 33.8 (2021), pp. 2163–2192.

- 591 [80] Martin Benning and Martin Burger. “Modern regularization methods for inverse problems”. In: *Acta*
592 *Numerica* 27 (2018), pp. 1–111.
- 593 [81] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. “Deep learning: a statistical viewpoint”.
594 In: *Acta numerica* 30 (2021), pp. 87–201.
- 595 [82] Claudio Filipi Gonçalves dos Santos and João Paulo Papa. “Avoiding Overfitting: A Survey on
596 Regularization Methods for Convolutional Neural Networks”. In: *ACM Computing Surveys (CSUR)*
597 (2022).
- 598 [83] David A Van Dyk and Xiao-Li Meng. “The art of data augmentation”. In: *Journal of Computational*
599 *and Graphical Statistics* 10.1 (2001), pp. 1–50.
- 600 [84] Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell. “Understanding data
601 augmentation for classification: when to warp?” In: *2016 international conference on digital image*
602 *computing: techniques and applications (DICTA)*. IEEE. 2016, pp. 1–6.
- 603 [85] Sima Behpour, Kris M Kitani, and Brian D Ziebart. “Ada: Adversarial data augmentation for object
604 detection”. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE.
605 2019, pp. 1243–1252.
- 606 [86] Hadi Keivan Ekbatani, Oriol Pujol, and Santi Segui. “Synthetic Data Generation for Deep Learning
607 in Counting Pedestrians.” In: *ICPRAM*. 2017, pp. 318–323.
- 608 [87] Agnieszka Mikołajczyk and Michał Grochowski. “Data augmentation for improving deep learning
609 in image classification problem”. In: *2018 international interdisciplinary PhD workshop (IIPhDW)*.
610 IEEE. 2018, pp. 117–122.
- 611 [88] Damien Dablain, Colin Bellinger, Bartosz Krawczyk, and Nitesh Chawla. “Efficient Augmentation
612 for Imbalanced Deep Learning”. In: *arXiv e-prints* (2022), arXiv–2207.
- 613 [89] Jing Zhou, Yanan Zheng, Jie Tang, Jian Li, and Zhilin Yang. “Flipda: Effective and robust data
614 augmentation for few-shot learning”. In: *arXiv preprint arXiv:2108.06332* (2021).
- 615 [90] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. “Generalizing from a few examples:
616 A survey on few-shot learning”. In: *ACM computing surveys (csur)* 53.3 (2020), pp. 1–34.
- 617 [91] Hari Prasanna Das, Ryan Tran, Japjot Singh, Xiangyu Yue, Geoffrey Tison, Alberto Sangiovanni-
618 Vincentelli, and Costas J Spanos. “Conditional synthetic data generation for robust machine learning
619 applications with limited pandemic data”. In: *Proceedings of the AAAI Conference on Artificial*
620 *Intelligence*. Vol. 36. 11. 2022, pp. 11792–11800.
- 621 [92] Katherina K Hauner, Richard E Zinbarg, and William Revelle. “A latent variable model approach
622 to estimating systematic bias in the oversampling method”. In: *Behavior Research Methods* 46.3
623 (2014), pp. 786–797.