# Synthetic data generation: A literature review

Joao Fonseca[1*], Fernando Bacao[1]

[1]NOVA 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 problems using tabular data. 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 the generation mechanism is still dependant on (1) the importance of the relationships found between the different features, (2) the ML task to be 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 7.

## 1.1 Motivation and Contributions

This literature review focuses on the generation mechanisms and generative models underlying the different techniques where synthetic data is generated. Specifically, we focus on techniques used in studies published since 2019. 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 reviews [1] is shown in Table 1.

[1]Results obtained using Google Scholar, limited to articles published since 2019, using the search

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. |

The different taxonomies established in the literature follow a similar philosophy, but vary in terminology and are often specific to the technique discussed. Regardless, it is possible to establish a broader taxonomy without giving up on specificity. This study provides a joint overview of the different data generation approaches, domains and ML techniques where data generation is being used, as well as a common taxonomy across domains. It extends the analyses found in these articles and uses the compiled knowledge to identify research gaps. We compare the strengths and weaknesses of the models developed within each of these fields. Finally, we identify possible future research directions to address some of the limitations found. The contributions of this paper are summarized below:

query `("synthetic data generation" OR "oversampling" OR "imbalanced learning" OR "data augmentation") AND ("literature review" OR "survey")`. Retrieved on August $11^{th}$, 2022. More articles were added later whenever found relevant.

- Bridge different ML concepts using synthetic data generation in its core (Algorithmic applications + Review of the State-of-the-art).

- Propose a synthetic data generation/data augmentation taxonomy to resolve the ambiguity in the literature (Data augmentation taxonomy).

- Characterize all relevant data generation methods using the proposed taxonomy.

- Discuss the ML techniques in which synthetic data generation/data augmentation is used, beyond regularization and consolidate the current data generation mechanisms across the different techniques (Algorithmic Applications).

- Bring to light the key challenges of synthetic data generation and put forward possible research directions in the future.

### 1.2 Paper Organization

This paper is organized as follows: Section 2 defines and formalizes the different concepts, goals, trade-offs and motivations related to synthetic data generation. Section 3 establishes the taxonomy used to categorize all the methods described in the paper. Section 4 reviews synthetic data generation mechanisms in the feature space. Section 5 reviews synthetic data generation mechanisms in the input space. Section 6 describes the applications of synthetic data in ML methods. Section 7 reviews performance evaluation methods of synthetic data generation mechanisms. Section 8 summarizes the main findings and discusses limitations and possible research directions in the state-of-the-art. Section 9 presents the main conclusions drawn from this study.

## 2 Background

In this section we define basics concepts, common goals, trade-offs and motivations regarding the generation of synthetic data in ML. We define synthetic data generation as the production of observations using a generative model (regardless of its nature) that resemble naturally occurring observations within a certain domain. It requires access to either a training dataset, a generative process, or a data stream. However, additional requirements might be imposed depending on the ML task being developed. For example, to generate artificial data for regularization purposes in supervised learning (*i.e.*, data augmentation) the training dataset must be annotated **[CITATION]**. The generation of synthetic data for anonymization purposes assumes synthetic datasets to be different from the original data, while following the same statistical properties **[CITATION]**. Domain knowledge may also be necessary to encode specific 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 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 6.1.

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 6.2.

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 6.3.

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 6.4 and 6.5.

The two other techniques reliant on synthetic data generation is both 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 6.6 and 6.7.

## 2.2 Problem Formulation

Original dataset, $\mathcal{D}$, which can be 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, $\mathcal{D}$, $\mathcal{D}^L$ and $\mathcal{D}^U$ consist of ordered collections with lengths $l+u$, $l$ and $u$, respectively. Synthetic data generation is performed using a generator, $f_{gen}(x; \tau) = \tilde{x}$, where $\tau$ defines the generation policy (*i.e.*, type of transformation and its hyperparameters), $x \in \mathcal{D}$ is an observation and $\tilde{x} \in \mathcal{D}^S$ is a synthetic observation. Analogous to $\mathcal{D}$, the synthetic dataset, $\mathcal{D}^S$, is also distinguished according to whether there is an assignment of a target feature, $\mathcal{D}_L^S = ((\tilde{x}_i, \tilde{y}_i))_{i=1}^{l'}$, or not, $\mathcal{D}_U^S = (\tilde{x}_i)_{i=1}^{u'}$.

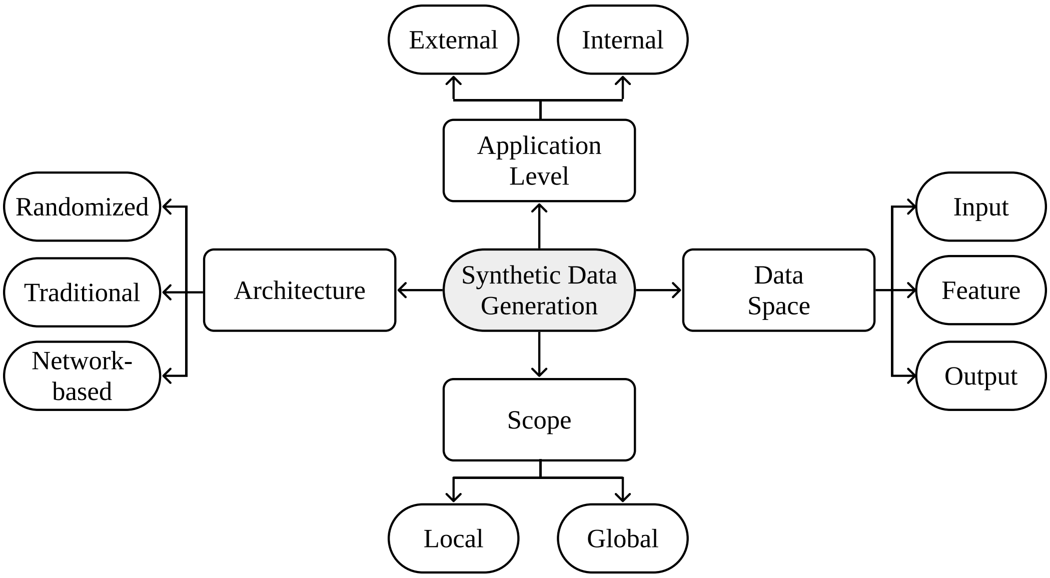


Figure 1: General taxonomy of data generation mechanisms proposed in this paper.

Depending on the ML task, it may be relevant to establish metrics to measure the quality of $\mathcal{D}^S$. In this case, a metric $f_{qual}(\mathcal{D}^s, \mathcal{D})$ is used to determine the level of similarity/dissimilarity

# 3 Data Generation Taxonomy

Image data augmentation taxonomy [28]

There is a distinction between semantic and traditional image data augmentation [38], also discussed in [23]

Synthetic data generation for medical records taxonomy [14] which is incomplete

Data generation mechanisms can be characterized in 4 properties: Architecture, Application level, Scope and Data space. The overall definition of the proposed taxonomy is shown in Figure 1.

1. Level of application (External or Internal)
2. Scope (Local or Global augmentation)
3. Architectural approach (heuristic, network-based or others)
4. Data space (Input, feature or output). Within feature and output: Domain

# 4 Data Generation in the Feature Space

The concept of data generation in the feature space was popularized with [9].

According to [1]. The generation of synthetic data should aim to fulfil the conditions below:

- Privacy preserving.
- Human readable.
- Compact.

Discuss Auto-augmentation (as mentioned in [22]) or meta learning (as mentioned in [23])

# 5 Data Generation in the Input Space

In this section, we describe some popular domain and data type-specific data generation techniques. For each data type we include a table with related literature reviews specific to different domains.

## 5.1 Tabular

## 5.2 Time series

Generative adversarial networks in time series

## 5.3 Image

Image-specific data generation mechanisms can be further divided into traditional and semantic techniques [38]. Traditional generation techniques comprise simple modifications such as translation, cropping or random erasing [39]. Semantic generation methods involve more complex tasks, such changing colors of specific attributes, backgrounds and visual angles **[CITATION]**.

Data generation by modifying specific attributes in data points with known perturbations [6]. For example, overlaying facial elements into a picture containing a human face (*e.g.*, adding sunglasses and different hairstyles), introducing perturbations in facial landmarks, different illumination and artificial misalignment are different approaches to generate artificial observations for facial recognition.

Generative Adversarial Networks in computer vision [40]

### 5.4 Text

NLP motivations [7]:

1. Low-resource languages (NLP)
2. Mitigate bias
3. Fixing class imbalance
4. Few-shot learning
5. Adversarial examples

NLP also benefit from data augmentation [7].

In NLP, there is the challenge of establishing universal rules for text transformations to provide new linguistic patterns [41]

https://github.com/styfeng/DataAug4NLP

### 5.5 Graphs

Another relevant paper [42]

Various graph data augmentation methods can be applied to related data types such as text data [18].

An analysis on different graph data augmentation techniques and a new graph data augmentation framework Zhao et al. [43]

List of papers about graph data augmentation: https://github.com/zhao-tong/graph-data-augmentation-papers

### 5.6 Audio

## 6 Algorithmic applications

### 6.1 Data Privacy

SynSys [44], Sensegen [45], The Synthetic Data Vault [30]

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 [46, 47]. 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 [48]. 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 [49]. 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 [50, 51].

The generation of synthetic datasets allow a more flexible approach to the successful implementation of ML tasks. However,

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

Guidelines for effective synthetic data generation [29]

## 6.2 Regularization in Supervised Learning

The performance of Machine Learning models is highly dependent on the quality of the training dataset used [52, 53]. The presence of imbalanced and/or small datasets, target labels incorrectly assigned, outliers and high dimensional input spaces reduce the prospects of a successful machine learning (ML) model implementation [53, 54, 55]. In the case of deep learning, for example, these models are often limited by a natural inclination to overfitting, label noise memorization and catastrophic forgetting [56]. Regularization methods are the typical approach to address these problems, but producing robust ML solutions is still a challenge [31].

It is frequently assumed that the training data is sampled from a fixed data source, it is balanced and does not contain label noise. Under these conditions, the resulting ML classifier is expected to achieve good generalization performance [57]. Although, in practical applications, this is rarely the case. When the training data is not representative of the true population, or the model is over-parametrized, it becomes particularly prone to overfitting [58]. Regularization methods attempt to address these limitations. They can be divided into three categories [59]:

1. Output level modifications. Transforms the labels in the training data.
2. Algorithmic level modifications. Modifies the classifier's architecture, loss function or other components in the training procedure.
3. Input level modifications. Modifies the training dataset by expanding it with synthetic data.

The last approach, input level modifications, is known as data augmentation. Data augmentation is used to increase the size and data variability of data in a training dataset, by producing synthetic observations [60, 61]. Since it is applied at the data level, it can be used for various types of problems and classifiers [62].

Problems such as fraud detection and healthcare are frequently tackled via synthetic data generation [63].

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

"Moreover, the current research about so called adversarial attacks on CNNs showed that deep neural networks can be easily fooled into misclassification of images just by partial rotations and image translation [1], adding the noise to images [5] and even changing one, skillfully selected pixel in the image [6]." Mikołajczyk et al. [64]

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

### 6.3 Oversampling

KernelADASYN [65]

The original author of SMOTE recently published the paper "Efficient Augmentation for Imbalanced Deep Learning" [66]

### 6.4 Active Learning

### 6.5 Few-shot Learning

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

FlipDA [67]

### 6.6 Semi-supervised Learning

Synthetic data generation for semi-supervised learning given limited labeled data regarding the COVID-19 pandemic [68].

Extensive literature review on semi-supervised learning [37]

### 6.7 Self-supervised Learning

## 7 Evaluating the Quality of Synthetic Data

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

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

1. Fidelity
2. Diversity
3. Generalization

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

### 7.1 Statistical Divergence Metrics

### 7.2 Precision/Recall Metrics

## 8 Discussion

### 8.1 Main Findings

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

### 8.1.1 RQ1: bla bla bla

### 8.1.2 RQ2: bla bla bla

### 8.1.3 RQ3: bla bla bla

## 8.2 Limitations

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

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

The evaluation of anonymization techniques lack standardized, objective and reliable performance metrics and benchmark datasets to allow an easier comparison across classifiers to evaluate key aspects of data anonymization (resemblance, utility, privacy and performance). These datasets should contain mixed data types (*i.e.*, a combination of categorical, ordinal, continuous and discrete features) and the metrics should evaluate the performance of different data mining tasks along with the anonymization reliability. This problem appears to be universal across domains. For example, Hernandez et al. [14] observed the lack of a universal method or metric to report the performance synthetic data generation algorithms for tabular health records.

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

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

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

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

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

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

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

oversampling does not seem to be a relevant source of bias in behavioral research and does not appear to have an appreciably different effect on results for directly versus indirectly oversampled variables [69]

## 8.3 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

# 9 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. 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 Wlliam 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] Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu. "Regularizing deep networks with semantic data augmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[39] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. "Random erasing data augmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 13001–13008.

[40] Zhengwei Wang, Qi She, and Tomas E Ward. "Generative adversarial networks in computer vision: A survey and taxonomy". In: *ACM Computing Surveys (CSUR)* 54.2 (2021), pp. 1–38.

[41] Markus Bayer, Marc-André Kaufhold, Björn Buchhold, Marcel Keller, Jörg Dallmeyer, and Christian Reuter. "Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers". In: *International Journal of Machine Learning and Cybernetics* (2022), pp. 1–16.

[42] Jiajun Zhou, Jie Shen, and Qi Xuan. "Data augmentation for graph classification". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 2341–2344.

[43] Tong Zhao, Yozen Liu, Leonardo Neves, Oliver Woodford, Meng Jiang, and Neil Shah. "Data augmentation for graph neural networks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 12. 2021, pp. 11015–11023.

[44] Jessamyn Dahmen and Diane Cook. "SynSys: A synthetic data generation system for healthcare applications". In: *Sensors* 19.5 (2019), p. 1181.

[45] 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.

[46] 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.

[47] 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).

[48] Jerome P Reiter. "New approaches to data dissemination: A glimpse into the future (?)" In: *Chance* 17.3 (2004), pp. 11–15.

[49] 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.

[50] Kalpana Singh and Lynn Batten. "Aggregating privatized medical data for secure querying applications". In: *Future Generation Computer Systems* 72 (2017), pp. 250–263.

[51] 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.

[52] 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.

[53] Alon Halevy, Peter Norvig, and Fernando Pereira. "The unreasonable effectiveness of data". In: *IEEE Intelligent Systems* 24.2 (2009), pp. 8–12.

[54] Pedro Domingos. "A few useful things to know about machine learning". In: *Communications of the ACM* 55.10 (2012), pp. 78–87.

[55] Shaeke Salman and Xiuwen Liu. "Overfitting mechanism and avoidance in deep neural networks". In: *arXiv preprint arXiv:1901.06566* (2019).

[56] 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.

[57] Martin Benning and Martin Burger. "Modern regularization methods for inverse problems". In: *Acta Numerica* 27 (2018), pp. 1–111.

[58] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. "Deep learning: a statistical viewpoint". In: *Acta numerica* 30 (2021), pp. 87–201.

[59] Claudio Filipi Gonçalves dos Santos and João Paulo Papa. "Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks". In: *ACM Computing Surveys (CSUR)* (2022).

[60] David A Van Dyk and Xiao-Li Meng. "The art of data augmentation". In: *Journal of Computational and Graphical Statistics* 10.1 (2001), pp. 1–50.

[61] Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell. "Understanding data augmentation for classification: when to warp?" In: *2016 international conference on digital image computing: techniques and applications (DICTA)*. IEEE. 2016, pp. 1–6.

[62] Sima Behpour, Kris M Kitani, and Brian D Ziebart. "Ada: Adversarial data augmentation for object detection". In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019, pp. 1243–1252.

[63] Hadi Keivan Ekbatani, Oriol Pujol, and Santi Segui. "Synthetic Data Generation for Deep Learning in Counting Pedestrians." In: *ICPRAM*. 2017, pp. 318–323.

[64] Agnieszka Mikołajczyk and Michał Grochowski. "Data augmentation for improving deep learning in image classification problem". In: *2018 international interdisciplinary PhD workshop (IIPhDW)*. IEEE. 2018, pp. 117–122.

[65] 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.

[66] Damien Dablain, Colin Bellinger, Bartosz Krawczyk, and Nitesh Chawla. "Efficient Augmentation for Imbalanced Deep Learning". In: *arXiv e-prints* (2022), arXiv–2207.

[67] Jing Zhou, Yanan Zheng, Jie Tang, Jian Li, and Zhilin Yang. "Flipda: Effective and robust data augmentation for few-shot learning". In: *arXiv preprint arXiv:2108.06332* (2021).

[68] Hari Prasanna Das, Ryan Tran, Japjot Singh, Xiangyu Yue, Geoffrey Tison, Alberto Sangiovanni-Vincentelli, and Costas J Spanos. "Conditional synthetic data generation for robust machine learning applications with limited pandemic data". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 11. 2022, pp. 11792–11800.

[69] Katherina K Hauner, Richard E Zinbarg, and William Revelle. "A latent variable model approach to estimating systematic bias in the oversampling method". In: *Behavior Research Methods* 46.3 (2014), pp. 786–797.