Machine Learning for Synthetic Data Generation: A Review

Yingzhou Lu*, Minjie Shen†, Huazheng Wang‡, Xiao Wang§, Capucine van Rechem*, Wenqi Wei¶‖,

***Abstract*—Machine learning heavily relies on data, but real-world applications often encounter various data-related issues. These include data of poor quality, insufficient data points leading to under-fitting of machine learning models, and difficulties in data access due to concerns surrounding privacy, safety, and regulations. In light of these challenges, the concept of *synthetic data generation* emerges as a promising alternative that allows for data sharing and utilization in ways that real-world data cannot facilitate. This paper presents a comprehensive systematic review of existing studies that employ machine learning models for the purpose of generating synthetic data. The review encompasses various perspectives, starting with the applications of synthetic data generation, spanning computer vision, speech, natural language processing, healthcare, and business domains. Additionally, it explores different machine learning methods, with particular emphasis on neural network architectures and deep generative models. The paper also addresses the crucial aspects of privacy and fairness concerns related to synthetic data generation. Furthermore, this study identifies the challenges and opportunities prevalent in this emerging field, shedding light on the potential avenues for future research. By delving into the intricacies of synthetic data generation, this paper aims to contribute to the advancement of knowledge and inspire further exploration in synthetic data generation.**

***Index Terms*—data synthesis, machine learning, data trustworthiness**

I. Introduction

MACHINE learning endows intelligent computer systems with the capacity to autonomously tackle tasks, pushing the envelope of industrial innovation [1]. By integrating high-performance computing, contemporary modeling, and simulations, machine learning has evolved into an indispensable instrument for managing and analyzing massive volumes of data [2], [3].

Nonetheless, it is important to recognize that machine learning does not invariably resolve problems or yield the optimal solution. Despite artificial intelligence is currently experiencing a golden age, numerous challenges persist in the development and application of machine learning technology [4]. As the field continues to advance, addressing these obstacles will be essential for unlocking the full potential of machine learning and its transformative impact on various industries. Machine learning endows intelligent computer systems with the capacity to autonomously tackle tasks, pushing the envelope of industrial innovation [1]. By integrating high-performance computing, contemporary modeling, and simulations, machine learning has evolved into an indispensable instrument for managing and analyzing massive volumes of data [2], [3].

The process of collecting and annotating data is both time-consuming and expensive [5], giving rise to numerous issues. As machine learning is heavily dependent on data, some of the key hurdles and challenges it faces include:

- **Data quality**. Ensuring data quality is one of the most significant challenges confronting machine learning professionals. When data is of subpar quality, models may generate incorrect or imprecise predictions due to confusion and misinterpretation [6] [7].
- **Data scarcity**. A considerable portion of the contemporary AI dilemma stems from inadequate data availability: either the number of accessible datasets is insufficient, or manual labeling is excessively costly [8].
- **Data privacy and fairness**. There are many areas in which datasets cannot be publicly released due to privacy ad fair issues. In these cases, generating synthetic data can be very useful, and we will investigate ways of creating anonymized datasets with differential privacy protections.

Addressing these challenges will be essential for unlocking the full potential of machine learning and its transformative impact on various industries [9]–[11]. Generally, synthetic data are defined as the artificially annotated information generated by computer algorithms or simulations [4], [12]. Synthetic data is generally defined as artificially annotated information generated by computer algorithms or simulations [4], [12]. In many cases, synthetic data is necessary when real data is either unavailable or must be kept private due to privacy or compliance risks [10], [13], [14]. This technology is extensively utilized in various sectors, such as healthcare, business, manufacturing, and agriculture, with demand growing at an exponential rate [15].

The objective of this paper is to offer a high-level overview of several state-of-the-art approaches currently being investigated by machine learning researchers for synthetic data generation. For the reader's convenience, we summarize the paper's main contributions as follows:

- We present pertinent ideas and background information on synthetic data, serving as a guide for researchers

*Department of Pathology, Stanford University, Stanford, CA, 94305.
† The Bradley Department of Electrical and Computer Engineering, Virginia Tech
‡School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, 97331.
§ School of Computer Science & Engineering University of Washington, Seattle, WA, 98105.
‖ Computer and Information Science Department, Fordham University, New York City, NY, 10023.
¶Corresponding author.
Contacting E-mail: lyz66@stanford.edu, wenqiwei@fordham.edu.
Manuscript received xxxx xx, xxxx; revised xxxxx xx, xxxx.

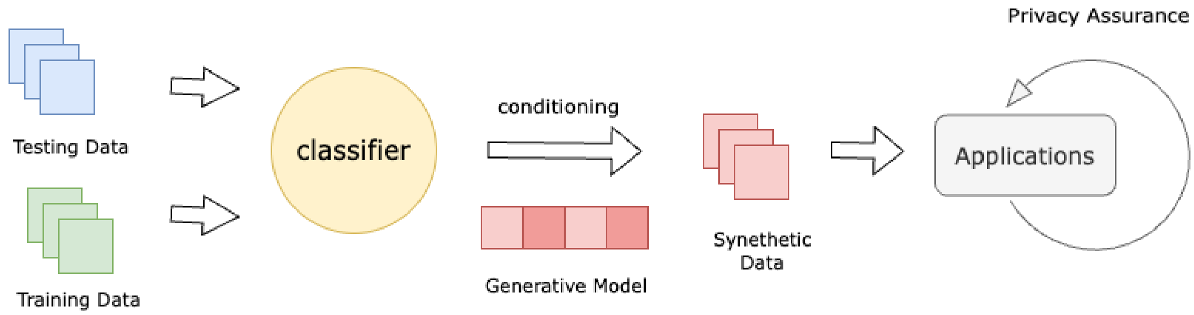

Fig. 1. Synthetic data generation.

interested in this domain.

- We explore different real-world application domains and emphasize the range of opportunities that GANs and synthetic data generation can provide in bridging gaps (Section II).
- We examine a diverse array of deep neural network architectures and deep generative models dedicated to generating high-quality synthetic data, present advanced generative models, and outline potential avenues for future research (Section III and IV).
- We address privacy and fairness concerns, as sensitive information can be inferred from synthesized data, and biases embedded in real-world data can be inherited. We review current technological advancements and their limitations in safeguarding data privacy and ensuring the fairness of synthesized data (Section V and VI).
- We outline several general evaluation strategies to assess the quality of synthetic data (Section VIII).
- We identify challenges faced in generating synthetic data and during the deployment process, highlighting potential future work that could further enhance functionality (Section IX).

II. Application

Synthetic data offers a multitude of compelling advantages, making it a highly appealing option for a wide range of applications. By streamlining the processes of training, testing, and deploying AI solutions, synthetic data facilitates more efficient and effective development. Furthermore, this cutting-edge technology reduces the risk of exposing sensitive information, thereby ensuring customer security and privacy [4].

As researchers transition synthetic data from the lab to practical implementations, its real-world applications continue to broaden. This section explores several notable domains where synthetic data generation substantially impacts addressing real-world challenges.

A. Vision

Supervised learning relies heavily on the availability of labeled data [51]. However, in many applications, particularly in computer vision, manual labeling is often necessary [52], [53]. Tasks such as segmentation, depth estimation, and optical flow estimation can be exceedingly challenging to label manually. Synthetic data has emerged as a transformative solution in this context, significantly improving the labeling process [54].

Sankaranarayanan et al. introduced a generative adversarial network (GAN) that narrows the gap between embeddings in the learned feature space, facilitating Visual Domain Adaptation [55]. This approach enables semantic segmentation across different domains. The GAN uses a generator to project features onto the image space, which the discriminator subsequently operates on. Adversarial losses can be derived from the discriminator's output [56]. Notably, applying adversarial losses to the projected image space has been shown to yield significantly better performance compared to applying them directly to the feature space [55].

In a recent study, a Microsoft research team demonstrated the effectiveness of synthetic data in face-related tasks by combining a parametric 3D face model with an extensive library of hand-crafted assets [57]. This approach rendered training images with remarkable realism and diversity. The researchers trained machine learning systems for tasks such as landmark localization and face parsing using synthetic data, showing that it can achieve comparable accuracy to real data. Furthermore, synthetic data alone proved sufficient for detecting faces in unconstrained settings [57].

B. Voice

The field of synthetic voice is at the forefront of technological advancement, and its evolution is happening at a breakneck pace. With the advent of machine learning and deep learning, creating synthetic voices for various applications such as video production, digital assistants, and video games [58] has become easier and more accurate. This field is an intersection of diverse disciplines, including acoustics, linguistics, and signal processing. Researchers in this area continuously strive to improve synthetic voices' accuracy and naturalness. As technology advances, we can expect to see synthetic voices become even more prevalent in our daily lives, assisting us in various ways and enriching our experiences in many fields [59].

The earlier study includes spectral modeling for statistical parametric speech synthesis, in which low-level, untransformed spectral envelope parameters are used for voice synthesis. The low-level spectral envelopes are represented by graphical models incorporating multiple hidden variables,

TABLE I
SUMMARIZATION OF REPRESENTATIVE WORKS IN SYNTHETIC DATA GENERATION.

Paper	application	generative AI	DNN	dataset
MedGAN [16]	healthcare	GAN	MLP	MIMIC/Sutter (Electronic health record)
MMCGAN [17]	healthcare & CV	GAN	CNN	chest CT images
DeepSynth [18]	healthcare & CV	GAN	CNN	rat kidney tissue (microscope image)
ChemSpaceE [19]	drug	VAE	GNN	ZINC (drug molecule) [20]
JTVAE [21]	drug	VAE	GNN	ZINC (drug molecules) [20]
REINVENT [22]	drug	RL	RNN	ZINC (drug molecules) [20]
CORE [23]	drug	VAE	GNN	ZINC (drug molecule) [20]
RGA [24]	drug	RL	geometric NN	ZINC and TDC [25]
CorGAN [26]	healthcare	GAN	CNN	MIMIC-III dataset, UCI Epileptic Seizure Recognition dataset
DAAE [27]	healthcare	VAE+GAN	recurrent autoencoder	MIMIC-III, UT Physicians clinical databases
HAPNEST [28]	healthcare	approximate Bayesian computation (not deep learning)	NA (w.o. DNN)	Genomes Project and HGDP datasets
synthpop [29]	healthcare	proper synthesis	Statistical hypothesis testing	SD2011
CycleGAN [30]	vision	GAN	CNN	pix2pix
DP-CGAN [31]	vision	GAN	deep CGAN	MNIST
BigGANs [32]	vision	GAN	large scale GAN	ImageNet
VideoDiff [33]	vision	diffusion	CNN	BAIR Robot Pushing, Kinetics-600
VQ-VAE [34]	vision	VAE	PixelCNN	ImageNet
GIRAFFE [35]	vision	GAN	CNN	CompCars, LSUN Churches, and FFHQ
Wavegrad [36]	TTS	diffusion	gradient-based sampling	LJ Speech
TTS-GAN [37]	TTS	GAN	auto-regressive model	Tacotron2
Seq-GAN [38]	NLP	GAN+RL	CNN	Nottingham dataset
BLEURT [39]	NLP	Language model	BERT	WebNLG Competition dataset
TextGen-RL [40]	NLP	RL	LSTM	
SynBench [41]	NLP	conditional Gaussian mixture		CIFAR10
RelGAN [42]	image and text	GAN	CNN	COCO Image Captions dataset
DPGM [43]	audio and text	generative artificial neural networks	differentially private kernel k-means	MNIST, anonymized Call Detail Record (CDR)
WaveGAN [44]	audio	GAN	DCGAN	Speech Commands Dataset
Wavenet [45]	audio	GAN	LSTM	CSTR voice corpula (multi-channel English audio)
Stutter-TTS [46]	audio	phonetic encoder and the decoder	CNN	recordings
Quant GANs [47]	business	GAN	MLP+ Temporal convolutional networks(TCN)	simulated data
CGAN [48]	business	GAN	CNN	Vector autoregressive (VAR) time series
PATE-GAN [49]	business	GAN	PRIVATE AGGREGATION OF TEACHER ENSEMBLES (PATE)	Kaggle
CollGAN [50]	physics (particle collision)	VAE/GAN	MLP	ATLAS

such as restricted Boltzmann machines and deep belief networks (DBNs) [60]. The proposed conventional hidden Markov model (HMM)-based speech synthesis system can be significantly improved in terms of naturalness and over-smoothing [61].

Synthetic data can also be applied to Text-to-Speech (TTS) to achieve near-human naturalness [62], [63]. As an alternative to sparse or limited data, synthetic speech (SynthASR) was developed for automatic speech recognition. The combination of weighted multi-style training, data augmentation, encoder freezing, and parameter regularization is also employed to address catastrophic forgetting. Using this novel model, the researchers were able to apply state-of-the-art techniques to train a wide range of end-to-end (E2E) automatic speech recognition (ASR) models while reducing the need for production data and the costs associated with it [62].

C. Natural Language Processing (NLP)

The increasing interest in synthetic data has spurred the development of a wide array of deep generative models in the field of natural language processing (NLP) [51]. In recent years, a multitude of methods and models have illustrated the capabilities of machine learning in categorizing, routing, filtering, and searching for relevant information across various domains [64].

Despite these advancements, challenges remain. For example, the meaning of words and phrases can change depending on their context, and homonyms with distinct definitions can pose additional difficulties [65]. To tackle these challenges, the

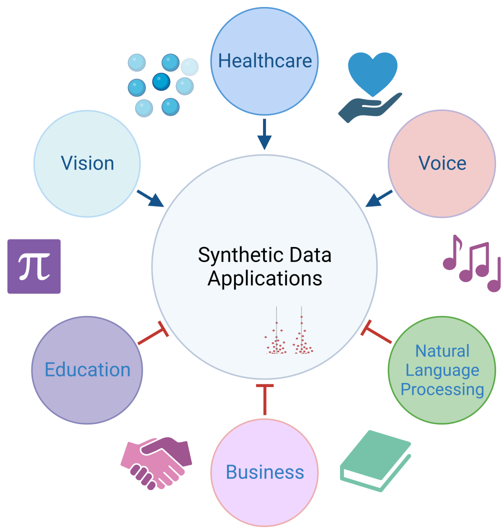

Fig. 2. Synthetic data applications

BLEURT model was proposed, which models human judgments using a limited number of potentially biased training examples based on BERT. The researchers employed millions of synthetic examples to develop an innovative pre-training scheme, bolstering the model's ability to generalize [66]. Experimental results indicate that BLEURT surpasses its counterparts on both the WebNLG Competition dataset and the WMT Metrics, highlighting its efficacy in NLP tasks [39].

Another significant breakthrough in text generation using GANs is RelGAN, developed by Rice University. This model is comprised of three main components: a relational memory-based generator, a Gumbel-Softmax relaxation algorithm, and multiple embedded representations within the discriminator. When benchmarked against several cutting-edge models, RelGAN demonstrates superior performance in terms of sampling quality and diversity. This showcases its potential for further investigation and application in a wide range of NLP tasks and challenges [42], [67].

D. Healthcare

In order to protect health information and improve reproducibility in research, synthetic data has drawn mainstream attention in the healthcare industry [68], [69]. Many labs and companies have harnessed the tools of big data and advanced computation tools to produce large quantities of synthetic data [70]. Modeled after patient data, synthetic data generation is essential to understanding diseases while maintaining patient confidentiality and privacy simultaneously [71]. Theoretically, synthetic data can reflect the original distribution of the data instead of revealing actual patient data [71]–[73].

Synthetic data generation can also be utilized to discover new scientific principles by grounding it in biological priors [68]. There have been a good number of models and software developed, such as SynSys, which uses hidden Markov models and regression models initially trained on real datasets to generate synthetic time series data consisting of nested sequences [69]; and corGAN, in which synthetic data is generated by capturing correlations between adjacent medical features in the data representation space [26].

Synthetic data generation has also been widely used in drug discovery, especially *de novo* drug molecular design. Drugs are essentially molecular structures with desirable pharmaceutical properties. The goal of *de novo* drug design is to produce novel and desirable molecule structures from scratch. The word "*de novo*" means from the beginning. The whole molecule space is around 10^{60} [25], [74], [75]. Most of the existing methods rely heavily on brute-force enumeration and are computationally prohibitive. Generative models are able to learn the distribution of drug molecules from the existing drug database and then draw novel samples (i.e., drug molecules) from the learned molecule distribution, including variational autoencoder (VAE) [21], [76], [77], generative adversarial network (GAN) [78], energy-based model (EBM) [79], [80], diffusion model [81], reinforcement learning (RL) [22], [24], [82], genetic algorithm [83], sampling-based methods [84], [85], etc.

In healthcare, patient information is often stored in electronic health records (EHR) format. [86]–[88]. Research in medicine has been greatly facilitated by the availability of information from electronic health records [89], [90]. MedGAN, an adversarial network model for generating realistic synthetic patient records, has been proposed by Edward Choi and other colleagues. With the help of an autoencoder and generative adversarial networks, medGAN can generate high-dimensional discrete variables (e.g., binary and count features) based on real patient records [16]. Based on their evaluations of medGAN's performance on a set of diverse tasks reported, including reporting distribution statistics, classification performance [91], and expert review, medGAN exhibits close-to-real-time performance [16], [92]–[95]. Using synthetic data can help reduce the regulatory barriers preventing the widespread sharing and integration of patient data across multiple organizations in the past [96], [97]. Researchers across the globe would be able to request access to synthetic data from an institution to conduct their own research using the data. Such capabilities can increase both the efficiency and scope of the study as well as reduce the likelihood of biases being introduced into the results [69], [98], [99].

E. Business

The inherent risk of compromising or exposing original data persists as long as it remains in use, particularly in the business sector, where data sharing is heavily constrained both within and outside the organization [100]. Consequently, it is crucial to explore methods for generating financial datasets that emulate the properties of "real data" while maintaining the privacy of the involved parties [100].

Efforts have been made to secure original data using technologies like encryption, anonymization, and cutting-edge privacy preservation [101]. However, information gleaned from the data may still be employed to trace individuals, thereby posing the risk [102]. A notable advantage of synthetic data lies in its ability to eliminate the exposure of critical data, thus ensuring privacy and security for both companies and their customers [103].

Moreover, synthetic data enables organizations to access data more rapidly, as it bypasses privacy and security protocols [104]. In the past, institutions possessing extensive data repositories could potentially assist decision-makers in resolving a broad spectrum of issues. However, accessing such data, even for internal purposes, was hindered by confidentiality concerns. Presently, companies are harnessing synthetic data to refresh and model original data, generating continuous insights that contribute to enhancing the organization's performance [4].

F. Education

Synthetic data is gaining increasing attention in the field of education due to its vast potential for research and teaching. Synthetic data refers to computer-generated information that mimics the properties of real-world data without disclosing any personally identifiable information [105]. This approach proves instrumental for educational settings, where ethical constraints often limit the use of real-world student data. Therefore, synthetic data offers a robust solution for privacy-concerned data sharing and analysis, enabling the creation of accurate models and strategies to improve the teaching-learning process.

A detailed example of synthetic data usage in education is the simulation of student performance data to aid in designing teaching strategies. Suppose an educational researcher wants to investigate the impact of teaching styles on student performance across different backgrounds and learning abilities. However, obtaining real student data for such studies can be ethically complex and potentially intrusive. In such a situation, synthetic data can be generated that mirrors the demographic distributions, learning patterns, and likely performance of a typical student population. This data can then be used to model the effects of various teaching strategies without compromising student privacy [106].

Furthermore, synthetic data can be a powerful tool in teacher training programs. For example, teacher candidates can use synthetic student data to practice data-driven instructional strategies, including differentiated instruction and personalized learning plans. They can analyze this synthetic data, identify patterns, determining student needs, and adjusting their instructional plans accordingly. By using synthetic data, teacher candidates gain practical experience in analyzing student data and adapting their teaching without infringing on the privacy of actual students [107]. Thus, synthetic data serves as a valuable bridge between theory and practice in education, driving innovation while safeguarding privacy.

G. Location and Trajectory Generation

Location and trajectory are a particular form of data that could highly reflect users' daily lives, habits, home addresses, workplaces, etc. To protect location privacy, synthetic location generation is introduced as opposed to location perturbation [108]. The main challenge of generating synthetic location and trajectory data is to resemble genuine user-produced data while offering practical privacy protection simultaneously. One approach to generating the location and trajectory data is to inject a synthetic point-based site within a user's trajectory [109]–[111]. Synthetic trajectory generation is frequently combined with privacy-enhancing techniques to further prevent sensitive inference from the synthesized data. For example, Chen et al. [112] introduces an N-gram-based method to predict the following position based on previous positions for publishing trajectory. They exploit the prefix tree to describe the n-gram model while combining it with differential privacy [113]. [114] proposes a synthetic trajectory strategy based on the discretization of raw trajectories using hierarchical reference systems to capture individual movements at differing speeds. Their method adaptively selects a small set of reference systems and constructs prefix tree counts with differential privacy. Applying direction-weighted sampling, the decrease in tree nodes reduces the amount of added noise and improves the utility of the synthetic data. By extracting multiple differential private distributions with redundant information, [115] the author generated a new trajectory with samples from these distributions. In addition to differential privacy, Bindschaedler and Shokri [116] enforce plausible deniability to generate privacy-preserving synthetic traces. It first introduces trace similarity and intersection functions that map a fake trace to a real hint under similarity and intersection constraints. Then, it generates one fake trace by clustering the locations and replacing the trajectory locations with those from the same group. If the fake trace satisfies plausible deniability, i.e., there exist k other real traces that can map to the fake trace, then it preserves the privacy of the seed trace. While existing studies mainly use the Markov chain model, [117] proposes to choose between the first-order and second-order Markov model adaptively. The proposed PrivTrace controls the space and time overhead by the first-order Markov chain model and achieves good accuracy for next-step prediction by the second-order Markov chain model.

H. AI-Generated Content (AIGC)

AI-Generated Content (AIGC) stands at the forefront of the technology and content creation industry, changing the dynamics of content production. A typical example of AIGC is OpenAI's ChatGPT, an AI-driven platform generating human-like text in response to prompts or questions. It leverages a vast corpus of internet text to generate detailed responses, often indistinguishable from those a human writer would produce. This capacity extends beyond simple question-answer pairs to crafting whole articles, stories, or technical explanations on a wide range of topics, thus creating a novel way of producing blog posts, articles, social media content, and more [118], [119].

Google's Project Bard focuses more on the creative aspects of text generation. It is designed to generate interactive fiction and assist in storytelling. Users can engage in an interactive dialogue with the model, directing the course of a narrative by providing prompts that the AI responds to, thus co-creating a story. This opens up fascinating possibilities for interactive entertainment and digital storytelling [120].

An innovative application of AIGC is in the field of news reporting. News agencies are increasingly using AI systems, such as the GPT series, to generate news content. For instance, the Associated Press uses AI to generate news articles about corporate earnings automatically. The AI takes structured data about company earnings and transforms it into a brief, coherent, and accurate news report. This automation allows the agency to cover a much larger number of companies than would otherwise be possible with human journalists alone [121].

Additionally, AIGC has found its place in the creative domain, with AI systems being used to generate book descriptions, plot outlines, and even full chapters of novels. For instance, a novelist could use ChatGPT to generate a synopsis for their upcoming book based on a few keywords or prompts related to the story. Similarly, marketing teams utilize AI to create compelling product descriptions for online marketplaces [122]. This not only increases efficiency but also provides a level of uniformity and scalability that would be challenging to achieve with human writers alone. Through these examples, it is clear that AIGC is profoundly impacting the landscape of content creation and will continue to shape it in the future [120].

III. Deep Neural Network

It is no secret that deep neural networks have become increasingly prominent in the field of computer vision and in other areas. Nevertheless, they require large amounts of annotated data for supervised training, which limits their effectiveness [123]. In this section, we review and compare various commonly-used deep neural network architectures as background knowledge, including multiple layer perception (MLP) in Section III-A, convolutional neural network (CNN) in Section III-B, recurrent neural network (RNN) in Section III-C, graph neural network (GNN) in Section III-D and transformer in Section III-E.

A. Multiple Layer Perception (MLP)

Multiple layer perception (MLP) is a classical (or vanilla) feedforward artificial neural network that uses a fully-connected connection in a single layer. It models all the possible interactions between the features. It is also the most used neural network. In the i-th layer, the propagation can be written as

$$h^{i+1} = \sigma(\mathbf{W}^T h^i + b), \tag{1}$$

where h^i is the input and h^{i+1} is the output (which is also the input of the $i+1$-th layer), $\mathbf{W}$ is weight matrix, b is the bias vector, both $\mathbf{W}$ and b are parameters of MLP. $\sigma()$ denotes the activation function. Popular activation functions incorporate the sigmoid, ReLU, tanh, etc. The goal of the activation function is to provide nonlinear transformation. MLP is the basis of many neural network architectures. Please refer to [124] for more details about MLP.

B. Convolutional Neural Network (CNN)

Convolutional neural network (CNN) was proposed to learn a better representation of images [124]. The core idea of the convolutional neural network is to design a two-dimensional convolutional layer that slides over the image to model the small-sized patch horizontally and vertically. Convolutional neural networks with one-dimensional convolutional layers can also model sequence data. Please refer to [124] for more details about CNN.

C. Recurrent Neural Network (RNN)

The sequence is one of the most popular data formats in real-world applications, e.g., natural language, speech signal, etc. Recurrent neural network (RNN) was designed to represent and generate sequence data [125]. Suppose we have a sequence of length T, $X = [x_1, x_2, ..., x_T]$, x_t denotes the t-th token, recurrent neural network designs a recursive (or recurrent) mechanism to reuse the neural network module. Specifically, at the t-th step, we have

$$o_t, h_t = \mathrm{RNN}(x_t, h_{t-1}), \tag{2}$$

where h_{t-1} denotes the hidden state at the $t-1$-th step and is used to represent all the historical memory from previous $t-1$ tokens $(x_1, x_2, ..., x_{t-1})$; o_t denotes the output of the RNN at the t-th step. The RNN module is recursively reused for each token. However, the vanilla RNN structure suffers from gradient vanishment issues, especially for long sequences. The state-of-the-art RNN architectures long short-term memory (LSTM) [126] and gated recurrent unit (GRU) [127] design a specialized gate to save the memory to alleviate the issue.

D. Graph Neural Network (GNN)

There are many graph-structured data in downstream applications, such as social networks, brain networks, chemical molecules, knowledge graphs, etc. Graph neural network was proposed to model graph-structured data and learn the topological structure from graph [128]–[130]. Specifically, graph neural networks build the interaction between the connected nodes and edges to model the topological structure of the graph [131]–[134]. The feedforward rule of graph neural network can be formulated as

$$\begin{aligned} & h_i^{(l+1)} = \mathrm{Aggregate}_{j\in\mathcal{N}(i)}(h_j^{(l)}, m_j, m_{ji}), \\ & l = 1, \cdots, L, \end{aligned} \tag{3}$$

where $h_i^{(l)}$ denotes the embedding of the i-th node at the l-th layer, m_j denotes the node feature of the node j, m_{ji} denotes the edge feature of the edge that connects nodes j and i, $\mathcal{N}(i)$ is the set of nodes that connect with i. Within each layer, we update the current representation via aggregating the information from (1) representation from the previous layer, (2) node feature (3) edge feature. For example, for chemical compounds, each node corresponds to an atom, the node feature is the category of the atom, e.g., Carbon atom, Nitrogen atom, Oxygen atom; each edge is a chemical bond, the edge feature is the type of the bond, including single bond, double bond, triple bond and aromatic bond.

E. Transformer

The Transformer architecture, introduced by Vaswani et al. in the groundbreaking paper "Attention is All You Need" in 2017, revolutionized the field of natural language processing and machine learning. Unlike traditional sequential models like RNNs (Recurrent Neural Networks) and LSTMs (Long Short-Term Memory), the Transformer model leverages a unique mechanism called self-attention, which allows it to capture long-range dependencies and relationships within the input data more effectively. This architecture consists of a stack of identical layers, each containing a multi-head self-attention mechanism followed by position-wise fully connected feed-forward networks. By eschewing recurrent or convolutional layers, the Transformer model is highly parallelizable and computationally efficient, leading to faster training times and improved performance on various NLP tasks.

The Generative Pre-trained Transformer (GPT) is a cutting-edge deep learning model that has revolutionized natural language processing (NLP) tasks [135]. Developed by OpenAI, GPT is an autoregressive transformer-based model that has displayed unparalleled performance in tasks such as text generation, translation, text summarization, and question-answering.

The model's architecture consists of multiple self-attention mechanisms and position-wise feedforward layers, enabling it to capture long-range dependencies and generate highly coherent and contextually relevant text. The key to GPT's success lies in its unsupervised pre-training on vast amounts of textual data, followed by fine-tuning on specific tasks. As the GPT model series progresses, with GPT-3 being the latest version at the time of this writing, the size and capabilities of the model continue to grow, paving the way for increasingly sophisticated NLP applications and opening up new possibilities for the generation of synthetic data.

By leveraging its pre-training on massive datasets and fine-tuning for specific tasks, GPT can produce artificial data samples that closely resemble real-world data. This capability is particularly valuable in scenarios where access to real data is limited due to privacy, regulatory, or resource constraints. GPT-generated synthetic data can be used to augment existing datasets, enabling researchers and practitioners to build more robust and accurate machine learning models while mitigating the risks associated with using sensitive or private data. Additionally, the synthetic data generated by GPT models can help address challenges related to data scarcity, class imbalance, or the need for domain-specific data, ultimately contributing to developing and deploying more effective AI solutions across various domains and applications.

IV. Generative AI

Generative AI models refer to a wide class of AI methods that could learn the data distribution from existing data objects and generate novel structured data objects, which fall into the category of unsupervised learning. Generative AI models, also known as deep generative models, or distribution learning methods, learn the data distribution and samples from the learned distribution to produce novel data objects. In this section, we investigate several generative AI models that are frequently used in synthetic data generation, including the language model in Section IV-A, variational autoencoder (VAE) in Section IV-C, generative adversarial network (GAN) in Section IV-D, reinforcement learning (RL) in Section IV-E, and diffusion model in Section IV-F. Table II compares various generative AI methods from several aspects.

A. Language Model

The language model was originally designed to model natural language. It is able to learn structured knowledge from massive unlabelled sequence data. Specifically, suppose the sequence has N tokens, denoted $X = [x_1, \cdots, x_N]$, then the probability distribution of the sequence can be decomposed as the product of a series of conditional probabilities,

$$p(X) = p\big([x_1, \cdots, x_N]\big) = \prod_{i=1}^{N} p(x_i | x_1, \cdots, x_{i-1}), \quad (4)$$

where a single conditional probability $p(x_i | x_1, \cdots, x_{i-1})$ denote the probability of the token x_i given all the tokens before x_i. The conditional probability can be modeled by the recurrent neural network (RNN). The language model can be used to generate all types of sequence data, such as natural language [136], electronic health records [142], etc. The language model can be combined with other deep learning models, such as variational autoencoder (VAE) and generative adversarial network (GAN), which will be described later.

B. Self-Supervised Learning (SSL)

Labeled data are expensive to acquire so the number of available labeled data is usually limited. To address this issue, self-supervised learning (SSL) was proposed, which is a learning paradigm that curates the supervision signal from the data itself. It is parallel to supervised learning and unsupervised learning. Different from supervised learning, self-supervised learning can learn from massive unlabeled data. Self-supervised learning is usually used as a pretraining strategy to learn the representation from massive unlabelled data [143]. The core idea of self-supervised learning is to mask a subset of the raw data feature and build a machine learning model to predict the masked data. then the pre-trained machine learning model (usually a neural network) is used as a "warm start", and is furtherly finetuned for the downstream applications.

C. Variational Autoencoder (VAE)

Variational autoencoder (VAE) [137] employs a continuous latent variable to characterize the data distribution. Specifically, it contains two neural network modules: encoder and decoder. The objective of the encoder is to convert the data object into a continuous latent variable. Then decoder takes

TABLE II
COMPARISON OF ALL THE GENERATIVE AI METHODS FROM DIFFERENT ASPECTS.

Method	supervision	NN architecture	MLE	with latent variable	paper
Language model (LM)	no	autoregressive model	yes	no	[136]
self-supervised learning (SSL)	no	encoder (representation)	yes	no	
variational autoencoder (VAE)	no	encoder-decoder	yes	yes	[137]
generative adversarial network (GAN)	no	generator & discriminator	yes	yes	[138], [139]
diffusion (score-based model) model	no	representation	yes	no	[140]
reinforcement learning (RL)	yes	policy network or Q-network	no	no	[141]

the latent variable as the input feature and reconstructs the data object.

Formally, suppose the data object is denoted x, the latent variable is a d-dimensional real-valued vector z, the encoder is $p(z|x)$, and the decoder is $q(x|z)$. The learning objective contains two parts: (1) reconstruct the data object x and (2) encourage the distribution of latent variables to be close to the normal distribution.

The Kullback-Leibler (KL) divergence measures the difference between two probability distributions. Given two probability distributions $p_1(x)$ and $p_2(x)$ on the same continuous domain, KL divergence between them is formally defined as

$$
\begin{aligned}
\mathrm{KL}(p_1||p_2) &= \int_x p_1(x) \log \frac{p_1(x)}{p_2(x)} dx \\
&= \int_x p_1(x) \big[\log p_1(x) - \log p_2(x) \big] dx. \\
\log p(x) &= \log \int_z p(z) p(x|z) dz \\
&\geq \mathbb{E}_{q(z|x)} \big[\log p(x|z) \big] - D_{KL}(q(z|x)||p(z)) \\
&\triangleq \mathrm{ELBO}.
\end{aligned}
$$

where $p(z)$ is the normal distribution and is used as the prior distribution. VAE encourages the distribution of latent variables to be close to normal distribution. Then during the inference phase, we sample latent variables from the normal distribution and generate the novel data objects. There are several VAE variants, such as disentangled VAE [144], hierarchical VAE [145], and sequence VAE [76].

D. Generative Adversarial Network (GAN)

Generative adversarial network (GAN) [138], [146], [147] formulates the generation problem into a supervised learning task. Specifically, it comprises two neural network modules: discriminator and generator. The objective of the generator is to generate data that are close to the real data, On the other hand, the objective of the discriminator is to discriminate the fake data (generated by the generator) from the real ones. It performs a binary classification task, where the real data from the training set are regarded as the positive samples; the generated data (by generator) are regarded as the negative samples. generator and discriminator are trained in a mini-max manner.

Formally, the generator is denoted $G(z)$, and the discriminator predicts a probabilistic score for a data object and is denoted $D(x)$. The learning objective is formulated as

$$
\begin{aligned}
\min_G \max_D \mathcal{L}(D, G) = \; & \mathbb{E}_{x\sim \text{training set}} [\log D(x)] \\
& + \mathbb{E}_{z\sim p(z)} \big[\log(1 - D(G(z)) \big],
\end{aligned} \tag{5}
$$

where z is the latent variable and is drawn from the normal distribution $p(z)$ to enhance the diversity of the generated data objects.

When learning GAN, the generator and discriminator are optimized alternatively.

- optimize generator and fix discriminator: the objective function becomes

$$\min \mathcal{L}(G) = \mathbb{E}_{z\sim p(z)} \big[\log(1 - D(G(z)) \big], \tag{6}$$

  where the generator is optimized to generate data that is close to the real data (with higher discriminator's scores).
- optimize discriminator and fix generator: the objective function reduces to a binary classification problem, that is,

$$
\begin{aligned}
\max \mathcal{L}(D) = \; & \mathbb{E}_{x\sim \text{training set}} [\log D(x)] \\
& + \mathbb{E}_{z\sim p(z)} \big[\log(1 - D(G(z)) \big],
\end{aligned} \tag{7}
$$

  which can be seen as a cross-entropy loss function, where the real data objects from the training set are seen as positive samples while the synthetic data objects $G(z)$ are seen as negative samples.

Then we discuss a popular variant of GAN. The Wasserstein Generative Adversarial Network (W-GAN) was proposed in 2017 and aims to enhance the stability of learning, accelerate the training process, and get rid of problems like mode collapse [148].

E. Reinforcement Learning (RL)

Reinforcement learning (RL) focuses on addressing sequential decision-making problems [149]. It can be used in synthesis data generation by growing a basic component at one time and generating data objects sequentially. It formulates sequential decision-making as a Markov decision process (MDP) [141]. Markov decision process assumes that given the current state, the future state of the stochastic process does not depend on the historical states. Suppose the state at the time t is x^t, Markov decision process satisfies

$$p(x^{t+1}|x^t, x^{t-1}, x^{t-2}, ...) = p(x^{t+1}|x^t). \tag{8}$$

At the time t, given the state x^t, the RL agent would generate an action a^t from action space, which is denoted $p_\theta(a^t|s^t)$, θ is the parameter of the RL agent. After performing the action, the system would jump into the next state x^{t+1}, i.e., $x^{t+1} = f(x^t, a^t)$. At the same time, the system would receive the reward $r(x^t)$ from the environment, where $r(\cdot)$ is called the reward function. The goal is to learn an agent that can receive the maximal expected reward in total.

$$\arg\max_{\theta} \quad \mathcal{L}(\theta) = \sum_{t=1}^{\infty} \mathbb{E}_{p_\theta(a^t|x^t)}[r(x^t)]. \tag{9}$$

The objective function is not differentiable with parameter θ. We use policy gradient to obtain an unbiased estimator of the objective gradient $\nabla_\theta \mathcal{L}(\theta)$ [141] and then use stochastic optimization methods to maximize the expected reward. Generating synthesis data can be viewed as sequential decision-making by sequentially generating one basic component (basic structure).

F. Diffusion Model

The diffusion model, also known as the score-based model or score matching method, was proposed in recent years [140] and is widely validated in many generative AI problems such as speech synthesis [36].

Specifically, suppose the data object is x, and the likelihood function is denoted $p(x)$. We are interested in estimating the gradient of the logarithm of the likelihood function.

Diffusion models [150], [151] are inspired by non-equilibrium thermodynamics and can be split into the forward and backward diffusion processes. During the forward diffusion process, diffusion models will gradually add Gaussian noise to the data, and the last-step data will follow an isotropic Gaussian. The reverse diffusion process will revert such a process and construct the data from noise distribution.

More rigorously, we can define the forward process as from the actual data $x_0 \sim p(x)$ to the random noise x_T with T diffusion steps. Let us first assume that for the forward process, the Gaussian distribution is

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I),$$

where $\beta_t \in (0, 1)$. Then, the corresponding backward process is

$$\begin{aligned} & p_\theta(x_{t-1}|x_t) \\ =& \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \\ =& \mathcal{N}(x_{t-1}; \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon), \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t), \end{aligned}$$

where $\epsilon \sim \mathcal{N}(0, I)$ follows the standard Gaussian, $\alpha_t = 1-\beta_t$, and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

The objective of diffusion models is to estimate the variational lower bound (VLB) of the negative log-likelihood of data distribution:

$$\log p(x) \geq -\mathbb{E}_{q(x_{1:T}|x_0)}[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}] = -\mathcal{L}_{\text{VLB}}.$$

The VLB can be rewritten as:

$$\begin{aligned} \mathcal{L}_{\text{VLB}} =& \underbrace{KL[q(x_T|x_0)||p_\theta(x_T)]}_{\mathcal{L}_T} \\ &+ \sum_{t=2}^{T} \underbrace{KL[q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)]}_{\mathcal{L}_{t-1}} \underbrace{-\mathbb{E}_q[\log p_\theta(x_0|x_1)]}_{\mathcal{L}_0}. \end{aligned}$$

Here $\mathcal{L}_T$ is a constant and can be ignored, and diffusion models [151] have been using a separate model for estimating $\mathcal{L}_0$. For $\{\mathcal{L}_{t-1}\}_{t=2}^{T}$, we model a neural network to approximate the conditionals during the reverse process, i.e.,, we want to train $\mu_\theta(x_t, t)$ to predict $\frac{1}{\sqrt{\alpha_t}}\big(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\big)$. If we plug this into the closed-form solution of the KL-divergence between two multivariate Gaussian distributions, we will have the following for $t = 1, \cdots, T-1$:

$$\mathcal{L}_t = \mathbb{E}_{x_0, z}\Big[\|\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t)\|^2\Big].$$

The diffusion model has achieved wide success in many downstream synthetic problems [80], [152]–[154]. As a summarization, Table II compares various generative AI methods from several aspects.

V. Privacy Risks and Prevention

Open release and free data exchange would benefit research and industry development. However, there are cases where datasets exist but cannot be publicly disclosed due to privacy concerns. In the meantime, the emergence and the enforcement of privacy regulations such as HIPAA[1], EU General Data Protection Regulation (GDPR[2]) and the California Consumer Privacy Act (CCPA[3]) make it infeasible for sensitive data with privacy implications. Regulated data such as clinical and genomics data in raw may not be shared and one solution is to share synthesized data instead.

A. Privacy Risks in Data Synthesis

Due to the utility goal of data synthesis, the synthesized data tend to preserve the distribution of the original data. Therefore, the deployment upon these models could be subject to privacy leakage. For deep neural network based approaches, membership inference attack [155], [156] would identify if an input is in the training data or not and thus can be used to determine how close the synthesized data is to the original data. At feature level, sensitive attributes such as skin color can be inferred from the behavior of the deep learning model [157] and even the single training instance can be reconstructed [158]–[160]. For generative AI models, the generative learning process and the high complexity of the model jointly encourage a distribution that is concentrated around training samples. By repeatedly sampling from the distribution, there is a considerable chance of recovering the training samples [161]–[163] or the membership of the training data [164].

[1]https://www.hhs.gov/hipaa/index.html
[2]https://gdpr-info.eu
[3]https://oag.ca.gov/privacy/ccpa

TABLE III
SUMMARIZATION OF PRIVACY PREVENTION STRATEGIES IN SYNTHETIC DATA GENERATION.

Paper	privacy-enhancing techniques	generative AI	DNN	data format
[165]	plausible deniability	probabilistical transform	NA (w.o. DNN)	attribute
[43]	differential privacy	generative artificial neural networks	kernel k-means	image and text
[166]	data replacement and item regularizer	latent space projection	MLP	attributes
[167]	differential privacy	GAN	CNN	image
[49]	differential privacy	GAN	Private Aggregation of Teacher Ensembles	attribute
[168]	differential privacy	GAN	CNN	EHR
[169]	differential privacy	GAN	CNN	image and EHR
[14]	differential privacy	autoencoder	autoencoder	attribute
[170]	differential privacy	MRF	NA (w.o. DNN)	attribute
[171]	compressive privacy	GAN	CNN	image
[172]	differential privacy	GAN	CNN	image
[173]	local differential privacy	Maximum Entropy estimation	NA (w.o. DNN)	attribute
[174]	differential privacy	Maximum Entropy estimation	NA (w.o. DNN)	attribute
[175]	differential privacy	GAN	CNN	image
[176]	differential privacy	Bayesian network	NA (w.o. DNN)	attribute
[177]	differential privacy	statistical database	NA (w.o. DNN)	attribute
[178]	differential privacy	Graduate Update Method	NA (w.o. DNN)	attribute
[179]	differential privacy	autoencoder	autoencoder	image and attribute

B. Privacy Protection in Data Synthesis

Solutions have been proposed in two broad categories. In the first category, different data anonymization-based approaches such as K-anonymity [180]–[182] and nearest marginal [183] to sanitize data so that it cannot be easily re-identified. However, these data anonymization approaches do not provide rigorous privacy guarantees [14]. In the second category, synthetic data generation approaches have been proposed to generate realistic synthetic data using rigorous differential privacy definitions [113], [165], [184] for various applications. In particular, Bindschaedler et al. [165] introduced the idea of plausible deniability instead of directly adding noise to the generative model. This mechanism results in input indistinguishability that means by observing the output set (i.e., synthetics) an adversary cannot make sure whether a particular data record was in the input set (i.e., real data). With the help of generative modeling, Acs et al. [43] clusters the original datasets into k clusters with differentially private kernel k-means and produce synthetic data for each cluster. By comparison, Liu et al [166] introduce two-level privacy-preserving synthetic data generation. At the data level, a selection module is used to select the items which contribute less to the user's preference. At the item level, a synthetic item generation module is developed to create the corresponding synthetic item.

Taking advantage of the GAN, several methods are proposed to generate synthetic data to get better effect [49], [167]–[169] which closely matches the distribution of the source data than the hidden Markov model-based approach [71], RBF based approach [170] and Auto-encoder based approach [14]. Xie et al [169] propose DPGAN by adding noise on the gradient of the Wasserstein distance with respect to the training data. This approach does not adopt the optimization strategy to improve the training stability and convergence speed. To address these problems, Zhang et al. [167] proposed dp-GAN, a general private data publishing framework for rich semantic data without the requirement of tag information compared to [168]. By comparison, Beaulieu-Jones et al. [168] trained the discriminator under differentially private SGD, which generates plausible individuals of clinical datasets. Tseng and Wu [171] apply compressive privacy [185] for CPGAN, which would generate compressing representations that retain high utility. Jordon et al. [49] modifies the Private Aggregation of Teacher Ensembles (PATE) framework and applies it to GANs. The proposed approach allows a tight bound on the influence of any individual sample on the model, resulting in tight differential privacy guarantees and thus an improved performance over models for data synthesis. While most of these approaches inject noise into the energy function, a differentially private GAN called GANobfuscator [172] achieve differential privacy by adding noise within the training procedure.

While centralized differential privacy assumes data aggregators are reliable, local differential privacy (LDP) [184] assumes that aggregators cannot be trusted and relies on data providers to perturb their own data and is used to generate private synthetic datasets that is similar to the private dataset. [173] is inspired by PriView [174] but for computing any k-way marginals under the LDP setting for the marginal table release problem. Apart from LDP in distributed setting, Triastcyn and Faltings [175] propose federated generative privacy that utilizes insuffient local data from multiple clients to train a GAN. The method shares only generators that do not come directly into contact with data and the discriminator remain private. This model can output artificial data, not belonging to any real user in particular, but coming from the common cross-user data distribution.

These privacy-preserving data synthesis methods mainly aim at structured data like tables, which cannot be applied to high dimensionality and complexity. To solve this problem, PriView [174] constructs the private k-way marginal tables for $k \geq 3$ by first extracting low-dimensional marginal views from the flat data and adding noise to the views and then applying a reprocessing technique to ensure the consistency of the noisy views. Zhang et al. [176] consider repetitive perturbation of the original data as a substitute to the original data with a synthetic data generation technique called PrivBayes.

PrivBayes decomposes high dimensional data into low dimensional marginals by constructing a Bayesian network and injects noise into these learned low dimensional marginals to ensure differential privacy and the synthetic data is inferred from these noised marginals. Instead of the Bayesian network, differentially private auto-encoder [14] significantly improves the effectiveness of differentially private synthetic data release. [177] applies data cleaning method [186] to fix the violations on the structure of the data in the synthetic data. Instead of using graphical models as the summarization/representation of a dataset [14], [165], [167], [187], [188], [178] proposes to use a set of large number of low-degree marginals to represent a dataset. The advantage of this approach is that it makes weak assumptions about the conditional independence among attributes, and simply tries to capture correlation relationships that are in the dataset. Meanwhile, the method is especially attractive under differential privacy for its straightforward sensitivity measurement, reduced noise variance, and efficient privacy cost.

C. Privacy Threats in Foundation Models

Entering the era of foundation models, recent research has demonstrated that training data can be exposed from large language models [189] as well as stable diffusion [190]. In both types of models, attackers can generate sequences from the trained model and identify those memorized from the training set. Studies have shown that a sequence that appears multiple times in the training data is more likely to be generated than a sequence that occurred only once [154], [191], [192]. Accordingly, Kandpal et al [193] propose to deduplicate the training data that appears multiple times such that the privacy risks in language models is mitigated.

Given that we are still at the very early stage of the generative foundational models, the potential of the foundation models for data synthesis has not been fully explored. While more possible privacy threats on the foundation models are yet to be discovered, existing privacy measures may be inadequate to meet its demands of privacy. Further investigation is needed to design countermeasures that would mitigate the memorization and generalization problems for privacy protection.

D. Other strategies

Other strategies include

1. Data Anonymization: This process removes personally identifiable information from data sets, ensuring that the individuals the data describe remain anonymous. This is crucial in industries such as healthcare, where patient data privacy is a legal requirement.

2. Data Masking: This technique involves replacing sensitive data with fictitious yet realistic data. It is often used to protect the data while maintaining its usability for testing or development purposes. For example, a developer might need to use customer data to test a new feature, but they don't need to know the actual personal details of the customer to do so.

3. Data Perturbation: This involves adding noise to the data to prevent the identification of individuals in the dataset while preserving the statistical properties of the data. This is particularly useful in research scenarios where data needs to be shared but individual privacy must be maintained.

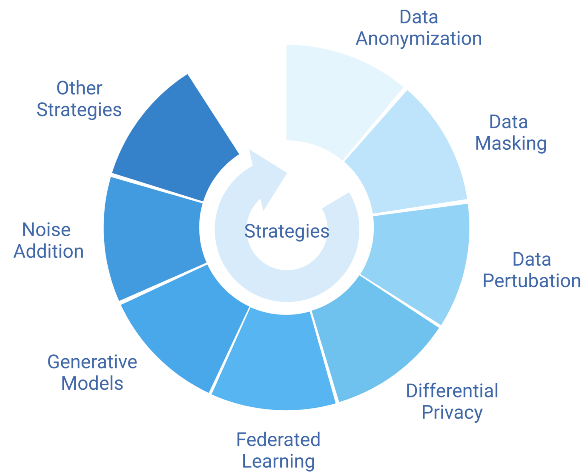

Fig. 3. Synthetic data applications

4. Differential Privacy: This system publicly shares information about a dataset by describing the patterns of groups within the dataset while withholding information about individuals in the dataset. It provides a mathematical guarantee of privacy and is becoming an increasingly popular method for enhancing privacy in machine learning.

5. Federated Learning: This is a machine learning approach where the model is trained across multiple decentralized edge devices or servers holding local data samples, without exchanging them. This approach allows for the utilization of a wide array of data sources, while also ensuring that sensitive data does not leave its original device.

These techniques are all part of the broader field of synthetic data generation, which aims to create data that can be used for a variety of purposes (such as training machine learning models) without compromising privacy.

VI. Fairness

Generating synthetic data that reflect the important underlying statistical properties of the real-world data may also inherit the bias from data preprocessing, collection, and algorithms [194]. The fairness problem is currently addressed by three types of methods [195]: (i) preprocessing, which revises input data to remove information correlated to sensitive attributes, usually via techniques like massaging, reweighting, and sampling. (ii) in-processing, which adds fairness constraints to the model learning process; and (iii) post-processing, which adjusts model predictions after the model is trained.

Most existing fairness-aware data synthesis methods leverage preprocessing techniques. The use of balanced synthetic datasets created by GANs to augment classification training has demonstrated the benefits for reducing disparate impact due to minoritized subgroup imbalance [196]–[198]. [199] models bias using a probabilistic network exploiting structural equation modeling as the preprocessing to generate a fairness-aware synthetic dataset. Authors in [200] leverage GAN as the pre-processing for fair data generation that ensures the generated data is discrimination free while maintaining high data utility. By comparison, [201] is geared towards high dimensional image data and proposes a novel auxiliary classifier GAN that strives for demographic parity or equality of opportunity. However, preprocessing would require the synthesized data provider to know all correlations, biases, and distributions of variables in the existing datasets as a priori. Compared to preprocessing, the latter two categories are less-developed for fair data synthesis.

In the meantime, differential privacy amplifies the fairness issues in the original data [202]. [203] demonstrate that differential privacy does not introduce unfairness into the data generation process or to standard group fairness measures in the downstream classification models, but does unfairly increase the influence of majority subgroups. Differential privacy also significantly reduces the quality of the images generated from the GANs, decreasing the synthetic data's utility in downstream tasks. To measure the fairness in synthesized data, [92] develops two covariate-level disparity fairness metrics for synthetic data. The authors analyze all subgroups defined by protected attributes to analyze the bias.

In the emerging AIGC using foundation models, the generated images and texts may also inherit the stereotypes, exclusion and marginalization of certain groups and toxic and offensive information in the real-world data. This would lead to discrimination and harm to certain social groups. The misuse of such data synthesis approaches by misinformation and manipulation would lead to further negative social impact [204]. Given that the quality of the data generated by foundation models is inextricably linked to the quality of the training corpora, it is essential to regulate the real-world data being used to form the data synthesis distribution. While reducing bias in data is important, the remaining bias in the data may also be amplified by the models [195] or the privacy-enhancing components [202]. With frequent inspection and sensitive and toxic information removal on both data and model, it will help govern the information generated from those foundation models and ensure the models would do no harm (we hope!).

VII. Trustworthiness

As data-driven decision making proliferates across various industries, the creation and use of synthetic data has become increasingly prevalent. Synthetic data, artificially generated data that simulates real-world scenarios, offers a way to bypass several problems associated with real data, such as privacy concerns, scarcity, or data collection difficulty. Nevertheless, the trustworthiness of synthetic data is a subject of ongoing debate, hinging on aspects such as data representativeness, privacy preservation, and potential biases.

For synthetic data to be trustworthy, it must offer a faithful statistical representation of the original data, while maintaining the inherent variability and structure. The risk lies in creating data that oversimplifies or misrepresents the complexities of real-world data, potentially leading to inaccurate conclusions or ineffective solutions when used in analysis or modelling.

Privacy preservation is another critical factor in synthetic data generation. Synthetic data is often utilized in situations where the use of real data may breach privacy regulations or ethical boundaries. While synthetic data promises a level of anonymity, there is an ongoing debate about the extent to which this data can be de-anonymized. If synthetic data could be traced back to the original contributors, it would undermine its trustworthiness and the privacy it promises to uphold.

Potential biases in synthetic data are a significant concern. Even though synthetic data is artificially generated, it often relies on real-world data to inform its creation. Thus, if the real-world data is biased, these biases could be unwittingly replicated in the synthetic data, perpetuating the same flawed patterns and undermining its trustworthiness.

Moreover, assessing the trustworthiness of synthetic data involves the evaluation of the synthetic data generation methods themselves. Transparency in the generation process, including a clear understanding of the underlying algorithms and parameters used, is crucial in judging the trustworthiness of the resultant synthetic data.

In conclusion, while synthetic data presents compelling benefits, the trustworthiness of such data depends on its representativeness, privacy preservation, and absence of bias. By recognizing and addressing these concerns, researchers and practitioners can make informed decisions about synthetic data's validity and ethical implications. Transparent and robust synthetic data generation methods are paramount in fostering this trust.

VIII. Evaluation Strategy

In this section, we discuss various approaches to evaluating the quality of synthesized data, which is essential for determining the effectiveness and applicability of synthetic data generation methods in real-world scenarios. We categorize these evaluation strategies as follows:

1) **Human evaluation**. This method is the most direct way to assess the quality of synthesized data. Human evaluation involves soliciting opinions from domain experts or non-expert users to judge the synthesized data's quality, similarity to real data, or usability in specific applications. For example, in speech synthesis, the human evaluator rates the synthesized speech and real human speech in a blind manner [44], [205]. However, human evaluation has several drawbacks, including being expensive, time-consuming, error-prone, and not scalable. Additionally, it struggles with high-dimensional data that cannot be easily visualized and evaluated by humans.
2) **Statistical difference evaluation**. This strategy involves calculating various statistical metrics on both the synthesized and real datasets and comparing the results. For example, [53], [206] use first-moment statistics of individual features (e.g., medical concept frequency/correlation, patient-level clinical feature) to evaluate the quality of generated electronic health record (EHR) data. The smaller the differences between the statistical properties of synthetic and real data, the better the quality of the synthesized data.

3) **Evaluation using a pre-trained machine learning model**. As mentioned in Section IV-D, in the generative adversarial network (GAN), the discriminator differentiates fake data (synthesized data) from real ones. Consequently, the output of the discriminator can measure how closely synthetic data resembles real data. The performance of the discriminator on the synthesized data can be used as an indicator of how well the generator produces realistic data. This strategy can be applied not only to GANs but also to other generative models where a pre-trained machine learning model is used for evaluation.
4) **Training on synthetic dataset and testing on the real dataset (TSTR)**. This strategy involves using synthetic data to train machine learning models and assessing their prediction performance on real test data in downstream applications. High performance on real test data indicates that the synthetic data has successfully captured essential characteristics of the real data, making it a useful proxy for training. For example, [207] employ synthetic data to train machine learning models and assess their prediction performance on real test data in downstream applications. TSTR can provide insights into the effectiveness of synthetic data for training machine learning models in a wide range of tasks and domains.
5) **Application-specific evaluation**. Depending on the specific use case or domain, tailored evaluation methods may be employed to assess the quality of synthesized data. These evaluation methods can consider the unique requirements or constraints of the application, such as regulatory compliance, privacy concerns, or specific performance metrics. By evaluating the synthesized data in the context of its intended use, a more accurate assessment of its quality and applicability can be obtained.

These evaluation strategies offer various ways to gauge the quality of synthesized data, helping researchers and practitioners determine the effectiveness of synthetic data generation methods and their applicability in real-world scenarios. Employing a combination of these strategies can provide a more comprehensive understanding of the strengths and weaknesses of the synthesized data, facilitating further improvements in synthetic data generation techniques [208].

IX. Challenges and Opportunities

The aim of this research is to present a comprehensive survey of synthetic data generation—a promising and emerging technique in contemporary deep learning. This survey outlines current real-world applications and identifies potential avenues for future research in this field. The utilization of synthetic data has been proven effective across a diverse array of tasks and domains [9]. In this section, we delve into the challenges and opportunities presented by this rapidly evolving area.

First and foremost, evaluation metrics for synthetic data are essential to determine the reasonableness of the generated data. In industries like healthcare, where data quality is of paramount importance, clinical quality measures and evaluation metrics are not always readily available for synthetic data. Clinicians often struggle to interpret existing criteria such as probability likelihood and divergence scores when assessing generative models [68]. Concurrently, there is a pressing need to develop and adopt specific regulations for the use of synthetic data in medicine and healthcare, ensuring that the generated data meets the required quality standards while minimizing potential risks.

Secondly, due to limited attention and the challenges associated with covering various domains using synthetic data, current methods might not account for all outliers and corner cases present in the original data. Investigating outliers and regular instances and their impact on the parameterization of existing methods could be a valuable research direction [209]. To enhance future detection methods, it may be beneficial to examine the gap between the performance of detection methods and a well-designed evaluation matrix, which could provide insights into areas that require improvement.

Thirdly, synthetic data generation may involve underlying models with inherent biases, which might not be immediately evident [92]. Factors such as sample selection biases and class imbalances can contribute to these issues. Typically, algorithms trained with biases in sample selection may underperform when deployed in settings that deviate significantly from the conditions in which the data was collected [68]. Thus, it is crucial to develop methods and strategies that address these biases, ensuring that synthetic data generation leads to more accurate and reliable results across diverse applications and domains.

In general, the use of synthetic data is becoming a viable alternative to training models with real data due to advances in simulations and generative models. However, a number of open challenges need to be overcome to achieve high performance. These include the lack of standard tools, the difference between synthetic and real data, and how much machine learning algorithms can do to exploit imperfect synthetic data effectively. Though this emerging approach is not perfect now, with models, metrics, and technologies maturing, we believe synthetic data generation will make a bigger impact in the future.

X. Conclusion

In conclusion, machine learning has revolutionized various industries by enabling intelligent computer systems to autonomously tackle tasks, manage and analyze massive volumes of data. However, machine learning faces several challenges, including data quality, data scarcity, and data governance. These challenges can be addressed through synthetic data generation, which involves the artificial annotation of information generated by computer algorithms or simulations. Synthetic data has been extensively utilized in various sectors due to its ability to bridge gaps, especially when real data is either unavailable or must be kept private due to privacy or compliance risks.

This paper has provided a high-level overview of several state-of-the-art approaches currently being investigated by machine learning researchers for synthetic data generation. We have explored different real-world application domains, and examined a diverse array of deep neural network architectures and deep generative models dedicated to generating high-quality synthetic data.

To sum up, synthetic data generation has enormous potential for unlocking the full potential of machine learning and its impact on various industries. While challenges persist in the development and application of machine learning technology, synthetic data generation provides a promising solution that can help address these obstacles. Future research can further enhance the functionality of synthetic data generation.

Acknowledgement. The authors acknowledge partial support by the xxx

REFERENCES

[1] A. Ng, "What artificial intelligence can and can't do right now," *Harvard Business Review*, vol. 9, no. 11, 2016.

[2] M. A. Boden, *Artificial intelligence*. Elsevier, 1996.

[3] M. Haenlein and A. Kaplan, "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence," *California management review*, vol. 61, no. 4, pp. 5–14, 2019.

[4] F. Lucini, "The real deal about synthetic data," *MIT Sloan Management Review*, vol. 63, no. 1, pp. 1–4, 2021.

[5] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[6] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, no. 4, pp. 211–218, 2002.

[7] M. Shen, Y.-T. Chang, C.-T. Wu, S. J. Parker, G. Saylor, Y. Wang, G. Yu, J. E. Van Eyk, R. Clarke, D. M. Herrington *et al.*, "Comparative assessment and novel strategy on methods for imputing proteomics data," *Scientific reports*, vol. 12, no. 1, p. 1067, 2022.

[8] R. Babbar and B. Schölkopf, "Data scarcity, robustness and extreme multi-label classification," *Machine Learning*, vol. 108, no. 8, pp. 1329–1351, 2019.

[9] S. I. Nikolenko, *Synthetic data for deep learning*. Springer, 2021, vol. 174.

[10] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and information systems*, vol. 34, no. 3, pp. 483–519, 2013.

[11] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," in *IEEE international symposium on biomedical imaging (ISBI)*, 2018.

[12] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *IEEE/CVF conference on computer vision and pattern recognition*, 2019.

[13] J. M. Abowd and L. Vilhuber, "How protective are synthetic data?" in *International Conference on Privacy in Statistical Databases*. Springer, 2008.

[14] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney, "Privacy preserving synthetic data release using deep learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019.

[15] T. E. Raghunathan, "Synthetic data," *Annual Review of Statistics and Its Application*, vol. 8, pp. 129–140, 2021.

[16] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," in *Machine learning for healthcare conference*. PMLR, 2017.

[17] J. D. Ziegler, S. Subramaniam, M. Azzarito, O. Doyle, P. Krusche, and T. Coroller, "Multi-modal conditional GAN: Data synthesis in the medical domain," in *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.

[18] K. W. Dunn, C. Fu, D. J. Ho, S. Lee, S. Han, P. Salama, and E. J. Delp, "DeepSynth: Three-dimensional nuclear segmentation of biological images using neural networks trained with synthetic data," *Scientific reports*, vol. 9, no. 1, pp. 1–15, 2019.

[19] Y. Du, X. Liu, N. Shah, S. Liu, J. Zhang, and B. Zhou, "Chemspace: Interpretable and interactive chemical space exploration," 2022.

[20] T. Sterling and J. J. Irwin, "Zinc 15–ligand discovery for everyone," *Journal of chemical information and modeling*, vol. 55, no. 11, pp. 2324–2337, 2015.

[21] W. Jin, R. Barzilay, and T. S. Jaakkola, "Junction tree variational autoencoder for molecular graph generation," in *International Conference on Machine Learning*, 2018.

[22] M. Olivecrona, T. Blaschke, O. Engkvist, and H. Chen, "Molecular de-novo design through deep reinforcement learning," *Journal of cheminformatics*, vol. 9, no. 1, p. 48, 2017.

[23] T. Fu, C. Xiao, and J. Sun, "CORE: Automatic molecule optimization using copy and refine strategy," *AAAI conference on artificial intelligence*, 2020.

[24] T. Fu, W. Gao, C. W. Coley, and J. Sun, "Reinforced genetic algorithm for structure-based drug design," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[25] K. Huang, T. Fu, W. Gao, Y. Zhao, Y. Roohani, J. Leskovec, C. W. Coley, C. Xiao, J. Sun, and M. Zitnik, "Artificial intelligence foundation for therapeutic science," *Nature Chemical Biology*, pp. 1–4, 2022.

[26] A. Torfi and E. A. Fox, "Corgan: Correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records," in *International Flairs Conference*, 2020.

[27] D. Lee, H. Yu, X. Jiang, D. Rogith, M. Gudala, M. Tejani, Q. Zhang, and L. Xiong, "Generating sequential electronic health records using dual adversarial autoencoder," *Journal of the American Medical Informatics Association*, vol. 27, no. 9, pp. 1411–1419, 2020.

[28] S. Wharrie, Z. Yang, V. Raj, R. Monti, R. Gupta, Y. Wang, A. Martin, L. J. O'Connor, S. Kaski, P. Marttinen *et al.*, "HAPNEST: an efficient tool for generating large-scale genetics datasets from limited training data," in *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.

[29] B. Nowok, G. M. Raab, and C. Dibben, "synthpop: Bespoke creation of synthetic data in R," *Journal of statistical software*, vol. 74, pp. 1–26, 2016.

[30] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE international conference on computer vision*, 2017.

[31] R. Torkzadehmahani, P. Kairouz, and B. Paten, "Dp-cgan: Differentially private synthetic data and label generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[32] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[33] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *arXiv preprint arXiv:2204.03458*, 2022.

[34] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *Advances in neural information processing systems*, vol. 32, 2019.

[35] M. Niemeyer and A. Geiger, "Giraffe: Representing scenes as compositional generative neural feature fields," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[36] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," *International Conference on Learning Representations (ICLR)*, 2021.

[37] H. Guo, F. K. Soong, L. He, and L. Xie, "A new GAN-based end-to-end TTS training algorithm," *arXiv preprint arXiv:1904.04775*, 2019.

[38] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in *AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[39] T. Sellam, D. Das, and A. P. Parikh, "Bleurt: Learning robust metrics for text generation," *arXiv preprint arXiv:2004.04696*, 2020.

[40] Z. Shi, X. Chen, X. Qiu, and X. Huang, "Toward diverse text generation with inverse reinforcement learning," in *International Joint Conference on Artificial Intelligence*, 2018.

[41] C.-Y. Ko, P.-Y. Chen, J. Mohapatra, P. Das, and L. Daniel, "Synbench: Task-agnostic benchmarking of pretrained representations using synthetic data," *arXiv preprint arXiv:2210.02989*, 2022.

[42] W. Nie, N. Narodytska, and A. Patel, "Relgan: Relational generative adversarial networks for text generation," in *International conference on learning representations*, 2018.

[43] G. Acs, L. Melis, C. Castelluccia, and E. De Cristofaro, "Differentially private mixture of generative neural networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 6, pp. 1109–1121, 2018.

[44] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *arXiv preprint arXiv:1802.04208*, 2018.

[45] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet:

A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[46] X. Zhang, I. Vallés-Pérez, A. Stolcke, C. Yu, J. Droppo, O. Shonibare, R. Barra-Chicote, and V. Ravichandran, "Stutter-tts: Controlled synthesis and improved recognition of stuttered speech," *arXiv preprint arXiv:2211.09731*, 2022.

[47] M. Wiese, R. Knobloch, R. Korn, and P. Kretschmer, "Quant GANs: deep generation of financial time series," *Quantitative Finance*, vol. 20, no. 9, pp. 1419–1440, 2020.

[48] R. Fu, J. Chen, S. Zeng, Y. Zhuang, and A. Sudjianto, "Time series simulation by conditional generative adversarial net," *arXiv preprint arXiv:1904.11419*, 2019.

[49] J. Jordon, J. Yoon, and M. Van Der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees," in *International conference on learning representations*, 2018.

[50] A. Collaboration *et al.*, "Deep generative models for fast photon shower simulation in atlas," *arXiv preprint arXiv:2210.06204*, 2022.

[51] C. Dewi, R.-C. Chen, Y.-T. Liu, and S.-K. Tai, "Synthetic data generation using dcgan for improved traffic sign recognition," *Neural Computing and Applications*, vol. 34, no. 24, pp. 21 465–21 480, 2022.

[52] Z. Zhao, K. Xu, S. Li, Z. Zeng, and C. Guan, "Mt-uda: Towards unsupervised cross-modality medical image segmentation with limited source labels," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 293–303.

[53] S. Yi, M. Lu, A. Yee, J. Harmon, F. Meng, and S. Hinduja, "Enhance wound healing monitoring through a thermal imaging based smartphone app," in *Medical Imaging: Imaging Informatics for Healthcare, Research, and Applications*. SPIE, 2018.

[54] Y. Chen, W. Li, X. Chen, and L. V. Gool, "Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[55] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *IEEE/CVF conference on computer vision and pattern recognition*, 2018.

[56] H.-W. Dong and Y.-H. Yang, "Towards a deeper understanding of adversarial losses," *arXiv preprint arXiv:1901.08753*, 2019.

[57] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, T. J. Cashman, and J. Shotton, "Fake it till you make it: face analysis in the wild using synthetic data alone," in *IEEE/CVF international conference on computer vision*, 2021.

[58] A. Werchniak, R. B. Chicote, Y. Mishchenko, J. Droppo, J. Condal, P. Liu, and A. Shah, "Exploring the application of synthetic audio in training keyword spotters," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[59] W. Li, H. You, J. Zhu, and N. Chen, "Feature sparsity analysis for i-vector based speaker verification," *Speech Communication*, vol. 80, pp. 60–70, 2016.

[60] Y. Qian, Y. Liu, and K. Yu, "Tandem deep features for text-dependent speaker verification," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[61] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 10, pp. 2129–2139, 2013.

[62] A. Fazel, W. Yang, Y. Liu, R. Barra-Chicote, Y. Meng, R. Maas, and J. Droppo, "Synthasr: Unlocking synthetic data for speech recognition," *arXiv preprint arXiv:2106.07803*, 2021.

[63] W. Li and J. Zhu, "An improved i-vector extraction algorithm for speaker verification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, pp. 1–9, 2015.

[64] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[65] X. Yue, H. A. Inan, X. Li, G. Kumar, J. McAnallen, H. Sun, D. Levitan, and R. Sim, "Synthetic text generation with differential privacy: A simple and practical recipe," *arXiv preprint arXiv:2210.14348*, 2022.

[66] X. Zheng, Y. Liu, D. Gunceler, and D. Willett, "Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[67] Z. Zhao, A. Zhu, Z. Zeng, B. Veeravalli, and C. Guan, "Act-net: Asymmetric co-teacher network for semi-supervised memory-efficient medical image segmentation," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 1426–1430.

[68] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. Williamson, and F. Mahmood, "Synthetic data in machine learning for medicine and healthcare," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 493–497, 2021.

[69] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles, "Generating high-fidelity synthetic patient data for assessing machine learning healthcare software," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–13, 2020.

[70] Y. Lu, C.-T. Wu, S. J. Parker, Z. Cheng, G. Saylor, J. E. Van Eyk, G. Yu, R. Clarke, D. M. Herrington, and Y. Wang, "Cot: an efficient and accurate method for detecting marker genes among many subtypes," *Bioinformatics Advances*, vol. 2, no. 1, p. vbac037, 2022.

[71] J. Dahmen and D. Cook, "Synsys: A synthetic data generation system for healthcare applications," *Sensors*, vol. 19, no. 5, p. 1181, 2019.

[72] Y. Lu, Y.-T. Chang, E. P. Hoffman, G. Yu, D. M. Herrington, R. Clarke, C.-T. Wu, L. Chen, and Y. Wang, "Integrated identification of disease specific pathways using multi-omics data," *bioRxiv*, p. 666065, 2019.

[73] Z. Wang, P. Myles, and A. Tucker, "Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy," *Computational Intelligence*, vol. 37, no. 2, pp. 819–851, 2021.

[74] R. S. Bohacek, C. McMartin, and W. C. Guida, "The art and practice of structure-based drug design: a molecular modeling perspective," *Medicinal research reviews*, vol. 16, no. 1, pp. 3–50, 1996.

[75] K. Huang, T. Fu, L. M. Glass, M. Zitnik, C. Xiao, and J. Sun, "DeepPurpose: a deep learning library for drug–target interaction prediction," *Bioinformatics*, vol. 36, no. 22-23, pp. 5545–5547, 2020.

[76] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS central science*, vol. 4, no. 2, pp. 268–276, 2018.

[77] B. Zhang, Y. Fu, Y. Lu, Z. Zhang, R. Clarke, J. E. Van Eyk, D. M. Herrington, and Y. Wang, "DDN2.0: R and python packages for differential dependency network analysis of biological systems," *bioRxiv*, pp. 2021–04, 2021.

[78] N. De Cao and T. Kipf, "MolGAN: An implicit generative model for small molecular graphs," *arXiv preprint arXiv:1805.11973*, 2018.

[79] T. Fu and J. Sun, "Antibody Complementarity Determining Regions (CDRs) design using constrained energy model," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[80] T. Fu, W. Gao, C. Xiao, J. Yasonik, C. W. Coley, and J. Sun, "Differentiable scaffolding tree for molecular optimization," *International Conference on Learning Representations*, 2022.

[81] M. Xu, L. Yu, Y. Song, C. Shi, S. Ermon, and J. Tang, "GeoDiff: A geometric diffusion model for molecular conformation generation," in *International Conference on Learning Representations*, 2021.

[82] Z. Zhou, S. Kearnes, L. Li, R. N. Zare, and P. Riley, "Optimization of molecules via deep reinforcement learning," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[83] J. H. Jensen, "A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space," *Chemical science*, vol. 10, no. 12, pp. 3567–3572, 2019.

[84] T. Fu, C. Xiao, X. Li, L. M. Glass, and J. Sun, "MIMOSA: Multi-constraint molecule sampling for molecule optimization," in *AAAI Conference on Artificial Intelligence*, 2021.

[85] T. Fu and J. Sun, "SIPF: Sampling method for inverse protein folding," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[86] C. S. Kruse, B. Smith, H. Vanderlinden, and A. Nealand, "Security techniques for the electronic health records," *Journal of medical systems*, vol. 41, no. 8, pp. 1–9, 2017.

[87] Q. Wen, Z. Ouyang, J. Zhang, Y. Qian, Y. Ye, and C. Zhang, "Disentangled dynamic heterogeneous graph learning for opioid overdose prediction," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[88] T. Fu, T. Gao, C. Xiao, T. Ma, and J. Sun, "Pearl: Prototype learning via rule learning," in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019, pp. 223–232.

[89] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," *BMC medical research methodology*, vol. 20, no. 1, pp. 1–40, 2020.

[90] D. Du, S. Bhardwaj, S. J. Parker, Z. Cheng, Z. Zhang, Y. Lu, J. E. Van Eyk, G. Yu, R. Clarke, D. M. Herrington *et al.*, "Abds: tool suite for analyzing biologically diverse samples," *bioRxiv*, pp. 2023–07, 2023.

[91] Y. Lu, “Multi-omics data integration for identifying disease specific biological pathways,” Ph.D. dissertation, Virginia Tech, 2018.
[92] K. Bhanot, M. Qi, J. S. Erickson, I. Guyon, and K. P. Bennett, “The problem of fairness in synthetic healthcare data,” *Entropy*, vol. 23, no. 9, p. 1165, 2021.
[93] T. Fu, K. Huang, C. Xiao, L. M. Glass, and J. Sun, “HINT: Hierarchical interaction network for clinical-trial-outcome predictions,” *Patterns*, vol. 3, no. 4, p. 100445, 2022.
[94] T. Fu, T. N. Hoang, C. Xiao, and J. Sun, “DDL: Deep dictionary learning for predictive phenotyping,” in *International Joint Conference on Artificial Intelligence*, 2019.
[95] L. Chen, Y. Lu, C.-T. Wu, R. Clarke, G. Yu, J. E. Van Eyk, D. M. Herrington, and Y. Wang, “Data-driven detection of subtype-specific differentially expressed genes,” *Scientific reports*, vol. 11, no. 1, pp. 1–12, 2021.
[96] P. Eigenschink, S. Vamosi, R. Vamosi, C. Sun, T. Reutterer, and K. Kalcher, “Deep generative models for synthetic data,” *ACM Computing Surveys*, 2021.
[97] C.-T. Wu, M. Shen, D. Du, Z. Cheng, S. J. Parker, Y. Lu, J. E. Van Eyk, G. Yu, R. Clarke, D. M. Herrington *et al.*, “Cosbin: cosine score-based iterative normalization of biologically diverse samples,” *Bioinformatics Advances*, vol. 2, no. 1, p. vbac076, 2022.
[98] R. Wang and X. Qu, “Eeg daydreaming, a machine learning approach to detect daydreaming activities,” in *Augmented Cognition: International Conference*. Springer, 2022.
[99] Y. Du, T. Fu, J. Sun, and S. Liu, “Molgensurvey: A systematic survey in machine learning models for molecule design,” *arXiv preprint arXiv:2203.14500*, 2022.
[100] K. El Emam, L. Mosquera, and R. Hoptroff, *Practical synthetic data generation: balancing privacy and the broad availability of data*. O’Reilly Media, 2020.
[101] M. Mannino and A. Abouzied, “Is this real? generating synthetic data that looks real,” in *ACM Symposium on User Interface Software and Technology*, 2019.
[102] S. A. Assefa, D. Dervovic, M. Mahfouz, R. E. Tillman, P. Reddy, and M. Veloso, “Generating synthetic data in finance: opportunities, challenges and pitfalls,” in *ACM International Conference on AI in Finance*, 2020.
[103] P.-H. Lu, P.-C. Wang, and C.-M. Yu, “Empirical evaluation on synthetic data generation with generative adversarial network,” in *International Conference on Web Intelligence, Mining and Semantics*, 2019.
[104] M. Hittmeir, A. Ekelhart, and R. Mayer, “On the utility of synthetic data: An empirical evaluation on machine learning tasks,” in *International Conference on Availability, Reliability and Security*, 2019.
[105] A. M. Berg, S. T. Mol, G. Kismihók, and N. Sclater, “The role of a reference synthetic data generator within the field of learning analytics.” *Journal of Learning Analytics*, vol. 3, no. 1, pp. 107–128, 2016.
[106] B. Howe, J. Stoyanovich, H. Ping, B. Herman, and M. Gee, “Synthetic data for social good,” *arXiv preprint arXiv:1710.08874*, 2017.
[107] P. Bautista and P. S. Inventado, “Protecting student privacy with synthetic data from generative adversarial networks,” in *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part II*. Springer, 2021, pp. 66–70.
[108] H. Jiang, J. Li, P. Zhao, F. Zeng, Z. Xiao, and A. Iyengar, “Location privacy-preserving mechanisms in location-based services: A comprehensive survey,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–36, 2021.
[109] R. Kato, M. Iwata, T. Hara, A. Suzuki, X. Xie, Y. Arase, and S. Nishio, “A dummy-based anonymization method based on user trajectory with pauses,” in *International Conference on Advances in Geographic Information Systems*, 2012.
[110] Y. Ye, Y. Fan, S. Hou, Y. Zhang, Y. Qian, S. Sun, Q. Peng, M. Ju, W. Song, and K. Loparo, “Community mitigation: A data-driven system for covid-19 risk assessment in a hierarchical manner,” in *ACM International Conference on Information & Knowledge Management*, 2020.
[111] Y. Du, S. Wang, X. Guo, H. Cao, S. Hu, J. Jiang, A. Varala, A. Angirekula, and L. Zhao, “GraphGT: Machine learning datasets for graph generation and transformation,” in *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.
[112] R. Chen, G. Acs, and C. Castelluccia, “Differentially private sequential data publication via variable-length n-grams,” in *ACM conference on Computer and communications security*, 2012.
[113] C. Dwork, A. Roth *et al.*, “The algorithmic foundations of differential privacy,” *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.
[114] X. He, G. Cormode, A. Machanavajjhala, C. M. Procopiuc, and D. Srivastava, “Dpt: differentially private trajectory synthesis using hierarchical reference systems,” *VLDB Endowment*, vol. 8, no. 11, pp. 1154–1165, 2015.
[115] M. E. Gursoy, L. Liu, S. Truex, L. Yu, and W. Wei, “Utility-aware synthesis of differentially private and attack-resilient location traces,” in *ACM SIGSAC conference on computer and communications security*, 2018.
[116] V. Bindschaedler and R. Shokri, “Synthesizing plausible privacy-preserving location traces,” in *IEEE Symposium on Security and Privacy (SP)*, 2016.
[117] H. Wang, Z. Zhang, T. Wang, S. He, M. Backes, J. Chen, and Y. Zhang, “Privtrace: Differentially private trajectory synthesis by adaptive markov model,” in *USENIX Security Symposium 2023*, 2023.
[118] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, “A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt,” *arXiv preprint arXiv:2303.04226*, 2023.
[119] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
[120] W. Tao, S. Gao, and Y. Yuan, “Boundary crossing: an experimental study of individual perceptions toward aigc,” *Frontiers in Psychology*, vol. 14, 2023.
[121] R. J. M. Ventayen, “Openai chatgpt generated results: Similarity index of artificial intelligence-based contents,” *Available at SSRN 4332664*, 2023.
[122] T. Yue, D. Au, C. C. Au, and K. Y. Iu, “Democratizing financial knowledge with chatgpt by openai: Unleashing the power of technology,” *Available at SSRN 4346152*, 2023.
[123] V. Seib, B. Lange, and S. Wirtz, “Mixing real and synthetic data to enhance neural network training–a review of current approaches,” *arXiv preprint arXiv:2007.08781*, 2020.
[124] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[125] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
[126] S. Hochreiter and J. Schmidhuber, “Lstm can solve hard long time lag problems,” *Advances in neural information processing systems*, vol. 9, 1996.
[127] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN Encoder–Decoder for statistical machine translation,” in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
[128] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *International Conference on Learning Representations (ICLR)*, 2016.
[129] Y. Qian, Y. Zhang, Y. Ye, C. Zhang *et al.*, “Distilling meta knowledge on heterogeneous graph for illicit drug trafficker detection on social media,” in *Advances in Neural Information Processing Systems*, 2021.
[130] W. Du, H. Zhang, Y. Du, Q. Meng, W. Chen, B. Shao, and T.-Y. Liu, “Equivariant vector field network for many-body system modeling,” *arXiv preprint arXiv:2110.14811*, 2021.
[131] Y. Qian, Y. Zhang, Y. Ye, and C. Zhang, “Adapting meta knowledge with heterogeneous information network for covid-19 themed malicious repository detection,” in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.
[132] Y. Qian, Y. Zhang, N. Chawla, Y. Ye, and C. Zhang, “Malicious repositories detection with adversarial heterogeneous graph contrastive learning,” in *ACM International Conference on Information & Knowledge Management*, 2022.
[133] Y. Qian, Y. Zhang, Q. Wen, Y. Ye, and C. Zhang, “Rep2vec: Repository embedding via heterogeneous graph adversarial contrastive learning,” in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.
[134] Y. Du, X. Guo, A. Shehu, and L. Zhao, “Interpretable molecular graph generation via monotonic constraints,” in *SIAM International Conference on Data Mining (SDM)*, 2022.
[135] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, 2020.
[136] S. Ghosh, M. Chollet, E. Laksana, L.-P. Morency, and S. Scherer, “Affect-LM: A neural language model for customizable affective text generation,” in *Annual Meeting of the Association for Computational Linguistics*, 2017.

[137] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *International Conference on Learning Representations (ICLR)*, 2014.
[138] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014.
[139] I. J. Goodfellow, "On distinguishability criteria for estimating generative models," *arXiv preprint arXiv:1412.6515*, 2014.
[140] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.
[141] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
[142] S. H. Lee, "Natural language generation for electronic health records," *NPJ digital medicine*, vol. 1, no. 1, pp. 1–7, 2018.
[143] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, "Strategies for pre-training graph neural networks," in *International Conference on Learning Representations*, 2019.
[144] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-vae," *arXiv preprint arXiv:1804.03599*, 2018.
[145] A. Vahdat and J. Kautz, "Nvae: A deep hierarchical variational autoencoder," *Advances in neural information processing systems*, vol. 33, pp. 19 667–19 679, 2020.
[146] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[147] Y. Zhang, Y. Qian, Y. Fan, Y. Ye, X. Li, Q. Xiong, and F. Shao, "dstyle-gan: Generative adversarial network based on writing and photography styles for drug identification in darknet markets," in *Annual Computer Security Applications Conference*, 2020.
[148] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017.
[149] Y. Zhang, Y. Qian, Y. Ye, and C. Zhang, "Adapting distilled knowledge for few-shot relation reasoning over knowledge graphs," in *SIAM International Conference on Data Mining (SDM)*, 2022.
[150] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015.
[151] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, 2020.
[152] M. Liu, K. Yan, B. Oztekin, and S. Ji, "Graphebm: Molecular graph generation with energy-based models," *arXiv preprint arXiv:2102.00546*, 2021.
[153] L. Weng, "What are diffusion models?" *lilianweng.github.io/lil-log*, 2021. [Online]. Available: https://lilianweng.github.io/lil-log/2021/07/11/diffusion-models.html
[154] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein, "Diffusion art or digital forgery? investigating data replication in diffusion models," *arXiv preprint arXiv:2212.03860*, 2022.
[155] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE symposium on security and privacy (SP)*, 2017.
[156] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Demystifying membership inference attacks in machine learning as a service," *IEEE Transactions on Services Computing*, vol. 14, no. 6, pp. 2073–2089, 2019.
[157] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *IEEE symposium on security and privacy (SP)*, 2019, pp. 691–706.
[158] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in neural information processing systems*, vol. 32, 2019.
[159] W. Wei, L. Liu, M. Loper, K.-H. Chow, M. E. Gursoy, S. Truex, and Y. Wu, "A framework for evaluating client privacy leakages in federated learning," in *European Symposium on Research in Computer Security*. Springer, 2020.
[160] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients-how easy is it to break privacy in federated learning?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 937–16 947, 2020.
[161] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "Logan: Membership inference attacks against generative models," *Proceedings on Privacy Enhancing Technologies*, no. 1, pp. 133–152, 2019.
[162] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *ACM SIGSAC conference on computer and communications security*, 2017, pp. 603–618.
[163] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *IEEE conference on computer communications*, 2019.
[164] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "Gan-leaks: A taxonomy of membership inference attacks against generative models," in *ACM SIGSAC conference on computer and communications security*, 2020, pp. 343–362.
[165] V. Bindschaedler, R. Shokri, and C. A. Gunter, "Plausible deniability for privacy-preserving data synthesis," *VLDB Endowment*, vol. 10, no. 5, 2017.
[166] F. Liu, Z. Cheng, H. Chen, Y. Wei, L. Nie, and M. Kankanhalli, "Privacy-preserving synthetic data generation for recommendation systems," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.
[167] X. Zhang, S. Ji, and T. Wang, "Differentially private releasing via deep generative model (technical report)," *arXiv preprint arXiv:1801.01594*, 2018.
[168] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, and C. S. Greene, "Privacy-preserving generative deep neural networks support clinical data sharing," *Circulation: Cardiovascular Quality and Outcomes*, vol. 12, no. 7, p. e005122, 2019.
[169] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," *arXiv preprint arXiv:1802.06739*, 2018.
[170] K. Cai, X. Lei, J. Wei, and X. Xiao, "Data synthesis via differentially private markov random fields," *VLDB Endowment*, vol. 14, no. 11, pp. 2190–2202, 2021.
[171] B.-W. Tseng and P.-Y. Wu, "Compressive privacy generative adversarial network," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2499–2513, 2020.
[172] C. Xu, J. Ren, D. Zhang, Y. Zhang, Z. Qin, and K. Ren, "Ganobfuscator: Mitigating information leakage under gan via differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 9, pp. 2358–2371, 2019.
[173] Z. Zhang, T. Wang, N. Li, S. He, and J. Chen, "Calm: Consistent adaptive local marginal for marginal release under local differential privacy," in *ACM SIGSAC Conference on Computer and Communications Security*, 2018.
[174] W. Qardaji, W. Yang, and N. Li, "Priview: practical differentially private release of marginal contingency tables," in *ACM SIGMOD international conference on Management of data*, 2014.
[175] A. Triastcyn and B. Faltings, "Federated generative privacy," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 50–57, 2020.
[176] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via bayesian networks," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 4, pp. 1–41, 2017.
[177] C. Ge, S. Mohapatra, X. He, and I. F. Ilyas, "Kamino: Constraint-aware differentially private data synthesis," *arXiv preprint arXiv:2012.15713*, 2020.
[178] Z. Zhang, T. Wang, N. Li, J. Honorio, M. Backes, S. He, J. Chen, and Y. Zhang, "{PrivSyn}: Differentially private data synthesis," in *USENIX Security Symposium*, 2021.
[179] Q. Chen, C. Xiang, M. Xue, B. Li, N. Borisov, D. Kaarfar, and H. Zhu, "Differentially private data generative models," *arXiv preprint arXiv:1812.02274*, 2018.
[180] L. Sweeney, "k-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.
[181] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," in *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 1998, pp. 10–1145.
[182] P. Samarati, "Protecting respondents identities in microdata release," *IEEE transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.
[183] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar, "Privacy, accuracy, and consistency too: a holistic solution to contingency table release," in *ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2007.
[184] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *IEEE Annual Symposium on Foundations of Computer Science*, 2013.
[185] S.-Y. Kung, "Compressive privacy: From information\/estimation theory to machine learning," *IEEE Signal Processing Magazine*, vol. 34, no. 1, pp. 94–112, 2017.

[186] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré, "Holoclean: Holistic data repairs with probabilistic inference," *arXiv preprint arXiv:1702.00820*, 2017.

[187] M. Gaboardi, E. J. G. Arias, J. Hsu, A. Roth, and Z. S. Wu, "Dual query: Practical private query release for high dimensional data," in *International Conference on Machine Learning*, 2014.

[188] M. Hardt, K. Ligett, and F. McSherry, "A simple and practical algorithm for differentially private data release," *Advances in neural information processing systems*, 2012.

[189] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, U. Erlingsson *et al.*, "Extracting training data from large language models." in *USENIX Security Symposium*, vol. 6, 2021.

[190] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace, "Extracting training data from diffusion models," *arXiv preprint arXiv:2301.13188*, 2023.

[191] C. Meehan, K. Chaudhuri, and S. Dasgupta, "A non-parametric test to detect data-copying in generative models," in *International Conference on Artificial Intelligence and Statistics*, 2020.

[192] Q. Feng, C. Guo, F. Benitez-Quiroz, and A. M. Martinez, "When do gans replicate? on the choice of dataset size," in *IEEE/CVF International Conference on Computer Vision*, 2021.

[193] N. Kandpal, E. Wallace, and C. Raffel, "Deduplicating training data mitigates privacy risks in language models," in *International Conference on Machine Learning*. PMLR, 2022, pp. 10 697–10 707.

[194] W. Wei and L. Liu, "Trustworthy distributed ai systems: Robustness, privacy, and governance," *Technical report*, 2023.

[195] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[196] A. Abusitta, E. Aïmeur, and O. A. Wahab, "Generative adversarial networks for mitigating biases in machine learning systems," *arXiv preprint arXiv:1905.09972*, 2019.

[197] F. H. K. d. S. Tanaka and C. Aranha, "Data augmentation using gans," *arXiv preprint arXiv:1904.09135*, 2019.

[198] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, "Bagan: Data augmentation with balancing gan," *arXiv preprint arXiv:1803.09655*, 2018.

[199] E. Barbierato, M. L. D. Vedova, D. Tessera, D. Toti, and N. Vanoli, "A methodology for controlling bias and fairness in synthetic data generation," *Applied Sciences*, vol. 12, no. 9, p. 4619, 2022.

[200] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan: Fairness-aware generative adversarial networks," in *IEEE International Conference on Big Data*, 2018.

[201] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, "Fairness gan: Generating datasets with fairness properties using a generative adversarial network," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 3–1, 2019.

[202] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," *Advances in neural information processing systems*, vol. 32, 2019.

[203] V. Cheng, V. M. Suriyakumar, N. Dullerud, S. Joshi, and M. Ghassemi, "Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness," in *ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[204] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359*, 2021.

[205] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[206] C. Yan, Y. Yan, Z. Wan, Z. Zhang, L. Omberg, J. Guinney, S. D. Mooney, and B. A. Malin, "A multifaceted benchmarking of synthetic electronic health record generation models," *Nature Communications*, vol. 13, no. 1, pp. 1–18, 2022.

[207] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional gans," *arXiv preprint arXiv:1706.02633*, 2017.

[208] Z. Zhao, F. Zhou, Z. Zeng, C. Guan, and S. K. Zhou, "Meta-hallucinator: Towards few-shot cross-modality cardiac image segmentation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, 2022.

[209] H. Huang, K. Mehrotra, and C. K. Mohan, "Rank-based outlier detection," *Journal of Statistical Computation and Simulation*, vol. 83, no. 3, pp. 518–531, 2013.