IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 March 2023, accepted 1 May 2023, date of publication 10 May 2023, date of current version 18 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3275134

SURVEY

Deep Generative Models for Synthetic Data: A Survey

PETER EIGENSCHINK[1], THOMAS REUTTERER[1], STEFAN VAMOSI[1], RALF VAMOSI[1,2], CHANG SUN[3], AND KLAUDIUS KALCHER[4]

[1]Department of Marketing, Vienna University of Economics and Business, 1020 Vienna, Austria
[2]High Performance Computing, Vienna University of Technology, 1040 Vienna, Austria
[3]Institute of Data Science, Maastricht University, 6200 MD Maastricht, The Netherlands
[4]Mostly AI GmbH, 1030 Vienna, Austria

Corresponding author: Thomas Reutterer (thomas.reutterer@wu.ac.at)

This work was supported by the "Information and Communication Technology (ICT) of the Future" Funding Programme of the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology.

ABSTRACT A growing interest in synthetic data has stimulated the development and advancement of a large variety of deep generative models for a wide range of applications. However, as this research has progressed, its streams have become more specialized and disconnected from one another. This is why models for synthesizing text data for natural language processing cannot readily be compared to models for synthesizing health records anymore. To mitigate this isolation, we propose a data-driven evaluation framework for generative models for synthetic sequential data, an important and challenging sub-category of synthetic data, based on five high-level criteria: *representativeness*, *novelty*, *realism*, *diversity* and *coherence* of a synthetic data-set relative to the original data-set regardless of the models' internal structures. The criteria reflect requirements different domains impose on synthetic data and allow model users to assess the quality of synthetic data across models. In a critical review of generative models for sequential data, we examine and compare the importance of each performance criterion in numerous domains. We find that realism and coherence are more important for synthetic data natural language, speech and audio processing tasks. At the same time, novelty and representativeness are more important for healthcare and mobility data. We also find that measurement of representativeness is often accomplished using statistical metrics, realism by using human judgement, and novelty using privacy tests.

INDEX TERMS Artificial intelligence, big data, deep learning, generative models, neural networks, synthetic data, privacy.

I. INTRODUCTION

In recent years, the adoption of deep generative models for synthetic data has spread to various domains. Such models can generate impressive high-quality synthetic images [96], text [54], and music [12] as well as sensory data [61], electronic health records [6], mobility trajectories [53], and financial time-series [85]. This significant progress was made possible by a facilitated accessibility of vast amounts of data and computing technologies capable of handling the data, both emerging from the continuing rise of "big data" and advances in deep learning. Models based on deep learning can handle large amounts of complex, highly correlated, high-dimensional data and generate synthetic data for many use-cases. Among others, applications of synthetic data approaches boosted progress in data augmentation [32], imputation of missing data [19], fairness in biased data-sets [87], and sharing of privacy-sensitive data-sets [91]. Today, deep generative data synthesis is a large and mature field that involves many streams of research across a wide range of domains. An overview is provided by a few review articles on deep generative data synthesis, for example, in molecular science [50], graph data [39], engineering design [71], in finance [3], and in the industrial Internet of Things area [21].

While the field has advanced in big leaps, research in the various (sub-)domains also tend to drift apart. This is particularly the case for domains that deal with processing

The associate editor coordinating the review of this manuscript and approving it for publication was Chi-Yuan Chen.

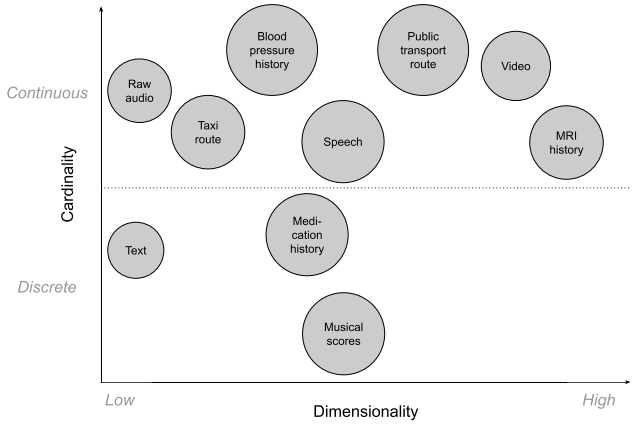

FIGURE 1. Illustration of the heterogeneity of sequential data based on cardinality and dimensionality.

of sequential data, such as geo-locations [75], shopping paths [47], text [54], video streaming [46], [78], music [12], [95], clickstreams [15], internet browsing behavior [28], financial transactions [60], electronic health records [6] or water treatment. These data stem from dynamic phenomena which are at the heart of many fields of research, but they pose significant challenges for modelers and analysts. While different types of sequential data share underlying serial correlational structures, they are also heterogeneous in terms of the dimensionality and cardinality of steps in a sequence (see Figure 1). And thus, indeed, it is difficult to compare models applied to problems in natural language processing (NLP) with models for the generation of synthetic health records. Still, some domains share common characteristics, and models applied in one field can be applied in others. Consider the recent success of so-called transformer models introduced in natural language generation (NLG) [14], [70] and now being applied in other domains to generate synthetic time-series data [45]. Because model transfer into other fields is not always straightforward, new insights can remain isolated to specific domains and fail to disseminate. The two most common barriers are (i) heterogeneity of the data and (ii) conflicting requirements for synthetic data in different use-cases. Because research in one domain can benefit from insights from other domains, a common basis for discussing generative models and guiding research is needed, especially in domains in which research to date is sparse.

To facilitate this discussion, we propose a framework for deep generative models designed to generate synthetic sequential data based on high-level evaluation criteria. This framework addresses the barriers of heterogeneity in the data and the data requirements via abstraction and allows researchers to put generative models into broader contexts. We present a critical review of publications on deep generative models in the context of synthetic sequential data and apply the proposed framework to those models.

The present article complements prior reviews in related fields, such as broad reviews on deep learning in general [67] and reviews of architectures of deep generative models

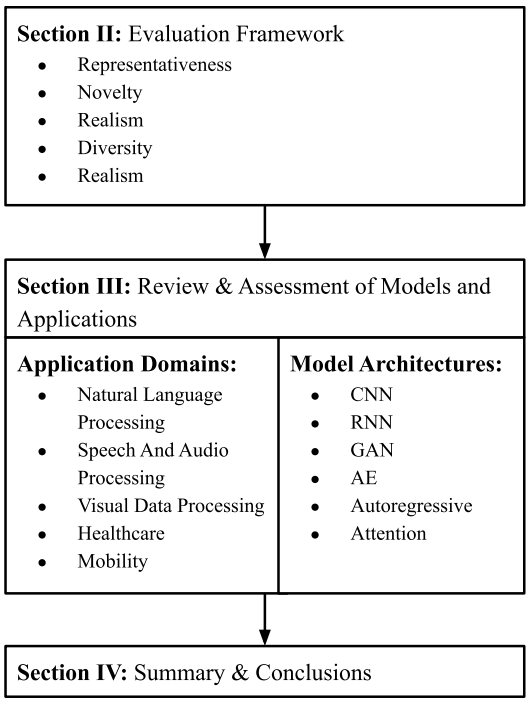

FIGURE 2. High-level structure of the article.

[35], [64], [81]. Furthermore, in addition to the above mentioned domain-specific reviews on synthetic data, a number of review articles have focused on specific model architectures, such as generative adversarial networks [44], normalizing flows [52]. However, the scope of those articles is narrow. They address specific model architectures or domains and disregard literature in other domains. The present article contributes to filling the gap between such broad methodological and narrow field-level reviews of deep generative models for synthetic data by proposing a domain- and model-agnostic framework to assess deep generative models for synthetic sequential data.

The remainder of the article is organized as summarized in Figure 2. Section II introduces the high-level evaluation framework for generative models. Then, in section III, we assess applications of synthetic data in different domains, compare strengths and weaknesses of the used models and their architectures and critically analyze them according to the proposed evaluation framework. Finally, section IV concludes the paper and provides directions for future research.

II. EVALUATION OF GENERATIVE MODELS FOR SYNTHETIC DATA

Metrics to evaluate the performance of deep generative models are as diverse as the models' objectives and specific data structures involved. General-purpose metrics, such as the commonly used negative log-likelihood (NLL), average log-likelihood (ALL) and maximum mean discrepancy (MMD) are rare and have limitations of their own [79]. Other metrics are specific particular model architectures. References [9]

and [10], for example, give a thorough overview of metrics commonly used to evaluate generative adversarial networks (GANs). Some metrics are domain-specific, such as the classifier-based inception score (IS) for synthetic images proposed by [74]. [79] reviews metrics used to evaluate generative models in the visual domain, and [39] for graph data. These metrics effectively measure progress in specific domains and compare models of a specific type, such as GANs; using them to compare different models and domains can be challenging. Even when considering only sequential data, heterogeneity is quickly apparent. The cardinality and dimensionality of the data illustrate this heterogeneity, being augmented only further by the lengths of sequences. For example, text is one-dimensional and discrete since it is made up of single words in a discrete vocabulary. Video data, on the other hand, is continuous and high-dimensional. At each step, there is a whole image that consists of many pixels, each is described by real numbers between 0 and 255. Figure 1 illustrates the heterogeneity in the landscape of sequential data by plotting the cardinality and dimensionality the data for several examples of sequential data relative to each other.

To tackle the numerous challenges associated with heterogeneous data and applications, we propose five high-level abstract criteria for evaluation of generative models: *representativeness*, *novelty*, *realism*, *diversity*, and *coherence*. The criteria are designed to compare the original data to the synthetically generated sample and can be applied to any generative model for synthetic data (see [66] for a recent example of a holdout-based framework for empirical assessment). They reflect requirements that are imposed on synthetic data in specific use-cases.

Because the criteria can be imposed on numerous types of sequential data, obtaining high scores on all five will rarely be the goal. Borji reviews qualitative and quantitative metrics for generative models in [9] and [10], but there is no one-to-one mapping between those criteria and ours. The two approaches share some aspects, reflected in what [9] defined as the desiderata of evaluation measures.

Our proposed criteria are abstract in nature but capture different concrete metrics depending on the use-case. Furthermore, some of our criteria conflict with each other. For example, we expect to see trade-offs between high representativeness of the synthetic data-set and it's novelty. Figure 3 illustrates synthetic data that have high and low scores on each criterion relative to a given data-set.

A. EVALUATION CRITERIA

1) REPRESENTATIVENESS

The representativeness of a generative model for synthetic data describes its ability to capture population-level properties of the original data. Ideally, generative models distill abstract structures from a set of training data. Consequently, the population-level properties of the synthetic and original data should be the same. For example, a data-set of face images is likely to have a certain distribution of hair colors and eye distances, and those distributions and the dependencies between the distributions (e.g., gray hair and the amount of wrinkles on a face) in the original and synthetic data should match. Depending on the type of data, there can be a multitude of ways to measure and quantify the similarity of the distributions.

Representativeness of synthetic data matters because statistical analyses and machine learning methods performed on synthetic data should result in the same statistical findings as analysis of the original data. A lack of representativeness despite all other criteria being fulfilled, indicates that the synthetic data provide a good representation only of a biased subspace of the actual data distribution and miss potentially critical information.

In many cases, representativeness is evaluated by statistical measures. Common methods are ALL, MMD and Kullback-Leibler divergence (KLD), which compare the probability distribution of the original data to the approximation of the distribution by the generative model. Recently, representativeness has also been evaluated by comparing the performance of classification models applied to the original and the synthetic data (see [17] for an example in healthcare).

2) NOVELTY

Evaluating the novelty of data from a generative model compares the original and synthetic data at an individual level. Novelty is sometimes overlooked in explicit quality evaluations, but the value of synthetic data without novelty is typically quite limited. The goal of using deep generative models usually is creation of entirely novel data-points. Novelty means that the synthetic data-points are entirely new observations of the latent distribution of the original data and should not closely resemble any original data-points.

Models that generate only novel data-points do not allow any individual-level information from the training data to leak into the synthetic data. Thus, novelty is tightly linked to privacy, and a high novelty score indicates that the ''inspiration'' behind the synthetic data-points is not identifiable at the individual level. The synthetic data records could just as well have been a holdout subset of the original data. The opposite of high novelty is a model that memorizes and exactly recreates the training data. Such synthetic data would fulfill the other four criteria (since a copy of the original training data is obviously indistinguishable in many respects from that data).

In some cases, such as in NLP, novelty of the synthetic data is irrelevant. In other cases, however, such as creative domains (e.g., music composition), the goal is to generate new creative content. For example, [23] used the average Euclidean distance of a synthetic data-point from its nearest neighbor in the original data-set to measure the novelty of synthetic music (see Section III-B for details). In other cases, such as healthcare, privacy is more important than novelty. The generative models used to produce private synthetic data

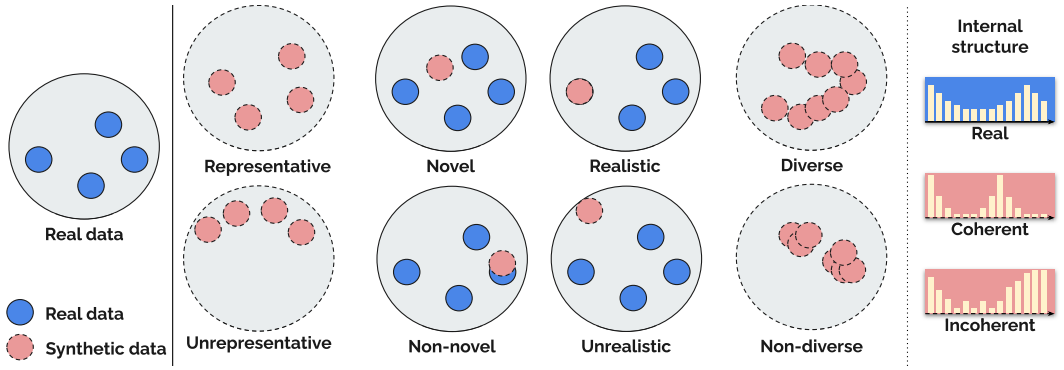

FIGURE 3. Illustration of synthetic data-sets that score high (top) and low (bottom) on the proposed criteria when compared to the original data-set on the left. *Coherence* only captures the internal structure of the data and is illustrated on the right.

must not leak any sensitive information (see Section III-D for more details).

3) REALISM

When considering an individual synthetic data-point generated by a highly realistic model on its own, it is difficult to know whether it is synthetic or original. Realism is similar to representativeness of the data, but at the individual subject level. A synthetic data-set can match all the statistics of the original data and still be unrealistic when individual data-points share characteristics that make them easily identifiable as synthetic. Consider a representative but unrealistic example obtained using a GAN trained on random cat images from the internet. Synthetic cat images can contain captions reminiscent of online memes that look plausible from a distance but actually consist solely of abstract symbols having shapes similar to letters.

Realism has been addressed in many publications in a variety of ways. The most common method is judgement of realism of the synthetic data by humans, either qualitatively (e.g. [59], [62]) or using empirical evaluations (e.g. [7], [63], [83]). Evaluation studies present individuals with the original data-point and the synthetic data-point and ask them to choose which is the most. In some publications, participants in the evaluation studies were restricted to experts (e.g., medical experts in [7] and [20] and music experts [11], [63]). In some cases, realism is quantitatively evaluated using objective measures. These evaluations are usually domain-specific and use metrics such as IS [73], [80] and the evaluation of synthetic music against theoretical music rules [48].

4) DIVERSITY

While representativeness, novelty and realism capture similarities between the original and the synthetic data, diversity measures similarities between each synthetic data-point and the whole synthetic data-set at the individual level. Therefore, models that score well on diversity generate unique data-points even when data-sets are large. Models that generate the same individual points over and over, such as some early versions of GANs, obviously lack diversity. For instance, generators sometimes create a single image that the discriminator cannot distinguish from an original image. Generating only that image is a local optimum, and the resulting effect is called mode collapse.

Many publications have not addressed the diversity of the generative models' synthetic data. In most cases, it is important that models achieve at least some diversity, and some models can generate only a small number of different samples (e.g., the aforementioned GAN with mode collapse).

There are several ways to measure the diversity of a model. Donahue et al. [23] used the average Euclidean distance of synthetic data-points to their respective nearest neighbors to evaluate the diversity of their model (see Section III-B). Others have used metrics based on classifiers. For example, to measure the diversity of their video-generation models, the authors of [73] and [80] used the IS [74] (see Section III-C). In other cases diversity has been captured only by subjective qualitative evaluations by humans.

5) COHERENCE

Unlike the first for criteria, which are based on the structure of the synthetic data at an individual or population level, coherence captures the internal structure of single synthetic data-points, specifically their consistency. Coherence is particularly relevant for sequential data, that reflect sequential orders of events and for data such as images. Coherence requirements depend on the use-case and the original data and, thus, can differ in terms of coarseness. Music, for example, should sound smooth and natural note-by-note and measure-by-measure, but also should stay within a certain genre overall. In images when multiple objects cast shadows from a single light source, the shadows must be coherent in terms of the direction in which they point and their length. While some incoherence in the data can lead to greater novelty or diversity, too much results in unrealistic data. A music sample that frequently changes its genres would certainly sound creative but would also sound unrealistic.

Some studies have measured the coherence of synthetic data implicitly when evaluating its realism. In the study

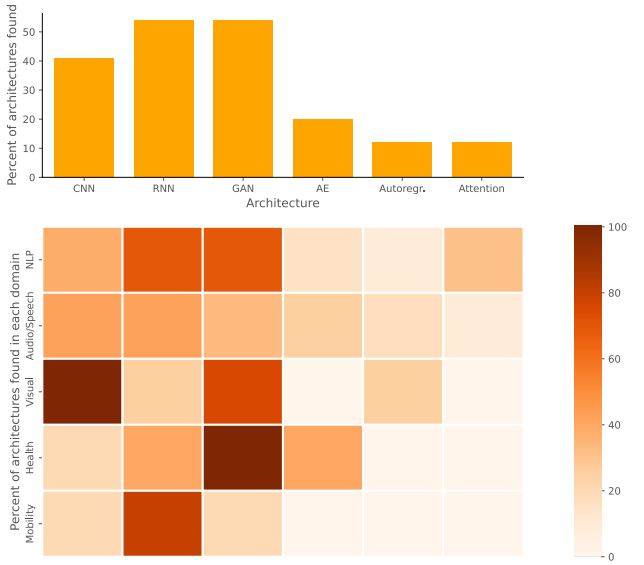

FIGURE 4. Popularity of various architectural elements in deep learning models used to generate synthetic sequential data. The top graph shows the percentage of each architecture found to be the basis for models used in the all reviewed studies. The heatmap in the bottom figure shows the prevalence in percent of each architectural element in five domains. The data underlying both graphics is included in the Web Appendix.

conducted by Bretran et al. [11] experts evaluated the naturalness of transitions in synthetic music. In many cases, however, domain-specific objective metrics have been used to judge coherence. In [80] coherence was computed using the average content distance between frames in synthetic videos (see Section III-C).

III. ASSESSMENT OF APPLICATIONS

We next review applications of deep generative models to generate synthetic sequential data in a variety of domains. We critically analyze the contributions to this fast-growing literature, evaluate them using our proposed criteria, and demonstrate that the criteria individually are not equally relevant in all domains and are not measured the same way. Each subsection discusses the applications in their focal domain and summarizes a few representative contributions in terms of the proposed assessment criteria. See Tables 1 to 5 for overviews of representative contributions in particular domains. Additionally, we analyze the architectures of the models used in the selected publications. Figure 4 summarizes the prevalence of architectural elements used in the reviewed articles in different domains. An excellent overview of deep neural network architectures is provided by [36].

A. NATURAL LANGUAGE PROCESSING

NLP is a broad field devoted to computers interacting with human language. Common tasks in NLP include language modeling, text translation [77], human-machine dialog generation, and natural language generation [26], [33]. Thanks to the widespread adoption of machine learning and deep neural networks in recent years, the research community has made significant progress in accomplishing these tasks. Today, highly capable language models can generate texts almost indistinguishable from human-generated text.

TABLE 1. Excerpts of studies of generative models for natural language processing and metrics used for evaluation.

Study	Represen-tativeness	Novelty	Realism	Diversity	Coherence
[40]	Perplexity	Qual.	Human eval.	Qual.	Human eval.
[76]	NLL, Perplexity		Human eval.	Self-BLEU, unique n-grams, 2-gram entropy	Human eval.
[30]	Perplexity		Human eval.	Unique 2,3,4-grams	Human eval.

Language data-sets are comprised of text, which can come in many different flavors - news articles, product reviews, medical diagnoses, and music lyrics. However, all text can be represented as a combination of tokens from a discrete vocabulary. The tokens are the most basic components of text, commonly single words complemented with punctuation.

Sentences, paragraphs, and longer texts are then merely sequences of such tokens. However, the sequences must obey certain grammatical, semantic, and logical rules. Moreover, since sentences are not just loosely strung together, later sentences and words in the text can be highly dependent on words that appeared multiple sentences before. For example, a character in a short story that disappears in the beginning, can reappear paragraphs later. The rules and contextual dependencies of a text pose significant challenges to language models and to the generation of synthetic text. A model must be capable of capturing the proper setting of various linguistic features such as syntax, semantics, pragmatics, and morphology. Otherwise, the resulting text can quickly become incoherent or unrealistic.

A fascinating language model is provided by [40]. It was inspired by how humans create complex texts, which rarely arise from scratch in a single pass. Instead, humans rather create initial drafts and revise the drafts incrementally. [40] adopted this idea in their *neural editor* model by sampling a prototype sentence from the training corpus, combining it with a random parameter for editing the sentence, and generating a modified, new sentence. Their edit parameter can lead to changes such as altered wording, shorter or longer sentences, and change from active to passive voice. Architecturally, the model is based on a VAE (variational autoencoder; [51], [72]) with an attention-based LSTM (long short-term memory; [34], [43]) encoder and an LSTM decoder. The prototype sentence and the edit parameters are

randomly sampled and then used to transform the sentence in a sequence-to-sequence fashion.

A metric commonly used to evaluate the quality of language models is the perplexity [49], which captures how ''surprised'' a language model is to see the words in the original training corpus in terms of probabilities it assigns to each word. Looking at language models as generative, perplexity measures the representativeness of the generative model. Reference [40] evaluated their neural editor using a data-set of restaurant reviews from Yelp and a more-general text data-set. They found that the synthetic texts were representative when measured by perplexity in both cases. Though they were able to generate novel sentences that were significantly different from the prototype sentences, each synthetic sentence still originated from a single prototype sentence and thus was somewhat close to the prototype.

The edit parameter can be used to perform similar edits on multiple sentences or to smoothly vary the degree to which editing is performed on a single sentence. Reference [40] used these properties to generate a variety of sentences, qualitatively suggesting that the generation of diverse data-sets is possible. Individuals deemed the synthetic sentences realistic and coherent according to their ratings of overall quality, grammaticality, and plausibility.

Though the neural editor effectively generates synthetic sentences, creating longer text samples composed of several coherent sentences that are non-repetitive, grammatically correct, and non-contradictory remains challenging. Models capable of that task require a greater capacity to capture the long-term dependencies in such texts. Shen et al. proposed such a model in [76] and, given the inherent hierarchical paragraph structure of longer texts, they chose a hierarchical VAE architecture. The encoder network consists of one low-level CNN (convolutional neural network) that maps each sentence to a latent variable and one high-level CNN that maps all the latent variables for each sentence into one latent variable for the entire text input. On the decoding side, two hierarchical LSTM networks operate the other way around at the sentence and on word level. The decoder obtains a latent variable for a text and transforms it via the sentence-level LSTM into latent sentence variables. The sentence-level latent variables are then passed down to the word-level LSTM, which generates the words for the synthetic sentences. The model can output longer synthetic paragraphs by putting all the words together into sentences and the sentences into paragraphs. Passing the latent variables down the LSTM hierarchy allows the decoder to capture relatively coarse characteristics of text and sentences, such as the topic and sentiment.

Shen et al. [76] evaluated their model using Yelp reviews and abstracts from arXiv papers and found that their multilevel-VAE (ml-VAE) model improved the representativeness of the output relative to a flat VAE model (the baseline). They evaluated representativeness by measuring the perplexity of the language model and calculating the corpus-level bilingual evaluation understudy (BLEU) score of the output. The BLEU score was originally developed for in-text translation and has proven to be a good metric for measuring translation quality that correlates well with human evaluations. It measures similarities between the generated text and a set of references by comparing their n-grams: n consecutive words/tokens in a text. When the set of references is the whole synthetic data-set, the BLEU score is called self-BLEU.

The average BLEU score of the ml-VAE model obtained by comparing generated text to the training corpus indicated that representativeness improved relative to the baseline. The authors also reported an acceptable diversity score. Diversity was especially important to them because VAEs used for NLG often suffer from mode collapse. They evaluated diversity by calculating self-BLEU scores, the percentage of unique n-grams, and the 2-gram entropy of a set of synthetic texts. They further evaluated the coherence and realism of the synthetic text by asking individuals to compare text generated by the baseline model to the ml-VAE synthetic text and choose the one that seemed most ''real'' to them. Individuals rated the texts' fluency, grammar, and consistency to measure their coherence. These human evaluations also showed that, in terms of realism and coherence, the ml-VAE yielded results that were superior to the results of the baseline model and acceptable when compared to human-generated text.

Likelihood-based models such as VAEs have their critics, who suggest that the models are well suited to optimizing perplexity and representativeness but lack the ability to generate realistic, coherent high-quality samples. In [30], Fedus et al. attempted to generate higher-quality samples using a GAN-based model that incorporated LSTM encoder-decoder networks in the generator and discriminator. To improve overall training, they masked the sentences by blanking words and asking the generator to predict the missing words based on the rest of the sentence. In that case, the networks knew the entire sentence context; most other models condition a word solely on the preceding words in the sentence. They found that their hybrid GAN model improved perplexity and thus representativeness relative to a likelihood-based baseline model. Still, they claim that low perplexity alone does not indicate high-quality synthetic text, their primary focus. Their human evaluations also showed that the hybrid GAN model produced more realistic samples than the baseline model in most cases. Distinguishing between the synthetic and human-generated texts seems relatively easy for the participants. Since mode collapse is a common issue in GANs, the authors also took a narrow look at the diversity of the synthetic results. They evaluated the percentages of unique 2-, 3-, and 4-grams. They found some mode collapse, indicating that the text generated by their model lacked diversity relative to the text generated by the baseline model. In addition, the synthetic sentences sometimes lacked coherence because they lost the global context. However, the authors expected to be able to improve coherence by increasing the capacity of the model.

TABLE 2. Excerpt of studies of synthetic speech and audio data and metrics used to evaluate the output.

Study	Representativeness	Novelty	Realism	Diversity	Coherence
[62]		Qual.	Qual.		Qual.
[22]			Human eval.		Qual.
[58]	ALL				Qual.
[24]	Music-theoretical measures		Human eval.		Tonal distance, Human eval.
[23]		Avg. Eucl. distance to orig. set	IS, Human eval.	Avg. Eucl. distance to synth. set	

NLP is a heavily researched domain that has produced a wide range of applications. The primary concern of most studies of generative models is the generation of representative and realistic synthetic texts with realism implicitly used as a metric for coherence in most cases. Diversity is also investigated in detail when models are prone to mode collapse. However, novelty is rarely addressed and could be of interest primarily in privacy-sensitive cases such as medical patients' chief complaints [55]. Interestingly, for most of our high-level evaluation criteria (Section II) some metrics have been established for NLP. NLL, BLEU, and perplexity are often used to measure representativeness. Realism and coherence are mainly evaluated together as parts of human evaluation studies, with participants choosing between synthetic and human-generated text based on various properties. Finally, to assess the diversity of synthetic results, studies used either self-BLEU or statistics such as the percentage of unique n-grams. Metrics employed in [30] and [76] to evaluate diversity stood out, especially when compared to the qualitative diversity evaluation used in [40].

B. SPEECH AND AUDIO PROCESSING

Generation of audio data has a long history. It originated in several quite different domains and relied on completely different theories. Most notable the generation of synthetic music and speech. Both ultimately make data audible by converting it to sound. As different as these origins and the rules used to generate synthetic sound are, both are specific types of digital audio data that eventually yield the same result.

Following the success of deep neural networks in generating content such as images, video, and text and the availability of vast quantities of audio data, researchers began to apply the techniques to audio-synthesis. The resulting models learned either from raw audio signals or from intermediary representations such as musical scores and linguistic speech parameters. The models can grasp the underlying structure of the data to create realistic-sounding synthetic audio data.

The most general representation of sound is the amplitude of sound waves over time sampled at a constant rate (i.e., raw audio). Consequently, the sound signal is continuous and one-dimensional. Still, because of the high frequencies of natural sounds, the sequences are long and complex. Typical sampling rates are at least 16kHz, resulting in signals with thousands of steps per second.

Models designed to work with raw audio generally are the most adaptable. Unlike models that use intermediary representations, their results do not have to go through one or more conversion steps before becoming audible [63]. The drawback of raw audio is the need for high-capacity models that can learn certain rules on their own instead of having to encode the rules in specific representation. To generate realistic speech, for example, models have to learn how intonation affects meaning to generate realistic speech. Speech parameters already encode intonation rules to some extent.

Deep learning models can leverage some aspects of audio data by choosing appropriate representations. But, as previously mentioned, there are drawbacks. Musical scores, for example, require multiple conversion steps to become audible. Additionally, representations can abstract away relevant nuances of music and speech. Timing and volume, for example, can be important when generating synthetic music, but often cannot be represented accurately in musical scores.

When generating music and speech, use of raw audio signals in generative models is in the minority. Applications such as WaveNet [62] show that raw audio models can succeed in multiple domains by leveraging the flexibility of deep learning models. WaveNet is an autoregressive model that predicts one step of a sequence at a time conditioned on previous steps. Multiple layers of causal convolutions incorporate causality into the network. These are one-dimensional convolutions that depend only on present and past time steps. A key problem of networks involving causal convolutions is that, when the convolutions depend on the present and previous time steps, the network has to be quite deep to capture long-term dependencies. WaveNet [62] overcomes this obstacle by dilating the convolutions in each layer. Therefore, instead of using the output of the preceding time step as input, WaveNet skips multiple time steps.

The WaveNet [62] model has been evaluated in numerous experiments. Most important for this review is the unconditional generation of polyphonic single-voice piano music and of speech for a single speaker. WaveNet made a significant leap forward in the ability to generate of synthetic audio data by adopting deep learning models and still serves as a baseline for evaluation of new models. Reference [62]

addressed novelty, realism, and coherence of the sample output of WaveNet only qualitatively and did not address representativeness or diversity. Qualitatively, the synthetic music was rated as harmonious and aesthetically pleasing. Their synthetic speech samples consisted of non-existent words that resembled actual words and were spoken with realistic intonations. The authors argue that conditioning on information such as a speaker's ID for speech and genre for music, yields better results. Additionally, because the input size was limited, WaveNet's synthetic outputs lacked long-term coherence and synthetic music samples sometimes changed genre and volume from one second to another.

The structure of raw audio makes generation of long coherent audio signals challenging. The signal at one time step can depend on the values of neighboring time steps and on the values of thousands of preceding time steps. WaveNet lacks this long-term coherence but yields short audio samples of good quality. To overcome this limitation, WaveNet has been incorporated into higher-level architectures (e.g., [22], [58]). Dieleman et al. [22] transformed raw audio signals into a more-abstract, higher-level representation and train WaveNet on the representation.

In their SampleRNN model, Mehri et al. [58] addressed the problem of coherence by hierarchically stacking networks that operated at different timescales. The lowest layer of the SampleRNN is a WaveNet network operating on the raw audio signal. Higher layers operate on coarser timescales by collating multiple time steps of the signal into the state of a recurrent neural network (RNN). As a result, the higher layers can capture long-term dependencies and pass that information down the network hierarchy, allowing WaveNet to obtain aggregated dependency information from numerous preceding time steps. The SampleRNN was evaluated on speech data, human sounds, and music data. The authors reported that it generated more-representative synthetic audio samples than a simple WaveNet, based on the NLL of the synthetic samples. Also, participants who evaluated the results of SampleRNN in an empirical study perceived the synthetic output more realistic than the output of WaveNet.

Donahue et al. [23] also worked with raw audio but applied an interesting approach. They transferred the DCGAN network [69], a model prominently known for its success in image synthesis, to audio generation. They created two models: WaveGAN for raw audio and SpecGAN for spectrograms of sound data. Both are GAN models with a convolutional generator and discriminator and a structure similar to DCGAN. However, since as raw audio is one-dimensional and images are two-dimensional, the convolutions are flattened. Two-dimensional filters sized 5×5 in DCGAN become one-dimensional filters of length 25 in WaveGAN and WaveGAN's output is a raw audio sample of length 16,384 instead of an image of size 128×128. SpecGAN, on the other hand, operates on the two-dimensional spectrograms of raw audio data. The raw audio samples are first transformed into intensity distributions of different frequencies at each timestep, creating spectrograms. SpecGAN then generates synthetic two-dimensional spectrograms that are inverted back to raw audio to obtain audible sound.

With a sampling rate of 16kHz WaveGAN and SpecGAN generate synthetic audio samples that have a duration of about one second. The models are applied to data-sets with similarly short sounds, such as intonations of the numbers zero through nine in speech, short drum and piano sounds, and bird vocalizations. The authors thoroughly evaluated the two models using IS, nearest-neighbor comparisons, and human judgement. Donahue et al. [23] used the IS, which was originally developed to evaluate of synthetic images, to determine the realism and diversity of their synthetic sounds. To evaluate diversity, they measured the mean Euclidean distance between a synthetic sound and its nearest neighbors. Novelty was determined by the mean Euclidean distance between a synthetic sound and nearest neighbors in the original data-set. Additionally, study participants evaluated the quality, diversity, and realism of the synthetic vocalizations of the numbers. The authors report better results in terms of novelty, diversity and realism than achieved using SampleRNN [58] and WaveNet [62].

The limitations associated with using raw audio data in terms of sequence length make use of higher-level representations of sound such as musical scores and the Musical Instrument Digital Interface (MIDI) standard for music beneficial in some scenarios. Higher-level representations can encode important information but abstract away some aspects of raw audio. Less capacity is needed for these models, but the representations cannot be made audible directly. Often, some interpretation is to musicians or to computer programs. Additionally, abstraction reduces sequence length while usually increasing dimensionality.

Piano rolls are an example of a higher-level representation of music. They were inspired by the rolls used in automated pianos that triggered playing of a note for a certain duration. Similarly, piano roll representations encode whether a note —or multiple notes in polyphonic cases —is played in a particular time step of a song. The duration of the time steps is constant for a single piano roll and across a data-set. The duration is much longer than in raw audio data so piano rolls can encode multiple seconds of melodies using shorter sequences and thus make it easier to capture intra-sequence dependencies. However, piano rolls slightly increase dimensionality because each note in a track is encoded instead of amplitudes of sound waves. There are several other representations used for music, and the literature on deep learning models for generating symbolic music is extensive [13], [31].

In [24], Dong et al. described a model designed to generate multi-voice polyphonic rock music called MuseGAN operating on piano rolls. Multi-voice music consists of multiple tracks for the instruments (e.g., piano, guitar, and bass). Each track is represented by a piano roll. The challenge in modelling multi-voice polyphonic piano rolls

is capturing the intra-dependencies of notes in a track and the inter-dependencies of notes played in different tracks.

MuseGAN [24] uses the intra- and inter-dependencies of tracks to compose synthetic music, further separating the dependencies into time-dependent and time-independent parts. The network is a GAN that uses a generator partly inspired by generative video models [73], [80], [83] (see Section III-C for details on these models). The synthetic music is sampled from the generator by track. Each track is generated from two random numbers representing all tracks and two random numbers representing individual tracks encoding time-dependent and time-independent intra- and inter-track dependencies. The track-generator captures dependencies in time and between notes played using a CNN structure. Similarly, the MuseGAN discriminator is a CNN that judges whether a melody is real or synthetic based on the structure of the notes played in a single track over time and in multiple tracks at the same time.

Dong et al. [24] leveraged symbolic representation to reduce the complexity of the problem and to assess the quality of the generated music samples. To evaluate the representativeness and coherence of the meusic they compared the training data and synthetic data based on music-theoretical measures. The authors computed the ratio of bars in which no notes were played, the number of pitch classes used in a bar, and the ratio of notes lasting longer than a 32nd note to evaluate representativeness. The model captured drum patterns observed in the training data fairly well, but the synthetic melodies were more fragmented and used a larger number of pitch classes than the original melodies, indicating noise in the synthetic data. The tonal distance [41] between tracks in the generated samples generally showed a strong harmonic relation, indicating strong coherence. In addition to these objective measures, the authors evaluated the synthetic samples' *harmonicity*, *rhythmicity*, *musical structure*, and *coherence* based on responses by study participants, who also gave the samples *overall ratings* that measured coherence and realism as defined in our proposed framework. The study participants rated the samples as 2.3 to 3.5 on a 1–5 scale; they did not compare the samples to baseline samples from other models or to the original music.

Speech also can be generated using representations. One of the most studied paradigms is statistical parametric speech synthesis (SPSS) [94], which uses linguistic features of speech such as phonemes, cadence, and word frequency to synthesize spoken words. Considerable research has been conducted on SPSS, but unconditional generation of synthetic speech data-sets is uncommon. Common tasks are text-to-speech, voice conversion, and vocoding (making speech parameters audible). In all three, cases speech is being generated from an input (text, speech fragments, speech parameters). Though these tasks fall outside the scope of our literature review it is important to note that deep learning based models for speech data are emerging (see, e.g., [68], [84], [93]).

TABLE 3. Excerpt of models for synthetic videos and metrics used to evaluate them.

Study	Representativeness	Novelty	Realism	Diversity	Coherence
[83]	Qual.		Human eval.		
[73]			IS	IS	
[80]		Qual.	IS, Human eval.	IS	ACD

Evaluation of synthetic audio data poses a challenge that cannot adequately be addressed in general: the significance of our proposed criteria and validity of metrics used to measure the criteria vary with the type of audio (e.g., speech versus music). Rhythms and harmony are highly relevant for music but only somewhat relevant for speech. Reasonable evaluations are often based on domain-knowledge. In the case of music, the relevant domain is music theory, for which metrics such as the ratio of pauses, fragmentation of a sample, and the tonal distance, as used in [24], are reasonable. For a review of objective metrics for evaluating synthetic music, see [92]. For all kinds of audio and for music and speech in particular, subjective evaluations of realism and coherence by humans are a significant part of evaluations. Reference [24] makes use of domain-specific metrics for representativeness and coherence and, thus, their assessment is better suited for synthetic music than, for example, the qualitative evaluation used in [22], [58], and [62]. Additionally, Donahue et al. introduced metrics for the evaluation of novelty and diversity of synthetic sound in [23], which is hardly analyzed in any of the other reviewed studies.

C. VISUAL DATA PROCESSING

Today, thanks to the prevalence of smartphones, images and videos are produced and consumed en masse. Access to such a vast amount of data has led to dramatic advances in processing and classification of existing images and in models to generate synthetic ones (see, e.g., [69], [96]). Since videos are merely sequences of images, the ability to generate synthetic videos also has advanced. The ongoing challenge is capturing a smooth dynamic motion in the transitions between images.

Models based on CNNs and GANs have been highly successful in generating images. Consequently, many successful generative models for synthetic videos have been based on them [73], [80], [83]. The primary challenge in designing such models is incorporation of the temporal dimension with videos' two spatial dimensions of the video.

The VGAN model proposed in [83] tackles this challenge by decomposing the dynamic foreground from the static

background, reducing the complexity of the problem. The dynamic foreground is captured by a three-dimensional spatio-temporal CNN, and the static background can be captured by a two-dimensional spatial CNN. Both CNNs are incorporated into the generator of a GAN that is then optimized against a three-dimensional spatio-temporal discriminator that judges the realism of the scene and the motion.

The VGAN model has been applied to small short videos of 64 $\times$ 64 pixels with 32 frames and duration of around one second from Flickr in different categories collected such as beaches, golf courses, and train stations. The authors of [83] assessed the representativeness of the resulting synthetic videos qualitatively and reported generally correct motion patterns for scenes in the various categories. Synthetic videos of beaches contained crashing waves and synthetic videos of trains contained train tracks and train cars with windows moving by quickly, as one would expect. The generated scenes were sharp overall, but individual objects such as people in the synthetic beach scenes tended to lack resolution. The realism of the resulting videos was evaluated by participants in an empirical study who were asked to view the synthetic and original videos and choose which seemed most realistic. Though the participants overwhelmingly chose the original videos, the synthetic scenes were chosen in 18% of the comparisons.

The VGAN architecture [83] is optimized for videos with static backgrounds. Saito et al. [73] relaxed this restriction in their TGAN model by decoupling the temporal dimension from the spatial dimensions. First, a one-dimensional temporal generator produces a sequence of temporal codes that are then mapped one-by-one to an image by a two-dimensional image generator. The discriminator, a three-dimensional spatio-temporal CNN, then distinguishes real videos from synthetic ones. According to the IS, the synthetic videos generated by TGAN are more diverse and realistic than those generated by VGAN.

Tulyakov et al. [80] argued that the straightforward decomposition of a video into temporal and spatial dimensions, as done in TGAN, unnecessarily increases the complexity of the problem by ignoring similar motion patterns. In [80], they proposed a decomposition of the content of a video and the motion therein, which they incorporated into a generative model called MoCoGAN. Consider the various facial expressions presented by a person in a video. In such a video, the person's face is the content and performance of an expression is the motion. This disentanglement allows the model to generate videos with the same content but different motions and vice versa —that is, videos of a person performing different facial expressions.

MoCoGAN incorporates this decomposition in the latent space of the generator. The input to the generator is split into a content variable and a sequence of motion codes. An RNN generates the motion codes and connects them to subsequent codes to ensure a coherent motion. Then, given the fixed content variable for all frames and a motion code, each frame in the video is synthesized by a two-dimensional CNN image generator. Similarly, the discriminator judges the realism of the content and motion of a video separately using a two-dimensional CNN for the content and a three-dimensional spatio-temporal CNN for the motion.

The authors evaluated the MoCoGAN's performance synthesizing small short videos of various scenes, including tai-chi movements and facial expressions. They qualitatively assessed the ability of the model to decompose content from motion by fixing a person as the content and generating videos of that person performing different motions. The results demonstrated MoCoGAN's ability to generate novel content by adjusting the input variables of the generator. They found that the synthetic videos generated by MoCoGAN were more diverse and realistic than synthetic videos generated by VGAN [83] and TGAN [73] based on the IS. Additionally, participants in an empirical study viewed the videos generated by MoCoGAN as more realistic than videos generated using VGAN [83] and TGAN [73]. Tulyakov et al. [80] also quantitatively evaluated the coherence of synthetic videos of facial expressions using the classifier-based average content distance (ACD), which quantifies the difference between two frames in a video in terms of content. OpenFace [5] is applied to each frame of a video presenting a facial expression to extract facial features that identify the person. Small differences (distances) in the features between frames indicate that the same person is displayed throughout the video and, therefore, a small ACD. The MoCoGAN obtained higher coherence scores than the VGAN [83] and TGAN [73] videos.

When generating synthetic videos, many concepts from image generation carry over. We see this in the prevalence of CNN and GAN models and in the metrics used to evaluate synthetic videos. Specifically, the IS is often used to measure realism and diversity of synthetic videos and [80] uses ACD to measure coherence; both rely on image classifiers. When evaluating reaslism, human studies are heavily used in addition to IS and ACD. Human evaluations of realism also capture coherence to some extent. The representativeness and novelty of synthetic videos are rarely evaluated explicitly. Altogether, the results so far are promising for synthetic videos that are short and relatively low resolution. The large number of dimensions associated with high-quality videos combined with the large number of frames needed even for short videos continue to thwart efforts to synthesize more complex videos. The introduction of the ACD as an objective measure for coherence in [80] is particularly noteworthy, since other studies such as [73] and [83] evaluate coherence only as a part of the realism assessment.

D. HEALTHCARE

Generative models for synthetic medical data have gained attention in recent years. The sensitivity of medical data and strict access restrictions make sharing of original medical data from patients extremely challenging [7].

TABLE 4. Excerpt of models to generate synthetic medical data and metrics used to evaluate them.

Study	Representativeness	Novelty	Realism	Diversity	Coherence
[27]	MMD, TSTR	NN distance	TRTS	Qual.	
[20]	Qual.	Disclosure risk	Human eval.	Qual.	Qual.
[6]	K–S test, Dim.-wise stats.	Qual.	ML pred., ARM	Qual.	ARM

A promising solution is to use generated synthetic data instead. Synthetic medical data can be shared and published for secondary analyses since the privacy of patients is guaranteed.

Data from intensive care units (ICUs), where patients with severe and life-threatening conditions first receive treatment, are especially valuable for clinical analysis [16]. The data can include real-valued monitoring information, such as measured oxygen saturation, heart rate, and respiratory rate.

Esteban et al. [27] generated such synthetic medical data based on information collected from the first four hours of patients' stays in an ICU. They employed an LSTM as the generator in a GAN and another LSTM as the discriminator of real and synthetic data sequences. They evaluated representativeness of the generated data using MMD and by training a classifier model on the synthetic data-set and testing it on a real holdout data-set (train on synthetic, test on real (TSTR)). They evaluated realism by training a classifier model on the real data-set and testing it on the synthetic data-set (train on real, test on synthetic (TRTS)). In both cases, the classifier models achieved results comparable to models trained and tested solely on original data.

Novelty is especially important in privacy contexts; that is, it must be impossible to reconstruct the original data-points from the synthetic ones. Overall, [27] found that the synthetic data-points were not close to original single data-points based on the evaluation of the distances between the synthetic data-points and their real nearest neighbors. Their qualitative exploration of the latent space —conducted by interpolating between generated points —also showed that the model yielded diverse results. To account for the importance of privacy, they adapted the training of the original model to incorporate differential privacy [1], [25]. Under the stricter privacy conditions, they reported that the synthetic data were highly representative and slightly less realistic.

The real-valued time-series data used by Esteban et al. [27] are important in healthcare but are one of many types of electronic health records (EHR). EHR data has been the main focus of recent studies [4] and turns out to be quite diverse. EHRs include patients' demographic information, diagnoses, laboratory test results, medication history, clinical notes, and medical images, and other medical records [86] and disclose discrete-valued codes for diagnoses, medications, and procedures.

Choi et al. [20] studied synthetic sequences of discrete-valued multi-label EHR data containing information on diagnoses and treatments. The sequences in the data were long and high-dimensional, thus presenting significant challenges for generation of synthetic data. The authors addressed these challenges by combining an autoencoder (AE) and a GAN in their generative model, medGAN. The AE was used to reduce the complexity of the output data of the generator, which learns salient features of the samples by projecting them to a lower dimensional space and then projecting them back to the original space [36], [82]. Thus, medGAN generates synthetic data in the lower dimensional space. Then, the pre-trained decoder converts the generated output to synthetic EHR data in the original space.

The authors evaluated medGAN and found that it outperformed several generative models, including random noise, independent sampling [20], stacked restricted Boltzmann machines [42] and VAEs. Representativeness and diversity are only evaluated qualitatively, but the authors argued that significant improvements were accomplished by applying the minibatch averaging method [20] to reduce overfitting and mode collapse. Novelty was evaluated by conducting two privacy risk evaluations. One measured the risk of disclosure of personally identifiable information and the other measured the risk of disclosure of personal sensitive medical data. The evaluations determined that medGAN can generate novel private synthetic data that reveal little information to potential attackers rather than simply reproducing the training samples. Overall, medGan's synthetic data were reported to be realistic, but qualitative evaluation by a single doctor is not entirely convincing.

Since introduction of medGAN, other researchers have extended it in different directions. Two that have outperformed medGAN in all experiments were proposed by Baowaly et al. [6]. The medWGAN model combines medGAN with the Wasserstein GAN model, which uses a gradient penalty [2], [38] to minimize divergences in Wasserstein distances. The medBGAN (medical boundary-seeking GAN) model trains the generator to obtain a distribution of samples located on the decision boundary of the discriminator. To evaluate the models' representativeness, the authors conducted the Kolmogorov-Smirnov (K-S) test and compare the dimension-wise probabilities and averages of the real and synthetic data. Realism was evaluated by comparing predictions made by machine learning models for the real and synthetic data. Association rule mining (ARM) is often used to identify associations and patterns in clinical concepts in EHR data [89] and was used by [6] to evaluate realism and coherence. Another extension of medGAN for generating real-valued time-series data, has been proposed by Yahi et al. [90].

TABLE 5. Excerpt of models for generating mobility data and applied metrics.

Study	Representativeness	Novelty	Realism	Diversity	Coherence
[65]	Absol. semantics, Marg. distributions				Rel. semantics
[53]	Absol. semantics, MMD	Location hiding			temporal dep. decay
[56]	Counting stat.				

Medical text and images also have attracted attention. Medical text consists of clinical notes and patients' chief complaints, which share characteristics of other types of text data (see Section III-A) but typically are short and are composed of a limited number of words from medical vocabularies. Lee [55] applied an encoder-decoder model to generate synthetic natural-language chief complaints using EHR data from around 5.5 million records of emergency department visits. Guan et al. [37] proposed a GAN model to generate Chinese EHR text data. Both models use demographic and disease features as inputs and generate corresponding EHR text data. However, they are conditional models that fall outside the scope of this survey.

In healthcare, synthetic EHR data is primarily used to protect patients' privacy while enabling data sharing and secondary data analyses. Thus, most studies in the field are mainly concerned with novelty, representativeness, and realism. Novelty is particularly important to privacy-protection and, thus, is often evaluated using privacy tests. Tests for representativesness and realism in EHRs are not necessarily domain-specific; TSTR and TRTS have most often been used to evaluate those criteria. [27] used particularly interesting evaluation procedures, compared to other reviewed studies such as [6] and [20]. They compared the original data with the synthetic data and evaluated representativeness using TSTR, realism using TRTS and novelty via the NN distance.

E. MOBILITY

Everyday, massive quantities of data on human mobility are collected. Mobile devices such as smartphones are equipped with GPS functionality and transportation systems (car sharing, logistics, public transports) usually incorporate automatic tracking. Mobility data are used in a wide range of tasks, including urban traffic predictions [57], shared mobility services [18], marketing services [47], and transportation of people and goods [29]. However, the risk of re-identification of individuals makes sharing of such data highly sensitive. The relevance of this risk has been demonstrated even for aggregated mobility data [88]. Synthetic mobility trajectories do not present this risk and thus enable sharing, by either obfuscating the original path data or generating completely synthetic trajectories that cannot be related to individuals.

Ouyang et al. [65] studied generation of synthetic realistic human location trajectories for privacy-sensitive secondary data analyses. Usually, mobility trajectories are represented as sequences of continuous coordinates (x, y) consisting of a longitudinal and latitudinal component over time t. Ouyang et al. converted this time-major representation into a location-major representation in the form of maps corresponding to times of stays at each coordinate (x, y). The maps were then fed into a GAN consisting of a deconvolutional generator and a convolutional discriminator.

The authors evaluate the model results primarily in terms of representativeness and coherence. Representativeness was evaluated by comparing geographical statistics describing the real data with the same statistics for the synthetic data. They compare the marginal probabilities of visiting a certain location at a certain time and of remaining there for a certain duration using Jensen-Shannon divergence (JSD).

The so called *semantics* of the trajectories play a key role in producing representativeness and coherence. The semantics give a trajectory intrinsic meaning, which can be difficult for generative models to capture. The path "home-bus-work-bus-home", for example, intuitively makes sense whereas "airport-home-work-train" does not make sense and semantically is unlikely to be true. Ouyang et al. further distinguished between absolute and relative semantics. Absolute semantics captures the meaning of each location in a trajectory; relative semantics capture the meaning of a location in a trajectory relative to other visited locations in the trajectory. To evaluate representativeness, the authors compared the absolute semantics of the real and synthetic data at the population level. Likewise, they measured coherence using a comparison of the relative semantics measured by the pair-wise semantic distance which was originally introduced by Bindschaedler and Shokri [8]. This metric accounts for trajectories of people who can live in geographically different locations but still share semantic patterns. Their results showed that the GAN-based approach preserved both the statistical characteristics of the original data and their relative semantics.

Ouyang et al. [65] did not conduct any privacy tests and limited the evaluation to one GAN-based model. In [53], Kulkarni et al. extended their study by testing the performance of seven generative models that used different architectures and conducting privacy tests to measure the novelty of the results. They compared deep generative models based on GANs, LSTMs, and other variations of RNNs with each other and with a statistical model, Copulas. Interestingly, Copulas and the GANs performed best in terms of representativeness, which was evaluated by comparing geographical statistics and absolute semantics (similar to [65]) and by measuring MMD. The RNNs and Copulas generated the most coherent synthetic trajectories. The long-range temporal dependencies throughout the generated trajectories, which measure coherence, decayed most slowly.

Interestingly, Kulkarni et al. [53] measured the novelty of the synthetic data by conducting two specific privacy tests. They applied a *location-sequence attack*, which determines the level of accuracy to which trajectories in the original data can be reconstructed, and a *membership interference attack*, which measures the accuracy of an inference that an individual contributed to a specific trajectory. In both tests, the RNN and GAN models outperformed the other models by a considerable margin.

The synthetically generated data in [53] and [65] were intended to be used in privacy-sensitive secondary data analyses. This is an important use, but the value of synthetic mobility data extends far beyond that. In [56], Lin et al. used labeled cellular geo-location data collected from mobile devices to generate synthetic mobility data for traffic volume simulations. Actual high-quality data on traffic volumes are difficult to collect. The simulations were applied to a super-district in the San Francisco Bay Area in California and were used to provide decision support for several transportation projects designed to improve urban transportation planning. The authors employed an LSTM model and evaluated its representativeness by comparing the vehicle traffic counts and public transit boarding and alighting counts of the simulated results and the actual counts. They argue that transportation policy-makers and planners can benefit from using synthetic location data to improve their understanding of urban mobility.

The literature on models for generating mobility data is not vast, and the quantitative approaches used to validate such models vary greatly. In reviewing different applications to mobility data, we observed that representativeness was particularly important in all of the studies. Consequently, the studies provide reliable metrics for representativeness, such as MMD, JSD, absolute semantics, and count variables. Coherence seems to be important in many cases but is evaluated in various ways, including relative semantics of the trajectories and observations of decays of temporal dependencies throughout the trajectory. Privacy, of course, is a significant issue. Reference [53] especially stands out its consideration of privacy as the authors conducted robust privacy tests on their synthetic data. We also find that all of the reviewed studies related to mobility presented a strong use-case for the generative models.

IV. SUMMARY AND CONCLUSION

Synthetic data allows governments, businesses, and researchers to easily access and share sensitive data without the risk of violating privacy regulations. The importance of having access to highly sensitive data was highlighted again through the COVID pandemic, where governments and researchers rely on high-quality sensitive medical data. Furthermore, the democratizing effect of accessible synthetic data mitigates the power of large data aggregators, such as Google and Facebook. It reduces the limitations of real-world data-sets, such as inherent biases, insufficient quantities, and class imbalance. With their ability to capture complex data and their relationships, deep generative models have boosted progress in synthetic data generation significantly.

This article discusses deep generative models for synthetic data and introduces a set of high-level evaluation criteria for a data-driven assessment of the quality of generated data. We examine their use and applicability to synthetic sequential data in the fields of natural language, speech, audio and visual data processing, healthcare and mobility.

The proposed evaluation framework allows for clear and easy communication of the requirements posed on synthetic data in different domains and use-cases. We find that synthetic texts in NLP applications are primarily evaluated for representativeness and realism. Synthetic music, speech, and video data must be realistic and coherent. Studies in healthcare are mainly concerned with generating private synthetic EHRs that still allow for secondary data analysis and, thus, assess the data's representativeness, novelty and realism. Synthetic mobility trajectories are generated for similar purposes with an additional focus on their coherence. However, not all mobility studies examine the synthetic data's novelty, potentially leading to privacy risks when sharing such data. Table 6 provides an overview of the assessment results in the reviewed domains.

The results show that in many domains the requirements posed on synthetic data do not conflict each other. For example, representativeness and realism in NLP applications or realism and coherence in video data go well together. However, there are domains where requirements do conflict each other. This can be observed in privacy sensitive domains such as healthcare and in creative domains such as music composition, where synthetic data have to be representative and novel at the same time. Finding an acceptable trade-off between those criteria can be challenging and usually involves a lot of tuning by the experimenter.

We also find that the nature of metrics used to evaluate the criteria can vary significantly. Some studies evaluate criteria only qualitatively by looking at synthetic text samples or listening to synthetic music samples. In most cases, only the individual-level criteria (i.e., novelty, realism, and coherence) are evaluated in this subjective way, but sometimes also, representativeness and diversity are. Other studies rely on human evaluations by laypeople or experts to judge realism and coherence of synthetic data. Human evaluations often also contain subjectivity either by the designer or the study participants. Thus, the most objective measures are formal computational metrics. Such metrics are primarily used to evaluate representativeness (e.g., by MMD or NLL) and diversity (e.g., self-BLEU or distances) of synthetic data. In many cases, novelty is not evaluated at all and coherence is assessed as part of evaluating the data's realism.

Our review highlights that generative architectures are used in a variety of applications and, in particular, GANs receive much attention. Most often, the architectural elements are used in conjunction with each other. In many cases, at the core of the networks, RNNs or CNNs are involved to ensure causally coherent generation of synthetic sequential data.

TABLE 6. Overview of reviewed application domains and metrics used to measure representativeness, novelty, realism, diversity, and coherence of deep generative models.

Domain	Studies	Representativeness	Novelty	Realism	Diversity	Coherence
Natural Language Processing	[30], [40], [76]	Perplexity, NLL	Qual.	Human eval.	Qual., Self-BLEU, Unique n-grams, 2-gram entropy, Unique 2,3,4-grams	Human eval., Self-BLEU
Speech & Audio Processing	[22]–[24], [58]	ALL, Music-theoretical measures	Qual., Avg. Eucl. distance to orig. set	Qual., Human eval., IS	Avg. Eucl. distance to synth. set	Qual., Human eval., Tonal distance
Visual Data Processing	[73], [80], [83]	Qual.	Qual.	Human eval., IS	IS	ACD
Healthcare	[6], [20], [27]	Qual., MMD, TSTR, K–S test, Dim.-wise stats.	Qual., NN distance, Disclosure risk	Human eval., TRTS, ML pred., ARM	Qual.	Qual., ARM
Mobility	[53], [56], [65]	Absol. semantics, Marg. distributions, MMD, Counting stat.	Location hiding			Rel. semantics, Temporal dep. decay

Autoregressive elements and attention mechanisms are also applied to some use-cases.

For the future, our proposed evaluation framework for unconditionally generated synthetic data has the potential to be extended for the evaluation of conditionally generated data. That kind of data is always generated within a given context, such as categories of videos or genres of songs. An evaluation framework for conditionally generated synthetic data has to account for that context. We expect conditionally generated synthetic data to need robustness within a context and variability to be more nuanced depending on the context.

With more jurisdictions passing privacy laws, in the future, we expect synthetic data to gain more attention. We expect more advanced and more objective metrics that allow a better and more objective assessment of synthetic data quality, particularly on the individual level. The development of the IS, used for synthetic images and videos, and other metrics that correlate well with the human judgement of realism, point in that direction. In-depth research of objective metrics allows systematic assessment of synthetic data quality with more robustness and less subjectivity in it. Meanwhile, we expect a continuation of the coexistence and combination of quality assessment based on expert judgment and formal computational metrics.

Another potentially interesting area worth further exploring is to complement a purely data-driven approach to assess the quality of synthetic data with a decision-oriented view. Credible decisions made on the basis of data can require certain properties of the data. Biased data with underrepresented minority groups can be a weak basis for decisions influencing all individuals, including the minority group. Other decisions can be sensitive to recent events in the data. The decisions made during the COVID pandemic, for example, were highly sensitive to the recency of the data. A decision-oriented evaluation approach could help improve decision-making (or avoid weak decisions) by contrasting the decisions derived from synthetic data scenarios with those based on the original, real-life data. Recent research into fairness and debiasing using synthetic data are promising starting points in this direction.

REFERENCES

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.

[2] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*.

[3] A. S. Assefa, D. Dervovic, M. Mahfouz, E. R. Tillman, P. Reddy, and M. Veloso, "Generating synthetic data in finance: Opportunities, challenges and pitfalls," in *Proc. 1st ACM Int. Conf. AI Finance (ICAIF)*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1–8.

[4] J. R. A. Solares, F. E. D. Raimondi, Y. Zhu, F. Rahimian, D. Canoy, J. Tran, A. C. P. Gomes, A. H. Payberah, M. Zottoli, M. Nazarzadeh, N. Conrad, K. Rahimi, and G. Salimi-Khorshidi, "Deep learning for electronic health records: A comparative review of multiple deep neural architectures," *J. Biomed. Informat.*, vol. 101, Jan. 2020, Art. no. 103337.

[5] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[6] M. K. Baowaly, C.-C. Lin, C.-L. Liu, and K.-T. Chen, "Synthesizing electronic health records using improved generative adversarial networks," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 3, pp. 228–241, Mar. 2019.

[7] 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," *Circulat., Cardiovascular Qual. Outcomes*, vol. 12, no. 7, Jul. 2019, Art. no. e005122.

[8] V. Bindschaedler and R. Shokri, "Synthesizing plausible privacy-preserving location traces," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 546–563.

[9] A. Borji, "Pros and cons of GAN evaluation measures," *Comput. Vis. Image Understand.*, vol. 179, pp. 41–65, Feb. 2019.

[10] A. Borji, "Pros and cons of GAN evaluation measures: New developments," 2021, *arXiv:2103.09396*.
[11] M. Bretan, G. Weinberg, and L. Heck, "A unit selection methodology for music generation using deep neural networks," in *Proc. 8th Int. Conf. Comput. Creativity (ICCC)*, 2017, pp. 72–79.
[12] J.-P. Briot, "From artificial neural networks to deep learning for music generation: History, concepts and trends," *Neural Comput. Appl.*, vol. 33, no. 1, pp. 39–65, Jan. 2021.
[13] J.-P. Briot, G. Hadjeres, and F.-D. Pachet, "Deep learning techniques for music generation—A survey," 2017, *arXiv:1709.01620*.
[14] T. B. Brown et al., "Language models are few-shot learners," 2020, *arXiv:2005.14165*.
[15] R. E. Bucklin and C. Sismeiro, "Click here for Internet insight: Advances in clickstream data analysis in marketing," *J. Interact. Marketing*, vol. 23, no. 1, pp. 35–48, Feb. 2009.
[16] L. Anthony Celi, R. G. Mark, D. J. Stone, and R. A. Montgomery, "'Big data' in the intensive care Unit. Closing the data loop," *Amer. J. Respiratory Crit. Care Med.*, vol. 187, no. 11, pp. 1157–1160, Jun. 2013.
[17] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, "Boosting deep learning risk prediction with generative adversarial networks for electronic health records," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 787–792.
[18] T. D. Chen, K. M. Kockelman, and J. P. Hanna, "Operations of a shared, autonomous, electric vehicle fleet: Implications of vehicle & charging infrastructure decisions," *Transp. Res. A, Policy Pract.*, vol. 94, pp. 243–254, Dec. 2016.
[19] Y. Chen, Y. Lv, and F. Wang, "Traffic flow imputation using parallel data and generative adversarial networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1624–1630, Apr. 2020.
[20] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," 2017, *arXiv:1703.06490*.
[21] S. De, M. Bermudez-Edo, H. Xu, and Z. Cai, "Deep generative models in the industrial Internet of Things: A survey," *IEEE Trans. Ind. Informat.*, vol. 18, no. 9, pp. 5728–5737, Sep. 2022.
[22] S. Dieleman, A. V. D. Oord, and K. Simonyan, "The challenge of realistic music generation: Modelling raw audio at scale," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2018, pp. 7989–7999.
[23] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *Proc. Int. Conf. Learn. Represent.*, 2019. [Online]. Available: https://arxiv.org/abs/1802.04208
[24] H. W. Dong, W. Y. Hsiao, L. C. Yang, and Y. H. Yang, "MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 34–41.
[25] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. Annu. Int. Conf. Theory Appl. Cryptograph. Techn.* Berlin, Germany: Springer, 2006, pp. 486–503.
[26] E. Erdem, M. Kuyu, S. Yagcioglu, A. Frank, L. Parcalabescu, B. Plank, A. Babii, O. Turuta, A. Erdem, I. Calixto, E. Lloret, E.-S. Apostol, C.-O. Truică, B. Šandrih, S. Martinčić-Ipšić, G. Berend, A. Gatt, and G. Korvel, "Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning," *J. Artif. Intell. Res.*, vol. 73, pp. 1131–1207, Apr. 2022.
[27] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional GANs," 2017, *arXiv:1706.02633*.
[28] X.-X. Fan, K.-P. Chow, and F. Xu, "Web user profiling based on browsing behavior analysis," in *Advances in Digital Forensics X*, G. Peterson and S. Shenoi, Eds. Berlin, Germany: Springer, 2014, pp. 57–71.
[29] E. Fatnassi, J. Chaouachi, and W. Klibi, "Planning and operating a shared goods and passengers on-demand rapid transit system for sustainable city-logistics," *Transp. Res. B, Methodol.*, vol. 81, pp. 440–460, Nov. 2015.
[30] W. Fedus, I. Goodfellow, and M. A. Dai, "MaskGAN: Better text generation via filling in the_____," Jan. 2018, *arXiv:1801.07736*.
[31] J. D. Fernández and F. Vico, "Ai methods in algorithmic composition: A comprehensive survey," *J. Artif. Intell. Res.*, vol. 48, no. 1, Oct. 2013, Art. no. 513582.
[32] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using GAN for improved liver lesion classification," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 289–293.
[33] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *J. Artif. Intell. Res.*, vol. 61, no. 1, p. 65170, Jan. 2018.
[34] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000.
[35] H. Gm, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "A comprehensive survey and analysis of generative models in machine learning," *Comput. Sci. Rev.*, vol. 38, Nov. 2020, Art. no. 100285.
[36] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[37] J. Guan, R. Li, S. Yu, and X. Zhang, "Generation of synthetic electronic medical record text," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 374–380.
[38] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 5767–5777.
[39] X. Guo and L. Zhao, "A systematic survey on deep generative models for graph generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5370–5390, May 2023.
[40] K. Guu, T. B. Hashimoto, Y. Oren, and P. Liang, "Generating sentences by editing prototypes," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 437–450, Dec. 2018.
[41] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proc. 1st ACM Workshop Audio Music Comput. Multimedia (AMCMM)*. New York, NY, USA: Association for Computing Machinery, 2006, p. 2126.
[42] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[43] S. Hochreiter and J. J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Dec. 1997.
[44] Y. Hong, U. Hwang, J. Yoo, and S. Yoon, "How generative adversarial networks and their variants work: An overview," *ACM Comput. Surv.*, vol. 52, no. 1, pp. 1–43, Feb. 2019.
[45] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," 2018, *arXiv:1809.04281*.
[46] L. Huang, B. Ding, Y. Xu, and Y. Zhou, "Analysis of user behavior in a large-scale VoD system," in *Proc. 27th Workshop Netw. Oper. Syst. Support Digit. Audio Video (NOSSDAV)*, New York, NY, USA, Jun. 2017, pp. 49–54.
[47] S. K. Hui, P. S. Fader, and E. T. Bradlow, "Path data in marketing: An integrative framework and prospectus for model building," *Marketing Sci.*, vol. 28, no. 2, pp. 320–335, Mar. 2009.
[48] N. Jaques, S. Gu, E. Richard Turner, and D. Eck, "Tuning recurrent neural networks with reinforcement learning," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017. [Online]. Available: https://arxiv.org/abs/1611.02796
[49] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker, "Perplexity—A measure of the difficulty of speech recognition tasks," *J. Acoust. Soc. Amer.*, vol. 62, no. S1, p. S63, Dec. 1977.
[50] P. B. Jørgensen, M. N. Schmidt, and O. Winther, "Deep generative models for molecular science," *Mol. Informat.*, vol. 37, nos. 1–2, Jan. 2018, Art. no. 1700133.
[51] P. D. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., Banff, AB, Canada, Apr. 2014. [Online]. Available: https://arxiv.org/abs/1312.6114
[52] I. Kobyzev, S. J. D. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," 2019, *arXiv:1908.09257*.
[53] V. Kulkarni, N. Tagasovska, T. Vatter, and B. Garbinato, "Generative models for simulating mobility trajectories," *CoRR*, vol. abs/1811.12801, 2018. [Online]. Available: https://arxiv.org/abs/1811.12801
[54] L. Kurup, M. Narvekar, R. Sarvaiya, and A. Shah, "Evolution of neural text generation: Comparative analysis," in *Advances in Computer, Communication and Computational Sciences*. Singapore: Springer, 2021, pp. 795–804.

[55] S. H. Lee, "Natural language generation for electronic health records," *npj Digit. Med.*, vol. 1, no. 1, p. 63, Dec. 2018.
[56] Z. Lin, M. Yin, S. Feygin, M. Sheehan, J.-F. Paiement, and A. Pozdnoukhov, "Deep generative models of urban mobility," *IEEE Trans. Intell. Transp. Syst.*, 2017.
[57] Z. Liu, Z. Li, K. Wu, and M. Li, "Urban traffic prediction from mobility data using deep learning," *IEEE Netw.*, vol. 32, no. 4, pp. 40–46, Jul. 2018.
[58] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–11.
[59] O. Mogren, "C-RNN-GAN: Continuous recurrent neural networks with adversarial training," 2016, *arXiv:1611.09904*.
[60] B. Noori, "An analysis of mobile banking user behavior using customer segmentation," in *Int. J. Global Bus.*, vol. 8, pp. 55–64, Dec. 2015.
[61] S. Norgaard, R. Saeedi, K. Sasani, and A. H. Gebremedhin, "Synthetic sensor data generation for health applications: A supervised deep learning approach," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 1164–1167.
[62] A. van den 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," 2016, *arXiv:1609.03499*.
[63] S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan, "This time with feeling: Learning expressive musical performance," *Neural Comput. Appl.*, vol. 32, no. 4, pp. 955–967, Feb. 2020.
[64] A. Oussidi and A. Elhassouny, "Deep generative models: Survey," in *Proc. Int. Conf. Intell. Syst. Comput. Vis. (ISCV)*, Apr. 2018, pp. 1–8.
[65] K. Ouyang, R. Shokri, D. S. Rosenblum, and W. Yang, "A non-parametric generative model for human trajectories," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3812–3817.
[66] M. Platzer and T. Reutterer, "Holdout-based empirical assessment of mixed-type synthetic data," *Frontiers Big Data*, vol. 4, p. 43, Jun. 2021.
[67] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–36, Sep. 2018.
[68] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3617–3621.
[69] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, Y. Bengio and Y. LeCun, Eds., May 2016. [Online]. Available: https://arxiv.org/abs/1511.06434
[70] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[71] L. Regenwetter, A. H. Nobari, and F. Ahmed, "Deep generative models in engineering design: A review," *J. Mech. Des.*, vol. 144, no. 7, Mar. 2022, Art. no. 071704.
[72] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. 31st Int. Conf. Mach. Learn.*, vol. 32, E. P. Xing and T. Jebara, Eds., Bejing, China, Jun. 2014, pp. 1278–1286.
[73] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2849–2858.
[74] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2016, Art. no. 22342242.
[75] Y. Shavitt and N. Zilberman, "A geolocation databases study," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 10, pp. 2044–2056, Dec. 2011.
[76] D. Shen, A. Celikyilmaz, Y. Zhang, L. Chen, X. Wang, J. Gao, and L. Carin, "Towards generating long and coherent text with multi-level latent variable models," 2019, *arXiv:1902.00154*.
[77] F. Stahlberg, "Neural machine translation: A review," *J. Artif. Intell. Res.*, vol. 69, pp. 343–418, Oct. 2020.
[78] A. M. Tekalp, *Digital Video Processing*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2015.
[79] L. Theis, A. V. D. Oord, and M. Bethge, "A note on the evaluation of generative models," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2016. [Online]. Available: https://arxiv.org/abs/1511.01844
[80] S. Tulyakov, M. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1526–1535.
[81] C. G. Turhan and H. S. Bilge, "Recent trends in deep generative models: A review," in *Proc. 3rd Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2018, pp. 574–579.
[82] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1096–1103.
[83] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2016, Art. no. 613621.
[84] X. Wang, S. Takaki, and J. Yamagishi, "Neural Source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5916–5920.
[85] M. Wiese, R. Knobloch, R. Korn, and P. Kretschmer, "Quant GANs: Deep generation of financial time series," *Quant. Finance*, vol. 20, no. 9, pp. 1–22, 2020.
[86] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1419–1428, Oct. 2018.
[87] D. Xu, S. Yuan, L. Zhang, and X. Wu, "FairGAN: Fairness-aware generative adversarial networks," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 570–575.
[88] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, "Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data," in *Proc. 26th Int. Conf. World Wide Web (WWW)*. Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017, Art. no. 12411250.
[89] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, "Mining electronic health records (EHRs): A survey," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–40, Jan. 2018.
[90] A. Yahi, R. Vanguri, N. Elhadad, and N. P. Tatonetti, "Generative adversarial networks for electronic health records: A framework for exploring and evaluating methods for predicting drug-induced laboratory test trajectories," 2017, *arXiv:1712.00164*.
[91] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. Bennett, "Privacy preserving synthetic health data," in *Proc. Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn. (ESANN)*, 2019, pp. 465–470.
[92] L.-C. Yang and A. Lerch, "On the evaluation of generative models in music," *Neural Comput. Appl.*, vol. 32, no. 9, pp. 4773–4784, May 2020.
[93] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7962–7966.
[94] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.
[95] B. Zhang, G. Kreitz, M. Isaksson, J. Ubillos, G. Urdaneta, J. A. Pouwelse, and D. Epema, "Understanding user behavior in spotify," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 220–224.
[96] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5908–5916.

PETER EIGENSCHINK received the B.S. and M.S. degrees in physics, with specialization in computational and gravitational physics from the University of Vienna, Austria. He is currently pursuing the Ph.D. degree in economics with the Vienna University of Economics and Business (WU Vienna).

Since 2017, he has been an independent IT consultant based in Vienna, Austria. From 2019 to 2021, he was a Research and Teaching Associate with WU Vienna. His research interests include synthetic data-based privacy in consumers analytics and algorithmic dynamic pricing of perishable products in grocery retailing.

THOMAS REUTTERER is currently a Professor in marketing and customer analytics with the Vienna University of Economics and Business (WU Vienna). His research interests include analyzing, modeling, and forecasting customer behavior in data-rich environments. In his research projects, he employs advanced statistical or machine learning methods to provide decision support for various business applications. His recent research interests include customer value and relationship management, customer base analysis, and content marketing supported by generative natural language models.

STEFAN VAMOSI received the B.S. and M.S. degrees in physics, with specialization in computational physics. He is currently pursuing the Ph.D. degree with the Vienna University of Economics and Business (WU Vienna).

He is also a Research and Teaching Associate with WU Vienna. During his master's thesis, he was based with CERN, where he developed a simulation software for an anti-hydrogen beam experiment. Prior to joining WU's Doctoral Program, in May 2018, he gained professional experience in a consulting firm. His research interests include time-series analysis, behavioral customer segmentation, and data prediction with deep learning approaches.

RALF VAMOSI received the B.S. degree in physics from the University of Vienna, Austria, where he is currently pursuing the Ph.D. degree in computer science.

Since 2017, he has been a Software Engineer with the Technical University of Vienna, Austria. From 2019 to 2020, he was a Researcher with the Vienna University of Economics and Business.

CHANG SUN received the Ph.D. degree in data science and the M.S. degree in artificial intelligence from the University of Maastricht, The Netherlands.

From 2017 to 2022, she was a Research and Teaching Associate with the University of Maastricht, where she has been a Postdoctoral Researcher, since 2022. Her research interests include privacy-preserving machine learning and synthetic data generation.

KLAUDIUS KALCHER received the B.S. and M.S. degrees in statistics from the Technical University of Vienna and the Ph.D. degree in medical physics from the Medical University of Vienna, Austria.

He co-founded the startup MostlyAI, focused on synthetic data generation for privacy applications, in 2017. From 2015 to 2016, he was a Postdoctoral Researcher with the Medical University of Vienna. His research interests include synthetic data generation and ethical artificial intelligence.