BENCHMARKING TABULAR DATA SYNTHESIS

FRAMEWORK FOR EVALUATING USE-CASE SUITABILITY AND PERFORMANCE ON COMMODITY HARDWARE

BENCHMARKING TABULAR DATA SYNTHESIS

FRAMEWORK FOR EVALUATING USE-CASE SUITABILITY AND PERFORMANCE ON COMMODITY HARDWARE

Dissertation

for the purpose of obtaining a doctoral degree at the
Faculty II - Department of Computing Science in the
Carl von Ossietzky University of Oldenburg

by

Maria Fernanda DAVILA RESTREPO

born in Barranquilla, Colombia

add date

Day of oral defense: XXXX

Pending Alex's habil

The following evaluators recommended the admission of the dissertation :

Prof. Dr. Wolfram Wingerath,	Carl von Ossietzky University of Oldenburg
Prof. Dr. Fabian Panse,	Augsburg University
Dr. -Ing. habil. Alexandra Pehlken,	OFFIS Institute for Information Technology

This document was created on February 1, 2025. Changes will be made available here.

What I cannot create, I do not understand

Richard Feynman

Contents

PREFACE

As Richard Feynman once said, “What I cannot create, I do not understand”. In the field of generative modeling, this idea transforms into a complementary perspective: What I understand, I can create. These ideas have guided my work and reflect the essence of this dissertationunderstanding the complexities of tabular data to build models that can recreate it.

My interest in tabular data synthesis began with the shared frustration of my nerdy friends in sciencefinding good data was always a struggle. That frustration turned into curiosity when I discovered the intriguing world of generative modeling, a method that seemed to hold answers to this problem. Over time, my growing interest in artificial intelligence combined with the practical challenges we faced daily at the research institute. This inspired me to work toward a goal: bridging the gap between scientists who need quality data and those who use data analysis and synthetic data in their work.

Throughout this journey, I have worked to understand not only how generative models function but also how to evaluate their effectiveness and relevance for real-world use. It has been a rewarding challenge to contribute to a field that holds immense potential for research and applications alike.

I am deeply grateful to the many people who supported me during this journey. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis.

Write my acknowledgments

May this work serve as a small step toward making data-driven modeling a little easier for everyone.

Maria Fernanda Davila Restrepo
Oldenburg, August 2025

Summary

Summary in English...

Zusammenfassung

Summary in German...

Acronyms

AI Artificial Intelligence

BN Bayesian Networks

CNN Convolutional Neural Networks

DPM Diffusion Probabilistic Models

EHR Electronic Health Records

GAN Generative Adversarial Network

GNN Graph Neural Networks

HMM Hidden Markov Models

LLM Large Language Models

LSTM Long Short Term Memory

ML Machine Learning

MRF Markov Random Fields

NF Normalizing Flows

PGM Probabilistic Graphical Models

RNN Recurrent Neural Networks

TDS Tabular Data Synthesis

VAE Variational Autoencoder

1

INTRODUCTION

[1]

This chapter includes first the motivation, starting with why data-driven decision-making has become important. Describe the challenges in data-driven techniques. Suggest synthetic data as a partial solution for those challenges. Describe the gap in data synthesis and our goal to address it.

Write this in a nice way

Next, the main concepts are defined, including data synthesis for different data types. Then, Tabular Data Synthesis in specific. Important to clearly state what I mean by tabular data, by synthetic data, what is outside of the scope...

Next, the challenges in Tabular Data Synthesis (TDS).

Finally the contributions I provide in this thesis.

Additionally I include the outline of the thesis and an overview of my published work.

[1] Parts of this chapter have been published in [**davila_navigating_2024??**].

1

SINCE the 19th century, society has increasingly relied on data and its analysis to inform decision-making across science and engineering. Historically, these fields relied heavily on physical models, mathematical representations based on first principles, to describe and predict the behavior of systems. Until the middle of the 20th century, these models were the primary tools for solving complex problems. However, they often required simplifying assumptions to make computations manageable, which limited their accuracy and applicability to real-world scenarios.

Already in the 19th century, a shift towards stochastic thinking began, as people realized that many real-world processes are influenced by random events and cannot be fully explained by deterministic models alone. A notable example is Ludwig Boltzmanns work in statistical mechanics [1], where he introduced probabilistic methods to describe the random motion of particles in a gas, marking a departure from purely deterministic approaches in physics.

During the early 20th century, Kolmogorov formalized probability theory, and concepts like Markov processes became essential tools for understanding stochastic systems [2]. World War II further accelerated the adoption of stochastic models, as they were used in operations research [3] to optimize supply chains and troop movements. The development of Monte Carlo methods [4] also emerged as a practical approach to solving problems involving randomness and uncertainty. Despite their growing importance, stochastic models faced opposition from proponents of deterministic models, who argued that the inclusion of randomness made them less rigorous or scientifically valid.

At the mid-20th century, the concept of Artificial Intelligence (AI) emerged as another approach to understanding the world and solving physical problems [5, 6]. Inspired by the human brain, AI sought to replicate human reasoning and decision-making. Early efforts focused on symbolic logic and rule-based systems, where researchers used explicit rules to encode knowledge and simulate thought processes. However, it soon became clear that many real-world problems involve uncertainty and variability, which deterministic methods could not fully address. This realization led to the integration of stochastic methods into AI, allowing for more robust handling of uncertain environments. Techniques such as probabilistic reasoning [7], Bayesian networks [8], and Markov decision processes [9] became foundational in AI, enabling systems to model complex environments, learn from data, and make predictions despite incomplete or noisy information.

The development of numerical methods, such as finite element analysis [10] and computational fluid dynamics [11], which became widely used in the 1970s and 1980s, marked a significant turning point. These methods, combined with advances in computing power, including the development of faster processors (e.g., microprocessors like Intel 386 [12]), the rise of parallel computing architectures (e.g., supercomputers like Cray [13]), and advancements in storage technologies (e.g., hard disk drives with increased capacity), allowed researchers to solve more complex equations and simulate systems with much greater precision. While this progress enhanced the utility of physical models, it was still largely constrained by the availability of domain knowledge and the inherent limitations of deterministic approaches.

The late 20th century saw another major shift. Innovations in sensors, communication systems, and storage technologies enabled the growth of the amount of data generated and collected. Around this time, Machine Learning (ML) [14] emerged as a key area within AI. ML focuses on enabling systems to learn patterns and make predictions directly from data, rather than relying on explicitly programmed rules. Its origins trace back to the mid-20th century, with early efforts including perceptrons [15] and simple neural networks [16]. However, progress stalled during the 1970s and 1980s due to limited computational power and the scarcity of large datasets. The field re-gained importance in the 1990s and early 2000s with the advances mentioned in computing power, the advent of the internet, and the exponential growth of data. These developments enabled the training of more sophisticated ML models, which have since become integral to modern AI systems. ML algorithms and statistical techniques became more powerful and accessible, enabling researchers to analyze large datasets and uncover patterns that were previously difficult to detect.

As a result, decision-making in many fields began to shift from being primarily model-driven to being increasingly data-driven. Instead of relying solely on theoretical equations, researchers and industry professionals started using data to train models that could learn complex relationships directly from observations. This approach has proven particularly effective in scenarios where the underlying system is too complex to be captured by physical models alone, or where the data provides additional insights that complement theoretical understanding.

However, data-driven models have limitations. For data-driven methods to be effective, they must maintain scientific validity and remain aligned with the physical problems [17]. In some cases, the required data may not exist at all, the available data might be incomplete, contain missing values, or exhibit imbalanced distributions that make analysis difficult. Furthermore, datasets often contain sensitive information, imposing restrictions on their accessibility and use [**davila_navigating_2024??**]. Addressing these issues by collecting extensive datasets is not only prohibitively expensive but also impractical, as it would require exhaustive experiments for every conceivable scenario.

To address these challenges, researchers are exploring several directions, for example hybrid models that use data-driven models with theoretical frameworks [17]. These hybrid models combine the strengths of both approaches. Another example is the use of synthetic data to improve the accuracy of data-driven models, which has been already widely implemented in domains, such as health [18], autonomous driving [19] and finance [20].

Synthetic data refers to artificially generated data that replicates real-world data distributions and characteristics. Early synthetic data generation used rule-based methods, using predefined heuristics and domain knowledge to replicate real-world processes [21, 22]. Over time, these methods included stochastic and statistical models. More recently, techniques use a mix of stochastic methods, statistical modeling, deep learning architectures, attention mechanisms, and probabilistic frameworks inspired by concepts from physics, such as diffusion models. These developments have significantly enhanced the quality and versatility of synthetic

data, making it an essential component of modern data science and engineering.

However, a significant gap exists: there is no universal tool that performs well in all use cases. Yet, the goal is not to develop such a universal tool. Instead, the objective is to enable users to understand which models and tools are suitable for their specific applications. This task is not trivial because the performance of a model can vary significantly depending on the evaluation metric and the intended purpose of the generated data [23]. Similarly, commonly used metrics often fail to capture important aspects like semantic consistency, and their effectiveness depends heavily on the target task [24]. This underscores the need for careful, context-driven evaluation methodologies tailored to the application at hand.

This thesis is motivated by the potential of synthetic data to address the challenges of data scarcity, insufficient quality, restricted accessibility and unsuitability. More specifically, this thesis aims to address the gap of the lacking benchmark, which allows to assess the models and tools' suitability for their specific use cases. By doing so, end users in research and industry would have a clear guidance on how to use these data synthesis tools effectively to generate the synthetic data needed for their data-driven models.

This thesis aims to make synthetic data generation accessible to end users outside the data synthesis community. Specifically, it focuses on evaluating and benchmarking the suitability of TDS models and tools for specific use-case requirements and their performance on commodity hardware. By developing a robust framework for this evaluation, the work seeks to provide practical insights into the use-case applicability, limitations, and efficiency of these methods, contributing to more effective and accessible synthetic data.

The following sections in this chapter define the scope of this thesis, including a clear definition of data synthesis, a specification of the type of data being addressed, tabular data, an overview of the challenges in TDS, and a summary of the contributions made by this research.

Distinction Between Models, Tools, and Algorithms in this Thesis

In this thesis, TDS models refer to the underlying architectures used to generate synthetic tabular data. These models define the theoretical and mathematical principles governing synthetic data generation.
Each model is based on a core algorithm, which provides the computational framework for learning and generating data. The algorithm defines the optimization process and mathematical operations that enable the model to solve the mathematical objective.
On the other hand, tools refer to specific implementations of these models. Tools transform theoretical models into practical applications by integrating them into software packages or libraries, making them accessible for real-world use.
One example of a model is a Generative Adversarial Network (GAN), which uses an adversarial training (minmax game) algorithm, and TGAN is a tool that implements a GAN model.

1.1. DATA SYNTHESIS

THE generation of synthetic data includes diverse data types, such as images, text, tabular datasets, time series, audio, and video. Each of these domains has seen remarkable progress over the last decade, caused by advances in synthesis models.

In image generation, GANs and diffusion[2] significantly improved synthetic image quality [25–28]. GANs consist of two competing networks, a generator and a discriminator, that work against each other to produce realistic outputs. Diffusion models, inspired by physical diffusion processes, generate images by iteratively denoising random noise until there is a realistic output. These models have enabled the generation of realistic images, transforming fields like entertainment [29], healthcare [30], and autonomous driving [31].

Similarly, synthetic text generation has advanced significantly, with models like OpenAI's GPT series [32] producing coherent and contextually relevant text, which has reached the general public. This is caused by the development of attention mechanisms [33]. Attention mechanisms and the Transformer architecture [34], allow models to focus on specific parts of the input data, significantly improving their ability to handle long-range dependencies. For example, in text generation, attention mechanisms enable models to produce contextually appropriate sentences by understanding the relationships between words across a sequence [34]. In image synthesis, attention has been incorporated into models to enhance the quality and coherence of generated images by focusing on important spatial regions [35].

Correspondingly, the synthesis of tabular data has also grown significantly in the past ten years, as it is shown in the timeline in Figure 1.1. Tabular data is one of the most widely used data formats in industries like finance, healthcare, and manufacturing [36, 37]. In this research, tabular data is defined according to the relational data model [38, 39]:

Definition of Tabular Data for this Thesis

A tabular dataset consists of one or several tables organized into rows (or records) representing individual data points and columns representing different features of those data points.
For instance, a customer database might have rows for each customer and columns for features like age, income, and purchase history.

Insert a timeline

The focus of this thesis is on tabular data synthesis (TDS). Despite its importance across many applications, generation of tabular data has not achieved the same level of accessibility among non-researchers as image and text synthesis. This thesis argues that this is caused by the fact that the inherent structure of tabular data presents unique challenges for synthesis, as described in Section 1.3. It also argues that these challenges have hindered the creation of a universal tool that is suitable for all tabular data applications, and understanding which tool is suitable for each application is not trivial [23, 24].

[2] The models will be described in detail in Section 2

1

Figure 1.1.: Caption

To the best of my knowledge, two main branches of research exist for tabular data generation, based on different applications: (i) generation of new records based on schema information, statistics, and domain knowledge, and (ii) capture and replication of the characteristics of a given (real-world) dataset as accurately as possible, in order to generate new records for different purposes (Section 1.2).

The first branch is typically used to benchmark database technology and, although the data should be realistic, the focus is more on efficient and scalable generation techniques. These techniques are mainly rule-based and tailored to a specific domain [**NeufeldML93??**, **GraySEBW94??**, **BrunoC05??**, **HoukjaerTW06??**, **HoagT07??**, **ChristenP09??**, **RablP11??**]. The second branch is TDS.

Definition of Tabular Data Synthesis for this Thesis

Tabular Data Synthesis (TDS) is defined as the process of generating artificial datasets that preserve the characteristics found in real-world tabular data. This includes capturing the column types, the dependencies across columns, maintaining realistic distributions, .

There is no reference that provides a classification of the main models and tools for TDS. For this reason, I propose my own in Chapters 2 and 3.

1.2. TABULAR DATA SYNTHESIS PURPOSES

Add some diagrams

BASED on our literature review, we pinpointed five purposes why users need to generate artificial tabular data:

- **Missing Values Imputation:** Datasets often have incomplete entries, which can distort analyses. For instance, Electronic Health Records (EHR), where the patient's smoking status is missing, which is vital information for predicting the risk of heart disease. TDS allows users to generate datasets that are free from these gaps, ensuring intentional completeness [40].

- **Dataset Balancing:** Some datasets have a few classes which significantly outnumber others, risking bias towards these dominant classes. For example, in a diabetes dataset, non-diabetic patient records may outnumber diabetic ones. This discrepancy risks biasing predictive models towards the dominant non-diabetic class. TDS can rectify this imbalance by generating additional synthetic diabetic patient records.

- **Dataset Augmentation:** TDS can be used for data augmentation, where the goal is to expand datasets for enhancing model robustness and generalization. In our EHR example, this would mean synthesizing records for new patients from all classes.

- **Customized Generation:** The generation of synthetic datasets must sometimes be directed by external factors, in order to create specific scenario data. For instance, in the EHR context, researchers might need to generate data for a scenario where there's a spike in a particular disease due to an environmental change. The data generation would be conditioned on new environmental factors.

- **Privacy Protection:** Many domains contain sensitive data, requiring effective measures to protect privacy when these data need to be shared or reused. In our EHR example, a hospital would like to share data for external analysis and therefore creates synthetic patient records that closely resemble real statistical patterns but do not correspond to actual patient data.

Privacy protection can be the sole reason for the synthesis but it can also be combined with any of the other purposes. While tools for class imbalance or missing values can most often be adapted for data augmentation or vice versa, it is important to choose the right one for the specific task at hand, as key relationships in the data might not be preserved. In addition it will likely increase the workload and computational costs.

1.3. TABULAR DATA SYNTHESIS CHALLENGES

FOR all domains of data synthesis, whenever privacy protection is of interest, one challenge is the *Privacy vs. Utility trade-off* [41]. Data utility refers to the ability of the data to serve its intended purpose effectively. Generating synthetic samples while preserving privacy is challenging because enhancing privacy often diminishes the utility of the data, and vice versa [42]. The level of privacy can be achieved using Differential Privacy (DP), a rigorous mathematical framework for guaranteeing privacy in statistical analysis [43].

TDS tools must capture and replicate the main characteristics of the original (real) dataset. This includes the column types and correlations between columns. This is challenging because of the following reasons:

- **Missing values:** Accurately capturing the characteristics of a dataset with information gaps is challenging, as these gaps represent a loss of information

and the generated data may poorly reflect the true correlations between the columns of the dataset.

- **Imbalanced datasets:** Capturing the characteristics of minority classes is particularly complicated when these classes are underrepresented. As a consequence of this under-representation, some algorithms may over fit the data, suffer from phenomena such as mode collapse, where some classes are not generated at all [23], or generate unrealistic samples of the minority classes [44]. However, for applications that aim to identify outliers, such as intrusion detection [45], it is very important to generate accurate samples of these minority classes.

- **Diversity of column types:** Unlike images, tabular datasets usually contain a mix of different column types, such as numerical, categorical, temporal, text, or even mixed types consisting of values from different basic types. Different column types might require distinct pre-processing or handling techniques.

- **Complex column distributions:** The distribution of a column contains its spread, tendencies, and patterns in the data, providing valuable insights into its characteristics and relationships with other columns. Capturing complex distributions is challenging because traditional methods such as simply modeling mean and standard deviation may not be sufficient to characterize non-Gaussian distributions.

- **Temporal Dependencies:** The temporal dimension of time series data introduces an additional layer of complexity. Two particular challenges for time series generation are discrete time series, because backpropagation presents problems [46], and long-term dependencies, because their discovery and modeling require extra memory [47].

1.3.1. THE UTILITY VS. PRIVACY TRADE-OFF

1.4. PRIMARY CONTRIBUTIONS

SINCE a di

1.4.1. CONTRIBUTION 1

asdasa

1.4.2. CONTRIBUTION 2

asdads

1.4.3. CONTRIBUTION 3

adsasd

1.4.4. CONTRIBUTION 4

asdada

1.5. THESIS OUTLINE

SINCE a di

1.6. PUBLISHED WORK

1.6.1. CORE RESEARCH PUBLICATIONS

5. **<empty citation>??** — Platform
4. **<empty citation>??** — Sensitivity?
3. **<empty citation>??** — Benchmark
2. **davila_navigating_2024??**
1. M. F. Davila R., W. Wingerath, and F. Panse. "Benchmarking Tabular Data Synthesis for User Guidance". In: EDBT/ICDT 2024 Joint Conference. Mar. 1, 2024. URL: https://ceur-ws.org/Vol-3651/PhDW-2.pdf

BEST POSTER AWARD

1. M. F. Davila R. Role of data synthesis in critical raw materials. International Round Table on Materials Criticality (IRTC). Feb. 2023. URL: Award

1.6.2. PROJECT-RELATED PUBLICATIONS - SYNTHETIC DATA APPLICATION

4. **<empty citation>??** — Book chapter Tobias
3. M. F. Davila R., T. Hoiten, P. Sander, N. Woltering, and A. Pehlken. "Container-Based Microservices Application for Product Carbon Footprint Calculation in Manufacturing Companies". In: *EnviroInfo 2024*. May 2025
2. A. Pehlken, M. Davila, L. Dawel, and O. Meyer. "Digital Twins: Enhancing Circular Economy through Digital Tools". In: *Procedia CIRP* 122 (Jan. 1, 2024), pp. 563–568. DOI: 10.1016/j.procir.2024.01.082
1. M. F. Davila R., F. Schwark, L. Dawel, and A. Pehlken. "Sustainability Digital Twin: a tool for the manufacturing industry". In: *Procedia CIRP*. 30th CIRP Life Cycle Engineering Conference 116 (Jan. 1, 2023), pp. 143–148. ISSN: 2212-8271. DOI: 10.1016/j.procir.2023.02.025. URL: https://www.sciencedirect.com/science/article/pii/S2212827123000252 (visited on 01/30/2025)

1

1.6.3. OTHER PUBLICATIONS

GREEN

4. **<empty citation>??**

3. F. Schwark, H. Garmatter, M. Davila R., L. Dawel, A. Pehlken, F. Cyris, and R. Scharf. "The application of image recognition methods to improve the performance of waste-to-energy plantsplants". In: EnviroInfo 2022. Gesellschaft für Informatik e.V., 2022, p. 167. ISBN: 978-3-88579-722-7. URL: https://dl.gi.de/handle/20.500.12116/39413 (visited on 01/30/2025)

2. H. Torio, A. Günther, M. F. Davila, and M. Knipper. "Paving the Way for Hybrid Teaching in Higher Education: Lessons from Students Perceptions and Acceptance of Different Teaching Modes during and after the Pandemic". In: *Creative Education* 14.5 (May 15, 2023), pp. 1029–1042. DOI: 10.4236/ce.2023.145066. URL: https://www.scirp.org/journal/paperinformation?paperid=125257 (visited on 01/30/2025)

1. F. Penaherrera, M. F. Davila R., A. Pehlken, and B. Koch. "Quantifying the Environmental Impacts of Battery Electric Vehicles from a Criticality Perspective". In: *2022 IEEE 28th International Conference on Engineering, Technology and Innovation (ICE/ITMC) & 31st International Association For Management of Technology (IAMOT) Joint Conference*. 2022 IEEE 28th International Conference on Engineering, Technology and Innovation (ICE/ITMC) & 31st International Association For Management of Technology (IAMOT) Joint Conference. June 2022, pp. 1–9. DOI: 10.1109/ICE/ITMC-IAMOT55089.2022.10033264. URL: https://ieeexplore.ieee.org/document/10033264 (visited on 01/30/2025)

1.7. SUMMARY

Definition / Problem Description (data-driven models limits) Gap Motivation Definition of data synthesis Definition of tabular data Definition of tabular data synthesis Main purposes why users use TDS Main challenges in TDS Contributions

2

BACKGROUND

This is an abstract of every chapter I will write to let people know what this chapter is about.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

2.1. THE FIVE TRIBES OF ARTIFICIAL INTELLIGENCE

SINCE a di [55]

2

2.2. DEEP LEARNING

SINCE a di [23]

2.2.1. IT'S ALL ABOUT JOINT PROBABILITIES

2.2.2. GENERATIVE MODELING

[56, 57]

3

Related Work

[1]

This is an abstract of every chapter I will write to let people know what this chapter is about.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

[1] Parts of this chapter have been published in [**davila_navigating_2024??**].

3.1. TABULAR DATA SYNTHESIS MODELS AND TOOLS

SINCE a di

3.1.1. IMPUTATION

Imputation-based data synthesis models were initially introduced to reduce the risk of disclosing sensitive data [58], as solutions for Statistical Disclosure Control (SDC) and Statistical Disclosure Limitation (SDL). Many imputation-based TDS tools use either multiple imputation or masking techniques [59–61]. Multiple imputation is originally a technique to handle missing values, where each missing value is replaced by two or more synthetic values. It has two steps: First, it constructs multiple synthetic populations. Then, it draws a random sample from each synthetic population and releases those samples.

Imputation is easy to understand and implement, and it does not require high computational resources. However, it is highly sensible to bias, it can produce extreme samples, and it can contain several repetitions of the observed records [62]. Finally, imputation-based approaches do not model the underlying joint distributions of the real dataset, which means they cannot preserve its semantic integrity [63].

3.1.2. SAMPLING

Data synthesis is often used to rebalance datasets. The straightforward solution to this problem is the augmentation with additional records of the minority class, known as *random over-sampling*. Accordingly, *random under-sampling* removes records from the majority class. Random over-sampling introduces an increased risk of over-fitting and random under-sampling frequently results in the loss of valuable information intrinsic to the original dataset [64]. Tools such as the Synthetic Minority Over-Sampling Technique (SMOTE) [44], can be used for class rebalancing, but also to generate complete synthetic datasets. The vanilla version only works for continuous data and the synthesized records are linearly dependent on the original minority class records, often leading to over-fitting [64]. Variations of SMOTE address these limitations and still generate samples with low computational resources [65]. The implementation presented in [66] combines the strengths of several other SMOTE variations [67–69], including the ability to handle categorical columns.

SMOTE adds data points to the minority class, not through mere duplication, but by employing the k-Nearest-Neighbors interpolation method [44]. Despite its advantages, SMOTE has some limitations: the vanilla version only works for continuous data, and the synthesized data points are linearly dependent on the original minority class data point, often leading to over-fitting [64]. Many variants of SMOTE address these limitations, such as borderline-SMOTE [67], SVM-SMOTE [68], and K-Means-SMOTE [69]. ADASYN [70], focuses on identifying minority class samples within spaces dominated by the majority class. This improvement reduces unnecessary over-sampling of points. Like SMOTE, ADASYN improves machine learning performance and requires relatively low computational cost [64].

Another SMOTE-based variant, ADASYN [70], focuses on identifying minority class samples within spaces dominated by the majority class. Consequently, ADASYN

generates more synthetic data points for minority classes with a higher proportion of majority class observations within the k-Nearest-Neighbors region. This improvement reduces unnecessary over-sampling of points. Like SMOTE, ADASYN improves machine learning performance and requires relatively low computational time [64].

3.1.3. DISCRIMINATIVE MODELS

In ML, a distinction is made between *discriminative* and *generative* models [**<empty citation>??**]. Discriminative models estimate the conditional probability of the output given the input. However, they do not learn the interdependence between all the columns (target and non-target). Discriminative models can be leveraged for TDS using the learned conditional probabilities. One example in privacy-preserving data mining are clustering-based algorithms that generate synthetic data while aiming to maintain certain properties of the original data [71]. However, these models have limitations in handling more intricate data characteristics, because they are built to estimate the probability that an observation belongs to a class and not to learn the complete distribution [57].

3.1.4. GENERATIVE MODELS

Generative models aim to learn the joint probability distribution of all columns [23]. Therefore, they are suitable for all the purposes introduced in Section 1.2.

SHALLOW LEARNING

Copulas [72] Tabular Copula [73] Gaussian Copula [74]

PROBABILISTIC GRAPHICAL MODELS (PGM)

PGM [75]

Bayesian Networks (BN) BN [8] PrivBayes [76]

Markov Random Fields (MRF) MRF [77] PrivMRF [78] PrivLava [79]

Hidden Markov Models (HMM) HMM

DEEP LEARNING

Deep learning generative models are typically called Deep Generative Models. They are composed of multiple neural layers that enable the model to learn hierarchical representations. They leverage deep learning techniques to model the joint probability distribution of a dataset.

The table below provides a structured distinction between models, tools, and their respective algorithms:

Check the information and also the text above and change for the tools I use. Find best placement, summary?

Variational Autoencoder (VAE) VAEs learn an encoder network that maps the input data to a latent space and a decoder network that reconstructs the original input from this latent space [80]. Synthetic datasets are generated by sampling new records from the latent space.

VAEs have demonstrated impressive results in synthesizing data across multiple domains, including images, text, and music [81–83]. In TDS, VAEs preserve the characteristics of the dataset better than sampling and shallow generative models [84]. Nevertheless, VAEs often over-simplify the distributions inherent in the original data because they use a standard Gaussian distribution for the latent space [85]. Additionally, VAEs struggle with discrete or categorical columns because they use a reparametrization trick for backpropagation, which only works well for continuous latent spaces [86].

Examples of TDS tools are tabular VAE (TVAE) [84], discrete VAE [87], which addresses the limitation with discrete columns, DP-VAEGM [88] for differential privacy, and TimeVAE [89] for generating multivariate time series data.

Generative Adversarial Networks GAN GANs [25] consist of two main neural networks: a generator and a discriminator. The generator uses random noise as input and generates synthetic data samples, while the discriminator aims to distinguish between real and synthetic samples. During training, the generator and the discriminator are trained in an adversarial manner, with the generator attempting to generate data that fools the discriminator, and the discriminator striving to correctly identify the generator's fake samples. Through this competitive process, and using their implicit modeling of the data distribution, GANs learn to generate realistic and high-quality synthetic data samples that closely resemble the distribution of the real training data [90].

For many years, most of the tools were based on GANs, even though applying GANs to tabular data synthesis requires some adaptations because of the challenges described in Section 1.3. Traditional GANs are unsupervised and generate data from random noise. However, for tabular data synthesis, it is beneficial to control the output in order to tackle data imbalance or impose domain constraints. Conditional GANs, introduced by Mirza and Osindero [91], extend GANs by conditioning the generator and discriminator on additional information, which increased semantic integrity.

Tools such as medGAN [92], DP-GAN [93], and PATE-GAN [94] were specifically developed for privacy-preserving TDS. However, they sacrifice data utility and report lower performance than a vanilla GAN for many ML tasks [94]. TableGAN [41], TGAN [95], and CTGAN [84] are able to achieve high data privacy with better data utility. CTGAN also uses a conditional vector to allow controlling the generated classes. Building upon them, the two predominant GAN tools nowadays are CTAB-GAN [96] and its successor CTAB-GAN+ [90]. They can both handle mixed data types, imbalanced datasets, and complex distributions.

GANBLR and GANBLR++ [97, 98] address the fact that GANs are not interpretable and do not exploit any prior knowledge on explicit feature interactions. C3-TGAN [99] introduces mechanisms to preserve explicit attribute correlations and

property constraints. Both approaches use Bayesian networks.

Most of the approaches for time series data use Recurrent Neural Networks (RNN)s [100], especially of the type Long Short Term Memory (LSTM) [101]. TimeGAN [102] combines a GAN model with Autoregressive models but it chunks the dataset into 24 epochs, which is not adequate for long-term dependencies [47]. DoppelGANger [47] is a custom workflow developed to address the key challenges of time series GAN approaches, such as long-term dependencies, complex multidimensional relationships, mode collapse, and privacy.

Normalizing Flows (NF) use invertible and differentiable transformations to convert simple distributions, such as Gaussians, into complex ones for probabilistic density modeling. This process is flexible and allows exact likelihood estimation but is computationally intensive. For this reason, there are not many NF TDS tools. Durkan et al. [103] demonstrated the effectiveness of NF on tabular and image data synthesis. Yet, Manousakas et al. [104] reports that NF underperform compared to models such as CTGAN [84] and TVAE [84]. Kamthe et al. [105] applied NF to learn the copula density for TDS, effectively capturing relations among columns. However, it performs worse than TVAE [84].

Graph Neural Networks (GNN) are neural network architectures for processing graph-structured data. They handle irregular data structures using relationships between entities (nodes) and their connections (edges) in a graph. For TDS, records are converted into graph nodes, connected by edges based on their similarity or domain-specific knowledge. This transformation allows GNNs to learn node representations that capture inter-record and inter-column correlations.

GOGGLE [106] is a TDS tool that replaces typical VAE decoder architectures with GNNs. It achieves realistic samples, highlighting the potential of GNNs in the synthesis of complex, domain-aware tabular data. However, GNNs can be computationally and memory-intensive, especially with large graphs [107].

Diffusion Probabilistic Models (DPM) [26] are inspired by non-equilibrium physics and have gained significant importance with the improvements introduced by Yang et al. [27] and Ho et al. [28]. They involve a two-step process where a backward denoising step is trained to remove the noise previously added by a forward diffusion step. DPMs model the data generation process as a reverse diffusion process, where noise is iteratively removed from a random initialization until a sample from the target distribution emerges [26, 28].

DPMs can be classified into three categories: Denoising Diffusion Probabilistic Models (DDPM), Score-based Generative Models (SGM), and Stochastic Differential Equations (SDE). They differ in how they transform noise into data records. DDPMs take a step-by-step approach, gradually refining noise. SGMs use the gradient of the data distribution to directly guide noise towards the outcome. Meanwhile, SDEs treat this transformation as a continuous process, modeling the addition and removal of noise through differential equations.

TDS diffusion tools, such as TabDDPM [65], SOS [108], and STaSy [109], are able to preserve complex distributions and correlations, and are reported to outperform simpler tools, such as TVAE [84] and TableGAN [41] in terms of ML utility. However, they are computationally more expensive than other deep generative alternatives [110]. TSGM [111] is an example for multivariate time series generation using an SGM. It generates records conditioned on past generated observations.

Transformer-based use an encoder-decoder architecture [34] that revolutionized the field of natural language processing by replacing traditional RNNs and Convolutional Neural Networks (CNN)s with attention mechanisms [33]. This allows each element of a sequence to focus on any other elements of the same sequence, effectively capturing long-range dependencies.

Currently, there are a few transformer-based TDS tools, with GReaT [112], REaLTabFormer [113], and TabuLa [114] being notable examples. They all use a pre-trained Large Language Models (LLM) consisting of only a decoder. GReaT uses the LLM GPT-2 and transforms tabular datasets into textual representations before providing them to the LLM (fine-tuning and inference). This step minimizes the required data pre-processing. REaLTabFormer also uses GPT-2 and addresses the generation of synthetic datasets with two tables being in a one-to-many relationship (i.e., one parent and one child table). They aim to reduce extensive fine-tuning, especially for child tables. The authors of TabuLa emphasize that LLM-based TDS tools provide two main advantages: elimination of the need to pre-define column types and elimination of the dimension explosion problem when synthesizing high-dimensional data. However, LLM-based tools have limitations on training efficiency and preserving column correlations [114].

PROBABILISTIC DATABASE-BASED

Ge et al. [42] remark that most (deep) generation models fail to preserve integrity constraints of the input data in the synthetic output data. To address this issue, they developed a constraint-aware differentially private data synthesis approach called KAMINO. KAMINO preserves denial constraints specified by the user. Similar to VAEs, it first uses the input dataset to learn a latent space and then uses this space to sample the synthetic dataset. The difference is that KAMINO represents this space by a factorized probabilistic database [115] and takes constraint violations into account when sampling the synthetic values one after another. By learning weights for the individual constraints, KAMINO also allows the modeling of soft constraints, which do not strictly have to be fulfilled. Experiments show that KAMINO produces much fewer constraint violations than other privacy-focused approaches, such as PrivBayes [76], DP-VAEGM [88], and PATE-GAN [94]. However, since KAMINO explicitly checks for constraint violations during sampling, it has longer execution times.

Hybrid Models & Other Approaches

Some state-of-the-art TDS tools use combinations of different TDS models. Examples of such hybrid tools are AutoDiff [116] and TabSyn [117], which combine VAEs with a diffusion model to improve the performance of typical DDPMs for different column types and distributions. They are also designed to reduce runtime compared to typical diffusion tools.

Gilad et al. [118] proposed a method that uses already-generated tables and connects them via foreign keys while considering cardinality and integrity constraints. In combination with an existing TDS approach, this allows the generation of datasets with complex schemas.

3.1.5. Summary & Discussion

3.2. Tabular Data Synthesis Platforms

Since a di

The Synthetic Data Vault (SDV) [119], Gretel AI [120], and Mostly AI [121] are platforms for the generation of tabular data. These platforms must choose from the available tools to address the widest range of use cases possible. However, these platforms do not report on the specific limitations of those tools. In contrast, our goal is to create a framework that allows to identify use-case specific requirements and determine those limitations.

3.3. Surveys

Since a di

Several surveys served as input to our work to identify the predominant TDS models and tools. Hernandez et al. [122], Fan et al. [123], Figueira et al, [124], and Brophy et al. [46] explored the use of Generative Adversarial Networks (GANs) for health records, categorical and numerical data types, and time series generation. Koo and Kim [125] reviewed generative diffusion models for tabular data, paralleling Lin et al. [110], who focused on time-series diffusion. Fonseca and Bacao [126] recently provide an extensive survey on tabular data synthesis including an evaluation of 70 tools across six different machine learning problems. However, while our focus is on various tools from the field of generative deep learning (including GANs, autoencoders, probabilistic diffusion, graph neural networks, and transformers), the only deep learning approaches they consider are GANs and autoencoders. Moreover, they do not address the problem of finding the most suitable tool for a specific use case and therefore do neither define functional and non-functional requirements for tabular data synthesis nor evaluate their tools in terms of those requirements.

In summary, none of these surveys provide a comparison of deep learning approaches (see Figure XX as we do in this paper. Additionally, they do not provide any insights into how users can assess a tool's fitness for use, or guide them in the process of choosing a suitable TDS tool for their specific use case.

3.4. EVALUATION OF TABULAR DATA SYNTHESIS TOOLS AND SYNTHETHIC DATA

SINCE a di

Evaluating generative models is challenging because the appropriateness of an evaluation metric depends heavily on the intended use of the generated data. As highlighted in Chapter 20.14 of Deep Learning by Ian Goodfellow, Yoshua Bengio, and Aaron Courville, different applications prioritize different aspects of generative performance, such as fidelity, diversity, or semantic relevance. A single metric often fails to capture all these aspects comprehensively. For example, a generative model intended for data augmentation in classification tasks might prioritize producing diverse samples that enhance classifier performance, while a model generating synthetic images for visualization might focus on high perceptual quality.

This complexity is further emphasized in Theis et al.s 2015 paper, A note on the evaluation of generative models. The authors argue that there is no universal evaluation metric for generative models because different applications demand different trade-offs between fidelity and diversity. Moreover, they highlight the limitations of commonly used metrics, such as the inability of some to detect mode collapse or their reliance on pre-trained discriminative models, which might not align with the target task. Thus, the evaluation of generative models must be carefully designed with the specific application in mind, making it a nuanced and context-dependent process.

3.5. HISTORICAL OVERVIEW & DISCUSSION

4

Functional and Non-Functional Requirements in Tabular Data Synthesis

[1]

This is an abstract of every chapter I will write to let people know what this chapter is about.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

[1] Parts of this chapter have been published in [**davila_navigating_2024??**].

Here the explanation that I did a thorough assessment of implementations and applications, in order to identify use case specific requirements. Why? Because it did not exist...

4.1. METHODOLOGY

SINCE a di

4.2. FUNCTIONAL REQUIREMENTS

SINCE a di

4

4.2.1. COLUMN TYPES

SINCE a di

4.2.2. COLUMN DISTRIBUTION

SINCE a di

4.2.3. CORRELATIONS

SINCE a di

INTER-TABLE CORRELATIONS

4.2.4. TIME DEPENDENCIES

SINCE a di

4.3. NON-FUNCTIONAL REQUIREMENTS

SINCE a di

4.4. SUMMARY & DISCUSSION

This is a table of the tools I am going to consider from here on as the state of the art of each model and their quick comparison.

5

EXPERIMENTS AND EVALUATION

[1]

Change reference when published

This is an abstract of every chapter I will write to let people know what this chapter is about.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

[1] Parts of this chapter have been published in [48].

5.1. EXPERIMENTAL SETUP

SINCE a di

5.2. METHODOLOGY

SINCE a di

5.2.1. TOOL SELECTION

5.2.2. DATASET SELECTION

5.2.3. METRICS

5.2.4. EXPERIMENTS

5.3. PROCEDURE

SINCE a di

5.3.1. DATASET PREPARATION

5.3.2. EXPERIMENT PIPELINE

5.4. RESULTS & DISCUSSION

SINCE a di

5.4.1. COLUMN DISTRIBUTIONS

5.4.2. CORRELATIONS

5.4.3. DATASET IMBALANCE

5.4.4. DATASET AUGMENTATION

5.4.5. MISSING VALUES

5.4.6. PRIVACY

5.4.7. MACHINE LEARNING UTILITY

5.4.8. PERFORMANCE

5.5. SUMMARY & DISCUSSION

6

TABULAR DATA SYNTHESIS MODEL INSIGHTS

[1]

Change reference when published

This is an abstract of every chapter I will write to let people know what this chapter is about.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

[1] Parts of this chapter have been published in [48].

6.1. EVALUATION FRAMEWORK

SINCE a di

6.2. RESULTS & DISCUSSION

SINCE a di

7

DECISION GUIDE PLATFORM

[1]

Change reference when published

This is an abstract of every chapter I will write to let people know what this chapter is about.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

[1] Parts of this chapter have been published in [48].

7.1. DATA GUIDE CRITERIA

SINCE a di

7.2. PLATFORM

SINCE a di

7.3. SUMMARY & DISCUSSION

8

CONCLUSION

This is a concluding chapter explaining the scientific and technical implications for society of the research findings in considerable detail.

BIBLIOGRAPHY

[1] L. Boltzmann. *Lectures on Gas Theory*. University of California Press, Nov. 15, 2023. ISBN: 978-0-520-32747-4. DOI: 10.1525/9780520327474. URL: https://www.degruyter.com/document/doi/10.1525/9780520327474/html?lang=en&srsltid=AfmBOop_9eUQK-LDeVsgF7ErKPr9ff53-loUv5ia7kAv4DW6P7z2oQAu (visited on 01/23/2025).

[2] A. N. Kolmogorov. *Foundations of the Theory of Probability: Second English: Second English Edition*. 2nd edition. Mineola, New York: Dover Publications Inc., Apr. 18, 2018. 96 pp. ISBN: 978-0-486-82159-7.

[3] P. M. Morse, G. E. Kimball, and S. I. Gass. *Methods of Operations Research*. Illustrated edition. Mineola, N.Y: DOVER PUBN INC, Oct. 17, 2003. 176 pp. ISBN: 978-0-486-43234-2.

[4] N. Metropolis and S. Ulam. "The Monte Carlo Method". In: *Journal of the American Statistical Association* 44.247 (Sept. 1, 1949). Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1949.10483310, pp. 335–341. ISSN: 0162-1459. DOI: 10.1080/01621459.1949.10483310. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1949.10483310 (visited on 01/23/2025).

[5] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955". In: *AI Magazine* 27.4 (Dec. 15, 2006). Number: 4, pp. 12–12. ISSN: 2371-9621. DOI: 10.1609/aimag.v27i4.1904. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904 (visited on 01/23/2025).

[6] A. M. Turing. "Computing Machinery and Intelligence". In: *Mind* 59.236 (1950). Publisher: [Oxford University Press, Mind Association], pp. 433–460. ISSN: 0026-4423. URL: https://www.jstor.org/stable/2251299 (visited on 01/23/2025).

[7] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Aug. 1988. 552 pp. ISBN: 978-1-55860-479-7.

[8] J. Pearl. *Bayesian Networks: A Model of Self-activated Memory for Evidential Reasoning*. Google-Books-ID: 1sfMOgAACAAJ. UCLA Computer Science Department, 1985. 20 pp.

[9] R. A. Howard. *Dynamic programming and Markov processes*. Google-Books-ID: fXJEAAAAIAAJ. Technology Press of Massachusetts Institute of Technology, 1960. 154 pp.

[10] "The Finite Element Method: Its Basis and Fundamentals". In: *The Finite Element Method: its Basis and Fundamentals (Seventh Edition)*. Ed. by O. C. Zienkiewicz, R. L. Taylor, and J. Z. Zhu. Oxford: Butterworth-Heinemann, Jan. 1, 2013, p. i. ISBN: 978-1-85617-633-0. DOI: 10.1016/B978-1-85617-633-0.00019-8. URL: https://www.sciencedirect.com/science/article/pii/B9781856176330000198 (visited on 01/23/2025).

[11] Anderson. *Computational Fluid Dynamics: The Basics With Applications*. International Ed edition. New York: McGraw-Hill Higher Education, Jan. 1, 1995. 672 pp. ISBN: 978-0-07-113210-7.

[12] Intel. *Explore Intels history- Raising the Bar with the 386*. URL: https://timeline.intel.com/1985/raising-the-bar-with-the-386 (visited on 01/23/2025).

[13] R. M. Russell. "The CRAY-1 computer system". In: *Commun. ACM* 21.1 (Jan. 1, 1978), pp. 63–72. ISSN: 0001-0782. DOI: 10.1145/359327.359336. URL: https://dl.acm.org/doi/10.1145/359327.359336 (visited on 01/23/2025).

[14] A. Samuel. *Some Studies in Machine Learning Using the Game of Checkers | IBM Journals & Magazine | IEEE Xplore*. URL: https://ieeexplore.ieee.org/document/5392560 (visited on 01/23/2025).

[15] F. Rosenblatt. "The perceptron: A probabilistic model for information storage and organization in the brain". In: *Psychological Review* 65.6 (1958). Place: US Publisher: American Psychological Association, pp. 386–408. ISSN: 1939-1471. DOI: 10.1037/h0042519.

[16] W. S. McCulloch and W. Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4 (Dec. 1, 1943), pp. 115–133. ISSN: 1522-9602. DOI: 10.1007/BF02478259. URL: https://doi.org/10.1007/BF02478259 (visited on 01/23/2025).

[17] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar. "Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data". In: *IEEE Transactions on Knowledge and Data Engineering* 29.10 (Oct. 2017). Conference Name: IEEE Transactions on Knowledge and Data Engineering, pp. 2318–2331. ISSN: 1558-2191. DOI: 10.1109/TKDE.2017.2720168. URL: https://ieeexplore.ieee.org/document/7959606 (visited on 01/23/2025).

[18] Z. Al-Ars, O. Agba, Z. Guo, C. Boerkamp, Z. Jaber, and T. Jaber. "NLICE: Synthetic Medical Record Generation for Effective Primary Healthcare Differential Diagnosis". In: *2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)*. 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE). ISSN: 2471-7819. Dec. 2023, pp. 397–402. DOI: 10.1109/BIBE60311.2023.00071. URL: https://ieeexplore.ieee.org/document/10431883/authors#authors (visited on 01/23/2025).

[19] Z. Song, Z. He, X. Li, Q. Ma, R. Ming, Z. Mao, H. Pei, L. Peng, J. Hu, D. Yao, and Y. Zhang. "Synthetic Datasets for Autonomous Driving: A Survey". In: *IEEE Transactions on Intelligent Vehicles* 9.1 (Jan. 2024). Conference Name: IEEE Transactions on Intelligent Vehicles, pp. 1847–1864. ISSN: 2379-8904. DOI: 10.1109/TIV.2023.3331024. URL: https://ieeexplore.ieee.org/document/10313052 (visited on 01/23/2025).

[20] V. K. Potluru, D. Borrajo, A. Coletta, N. Dalmasso, Y. El-Laham, E. Fons, M. Ghassemi, S. Gopalakrishnan, V. Gosai, E. Kreai, G. Mani, S. Obitayo, D. Paramanand, N. Raman, M. Solonin, S. Sood, S. Vyetrenko, H. Zhu, M. Veloso, and T. Balch. *Synthetic Data Applications in Finance*. Mar. 20, 2024. DOI: 10.48550/arXiv.2401.00081. arXiv: 2401.00081[cs]. URL: http://arxiv.org/abs/2401.00081 (visited on 01/23/2025).

[21] J. Gray, P. Sundaresan, S. Englert, K. Baclawski, and P. J. Weinberger. "Quickly generating billion-record synthetic databases". In: *SIGMOD Rec.* 23.2 (May 24, 1994), pp. 243–252. ISSN: 0163-5808. DOI: 10.1145/191843.191886. URL: https://dl.acm.org/doi/10.1145/191843.191886 (visited on 01/23/2025).

[22] A. Neufeld, G. Moerkotte, and P. C. Loekemann. "Generating consistent test data: Restricting the search space by a generator formula". In: *The VLDB Journal* 2.2 (Apr. 1993), pp. 173–213. ISSN: 1066-8888, 0949-877X. DOI: 10.1007/BF01232186. URL: http://link.springer.com/10.1007/BF01232186 (visited on 01/23/2025).

[23] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Cambridge, Massachusetts: The MIT Press, Nov. 18, 2016. 800 pp. ISBN: 978-0-262-03561-3.

[24] L. Theis, A. v. d. Oord, and M. Bethge. *A note on the evaluation of generative models*. Apr. 24, 2016. DOI: 10.48550/arXiv.1511.01844. arXiv: 1511.01844[stat]. URL: http://arxiv.org/abs/1511.01844 (visited on 01/24/2025).

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial networks". In: *Commun. ACM* 63.11 (Oct. 22, 2020), pp. 139–144. ISSN: 0001-0782. DOI: 10.1145/3422622. URL: https://dl.acm.org/doi/10.1145/3422622 (visited on 01/28/2025).

[26] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *Proceedings of the 32nd International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 1938-7228. PMLR, June 1, 2015, pp. 2256–2265. URL: https://proceedings.mlr.press/v37/sohl-dickstein15.html (visited on 01/28/2025).

[27] Y. Song and S. Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 1067. Red Hook, NY, USA: Curran Associates Inc., Dec. 8, 2019, pp. 11918–11930. (Visited on 01/28/2025).

[28] J. Ho, A. Jain, and P. Abbeel. “Denoising diffusion probabilistic models”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., Dec. 6, 2020, pp. 6840–6851. ISBN: 978-1-7138-2954-6. (Visited on 01/28/2025).

[29] T. Karras, S. Laine, and T. Aila. “A Style-Based Generator Architecture for Generative Adversarial Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.12 (Dec. 2021), pp. 4217–4228. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2020.2970919. URL: https://ieeexplore.ieee.org/document/8977347 (visited on 01/30/2025).

[30] O. Ronneberger, P. Fischer, and T. Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: ed. by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 9783319245744. DOI: 10.1007/978-3-319-24574-4_28.

[31] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. “CARLA: An Open Urban Driving Simulator”. In: *Proceedings of the 1st Annual Conference on Robot Learning*. Conference on Robot Learning. PMLR, Oct. 18, 2017, pp. 1–16. URL: https://proceedings.mlr.press/v78/dosovitskiy17a.html (visited on 01/30/2025).

[32] OpenAI. *GPT-4 Technical Report*. Version 2023. URL: https://arxiv.org/abs/2303.08774.

[33] D. Bahdanau, K. Cho, and Y. Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. May 19, 2016. DOI: 10.48550/arXiv.1409.0473. arXiv: 1409.0473[cs]. URL: http://arxiv.org/abs/1409.0473 (visited on 01/30/2025).

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser, and I. Polosukhin. “Attention is all you need”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., Dec. 4, 2017, pp. 6000–6010. ISBN: 978-1-5108-6096-4. (Visited on 01/28/2025).

[35] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. *Self-Attention Generative Adversarial Networks*. June 14, 2019. DOI: 10.48550/arXiv.1805.08318. arXiv: 1805.08318[stat]. URL: http://arxiv.org/abs/1805.08318 (visited on 01/30/2025).

[36] A. Khang, V. Abdullayev, A. V. Alyar, M. Khalilov, and B. Murad. “AI-Aided Data Analytics Tools and Applications for the Healthcare Sector”. In: IGI Global Scientific Publishing, 2023, pp. 295–313. ISBN: 9798369308769. DOI: 10.4018/979-8-3693-0876-9.ch018. URL: https://www.igi-global.com/chapter/ai-aided-data-analytics-tools-and-applications-for-the-healthcare-sector/www.igi-global.com/chapter/ai-aided-data-analytics-tools-and-applications-for-the-healthcare-sector/332841 (visited on 01/30/2025).

[37] N. L. Rane, S. K. Mallick, Ö. Kaya, and J. Rane. *Applications of machine learning in healthcare, finance, agriculture, retail, manufacturing, energy, and transportation: A review*. Deep Science Publishing, Oct. 13, 2024. DOI: 10.70593/978-81-981271-4-3. URL: https://deepscienceresearch.com/index.php/dsr/catalog/book/6/chapter/63 (visited on 01/30/2025).

[38] E. F. Codd. "A relational model of data for large shared data banks". In: *Commun. ACM* 13.6 (June 1, 1970), pp. 377–387. ISSN: 0001-0782. DOI: 10.1145/362384.362685. URL: https://dl.acm.org/doi/10.1145/362384.362685 (visited on 01/30/2025).

[39] H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database Systems: The Complete Book*. 2nd ed. USA: Prentice Hall Press, May 2008. 1248 pp. ISBN: 9780131873254.

[40] F. Naumann, J.-C. Freytag, and U. Leser. "Completeness of integrated information sources". In: *Information Systems*. Data Quality in Cooperative Information Systems 29.7 (Oct. 1, 2004), pp. 583–615. ISSN: 0306-4379. DOI: 10.1016/j.is.2003.12.005. URL: https://www.sciencedirect.com/science/article/pii/S0306437904000043 (visited on 01/28/2025).

[41] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim. "Data synthesis based on generative adversarial networks". In: *Proc. VLDB Endow.* 11.10 (June 1, 2018), pp. 1071–1083. ISSN: 2150-8097. DOI: 10.14778/3231751.3231757. URL: https://doi.org/10.14778/3231751.3231757 (visited on 01/28/2025).

[42] C. Ge, S. Mohapatra, X. He, and I. F. Ilyas. "Kamino: constraint-aware differentially private data synthesis". In: *Proc. VLDB Endow.* 14.10 (June 1, 2021), pp. 1886–1899. ISSN: 2150-8097. DOI: 10.14778/3467861.3467876. URL: https://doi.org/10.14778/3467861.3467876 (visited on 01/28/2025).

[43] C. Dwork. "Differential Privacy". In: *Automata, Languages and Programming*. Ed. by M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener. Berlin, Heidelberg: Springer, 2006, pp. 1–12. ISBN: 978-3-540-35908-1. DOI: 10.1007/11787006_1.

[44] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". In: *J. Artif. Int. Res.* 16.1 (June 1, 2002), pp. 321–357. ISSN: 1076-9757.

[45] M. S. Milosevic and V. M. Ciric. "Extreme minority class detection in imbalanced data for network intrusion". In: *Computers & Security* 123 (Dec. 1, 2022), p. 102940. ISSN: 0167-4048. DOI: 10.1016/j.cose.2022.102940. URL: https://www.sciencedirect.com/science/article/pii/S0167404822003327 (visited on 01/28/2025).

[46] E. Brophy, Z. Wang, Q. She, and T. Ward. "Generative Adversarial Networks in Time Series: A Systematic Literature Review". In: *ACM Comput. Surv.* 55.10 (Feb. 2, 2023), 199:1–199:31. ISSN: 0360-0300. DOI: 10.1145/3559540. URL: https://dl.acm.org/doi/10.1145/3559540 (visited on 01/28/2025).

[47] Z. Lin, A. Jain, C. Wang, G. Fanti, and V. Sekar. “Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions”. In: *Proceedings of the ACM Internet Measurement Conference*. IMC ’20. New York, NY, USA: Association for Computing Machinery, Oct. 27, 2020, pp. 464–483. ISBN: 978-1-4503-8138-3. DOI: 10.1145/3419394.3423643. URL: https://dl.acm.org/doi/10.1145/3419394.3423643 (visited on 01/28/2025).

[48] M. F. Davila R., W. Wingerath, and F. Panse. “Benchmarking Tabular Data Synthesis for User Guidance”. In: EDBT/ICDT 2024 Joint Conference. Mar. 1, 2024. URL: https://ceur-ws.org/Vol-3651/PhDW-2.pdf.

[49] M. F. Davila R., T. Hoiten, P. Sander, N. Woltering, and A. Pehlken. “Container-Based Microservices Application for Product Carbon Footprint Calculation in Manufacturing Companies”. In: *EnviroInfo 2024*. May 2025.

[50] A. Pehlken, M. Davila, L. Dawel, and O. Meyer. “Digital Twins: Enhancing Circular Economy through Digital Tools”. In: *Procedia CIRP* 122 (Jan. 1, 2024), pp. 563–568. DOI: 10.1016/j.procir.2024.01.082.

[51] M. F. Davila R., F. Schwark, L. Dawel, and A. Pehlken. “Sustainability Digital Twin: a tool for the manufacturing industry”. In: *Procedia CIRP*. 30th CIRP Life Cycle Engineering Conference 116 (Jan. 1, 2023), pp. 143–148. ISSN: 2212-8271. DOI: 10.1016/j.procir.2023.02.025. URL: https://www.sciencedirect.com/science/article/pii/S2212827123000252 (visited on 01/30/2025).

[52] F. Schwark, H. Garmatter, M. Davila R., L. Dawel, A. Pehlken, F. Cyris, and R. Scharf. “The application of image recognition methods to improve the performance of waste-to-energy plantsplants”. In: EnviroInfo 2022. Gesellschaft für Informatik e.V., 2022, p. 167. ISBN: 978-3-88579-722-7. URL: https://dl.gi.de/handle/20.500.12116/39413 (visited on 01/30/2025).

[53] H. Torio, A. Günther, M. F. Davila, and M. Knipper. “Paving the Way for Hybrid Teaching in Higher Education: Lessons from Students Perceptions and Acceptance of Different Teaching Modes during and after the Pandemic”. In: *Creative Education* 14.5 (May 15, 2023), pp. 1029–1042. DOI: 10.4236/ce.2023.145066. URL: https://www.scirp.org/journal/paperinformation?paperid=125257 (visited on 01/30/2025).

[54] F. Penaherrera, M. F. Davila R., A. Pehlken, and B. Koch. “Quantifying the Environmental Impacts of Battery Electric Vehicles from a Criticality Perspective”. In: *2022 IEEE 28th International Conference on Engineering, Technology and Innovation (ICE/ITMC) & 31st International Association For Management of Technology (IAMOT) Joint Conference*. 2022 IEEE 28th International Conference on Engineering, Technology and Innovation (ICE/ITMC) & 31st International Association For Management of Technology (IAMOT) Joint Conference. June 2022, pp. 1–9. DOI: 10.1109/ICE/ITMC-IAMOT55089.2022.10033264. URL: https://ieeexplore.ieee.org/document/10033264 (visited on 01/30/2025).

[55] P. Domingos. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Google-Books-ID: pjRkCQAAQBAJ. Penguin UK, Sept. 22, 2015. 353 pp. ISBN: 978-0-241-00455-5.

[56] J. M. Tomczak. *Deep Generative Modeling*. Google-Books-ID: uidgEAAAQBAJ. Springer Nature, Feb. 18, 2022. 210 pp. ISBN: 978-3-030-93158-2.

[57] D. Foster. *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play*. Google-Books-ID: RqegDwAAQBAJ. "O'Reilly Media, Inc.", June 28, 2019. 301 pp. ISBN: 978-1-4920-4189-4.

[58] D. B. Rubin. "MULTIPLE IMPUTATIONS IN SAMPLE SURVEYS-A PHENOMENOLOGICAL BAYESIAN APPROACH TO NONRESPONSE". In: 2002. URL: https://www.semanticscholar.org/paper/MULTIPLE-IMPUTATIONS-IN-SAMPLE-SURVEYS-A-BAYESIAN-Rubin/f0034e9045688520e87769d39bf7b8c69c26612c (visited on 01/28/2025).

[59] S. v. Buuren and K. Groothuis-Oudshoorn. "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45 (Dec. 12, 2011), pp. 1–67. ISSN: 1548-7660. DOI: 10.18637/jss.v045.i03. URL: https://doi.org/10.18637/jss.v045.i03 (visited on 01/28/2025).

[60] J. Honaker, G. King, and M. Blackwell. "Amelia II: A Program for Missing Data". In: *Journal of Statistical Software* 45 (Dec. 12, 2011), pp. 1–47. ISSN: 1548-7660. DOI: 10.18637/jss.v045.i07. URL: https://doi.org/10.18637/jss.v045.i07 (visited on 01/28/2025).

[61] T. Raghunathan, J. P. Reiter, and D. Rubin. "Multiple Imputation for Statistical Disclosure Limitation". In: *Journal of Official Statistics* (2003). URL: https://www.semanticscholar.org/paper/Multiple-Imputation-for-Statistical-Disclosure-Raghunathan-Reiter/69053e006f419362b308b07f204f22519db (visited on 01/28/2025).

[62] J. P. Reiter. "Satisfying Disclosure Restrictions With Synthetic Data Sets". In: 2002. URL: https://www.semanticscholar.org/paper/Satisfying-Disclosure-Restrictions-With-Synthetic-Reiter/8b2938329330be136efc6448878e02b (visited on 01/28/2025).

[63] P. Kowalczyk, G. Welsch, and F. Thiesse. "Towards a Taxonomy for the Use of Synthetic Data in Advanced Analytics". In: (2022). Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.2212.02622. URL: https://arxiv.org/abs/2212.02622 (visited on 01/28/2025).

[64] J. Brandt and E. Lanzén. *A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification*. 2021. URL: https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-432162 (visited on 01/28/2025).

[65] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko. "TabDDPM: modelling tabular data with diffusion models". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. ICML'23. Honolulu, Hawaii, USA: JMLR.org, July 23, 2023, pp. 17564–17579. (Visited on 01/28/2025).

[66] *SMOTE Version 0.13.0*. URL: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html (visited on 01/28/2025).

[67] H. Han, W.-Y. Wang, and B.-H. Mao. "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning". In: *Advances in Intelligent Computing*. Ed. by D.-S. Huang, X.-P. Zhang, and G.-B. Huang. Berlin, Heidelberg: Springer, 2005, pp. 878–887. ISBN: 978-3-540-31902-3. DOI: 10.1007/11538059_91.

[68] H. Sain and S. W. Purnami. "Combine Sampling Support Vector Machine for Imbalanced Data Classification". In: *Procedia Computer Science*. The Third Information Systems International Conference 2015 72 (Jan. 1, 2015), pp. 59–66. ISSN: 1877-0509. DOI: 10.1016/j.procs.2015.12.105. URL: https://www.sciencedirect.com/science/article/pii/S1877050915035668 (visited on 01/28/2025).

[69] G. Douzas, F. Bacao, and F. Last. "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE". In: *Information Sciences* 465 (Oct. 1, 2018), pp. 1–20. ISSN: 0020-0255. DOI: 10.1016/j.ins.2018.06.056. URL: https://www.sciencedirect.com/science/article/pii/S0020025518304997 (visited on 01/28/2025).

[70] H. He, Y. Bai, E. A. Garcia, and S. Li. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). ISSN: 2161-4407. June 2008, pp. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969. URL: https://ieeexplore.ieee.org/document/4633969 (visited on 01/28/2025).

[71] K. Liu, C. Giannella, and H. Kargupta. "An Attackers View of Distance Preserving Maps for Privacy Preserving Data Mining". In: *Knowledge Discovery in Databases: PKDD 2006*. Ed. by J. Fürnkranz, T. Scheffer, and M. Spiliopoulou. Berlin, Heidelberg: Springer, 2006, pp. 297–308. ISBN: 978-3-540-46048-0. DOI: 10.1007/11871637_30.

[72] B. Van Vliet. *Abe Sklar's*. Rochester, NY, Mar. 3, 2023. DOI: 10.2139/ssrn.4198458. URL: https://papers.ssrn.com/abstract=4198458 (visited on 01/27/2025).

[73] *TabularCopula*. Tabular Copula. 2023. URL: https://biomeddar.github.io/copula-tabular//copula-tabular/ (visited on 01/27/2025).

[74] D. X. Li. *On Default Correlation: A Copula Function Approach*. Rochester, NY, Sept. 1, 1999. DOI: 10.2139/ssrn.187289. URL: https://papers.ssrn.com/abstract=187289 (visited on 01/27/2025).

[75] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Google-Books-ID: 7dzpHCHzNQ4C. MIT Press, July 31, 2009. 1268 pp. ISBN: 9780262013192.

[76] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. "PrivBayes: Private Data Release via Bayesian Networks". In: *ACM Trans. Database Syst.* 42.4 (Oct. 27, 2017), 25:1–25:41. ISSN: 0362-5915. DOI: 10.1145/3134428. URL: https://dl.acm.org/doi/10.1145/3134428 (visited on 01/27/2025).

[77] J. Besag. "Spatial Interaction and the Statistical Analysis of Lattice Systems". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 36.2 (1974), pp. 192–236. ISSN: 0035-9246. URL: https://www.jstor.org/stable/2984812 (visited on 01/27/2025).

[78] K. Cai, X. Lei, J. Wei, and X. Xiao. "Data synthesis via differentially private markov random fields". In: *Proc. VLDB Endow.* 14.11 (July 1, 2021), pp. 2190–2202. ISSN: 2150-8097. DOI: 10.14778/3476249.3476272. URL: https://doi.org/10.14778/3476249.3476272 (visited on 01/27/2025).

[79] K. Cai, X. Xiao, and G. Cormode. "PrivLava: Synthesizing Relational Data with Foreign Keys under Differential Privacy". In: *Proc. ACM Manag. Data* 1.2 (June 20, 2023), 142:1–142:25. DOI: 10.1145/3589287. URL: https://dl.acm.org/doi/10.1145/3589287 (visited on 01/27/2025).

[80] D. P. Kingma and M. Welling. "Auto-Encoding Variational Bayes". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2014. URL: http://arxiv.org/abs/1312.6114 (visited on 01/28/2025).

[81] C. Doersch. *Tutorial on Variational Autoencoders*. Jan. 3, 2021. DOI: 10.48550/arXiv.1606.05908. arXiv: 1606.05908[stat]. URL: http://arxiv.org/abs/1606.05908 (visited on 01/28/2025).

[82] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. *Generating Sentences from a Continuous Space*. May 12, 2016. DOI: 10.48550/arXiv.1511.06349. arXiv: 1511.06349[cs]. URL: http://arxiv.org/abs/1511.06349 (visited on 01/28/2025).

[83] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck. "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music". In: *ArXiv* (Mar. 13, 2018). URL: https://www.semanticscholar.org/paper/A-Hierarchical-Latent-Vector-Model-for-Learning-in-Roberts-Engel/2b050df9e24eb65b0d37f13c6eea1d29b4e316ce (visited on 01/28/2025).

[84] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. "Modeling tabular data using conditional GAN". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 659. Red Hook, NY, USA: Curran Associates Inc., Dec. 8, 2019, pp. 7335–7345. (Visited on 01/28/2025).

[85] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. *Hybrid Models with Deep and Invertible Features*. May 29, 2019. DOI: 10.48550/arXiv.1902.02767. arXiv: 1902.02767[cs]. URL: http://arxiv.org/abs/1902.02767 (visited on 01/28/2025).

8

[86] E. Jang, S. Gu, and B. Poole. "Categorical Reparameterization with Gumbel-Softmax". In: International Conference on Learning Representations. Feb. 6, 2017. URL: https://openreview.net/forum?id=rkE3y85ee (visited on 01/28/2025).

[87] J. T. Rolfe. *Discrete Variational Autoencoders*. Apr. 22, 2017. DOI: 10.48550/arXiv.1609.02200. arXiv: 1609.02200[stat]. URL: http://arxiv.org/abs/1609.02200 (visited on 01/28/2025).

[88] Q. Chen, C. Xiang, M. Xue, B. Li, N. Borisov, D. Kaarfar, and H. Zhu. *Differentially Private Data Generative Models*. Dec. 6, 2018. DOI: 10.48550/arXiv.1812.02274. arXiv: 1812.02274[cs]. URL: http://arxiv.org/abs/1812.02274 (visited on 01/28/2025).

[89] A. Desai, C. Freeman, Z. Wang, and I. Beaver. *TimeVAE: A Variational Auto-Encoder for Multivariate Time Series Generation*. Dec. 7, 2021. DOI: 10.48550/arXiv.2111.08095. arXiv: 2111.08095[cs]. URL: http://arxiv.org/abs/2111.08095 (visited on 01/28/2025).

[90] Z. Zhao, A. Kunar, R. Birke, H. Van der Scheer, and L. Y. Chen. "CTAB-GAN+: enhancing tabular data synthesis". In: *Frontiers in Big Data* 6 (2023), p. 1296508. ISSN: 2624-909X. DOI: 10.3389/fdata.2023.1296508.

[91] M. Mirza and S. Osindero. *Conditional Generative Adversarial Nets*. Nov. 6, 2014. DOI: 10.48550/arXiv.1411.1784. arXiv: 1411.1784[cs]. URL: http://arxiv.org/abs/1411.1784 (visited on 01/28/2025).

[92] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun. "Generating Multi-label Discrete Patient Records using Generative Adversarial Networks". In: *Proceedings of the 2nd Machine Learning for Healthcare Conference*. Machine Learning for Healthcare Conference. ISSN: 2640-3498. PMLR, Nov. 6, 2017, pp. 286–305. URL: https://proceedings.mlr.press/v68/choi17a.html (visited on 01/28/2025).

[93] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou. *Differentially Private Generative Adversarial Network*. Feb. 19, 2018. DOI: 10.48550/arXiv.1802.06739. arXiv: 1802.06739[cs]. URL: http://arxiv.org/abs/1802.06739 (visited on 01/28/2025).

[94] J. Jordon, J. Yoon, and M. v. d. Schaar. "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees". In: International Conference on Learning Representations. Sept. 27, 2018. URL: https://openreview.net/forum?id=S1zk9iRqF7 (visited on 01/28/2025).

[95] L. Xu and K. Veeramachaneni. *Synthesizing Tabular Data using Generative Adversarial Networks*. Nov. 27, 2018. DOI: 10.48550/arXiv.1811.11264. arXiv: 1811.11264[cs]. URL: http://arxiv.org/abs/1811.11264 (visited on 01/28/2025).

[96] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen. "CTAB-GAN: Effective Table Data Synthesizing". In: *Proceedings of The 13th Asian Conference on Machine Learning*. Asian Conference on Machine Learning. ISSN: 2640-3498. PMLR, Nov. 28, 2021, pp. 97–112. URL: https://proceedings.mlr.press/v157/zhao21a.html (visited on 01/28/2025).

[97] Y. Zhang, N. A. Zaidi, J. Zhou, and G. Li. "GANBLR: A Tabular Data Generation Model". In: *2021 IEEE International Conference on Data Mining (ICDM)*. 2021 IEEE International Conference on Data Mining (ICDM). ISSN: 2374-8486. Dec. 2021, pp. 181–190. DOI: 10.1109/ICDM51629.2021.00103. URL: https://ieeexplore.ieee.org/document/9679177 (visited on 01/28/2025).

[98] Y. Zhang, N. Zaidi, J. Zhou, and G. Li. "GANBLR++: Incorporating Capacity to Generate Numeric Attributes and Leveraging Unrestricted Bayesian Networks". In: *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*. Proceedings. Society for Industrial and Applied Mathematics, Jan. 2022, pp. 298–306. DOI: 10.1137/1.9781611977172.34. URL: https://epubs.siam.org/doi/10.1137/1.9781611977172.34 (visited on 01/28/2025).

[99] P. Han, W. Xu, W. Lin, J. Cao, C. Liu, S. Duan, and H. Zhu. "C3-TGAN-Controllable Tabular Data Synthesis with Explicit Correlations and Property Constraints". In: (). URL: https://www.authorea.com/doi/full/10.36227/techrxiv.24249643.v1?commit=6222832164e697238a1ba1593e1184bb59023abb (visited on 01/28/2025).

[100] D. E. Rumelhart and J. L. McClelland. "Learning Internal Representations by Error Propagation". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. Conference Name: Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations. MIT Press, 1987, pp. 318–362. ISBN: 978-0-262-29140-8. URL: https://ieeexplore.ieee.org/document/6302929 (visited on 01/28/2025).

[101] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997). Conference Name: Neural Computation, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://ieeexplore.ieee.org/abstract/document/6795963 (visited on 01/28/2025).

[102] J. Yoon, D. Jarrett, and M. van der Schaar. "Time-series Generative Adversarial Networks". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. URL: https://papers.nips.cc/paper_files/paper/2019/hash/c9efe5f26cd17ba6216bbe2a7d26d490-Abstract.html (visited on 01/28/2025).

[103] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios. “Neural spline flows”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 675. Red Hook, NY, USA: Curran Associates Inc., Dec. 8, 2019, pp. 7511–7522. (Visited on 01/28/2025).

[104] D. Manousakas and S. Aydöre. *On the Usefulness of Synthetic Tabular Data Generation*. June 27, 2023. DOI: 10.48550/arXiv.2306.15636. arXiv: 2306.15636[cs]. URL: http://arxiv.org/abs/2306.15636 (visited on 01/28/2025).

[105] S. Kamthe, S. Assefa, and M. Deisenroth. *Copula Flows for Synthetic Data Generation*. Jan. 3, 2021. DOI: 10.48550/arXiv.2101.00598. arXiv: 2101.00598[stat]. URL: http://arxiv.org/abs/2101.00598 (visited on 01/28/2025).

[106] T. Liu, Z. Qian, J. Berrevoets, and M. v. d. Schaar. “GOGGLE: Generative Modelling for Tabular Data by Learning Relational Structure”. In: The Eleventh International Conference on Learning Representations. Sept. 29, 2022. URL: https://openreview.net/forum?id=fPVRcJqspu&referrer=%5BAuthor%20Console%5D(%2Fgroup%3Fid%3DICLR.cc%2F2023%2FConference%2FAuthors%23your-submissions) (visited on 01/28/2025).

[107] S. Iwata, R. H. Arpaci-Dusseau, and A. Kasagi. “An Analysis of Graph Neural Network Memory Access Patterns”. In: *Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*. SC-W '23. New York, NY, USA: Association for Computing Machinery, Nov. 12, 2023, pp. 915–921. ISBN: 979-8-4007-0785-8. DOI: 10.1145/3624062.3624168. URL: https://doi.org/10.1145/3624062.3624168 (visited on 01/28/2025).

[108] J. Kim, C. Lee, Y. Shin, S. Park, M. Kim, N. Park, and J. Cho. “SOS: Score-based Oversampling for Tabular Data”. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '22. New York, NY, USA: Association for Computing Machinery, Aug. 14, 2022, pp. 762–772. ISBN: 978-1-4503-9385-0. DOI: 10.1145/3534678.3539454. URL: https://doi.org/10.1145/3534678.3539454 (visited on 01/28/2025).

[109] J. Kim, C. Lee, and N. Park. “STaSy: Score-based Tabular data Synthesis”. In: The Eleventh International Conference on Learning Representations. Sept. 29, 2022. URL: https://openreview.net/forum?id=1mNssCWt_v (visited on 01/28/2025).

[110] L. Lin, Z. Li, R. Li, X. Li, and J. Gao. “Diffusion models for time-series applications: a survey”. In: *Frontiers of Information Technology & Electronic Engineering* 25.1 (Jan. 1, 2024), pp. 19–41. ISSN: 2095-9230. DOI: 10.1631/FITEE.2300310. URL: https://doi.org/10.1631/FITEE.2300310 (visited on 01/28/2025).

[111] H. Lim, M. Kim, S. Park, J. Lee, and N. Park. "TSGM: Regular and Irregular Time-series Generation using Score-based Generative Models". In: (Oct. 13, 2023). URL: https://openreview.net/forum?id=nFG1YmQTqi (visited on 01/28/2025).

[112] V. Borisov, K. Sessler, T. Leemann, M. Pawelczyk, and G. Kasneci. "Language Models are Realistic Tabular Data Generators". In: The Eleventh International Conference on Learning Representations. Sept. 29, 2022. URL: https://openreview.net/forum?id=cEygmQNOeI (visited on 01/28/2025).

[113] A. V. Solatorio and O. Dupriez. *REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers*. Feb. 4, 2023. DOI: 10.48550/arXiv.2302.02041. arXiv: 2302.02041[cs]. URL: http://arxiv.org/abs/2302.02041 (visited on 01/28/2025).

[114] Z. Zhao, R. Birke, and L. Chen. *TabuLa: Harnessing Language Models for Tabular Data Synthesis*. Jan. 10, 2025. DOI: 10.48550/arXiv.2310.12746. arXiv: 2310.12746[cs]. URL: http://arxiv.org/abs/2310.12746 (visited on 01/28/2025).

[115] C. De Sa, I. F. Ilyas, B. Kimelfeld, C. Ré, and T. Rekatsinas. "A Formal Framework for Probabilistic Unclean Databases". In: *LIPIcs, Volume 127, ICDT 2019* 127 (2019). In collab. with P. Barcelo and M. Calautti. Artwork Size: 18 pages, 743419 bytes ISBN: 9783959771016 Medium: application/pdf Publisher: Schloss Dagstuhl Leibniz-Zentrum für Informatik Version Number: 1.0, 6:1–6:18. ISSN: 1868-8969. DOI: 10.4230/LIPICS.ICDT.2019.6. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICDT.2019.6 (visited on 01/28/2025).

[116] N. Suh, X. Lin, D.-Y. Hsieh, M. Honarkhah, and G. Cheng. *AutoDiff: combining Auto-encoder and Diffusion model for tabular data synthesizing*. Nov. 17, 2023. DOI: 10.48550/arXiv.2310.15479. arXiv: 2310.15479[stat]. URL: http://arxiv.org/abs/2310.15479 (visited on 01/28/2025).

[117] H. Zhang, J. Zhang, Z. Shen, B. Srinivasan, X. Qin, C. Faloutsos, H. Rangwala, and G. Karypis. "Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space". In: The Twelfth International Conference on Learning Representations. Oct. 13, 2023. URL: https://openreview.net/forum?id=4Ay23yeuz0 (visited on 01/28/2025).

[118] A. Gilad, S. Patwa, and A. Machanavajjhala. "Synthesizing Linked Data Under Cardinality and Integrity Constraints". In: *Proceedings of the 2021 International Conference on Management of Data*. SIGMOD '21. New York, NY, USA: Association for Computing Machinery, June 18, 2021, pp. 619–631. ISBN: 978-1-4503-8343-1. DOI: 10.1145/3448016.3457242. URL: https://dl.acm.org/doi/10.1145/3448016.3457242 (visited on 01/28/2025).

[119] SDV. *The Synthetic Data Vault. Put synthetic data to work!* URL: https://sdv.dev/ (visited on 01/28/2025).

[120] GRETEL. *The synthetic data platform purpose-built for AI Gretel.ai*. URL: https://gretel.ai/ (visited on 01/28/2025).

[121] MOSTLY. *Synthetic Data Generation with the Highest Accuracy for Free*. May 31, 2024. URL: https://mostly.ai (visited on 01/28/2025).

[122] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin. "Synthetic data generation for tabular health records: A systematic review". In: *Neurocomputing* 493 (July 7, 2022), pp. 28–45. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2022.04.053. URL: https://www.sciencedirect.com/science/article/pii/S0925231222004349 (visited on 01/28/2025).

[123] J. Fan, J. Chen, T. Liu, Y. Shen, G. Li, and X. Du. "Relational data synthesis using generative adversarial networks: a design space exploration". In: *Proc. VLDB Endow.* 13.12 (July 1, 2020), pp. 1962–1975. ISSN: 2150-8097. DOI: 10.14778/3407790.3407802. URL: https://doi.org/10.14778/3407790.3407802 (visited on 01/28/2025).

[124] A. Figueira and B. Vaz. "Survey on Synthetic Data Generation, Evaluation Methods and GANs". In: *Mathematics* 10.15 (Jan. 2022). Number: 15 Publisher: Multidisciplinary Digital Publishing Institute, p. 2733. ISSN: 2227-7390. DOI: 10.3390/math10152733. URL: https://www.mdpi.com/2227-7390/10/15/2733 (visited on 01/28/2025).

[125] H. Koo and T. E. Kim. *A Comprehensive Survey on Generative Diffusion Models for Structured Data*. July 8, 2023. DOI: 10.48550/arXiv.2306.04139. arXiv: 2306.04139[cs]. URL: http://arxiv.org/abs/2306.04139 (visited on 01/28/2025).

[126] J. Fonseca and F. Bacao. "Tabular and latent space synthetic data generation: a literature review". In: *Journal of Big Data* 10.1 (July 10, 2023), p. 115. ISSN: 2196-1115. DOI: 10.1186/s40537-023-00792-7. URL: https://doi.org/10.1186/s40537-023-00792-7 (visited on 01/28/2025).

A

CODE, VALIDATION OR SOMETHING

Parts of this chapter have been published in Annalen der Physik **324**, 289 (1906) [**Einstein1906??**].

This is an abstract of every chapter I will write to let people know what this chapter is about.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

A

A.1. DATA GUIDE CRITERIA

SINCE a di

LIST OF FIGURES

List of Tables

Statutory Declaration / Eidesstattliche Erklärung

English:
I hereby declare, on oath, that I have written the present dissertation entitled

Benchmarking Tabular Data Synthesis: Framework for Evaluating Use-Case Suitability and Performance on Commodity Hardware

by myself and have not used sources or means without declaration in the text. Any thoughts or quotations which were inferred from these sources are clearly marked as such.
This thesis was not submitted in the same or in a substantially similar version, not even partially, to any other authority to achieve an academic grading and was not published elsewhere.
I agree that a copy of this thesis may be made available in the Informatics Library of the University of Oldenburg.

German:
Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertation mit dem Titel

Benchmarking Tabular Data Synthesis: Framework for Evaluating Use-Case Suitability and Performance on Commodity Hardware

selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht.
Diese Dissertation wurde weder in gleicher noch in wesentlicher ähnlicher Form, auch nicht auszugsweise, einer anderen Prüfungsbehörde zur Erlangung eines akademischen Grades vorgelegt oder anderweitig veröffentlicht.
Ich bin damit einverstanden, dass eine Kopie dieser Dissertation in der Informatik-Bibliothek der Universität Oldenburg verfügbar gemacht wird.

Maria Fernanda Davila Restrepo

Date: February 1, 2025