Karst, Fabian Sven; Li, Mahei Manhai; Leimeister, Jan Marco



# SynDEc: A Synthetic Data Ecosystem

Electronic Markets

**Provided in Cooperation with:**
Springer Nature

*Suggested Citation:* Karst, Fabian Sven; Li, Mahei Manhai; Leimeister, Jan Marco (2025) : SynDEc: A Synthetic Data Ecosystem, Electronic Markets, ISSN 1422-8890, Springer, Berlin, Heidelberg, Vol. 35, Iss. 1,
https://doi.org/10.1007/s12525-024-00746-8

Electronic Markets (2025) 35:7
https://doi.org/10.1007/s12525-024-00746-8

RESEARCH PAPER

# SynDEc: A Synthetic Data Ecosystem

**Fabian Sven Karst**[1] · **Mahei Manhai Li**[1,2] · **Jan Marco Leimeister**[1,2]

Received: 22 March 2024 / Accepted: 2 December 2024 / Published online: 25 January 2025


**Abstract**
Given the critical role of data availability for growth and innovation in financial services, especially small and mid-sized banks lack the data volumes required to fully leverage AI advancements for enhancing fraud detection, operational efficiency, and risk management. With existing solutions facing challenges in scalability, inconsistent standards, and complex privacy regulations, we introduce a synthetic data sharing ecosystem (SynDEc) using generative AI. Employing design science research in collaboration with two banks, among them UnionBank of the Philippines, we developed and validated a synthetic data sharing ecosystem for financial institutions. The derived design principles highlight synthetic data setup, training configurations, and incentivization. Furthermore, our findings show that smaller banks benefit most from SynDEcs and our solution is viable even with limited participation. Thus, we advance data ecosystem design knowledge, show its viability for financial services, and offer practical guidance for privacy-resilient synthetic data sharing, laying groundwork for future applications of SynDEcs.

**Keywords** Synthetic data · Data sharing platform · Data ecosystem · Financial services · Data scarcity

**JEL classification** M15

## Motivation

In the wake of recent global crises, the enhancement of financial services has become a crucial driver for accelerating economic recovery, particularly in developing economies where these services are essential for expanding financial inclusion and fostering socioeconomic growth (Demirgüç-Kunt et al., 2022; Pazarbasioglu et al., 2020; White et al., 2021). However, given the financial services industry's reliance on information, increasing data availability is key to success. This is especially true for smaller financial institutions, which lack the necessary volume of high-quality data to leverage current AI model advancements. This lack of data results in missed opportunities, with developing countries potentially losing out on up to 5% of GDP through improvements in fraud protection, operational efficiency, and workforce allocation (White et al., 2021; Zachariadis, 2020).

Although the sharing of financial transaction data could reduce risks and improve transparency (Brodsky & Oakes, 2017), thereby driving economic growth (O'Leary et al., 2021), it faces significant obstacles related to privacy regulation and information security. Existing solutions such as open banking and federated learning have significant limitations. Open banking, which enables customer-approved data exchange between financial institutions, often produces unreliable data due to selective participation (He et al., 2023) and lacks coverage of B2B transactions (Preziuso et al., 2023). Federated learning, an approach for training a model without direct data exchange, faces scalability issues, restricts

Responsible Editor: Gero Strobel

✉ Fabian Sven Karst
Fabian.Karst@unisg.ch

Mahei Manhai Li
Mahei.Li@unisg.ch; Mahei.Li@uni-kassel.de

Jan Marco Leimeister
JanMarco.Leimeister@unisg.ch; Leimeister@uni-kassel.de

[1] University of St.Gallen, Institute of Information Systems and Digital Business, Dufourstrasse 50, 9000 St.Gallen, Switzerland

[2] University of Kassel, Information Systems, Pfannkuchstraße 1, 34121 Kassel, Germany

participants to a single shared model, and lacks adaptability (Baabdullah et al., 2024; Chatterjee et al., 2024). Therefore, research is required to explore data ecosystems that facilitate the exchange of data between financial institutions and regulatory bodies while safeguarding the privacy of individual users' information (Assefa, 2020).

In the pursuit of establishing such an ecosystem enabling financial data sharing, the application of synthetic data generation emerges as a promising solution. Synthetic data, currently primarily used in financial services to tackle class imbalance in fraud detection models by synthesizing new fraudulent samples (Charitou et al., 2021), produces artificial data that if done correctly maintains privacy while capturing and generalizing the patterns and attributes essential for the training of machine learning models. Combining this with data sharing enables the creation of a secure and robust data ecosystem.

While plenty of research on synthetic data generation exists, significant gaps remain for its practical application within data ecosystems. Research has largely focused on algorithm development, leaving critical questions unanswered about how to design an ecosystem for privacy-preserving data exchange with the capability to handle complex data and achieve interoperability across institutions (Oliveira et al., 2019). Additionally, there is limited guidance on which algorithms are most effective in a context where synthetic data is leveraged to be shared between institutions and not merely used to increase the amount of training data (Langevin et al., 2022). Practical strategies for integrating shared synthetic data within machine learning models are also sparse, though such strategies are essential for realizing synthetic data's potential in AI applications (Sattarov et al., 2023; Strelcenia & Prakoonwit, 2023). Finally, incentives, for big as well as small players, necessary to encourage participation in a synthetic data-sharing ecosystem remain underexplored, despite being vital for fostering the cooperative engagement on which such ecosystems depend (Gelhaar & Otto, 2020). In response, our research seeks to answer the following questions: What architecture is best suited for secure data exchange? Which algorithms are most effective for data generation? What are the optimal strategies for utilizing shared synthetic data within individual institutions? And do the incentives within such an ecosystem effectively encourage participation? Furthermore, there is a need for specialized engineering and management methodologies tailored to the unique demands of financial services, where stringent privacy regulations and the complex nature of transaction data introduce distinct challenges (Oliveira et al., 2019).

Our research goal is to provide design knowledge for a synthetic data ecosystem that enables financial institutions to share financial transaction data and generate utility from doing so. Our study contributes to the existing literature in two significant ways. First, it advances the field of data ecosystems by addressing privacy challenges and exploring the use of data from multiple institutions for machine learning (Brée et al., 2024). Second, it offers practical guidance for financial institutions on generating and utilizing synthetic data, including benchmarking different algorithms, setups, and training schemes. Given the current lack of guidance on the conceptualization and implementation of such systems, this leads us to the following research question:

*RQ: How to design a financial data ecosystem (SynDEc) based on synthetic data sharing?*

To address the RQ, the paper adopts a multifaceted approach to investigate architectural design decisions. It encompasses an examination of synthetic data generation techniques within the ecosystem, explores its implications for training predictive models, and seeks to identify and mitigate potential challenges to the ecosystem's stability and functionality. Additionally, it assesses the generalizability of the derived principles beyond the domain of financial fraud detection.

The paper is organized as follows: In the next section, we present an overview of data ecosystems in financial services and synthetic data generation. Next, we outline, the Design Science Research Methodology by Peffers et al. (2007), combining context-driven innovation and iterative development, which we use as our methodological foundation. In the first of our four design cycles, we diagnose the problem space through the meta (MR) and design requirements (DR) based on both literature and expert interviews. Based on this, our initial set of design principles (DP) is derived and instantiated as a system architecture. Building on this the second design cycle evaluates the feasibility of different synthetic data generation and integration methods. The following design cycle extends this by evaluating the proposed approach in new domains while also investigating improvements to the ecosystem based on data generation and exchange. Lastly, design cycle four takes a network view, investigating design elements to ensure early challenges frequently seen in data ecosystems can be overcome. Finally, we discuss the findings, outline limitations, provide a perspective for future work, and conclude with a brief summary.

## Related work

### Data ecosystems in financial services

The growing recognition of data as a critical asset for innovation, growth, and value creation has led firms to increasingly seek external sources to enhance their data capabilities (Bagad et al., 2021; Gelhaar & Otto, 2020). One promising

approach is the formation of inter-organizational networks, where organizations collaborate to share resources and knowledge (Gray & Sites, 2013). Within this context, data ecosystems have emerged as an effective framework for data exchange (Abbas et al., 2021; Heinz et al., 2022; Zuiderwijk et al., 2014). Defined as "a set of networks composed of autonomous actors that directly or indirectly consume, produce, or provide data and other related resources" (Oliveira & Lóscio, 2018, p. 4), data ecosystems are built around four key constructs: (1) actors, (2) their roles, (3) relationships among them, and (4) the resources they require. Actors in these ecosystems—whether organizations, individuals, or institutions—take on roles such as data consumers, providers, and intermediaries, each contributing uniquely to the ecosystem's function (Oliveira & Lóscio, 2018; van Schalkwyk et al., 2016). The roles they assume drive specific tasks, such as data intermediaries connecting various actors and data consumers analyzing and providing feedback to data providers. These interactions, and the dependencies that arise from them, form the relationships that underpin the ecosystem (Heimstädt et al., 2014; Oliveira & Lóscio, 2018). At the core of a data ecosystem, data platforms provide the technical infrastructure for processing and managing data from diverse sources, enabling various data applications. These platforms often incorporate data marketplaces, which serve as self-service platforms that connect data producers and consumers (Gröger, 2021). Another closely related concept is data spaces, which are frequently used to describe data-sharing ecosystems across organizations and thus will be used as synonyms in this paper (Otto et al., 2019).

Building on this foundation, recent research has shifted its focus to the governance and operationalization of data ecosystems, particularly in the areas of data sovereignty (Jarke, 2017) and trust (Gelhaar & Otto, 2020; Schäfer et al., 2023), which are critical for ensuring secure and reliable data exchange. However, in their comprehensive review of data ecosystems, Brée et al. (2024) identified several gaps within the literature that are currently under-researched, among them data security and the integration of artificial intelligence and machine learning within data ecosystems. On the one hand, data security deals with ways data can be stored and shared within data ecosystems while remaining protected as well as the influence of such measures on the utility of data ecosystems (Brée et al., 2024). On the other hand, machine learning and artificial intelligence have become central to the formation of data ecosystems, yet there is a need for a deeper understanding of the requirements for sharing AI training data and how training on shared data should be conducted (Brée et al., 2024). Our research seeks to address these challenges by proposing a new type of data ecosystem centered on synthetic data, which offers a means to mitigate privacy risks while maintaining the benefits of data sharing. Additionally, we investigate strategies for maximizing the utility of shared data to enhance individual organizational performance, thereby contributing to both the theoretical and practical development of data ecosystems.

With current research on data ecosystems, predominantly concentrating on applications within healthcare, Industry 4.0, and smart cities (Cappiello et al., 2020), this study tries to extend this focus to the financial services industry. Given the sector's significant dependence on highly sensitive data and its advanced application of machine learning technologies, this context provides a suitable setting to address previously identified research gaps in data security and the implementation of AI models within data ecosystems. Current research on data ecosystems within the financial services industry can be broadly categorized into two research streams. The first stream centers on open banking, a customer-focused ecosystem where established standards facilitate the secure sharing of banking data with various actors within the financial services ecosystem, based on customer requests (Cosma et al., 2023). While this approach grants consumers greater control over their data, it also raises significant data security concerns due to the decentralized nature of data storage across multiple providers—a critical issue given the heightened sensitivity of financial transaction data (Y. Wang et al., 2018). Furthermore, open banking does not provide institutions with an efficient and secure mechanism for large-scale data exchange, which is essential for applications such as fraud detection and anti-money laundering (Asrow, 2021). The second stream of research revolves around federated learning, a methodology that completely eliminates data sharing by enabling distributed training of shared models, thereby ensuring compliance with privacy protection regulations (Awosika et al., 2024; Lei et al., 2023; Perez et al., 2023). However, federated learning presents significant challenges, including computational overhead, scalability issues, and still privacy risks, as malicious actors might be able to infer sensitive data from the model parameters shared during the training process (Baabdullah et al., 2024; Chatterjee et al., 2024). Additionally, the necessity for participants in a federated learning ecosystem to agree on a single model architecture, which is difficult to modify once established, further complicates its implementation. The constraints of existing solutions, coupled with the fact that data ecosystems do not emerge organically but instead necessitate strategic planning around a shared value proposition, have resulted in the lack of a comprehensive financial data ecosystem to date (Adner, 2017; Immonen et al., 2014). This is aggravated by a research gap in the development of specialized engineering and management methodologies tailored to the needs of such an ecosystem (Oliveira et al., 2019) which are especially critical in the financial services sector, where stringent privacy requirements and the complex nature of financial transaction data introduce distinct challenges. Consequently, further research is essential to

address these challenges and to delineate the architectural frameworks necessary for the creation of robust and secure data ecosystems within the financial industry.

## Synthetic data generation and its application

Synthetic data can be defined as "data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)" (Jordon et al., 2022, p. 5). This generation process can take many forms as comprehensively categorized by Bauer et al. (2024) into 20 distinct method types. Among these, generative adversarial networks (GANs) are the most popular. GANs learn by pitting a generator (synthesizes data from random noise) and a discriminator (classifies samples as real or fake) against each other, resulting in two highly skilled networks (Goodfellow et al., 2014). This architecture is highly adaptable, as discriminator and generator can be easily adjusted to new tasks (e.g., time series or graph generation) while being frequently the best-performing synthetic data generation method (Bauer et al., 2024). Another commonly employed synthetic data generation method is autoencoder-based architectures, especially variational autoencoder (VAE) (Kingma & Welling, 2013). VAEs are trained by mapping an input sample to a hidden representation, which is then mapped back to the original vector, thus creating a model that synthesizes valid data from a lower dimensional representation. This decoder model is then used to generate data from random noise which makes it especially useful for learning from data with disentangled features (Bauer et al., 2024). Third, recurrent neural networks, feedforward neural networks which include recurrent edges, are able to generate sequential data of arbitrary length. This makes them ideal for sequence generation tasks such as speech synthesis, music, and time series generation (Lipton et al., 2015). Finally, virtual environments are computer simulations in which algorithms interact with each other based on predefined rules, generating synthetic data in the process (Bonabeau, 2002).

In the context of machine learning, synthetic data is primarily utilized in three key areas: (i) private data release, (ii) data de-biasing and fairness, and (iii) data augmentation for robustness (Jordon et al., 2022). As the focus of this paper is employing synthetic data for private data release, it will be investigated in more detail. Hereby, private data release describes the case where synthetic data is used to mitigate disclosure risk, allowing privacy concerns and regulatory issues to be circumvented by substituting real data with synthetic data (Esteban et al., 2017; Jordon et al., 2018). However, this comes with certain risks of disclosure, which users need to be aware of. While multiple risks exist, the most relevant is membership inference which seeks to determine if an individual was part of the original dataset (Bun et al., 2021; Jordon et al., 2022). This risk is particularly critical in the context of financial transaction data, as revealing a user's membership in a specific bank's dataset could enable malicious actors to carry out more targeted fraudulent activities, making fraud prevention more difficult. Research on dealing with membership inference risks in synthetic data, primarily drawn from the healthcare domain, can be divided into two major streams. The first stream focuses on achieving guaranteed privacy by modifying models to conform to differential privacy principles, ensuring both the data and the model are protected. Algorithms implementing this are the PATE-GAN (Jordon et al., 2018) or DP2-VAE (Jiang et al., 2022) architectures. The second research stream focuses on evaluating and managing privacy risks within acceptable limits for a given volume of published synthetic data, providing various metrics and thresholds for guidance (H. Chen et al., 2023; Yan et al., 2022). Popular measures are the nearest neighbor adversarial accuracy risk (Yale et al., 2020), the membership inference risk (Choi et al., 2018), and the meaningful identity disclosure risk (Emam et al., 2020). Furthermore, these measures have also been adopted by regulators such as the European Medicines Agency and Health Canada which both provide thresholds for identifying disclosure risk (Yan et al., 2022).

As the complexity of models continues to grow, necessitating larger datasets, synthetic data has been applied in a variety of fields, where it is used to facilitate more efficient and effective development of AI solutions (Lu et al., 2023). In financial services, these have been mainly use cases that inhibit a strong class imbalance such as anti-money laundering and financial fraud detection. Here, synthetic data generation is used to increase the amount of data within the minority class, thereby increasing training efficiency (E. Altman et al., 2024; Hilal et al., 2022). The current landscape is largely dominated by GAN-based architectures especially Wasserstein GANs due to their superior training stability (Hilal et al., 2022; Sethia et al., 2018; Strelcenia & Prakoonwit, 2023). However, recent advancements have seen transformer-based architectures (Nickerson et al., 2023) and diffusion-based models (Sattarov et al., 2023) emerging as competitive alternatives to GANs. Due to the internal usage of this synthetic data, data privacy has not been a main consideration when building these models. Data privacy considerations have mostly been explored in academic studies that aim to make their synthetic data publicly available. These studies typically employ virtual environment-based systems, such as multi-agent simulations, which simulate financial transaction data by modeling interactions between known actors and behaviors (E. Altman et al., 2024; Jensen et al., 2023; Lopez-Rojas et al., 2016). While these approaches are very secure from a privacy perspective as real data is only used during model evaluation of the synthetic data, they require significant manual work to identify patterns and changing behaviors need to be detected first, before they can

be integrated into the simulation (Bauer et al., 2024). However, the automatic generation and sharing of synthetic data derived from real data have not been extensively explored. As privacy concerns intensify due to regulatory pressure and customer expectations, as well as a growing necessity for extensive datasets to support cutting-edge machine learning models (Hittmeir et al., 2019), employing synthetic data has the potential to address privacy challenges in data ecosystems. Recent studies by Sattarov et al. (2023) and Langevin et al. (2022) have begun to investigate this potential for financial services. However, these studies primarily focus on comparing different data generation methods and present synthetic data sharing as merely one potential application. This leaves significant research gaps regarding the mechanisms for data exchange, the optimal strategies for learning from cross-institutional synthetic data, and the incentives for participating institutions, reaching beyond financial services and tackling current challenges in data ecosystems in general. Moreover, these studies offer little guidance on the design of such an ecosystem, highlighting a clear need for establishing design principles and best practices.

## Research approach

A design science research project was initiated to address a research gap in approaches to enhance privacy protection within data ecosystems while preserving data utility for machine learning applications. This need, combined with the financial services industry's demand for solutions to address the limitations of inter-organizational collaboration in tackling financial fraud and anti-money laundering detection, prompted the research effort. This project is aimed at designing an innovative artifact that provides financial institutions with a tool to easily exchange high-quality data with each other enabling them to increase their fraud and anti-money laundering detection performance, creating guidance on how to implement such a system, as well as to evaluate its benefits and the associated privacy risks (Gregor & Hevner, 2013; Peffers et al., 2007). To achieve these objectives, we adopted design science research (DSR), a framework particularly suited for the iterative development of novel artifacts addressing solution spaces with broad implications for both theoretical and practical problem domains (Peffers et al., 2007) and providing theoretically justified prescriptive knowledge (Gregor et al., 2020). Following this paradigm, we focus on creating artifacts that serve organizational purpose, in our case enabling data sharing despite privacy restrictions, through a structured research process that rigorously builds and evaluates viable solutions (A. R. Hevner et al., 2004; March & Smith, 1995). Following Scheider et al. (2023), our artifact is a "model" (March & Smith, 1995), a type of DSR artifact that serves as a simplified representation of reality and accumulates specific design knowledge (March & Smith, 1995); thus, DSR provides a suitable framework for our study (A. R. Hevner, 2007; Iivari, 2007). Our model presents a structured approach to designing a data ecosystem under privacy and data complexity constraints, exemplifying a solution to the problem discussed in the earlier sections. Our methodological approach to DSR—the design science research methodology (DSRM) by Peffers et al. (2007) has six steps, arranged in sequential order, and incorporates an iterative research procedure by design. The process typically starts with the identification of a research problem with practical relevance, in our case, the challenge of data scarcity within financial fraud detection. Next, the solution objectives are designed to address the stated challenges and to create a meaningful artifact. In line with DSR, the insights gained from the build-and-evaluate process must be generalizable and therefore applicable in more generic settings (Jones & Gregor, 2007). Also, the design artifacts should result in profound disruptions to traditional ways of doing business (A. Hevner & Gregor, 2022). Based on these objectives and on theory, the artifact is designed and developed in the next research process step. Phase 5 comprises evaluation, which is necessary to test whether an artifact achieves the purpose of its creation and to prove this achievement using rigorous methods (Venable et al., 2016). The evaluation phase also helps one to better understand the problem at hand and thus to realize improved outcomes (A. R. Hevner et al., 2004). Due to the iterative nature of this process, it can be repeated until a suitable artifact is derived. The design knowledge in the form of DPs with their DRs and MRs generated during this process can be seen as a nascent design theory, capturing a general solution in a class of artifacts (Baskerville et al., 2018). While MRs are high-level, generalized goals that an artifact must satisfy to address a class of problems, providing the foundational objectives for artifact design (Walls et al., 1992), DRs are specific, actionable specifications that detail the necessary features and characteristics an artifact must have to fulfill the meta-requirements (Gregor & Hevner, 2013). Lastly, DPs are prescriptive, actionable guidelines derived from design requirements and grounded in both theoretical foundations and empirical evidence, providing clear instructions for creating artifacts that meet the specified requirements and address the underlying problem space (Gregor et al., 2020). Thus, especially the DPs can be used to guide actions in a wider range of problems, in particular, data ecosystems where data with a complex structure needs to be shared under privacy restrictions (A. R. Hevner et al., 2004). They contribute to the theoretical advancement of the information systems (IS) community and provide valuable guidance for practitioners in designing similar artifacts (Baskerville et al., 2018; Sein et al., 2011). Since the DSR approach requires integration into an organizational

context, the project was conducted in collaboration with the UnionBank of the Philippines, a rapidly growing digital bank, as well as a European neo bank with a focus on wholesale transaction banking. Both banks rapidly scaled their digital transaction infrastructure in recent years and are now looking for new ways to tackle transaction fraud and money laundering. While the banks granted us deep insights into the problem of limited transaction data and provided invaluable feedback through all cycles, it was decided that prototyping and evaluation would be conducted on publicly available datasets instead of real bank data to reduce risks and allow fast iterations to create a solid understanding of potential pitfalls.

Within this DSRM framework, four iterative design cycles were conducted, thus allowing for continuous refinement of the artifact's design based on feedback and derive insights (Mullarkey & Hevner, 2019; Sein et al., 2011). In the next paragraph, the activities in each cycle are introduced which are outlined in the following graphic (Fig. 1).

First, the DSRM project starts with problem identification and motivation, focusing on stakeholder problems and challenges. This was done by conducting a systematic literature review on data ecosystems, synthetic data, and financial fraud detection as well as semi-structured interviews with employees at different levels at our partner banks, who are engaged in data sharing initiatives, fraud detection or data analytics, and machine learning projects. Furthermore, these interviews were used to identify the objectives of our solution by deriving DRs and MRs. Next, we iterated the first "Design—Demonstrate—Evaluate" cycle. In the design phase, we formulated the initial set of DPs. These principles were then translated into a system architecture during the demonstration phase, specifying its material properties like algorithms and interaction layers. Subsequently, an evaluation was conducted, involving feedback from academics and industry experts through four semi-structured interviews. The outcomes helped evaluate the feasibility of the initial design and led to the refinement of selected DPs in the second iteration. In cycle 2, we conducted a literature review identifying suitable algorithms for synthetic financial transaction data generation and based on them, instantiated a prototype which was subsequently evaluated on a publicly available real-world credit card transaction dataset to identify the most suitable synthetic data generation algorithm, establish the feasibility of the solution, and demonstrate the privacy-preserving properties of synthetic data. Based on additional expert feedback as well as two large simulated financial transaction data sets, cycles 3 and 4 refine the existing DPs and introduce new ones where needed. While cycle 3 explores the local level of the ecosystem in more detail, cycle 4 focuses on the global level and cooperative challenges within the ecosystem. Throughout the DSRM cycles,

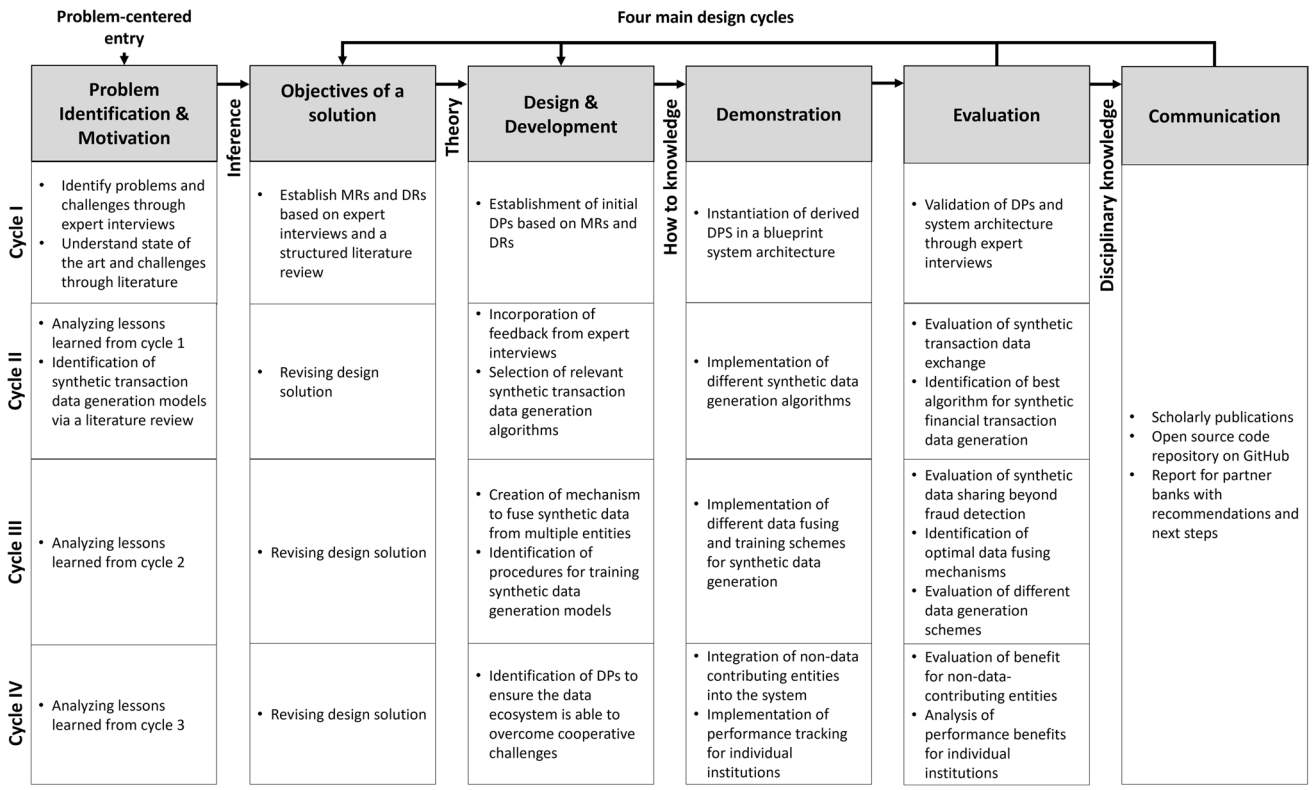


**Fig. 1** Steps and design cycles within our design science research study based on Peffers et al. (2007)

we iteratively abstracted the requirements, DPs, and system features. Thus, our main theoretical contributions lie in the abstracted artifacts, particularly the DPs, which are first derived in "Design of initial DPs" and continuously refined throughout the paper.

## Problem identification and motivation

The diagnosis phase consists of two tasks: understanding the problem and solution domain and defining the ecosystem's requirements. First, we positioned our DSRM project within the domain of inter-institutional collaboration within financial services. With a major focus of such collaboration being financial fraud detection, a first literature review on data ecosystems, synthetic data, and financial fraud detection was conducted. Following the methodology by Webster and Watson (2002), four search strings were established (Table 1) and the following databases: ScienceDirect, EBSCOhost, SpringerLink, IEEE Xplore, and AISeL, were queried for articles containing the previously defined search string in title, abstract, or the author keywords. Furthermore, only papers written in the English language and published within the past 5 years were included. This initial query resulted in a total of 3794 papers, which were then filtered based on a screening of titles and abstracts. While for papers identified by the "Fraud Detection" query strings only papers were included that deal with financial transaction fraud and either focus on privacy or a multi-organizational context, for papers selected by the "Data Ecosystem" string the only inclusion criteria were a focus on data ecosystems. After adding more relevant papers through a forward and backward search a total of 61 papers were selected for inclusion in the literature review.

The analysis of the first part of our literature review focusing on fraud detection revealed that the limited availability of data is a significant challenge, especially for smaller organizations (Kulatilleke, 2022; Pranto et al., 2022). Especially with increasingly sophisticated adversaries (Qiao et al., 2024) and thus, more complex fraud detection models, frequently built based on deep learning architectures, more data is needed for model training (Aurna et al., 2023; Hilal et al., 2022). This need for increasing amounts of training data is further aggregated by the extreme class imbalance of datasets (large datasets are needed for a sufficient number of samples in the minority class) as well as the fast-changing nature of fraudulent patterns (Abdul Salam et al., 2024; Ryman-Tubb et al., 2018). Tackling this, frequently, the proliferation of cross-institutional data is presented as a potential solution, to increase the amount of available data and train better and more robust models (Kong et al., 2024; Myalil et al., 2021; Qiao et al., 2024). However, due to the high sensitivity of financial transactions and the connected risk of privacy leakage, this exchange is usually prohibited by external regulation or internal guidelines (Bian & Zheng, 2023; Pranto et al., 2022; Ryman-Tubb et al., 2018). To overcome this problem, frequently federated-learning-based solutions are proposed, allowing the raw data to remain local, while a joined model is trained (Kong et al., 2024; Lei et al., 2023; Pranto et al., 2022). While these approaches show some promise, they retain significant drawbacks such as the computational overhead, scalability issues, and the necessity to agree on a single model architecture, which is difficult to modify once established (Baabdullah et al., 2024; Chatterjee et al., 2024). This leads us to the conclusion that there is a need for a data ecosystem that allows financial institutions to exchange data with one another while staying compliant with laws and internal regulations on data privacy and giving them the freedom to use this data to fulfill their specific needs.

## Definition of solution objectives

Looking for potential solutions, we drew on the second part of our literature review focusing on data ecosystems providing relevant insights on how such challenges can be navigated and potentially overcome in the context of financial data. Particularly papers from the healthcare domain (H. Chen et al., 2023; Morley-Fletcher, 2022), investigations into the emergence (Gelhaar & Otto, 2020) and organization (Langer & Mukherjee, 2023) of data ecosystems as well as

**Table 1** Results of systematic literature search

| ID | Search string | Hits | Filter: title[a] | Remove duplicates | Filter: abstract[a] | Fwd and Bwd search | Total |
|---|---|---|---|---|---|---|---|
| I | "Financial" AND "Fraud Detection" | 2471 | 336 | 449 | 30 | 5 | 35 |
| | "Transaction" AND "Fraud Detection" | 990 | 139 | | | | |
| II | "Financial" AND "Data Ecosystem" | 164 | 13 | 19 | 18 | 8 | 26 |
| | "Synthetic Data" AND "Data Ecosystem" | 169 | 6 | | | | |

[a]Detailed filter criteria can be found at https://anonymous.4open.science/r/SyntheticDataEcosystems-801C/Cycle1_InitialDesignPrinciples/README.MD

**Table 2** Overview interviewees for solution requirements

| ID | Job title | Expertise | Years of experience | Length of interview |
|---|---|---|---|---|
| Interviewee 1 | Chief data scientist | Data science | 10 years | 00:51:10 |
| Interviewee 2 | Senior data scientist | Data science | 5 years | 00:37:34 |
| Interviewee 3 | Data scientist | Data science | 5 years | 00:36:30 |
| Interviewee 4 | Chief financial officer | Fraud detection | > 20 years | 00:38:36 |
| Interviewee 5 | Senior compliance officer | Fraud detection | > 20 years | 00:59:08 |
| Interviewee 6 | Junior compliance officer | Fraud detection | 4 years | 00:44:27 |
| Interviewee 7* | Head of the AI center of excellence | Data science | > 20 years | 00:19:25 |
| Interviewee 8* | Head of data science ventures | Data science | 10 years | 00:31:03 |

*Interviewee from UnionBank of the Philippines

the preconditions for data sharing (Fassnacht et al., 2023), were detrimental in deriving the design requirements presented in the following section.

To extend our insights into the domain beyond academic literature next, nine semi-structured interviews with employees at various levels at our project partners, with a focus on fraud detection or data science, were conducted (for details, see Table 2). Querying them for challenges as well as potential solutions for tackling data scarcity within their domain.

Based on this, we formulated two meta-requirements (MR) that any solution must adhere to. MR1 emphasizes the ease of data sharing between financial institutions, encompassing both technical, legal, and collaboration aspects. The need for technical ease of use was informed by insights drawn from the medical field, where challenges related to tool availability and varying data standards were identified as hindrances to data sharing (van Panhuis et al., 2014). The legal dimension in ecosystem usability was motivated by diverse regulatory requirements across jurisdictions, as observed in existing approaches to sharing financial transaction data (Blake et al., 2019). Lastly, ease of collaboration was drawn from the ecosystem literature, where cooperative challenges were outlined as a major hurdle to data ecosystem development (Gelhaar & Otto, 2020). MR2 highlights the necessity of increased utility as a result of sharing data. This requirement emanated from discussions with our partners regarding their goal of establishing a data-sharing ecosystem and from the literature describing incentives for participation in data ecosystems (Gelhaar et al., 2021).

Next, we refined the MRs into more specific DRs, drawing from literature as well as the knowledge of our project partners.[1] To incentivize users to participate in data-sharing, setup as well as reoccurring costs need to be as low as possible, which is reflected in MR1 and propagates into DR1 and DR2. This is important because while a data standard for financial transaction data exists, different banks diverge from it (Major & Mangano, 2020), which was also confirmed during our interviews ("Different data providers have different schemas and transaction languages."—Interviewee 2); thus, a data ecosystem needs to be flexible enough to accommodate various input data structures (DR1). This is particularly important as data needs to be regularly updated and the cost for these updates should be as low as possible. Furthermore, data privacy standards imposed by regulators and internal policies must be upheld ("In terms of data sharing we do not engage in anything, because this is the pain with financial institutions, we are really protective of our data"—Interviewee 8–1). Our interviews revealed that in the context of our partner institutions, this means that all real data must be processed locally within the financial institution (DR2). From a data-centric perspective, the performance of machine learning methods can be enhanced by increasing the volume of training data available (Sun et al., 2017). Thus, MR2 can be achieved by enabling the combination of data from multiple sources through the data ecosystem and making it accessible as a unified data source (DR3). Given the goal of creating an ecosystem that is applicable to multiple tasks, the absence of a dominant algorithm in many fields (e.g., fraud detection), and the insight from our interviews that banks prefer to build and exclusively own their solutions ("One model will not be enough, it will be a collection of models which answer different questions …"—Interviewee 8–1), the data ecosystem must support diverse types of algorithms (DR4). Additionally, the imbalanced nature of fraud data necessitates tools on the ecosystem to address data imbalances through filtering, oversampling, and undersampling (DR7), as most machine learning algorithms perform better on balanced datasets (Longadge & Dongre, 2013). As fraud patterns change quickly when discovered, the timely integration of recent fraud patterns into fraud detection algorithms is crucial (Benchaji et al., 2021; Zhu et al., 2021). As this is utterly important, two DRs were dedicated to achieving this. First, institutions should have the capability to automatically

[1] A detailed mapping from interview quotes to DRs can be found on: https://github.com/Faruman/SyntheticDataEcosystems/blob/master/Cycle1_InitialDesignPrinciples/README.MD

update the data ("fraud, money laundering patterns will change, behavior patterns will change and that's why you need to establish this relationship where there is a continuous flow of information"—Interviewee 4), ensuring that the dataset incorporates the most recent fraud patterns (DR5). This not only aligns with MR1 by enhancing user convenience and reducing the need for frequent user inputs but also guards against model drift (Zhang, 2022). However, even with automatic updates, the dataset may still be dominated by outdated fraud patterns, posing a risk to the algorithms (Paleyes et al., 2023). Therefore, users should be able to incorporate pattern-based artificial data into the ecosystem ("…[the] machine has the benefit of learning the patterns you, as a human, identify as problematic. In the current world, such patterns are the key to everything because criminals will always evolve."—Interviewee 6) (DR6). Allowing the data ecosystem to benefit from expert domain knowledge is not yet reflected in the data (Richhariya, 2012). After having defined the problem as well as the solution space and outlined our requirements, we can now commence the first design, implementation, and evaluation cycle.

## Cycle 1: DPs and system architecture for synthetic data sharing

During the initial phase of the DSRM project, foundational DPs were established, integrating expert insights, relevant literature, and domain requirements, to develop a synthetic data ecosystem for financial institutions. Building on these insights an architecture for such an ecosystem was proposed.

### Design of initial DPs

In our first design phase, our primary emphasis was on identifying the foundational DPs. Building on the DRs derived in the previous section and following the recommendations of Chandra et al. (2015), we created DPs that followed the structure "Provide the system with [material property—in terms of form and function] in order for users to [activity of user/group of users—in terms of action], given that [boundary conditions—user group's characteristics or implementation settings]" (Chandra et al., 2015, p. 4045). Furthermore, to ground these artifacts in practical relevance, expert interviews with our partners were conducted to justify the DPs derived from the literature. Figure 2 depicts the relationship between MRs, DRs, and DPs.

DP1—Provide the system with modular systems design in order to ensure independence of local data and cross-institutional proliferation of synthetic data given that the raw data is sensitive: To address DR1 and DR3, the data ecosystem must possess the capability to process data from diverse sources while enabling the integration of this data for synthetic data generation. Drawing upon the principles of modular systems theory (Tiwana et al., 2010), institutions are granted flexibility in designing their module structures while adhering to a standardized representation, thereby ensuring that the data can be exchanged with the ecosystem. Additionally, once the initial setup is complete, automated data updating becomes straightforward, as all computations can be performed locally, without the need for sensitive data to be transmitted outside the local system. This capability fulfills the requirements outlined in DR5.

DP2—Provide the system with the ability to generate synthetic transaction data using generative adversarial

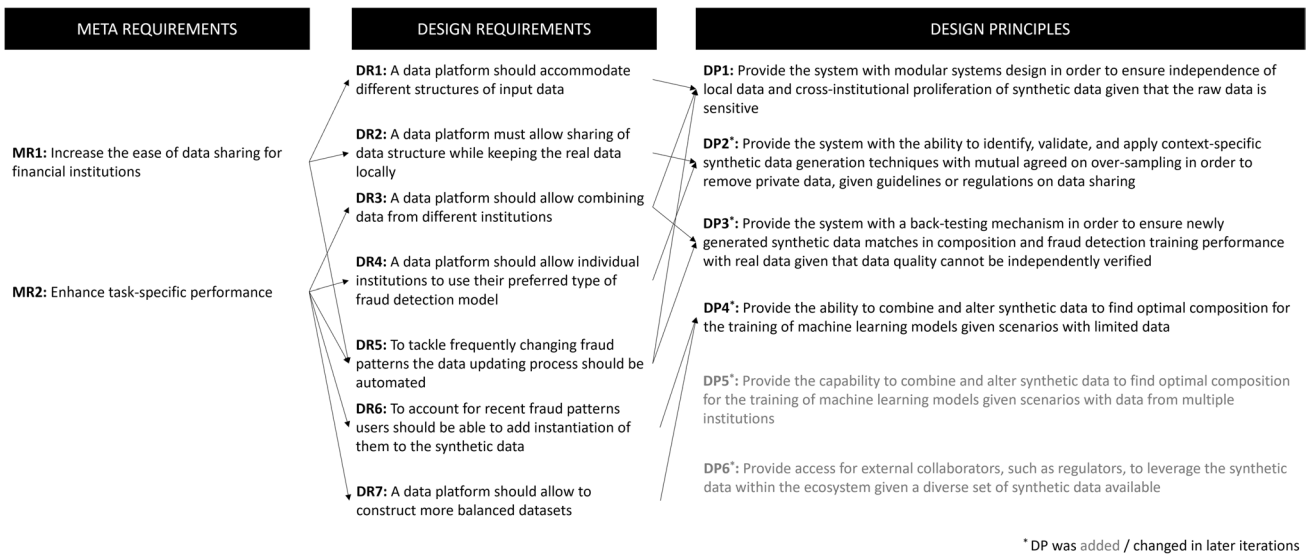


**Fig. 2** Relationship between MRs, DRs, and DPs using the final set of DPs

networks (GANs) in order to remove private data, given guidelines, or regulations on data sharing: Most models created in financial institutions, such as fraud detection algorithms, need to be trained on transaction-level data as its granularity and connectedness over time allows for complex patterns to emerge (Hilal et al., 2022) This combined with DR4, which requires users to train different types of algorithms and mandates a data ecosystem to provide the user with access to such low-level data. However, sharing transaction-level data poses challenges due to regulatory constraints (Blake et al., 2019) and internal policies mandating its local storage (DR2). As anonymization is not able to preserve both data utility and privacy for heavily interconnected data (Loukides et al., 2010), we propose to solve this challenge by using GANs, due to their unique ability to learn patterns in data and generate synthetic data nearly indistinguishable from the original (Walia et al., 2020). This enables us to preserve real data locally while sharing only the privacy-preserving GAN-generated data within the data ecosystem. This data can then be merged with synthetic data from other institutions and allows the training of machine learning models on the combined dataset. Therefore, ensuring the confidentiality of sensitive data while empowering the ecosystem to enhance fraud detection capabilities by training algorithms with substantial volumes of high-quality data.

DP3—Provide the system with a back-testing mechanism in order to ensure newly generated synthetic data matches in composition and fraud detection training performance with real data given that data quality cannot be independently verified: To facilitate the seamless integration of data from multiple institutions (DR3) and enable frequent system updates without human intervention (DR5), it is essential to establish a robust quality control mechanism. This mechanism serves to uphold the integrity of the data introduced into the ecosystem, as only a few bad data points can have tremendous effects on machine learning models (Chakravarty et al., 2020). One approach to achieve this is by implementing a back-testing procedure, which ensures that the synthetic data accurately captures the underlying patterns of the local real data (Dankar et al., 2022).

DP4—Provide the ability to alter synthetic data to give it the optimal composition for the training of machine learning models given that data in fraud detection is highly skewed: To further enhance model performance, a data-sharing ecosystem should be designed to provide users with the ability to alter and extend the existing data to create the right data for their use case. In financial services use cases, such as money laundering or fraud detection, the balance between the classes often is a challenge (Al-Hashedi & Magalingam, 2021), resulting in the requirement, that a data ecosystem should be able to provide more balanced datasets (DR7). This can be accomplished by equipping users with advanced filtering options or enabling them to manipulate the existing data through techniques such as under- or oversampling (Lopez-Rojas & Axelsson, 2012).

## Demonstration of DPs by instantiation in a system architecture

Based on the DRs and DPs, we present a multi-layered platform architecture for a synthetic data ecosystem. While the local processing layer is implemented at every institution, the synthetic data generation as well as the fraud detection layer are centralized. An overview of this architecture mapped with corresponding DPs can be seen in Fig. 3.

**Local processing layer** The local processing layer is modular and situated at every financial institution (DP1). Here, the GAN models are trained on sensitive transaction data to produce accurate synthetic representations of this data (DP3). Furthermore, the conversion to the data standard the synthetic data needs to conform to is enforced. Moreover, back-testing is done to ensure data quality while guaranteeing that the real data never leaves the local environment (DP2).

**Global data layer** Contrary to the previous layer, the synthetic data layer is not situated at a specific institution. Instead, this layer is where synthetic data is merged and modifications to the data composition through the addition of pattern-based data generators or the artificial rebalancing of different classes can be achieved (DP4).

**Fraud detection layer** This layer is accessible to any participating company allowing them to access the synthetically generated data and modify it to fit their models by providing capabilities to subsegment and alter data, making it optimal for their custom fraud detection models.

## Evaluation of derived DPs and system architecture

After deriving the system architecture from our DPs, we presented both to two experts from our partner institution as well as 2 academics (for details, see Table 3).

The feedback gathered from the experts was overall positive and especially the use of modular system design (DP1) to ensure reduced complexity of the eco-system and complete control of the local layer by the single institutions was highly appreciated. Furthermore, DP4 was approved by experts stating that "balancing data is a major concern when training ML models and a system providing smart

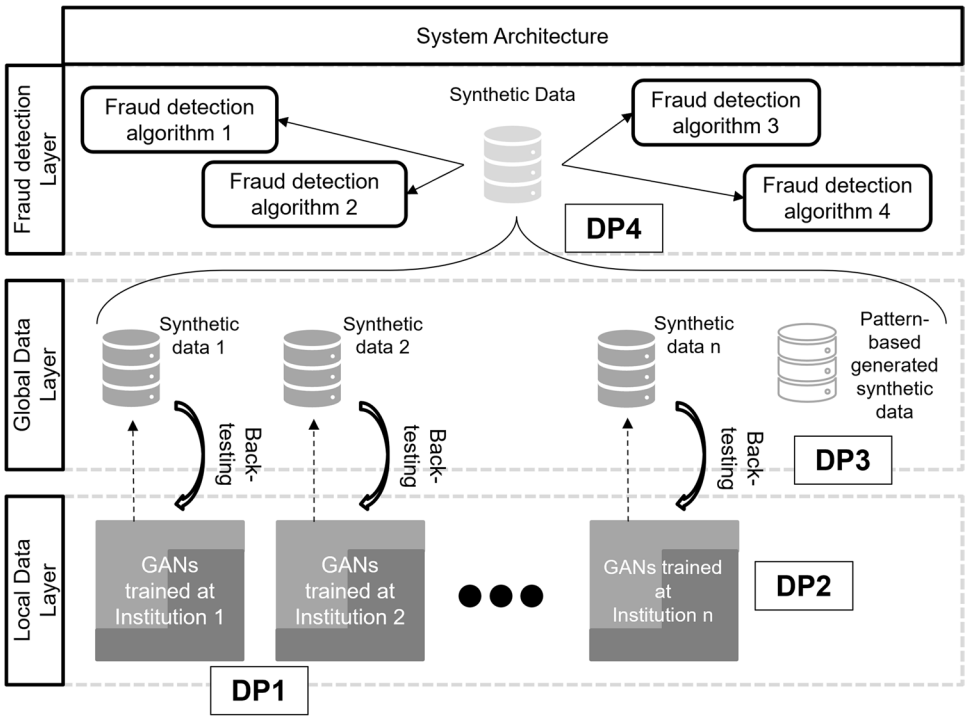


**Fig. 3** System architecture (version 1)

**Table 3** Interviewees for validation of DPs and platform architecture

| ID | Job title | Expertise | Years of experience | Length of interview |
|---|---|---|---|---|
| Interviewee 7* | Head of the AI center of excellence | Data science | >20 years | 00:23:16 |
| Interviewee 8* | Head of data science ventures | Data science | 10 years | 00:20:27 |
| Interviewee 9 | Research assistant | Statistical modeling | 5 years | 00:31:26 |
| Interviewee 10 | Research assistant | Design science research | 5 years | 00:22:01 |

*Interviewee from UnionBank of the Philippines

support for that could be particularly helpful" (Interviewee 10). Lastly, the proposed architecture was seen as a good first outline to create a prototype; however, the computational resources required to train the synthetic data generation models for frequent updates were raised as a concern. When discussing the proposed DPs as well as architecture with academic experts from the field of design science research, data sharing, and fraud detection, DP2 was criticized for multiple reasons. First, the limitation to a single technology for data generation (GANs) was seen as being too restrictive and limiting the system's adaptability to different domains ("Why do you limit yourself to a single data generation algorithm?"—Interviewee 10). Furthermore, concerns emerged about the feasibility of generating financial transaction data from limited local data and the utility of synthetic data to benefit fraud detection performance ("I doubt that abstracted data from other institutions with different data distributions can improve fraud detection performance."—Interviewee 9).

## Cycle 2: Synthetic financial transaction data generation and privacy

In the second cycle of the DSRM project, different methods for synthetic data generation were evaluated, thus tackling one of the limitations identified by expert feedback. This is done by testing the insights from a systematic literature review on synthetic data generation on a real-world financial fraud detection dataset, leading to the refinement of DP2.

### Design of synthetic data generation

Addressing the expert feedback, the second design cycle focuses on the refinement and extension of DP2. Based on the comments, it was adjusted to DP2—Provide the system with the ability to identify, validate, and apply context-specific synthetic data generation techniques with mutually agreed on over-sampling in order to remove private data, given guidelines or regulations on data sharing so that it is

**Table 4** Results of systematic literature search

| Search string | Hits | Selected | Fwd and Bwd search | Total |
|---|---|---|---|---|
| ("synthetic data generation" OR "artificial data generation") AND ("transaction data" OR "time series data") | 289 | 47 | 8 | 55 |

no longer restricted to a single method for generating synthetic data and includes the necessary validation of selected techniques to obtain optimal data generation performance.

To validate DP2 and identify suitable methods to generate synthetic financial transaction data, a literature review following vom Brocke et al. (2009) was conducted. In the first step, top publications regarding synthetic data generation were reviewed, resulting in our search string which was then used to identify journal articles and conference papers written in English and published after 2020 in the following databases: ScienceDirect, EBSCOhost, SpringerLink, IEEE Xplore, and AISeL. The results can be seen in Table 4.

From these papers, 46 distinct algorithms were extracted and grouped by their underlying algorithm type. Consequently, GANs emerge as the primary underlying mechanism (used by 55.3% of algorithms) for generating synthetic transaction data. GAN models work by creating two neural networks that learn by competing in synthesizing and identifying synthetic data and thus, once trained, can generate synthetic data that is indistinguishable from real one (Goodfellow et al., 2014). However, different implementations exist. To allow for variations between the algorithms tested and address the high degree of similarity between the different GAN architectures, we decided to only include two of them in our comparison: CTGAN (L. Xu et al., 2019), which was the most mentioned algorithm and is a representative of GANs taking only dependencies between attributes, but not samples, into account and TimeGAN (Yoon et al., 2019) (ranked third by mentions) which incorporates the temporal dimension between samples. To tackle the criticism from cycle one, we extended our overview beyond GAN-based architectures. The most frequently mentioned implementations using other algorithm types were Gaussian mixture models, which learn the distribution for each attribute and then generate new samples by drawing from these (S. Xu et al., 2021) and TVAE (Ishfaq et al., 2023), a variational autoencoder (VAE), which works by learning to compress and decompress data into a low-dimensional space and then use the decompress module in combination with random noise to synthesize new data. The literature predominantly focuses on applying these algorithms to health records (Xing et al., 2022), with limited exploration in other domains such as traffic data (S. Xu et al., 2021) and IoT data (Liu et al., 2019); however, none of the papers identified has examined the application of these methods for the cross-institutional proliferation of financial transaction data. Furthermore, while Weldon et al. (2021) found that using only synthetic data can achieve performance gains, others, such as Frid-Adar et al. (2018), show that mixing synthetic and real-world data is more beneficial. Thus, the optimal algorithm for generating financial transactions in the context of synthetic data sharing as well as the necessity of combining synthetic with real data remains unclear. Lastly, by employing algorithms that do not provide privacy guarantees by themselves, it remains unclear how safe it is to share the generated data. To tackle these two privacy measures frequently used in the literature, nearest neighbor adversarial accuracy and membership inference risk precision were used to ensure the evaluated algorithms do not leak information (Yan et al., 2022). While nearest neighbor adversarial accuracy measures if a classifier is able to distinguish between real (holdout set) and synthetic data and thus is a good indicator for privacy leakage through overfitting (Yale et al., 2020), membership inference risk precision measures how easy it is for an attacker to predict if a record is part of the train dataset or not based on the synthetic data (Choi et al., 2018). As no thresholds for these measures for financial transaction data exist, the ones for medical data were employed, which can be seen below (Table 5).

## Demonstration of synthetic financial data generation

In this section, we operationalized the derived DPs into a prototype system in Python using a modified version of the synthetic data vault library (Patki et al., 2016). Looking at the system architecture from design cycle one, the

**Table 5** Thresholds for privacy measures in medical synthetic data generation literature

| Measure | Threshold | Literature |
|---|---|---|
| Nearest neighbor adversarial accuracy | 0.030 | Yale et al. (2020) |
| Membership inference risk precision | close to 0.5 | Zhang et al., (2019, Appendix D)<br>Choi et al., (2018, Appendix F) |

local and global data layers were implemented, resulting in an ecosystem that allows data ingestion, synthetic data generation, and data sharing. Furthermore, the ecosystem was created in a way that allows to switch between different synthetic data generation methods, thus enabling the evaluation of different algorithms for financial transaction data generation.[2]

## Evaluation of synthetic financial data generation algorithms

This evaluation compares the different synthetic data generation approaches outlined before. As a real-world source for performance comparison, the credit card transaction dataset from the IEEE-CIS Kaggle competition[3] was chosen. This dataset was selected because credit card transactions, reflecting user spending patterns, are closely comparable to bank transactions. Furthermore, it was the only real dataset identified, which allowed matching transactions to users, allowing for models expecting time series data to be trained. However, limitations exist, such as the limited observation period (6 months), many obscured features as well as the inability to identify senders of payments but only receivers. As we aim to analyze the benefits of sharing synthetic data across financial institutions, we split the dataset by credit card provider, creating four distinct subsets. An analysis across subsets showed significant differences, aligning with anticipated variations in multi-institutional bank datasets. After obtaining a suitable dataset, we defined our evaluation process. For this, first, a Bayesian parameter search was used to tune the hyperparameters of the different synthetic data generation models using a subsample of 100,000 data points for each institution.[4] After selecting the best hyperparameter combination for each generation model, an XGBoost classifier (commonly used in fraud detection as per Interview with Interviewee 5 as well as Al-Hashedi and Magalingam (2021)) was trained on either real data, synthetic data, or combination of both (hyperparameter where tuned using threefold cross-validation). The results of this process were assessed using the ROC AUC score on a hold-out dataset (30% of the total data). The ROC AUC score was chosen as it provides a comprehensive evaluation of the classifier's performance across different levels of sensitivity and specificity and is frequently used in the literature (Sun et al., 2023). Furthermore, the evaluation was conducted in two stages. The first one covered the performance of individual synthetic data generation algorithms, thus helping us to validate DP2, while the second one looked at the overall benefit of the proposed synthetic data ecosystem. In the first stage, the focus was on evaluating the performance of different generation algorithms (Fig. 4), revealing that GMMs (ROC AUC score 0.52) and TimeGANs (ROC AUC score 0.5) underperformed expectations. This can be explained by the composition of the data. While GMMs struggled with the high dimensionality of the data (148 features), TimeGANs had problems with short transaction chains (below 2 transactions per user) due to the short observation period. While CTGAN (ROC AUC score 0.59) performed a little better, TVAE (ROC AUC score 0.89) excelled, particularly thriving in scenarios with limited training data, notably in datasets for "Discover" and "American Express," which had fewer than 10,000 transactions. Thus, confirming that the selection of the right algorithm is crucial and therefore validating DP2.

Next, we analyzed the privacy implications of the proposed algorithms, ensuring that the tested algorithms meet the previously defined privacy objectives and thus can be used in our proposed synthetic data ecosystem. As can be seen in Table 6, apart from TIMEGAN, all of the proposed algorithms stay within our previously defined privacy thresholds, leading us to the conclusion that, for the proposed dataset, GMM, CTGAN, and TVAE are able to sufficiently obscure the data and can thus be used in our ecosystem.

The second-stage evaluation assessed the advantage of training on shared synthetic data versus isolated real data. Figure 5 compares the performance of models trained on isolated real data, isolated synthetic data, shared synthetic data, and shared synthetic data combined with isolated real data. Models trained solely on synthetic data from one

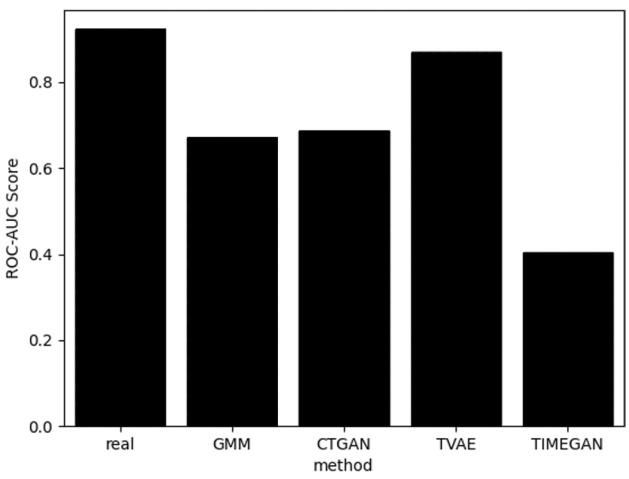


**Fig. 4** Comparison between different synthetic data generation algorithms

[2] The full implementation of Cycle 2 can be found on https://github.com/Faruman/SyntheticDataEcosystems/blob/master/Cycle2_AlgorithmComparison/README.MD

[3] https://www.kaggle.com/c/ieee-fraud-detection

[4] A detailed description of the hyperparameter tuning procedure can be found here:https://github.com/Faruman/SyntheticDataEcosystems/blob/master/Cycle2_AlgorithmComparison/02_paramSearch/README.MD

**Table 6** Privacy measures per algorithm

| Measure | GMM | CTGAN | TVAE | TIMEGAN |
|---|---|---|---|---|
| Nearest neighbor adversarial accuracy | 0.000554 | 0.00189 | 0.001499 | 0.000241 |
| Membership inference risk precision | 0.485238 | 0.489603 | 0.469872 | 0.130435 |

source underperformed compared to those trained on real data. Yet, combining synthetic data from multiple sources led to a further performance drop, likely due to varying fraud cases across providers, which dilutes relevant patterns. However, merging synthetic with real data for each institution boosted performance, increasing the ROC AUC score by 1%.

To better understand the impact of this improvement, we can look at the recall or what percentage of fraudulent cases are identified. Using synthetic and real data combined, we find that 2.14% more true positives are detected. Combining this with an estimated number of 24.16 million fraudulent card transactions per year only in the EU (European Central Bank, 2021), the improved model would have detected about half a million additional transactions. Thus, showing the benefit of our ecosystem. However, this fusion of shared synthetic data with local real data is not yet reflected in any DP; however, the evaluation showed it to be a critical principle of our proposed design. Thus, a new DP: DP5—Provide the capability to combine synthetic data to find an optimal composition for the training of machine learning models given scenarios with data from multiple institutions was created, incorporating this important design criterion. Based on this the proposed system architecture was revised, which can be seen below (Fig. 6).

Moreover, the outcomes of this design cycle were presented to additional experts in the field and two primary critiques emerged: the constraint that the design was only validated in a singular context on a single dataset, which poses questions about its generalizability, and the inherent challenges in establishing such an ecosystem, particularly concerning the incentivization mechanisms required to encourage active participation among the financial institutions.

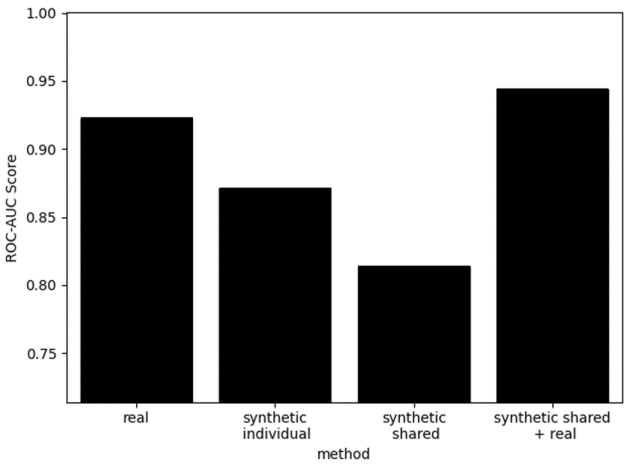


**Fig. 5** Comparison between synthetic and real data combinations

## Cycle 3: Local synthetic data recombination and usage

In the third cycle, we expanded the scope of our data ecosystem design to address a broader range of applications beyond fraud detection, aiming to validate the DPs' versatility and robustness in two contexts. Furthermore, the design elements of the local data level were investigated in more detail, resulting in the refinement and validation of DP5 and DP2.

### Design of mechanisms at the local data level

Building on the expert feedback, in this iteration of our research, we broaden the scope of our data ecosystem design to encompass a wider range of applications, aiming to demonstrate the versatility and robustness of our DPs in various contexts. Furthermore, this iteration focuses on investigating the design elements on the local data level, thus providing design knowledge for the individual institutions within the ecosystem. On the one hand, we focus on the validation and refinement of DP5—Provide the capability to combine synthetic data to find an optimal composition for the training of machine learning models given scenarios with data from multiple institutions by exploring the effect of the mixing percentage between synthetic and real data. On the other hand, we investigate DP2—Provide the system with the ability to identify, validate, and apply context-specific synthetic data generation techniques with mutually agreed on over-sampling in order to remove private data, given guidelines or regulations on data sharing in more detail by developing design recommendations on how to train the synthetic data generation models.

To extend our investigation to new domains, we consulted the literature and solicited input from our partner institutions, identifying money laundering detection as a significant use case that heavily relies on machine learning (Z. Chen et al., 2018) and often lacks sufficient training data (Jensen et al., 2023). Subsequently,

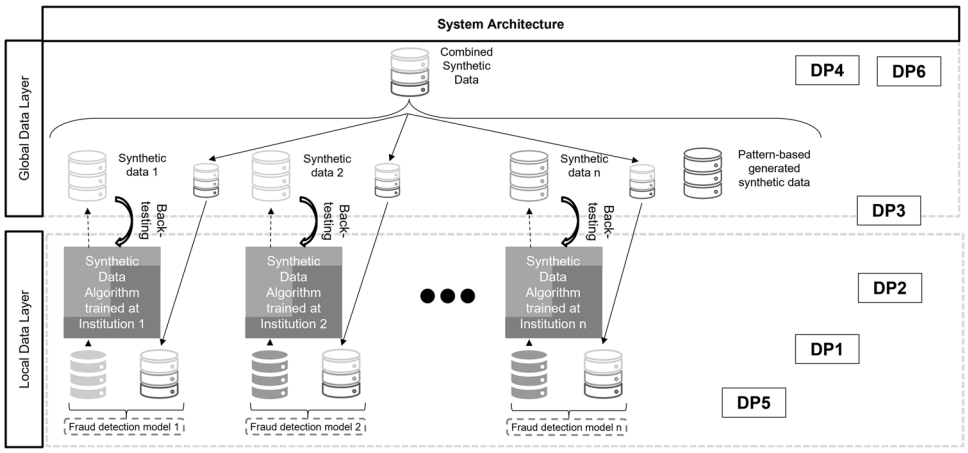


**Fig. 6** Updated system architecture (version 2)

an examination of the literature regarding the enhancement of machine learning performance through the incorporation of synthetic data was conducted, aiming to determine an optimal ratio of real to synthetic data (mix-in percentage). While some researchers only oversample the minority class using synthetic data (Charitou et al., 2021; Strelcenia & Prakoonwit, 2023), others train models exclusively on synthetic data (Sattarov et al., 2023) or combine real with synthetic data (Dahmen & Cook, 2019). Thus, it remains unclear if there is an optimal mix-in percentage that individual institutions should incorporate into their design.

To find the optimal way to generate synthetic data for our ecosystem, this section investigates data generation configurations that utilize the entire dataset as well as those trained on distinct data subsets and further analyzes the benefit of different pre-processing steps during the synthesizing process. Due to the challenge of dataset imbalance, models tend to be biased towards the majority class, decreasing the quality of data in the minority class; mitigating this issue, oversampling can be applied during the generation process to enhance generator robustness, albeit at the risk of distorting dataset composition (too many positive samples) (Kiran & Kumar, 2024). Second, the construction of distinct synthetic data generators for each class has been proposed as an alternative solution. Enabling the generator to better capture the characteristics of each individual class. However, this results in the problem that the minority class generator is only trained with a small dataset, which might harm its generalizability (Eilertsen et al., 2021). To remedy this, Fan et al. (2022) have suggested a novel methodology where the generator for the minority class is pre-trained using samples from the majority class, thus circumventing the problem.

## Demonstration through implementation of different training and data-fusing schemes

In this section, we operationalized the derived DPs into a prototype system in Python using a modified version of the synthetic data vault library (Patki et al., 2016). Building upon the architecture from design cycle two, the local layer was modified to accommodate for different generation schemes with and without oversampling as well as pre-training on the local level. Furthermore, the training scheme of the prediction model was modified so that the system was able to accommodate training with different mix-in percentages.[5]

## Evaluation of different training and data-fusing schemes

One challenge in evaluating the broader feasibility of our synthetic data sharing ecosystem is the lack of publicly available financial transaction data (Jensen et al., 2023). However, multiple researchers have shown that simulated financial transactions can be suitable for validating new models or even evaluating interventions (Langevin et al., 2022; Sattarov et al., 2023). Therefore, in this as well as the next cycle, we will use two datasets, one for anti-money laundering (IBM-AML[6]) and one for fraudulent transactions (IBM-CCF[7]), which were generated by using a multi-agent-based approach, simulating actors that act according to predefined rules, thus creating a stream of transactions (E. Altman et al., 2024; E. R. Altman, 2019). The resulting datasets have the advantage of being magnitudes larger in size (IBM-AML: 31898238/ IBM-CCF: 24386900) than the data used in the previous cycle (IEEE-CIS: 1097231) and have a network structure more similar to the one in real data. However, due to its simulation-based nature, it might not

[5] The full implementation of Cycle 3 can be found here: https://github.com/Faruman/SyntheticDataEcosystems/blob/master/Cycle3-4_EcosystemEvaluation/README.MD

[6] https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml

[7] https://www.kaggle.com/datasets/ealtman2019/credit-card-transactions

inhibit all characteristics found in real data. As the selected datasets do not include a financial institution (IBM-CFF) or the number of financial institutions present in the data is too big (IBM-AML, 122333 different banks), the data was artificially grouped. This was done by segmenting the data based on the location of the individual (IBM-CCF)/bank (IBM-AML) connected to a transaction, creating clusters that simulate the transactional networks of hypothetical financial institutions. As a result, the IBM-CCF dataset included four financial institutions with a relatively even data distribution, while the IBM-AML dataset emerged with seven banks of which two banks held over 75% of the data. This contrast in dataset composition affords a unique chance to explore the synthetic data sharing ecosystem's functionality under a broad array of conditions. Moreover, an analysis of client distribution post-split for each provider highlighted significant disparities, aligning with the anticipated diversity suspected within multi-institutional datasets. Details on the specific distributions are outlined in Table 7.

To limit the variables of this investigation, the synthetic data generation model and the fraud prediction model were kept constant. For the synthetic data generation model, the previously superior TVAE-based generator[8] with hyper parameters tuned to the individual institution was used. Similar to Cycle 2, a XGBoost classifier with hyperparameter selection using threefold-cross validation was chosen as the prediction model and the performance comparisons were done using the ROC AUC score on a holdout dataset (30% of the data).

Before, investigating the approaches modifying the local layer, the transferability of synthetic transaction data sharing beyond transaction fraud detection was evaluated (Fig. 7). To do this, we compared the average ROC AUC score between the dataset constructed for financial fraud detection (IBM-CFF) and the one constructed for anti-money laundering detection (IBM-AML).

Figure 7 demonstrates that across both datasets, models trained with synthetic shared data surpassed those trained without it, enhancing the ROC AUC score by 3.6% in the transaction fraud dataset (IBM-CCF) and 6.6% in the anti-money laundering dataset (IBM-AML). This effect can be considered substantial within this context as even recently introduced fraud detection algorithms often only increase the ROC-AUC score by a few percentage points (Hashemi et al., 2023; Lebichot et al., 2021). This performance gain suggests that the data ecosystem's effectiveness extends beyond merely detecting financial fraud but is also suitable for other use cases utilizing financial transaction data such as money laundering detection. Thus, confirming the versatility and potential of the synthetic data ecosystem in addressing a broad range of data challenges in financial services. Subsequently, we explore whether a specific mix-in percentage of real and synthetic data yields optimal results for machine learning performance. To accomplish this, we systematically assess the impact on model performance by varying the proportion of real and synthetic data used in training the models, exploring a spectrum from 0% (no synthetic data) to 300% (3 times as much synthetic as real data). Figure 8 visualizes this experiment.

**Table 7** Distribution of data across the different banks

| IBM-CCF | | IBM-AML | | | |
|---|---|---|---|---|---|
| Bank | Pct of data | Bank | Pct of data | Bank | Pct of data |
| 0 | 21.54% | 0 | 5.93% | 4 | 29.94% |
| 1 | 18.58% | 1 | 11.90% | 5 | 45.35% |
| 2 | 39.16% | 2 | 2.87% | 6 | 2.12% |
| 3 | 20.72% | 3 | 1.90% | | |

Observing the modest upward trajectory of the aggregated performance line (black), we can conclude that there is a positive effect of adding synthetic data. However, in contrast to the more volatile performance trends of individual banks (grey), it appears there is not a universally optimal mix-in percentage. Instead, distinct peaks in performance suggest that the most effective mix-in ratios vary by bank. Consequently, we infer that allowing banks to adjust the mix-in percentage independently is most beneficial. This insight has been integrated into DP5, which mandates that banks have the autonomy to determine their mix-in ratios, leading to the updated principle: DP5—Provides the capability to combine synthetic data to find optimal composition for the training of machine learning models given scenarios with data from multiple institutions.

Finally, we explored various configurations and preprocessing methods for synthetic data generation to offer optimal guidance for setting up these processes at the local level. Essentially, there are two primary setups. The first, referred to as "full," involves training the synthetic data generation model on the entire dataset. To mitigate the risk of the model predominantly generating samples from the majority class, versions that randomly oversample the minority class to a specified percentage of the data ("_OS{X}") while training the synthetic data generator have been implemented. The second setup, "sep" entails training distinct generation models for each class. An extension of this approach, "sepPre" utilizes separate generators for each class but pre-trains the minority class generator with majority class data. The outcomes of these varied approaches are detailed in Table 8.

The analysis of the data presented in Table 8 yields several key findings. Initially, the "full" model demonstrates its ability to surpass the baseline performance, yet models built on the same training scheme but utilizing oversampled data

[8] A detailed description of the hyperparameter tuning procedure can be found here: https://github.com/Faruman/SyntheticDataEcosystems/blob/master/Cycle3-4_EcosystemEvaluation/02_paramSearch/README.MD

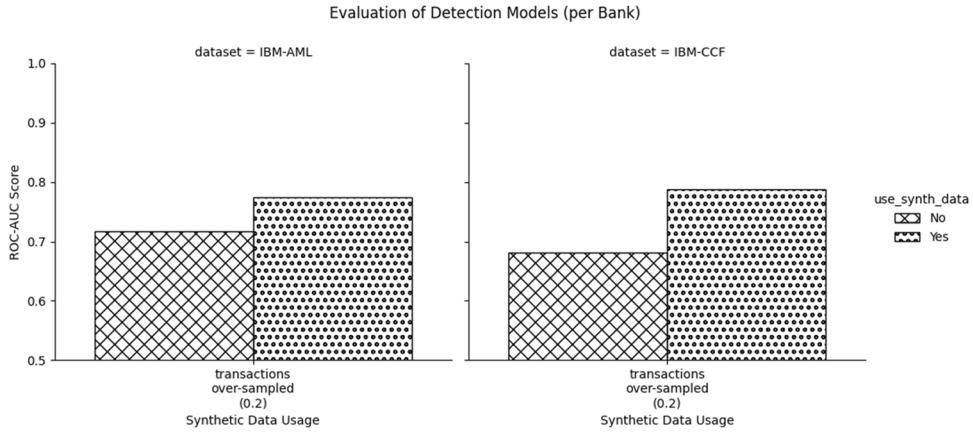


**Fig. 7** Comparison between models trained with and without synthetic data for both datasets

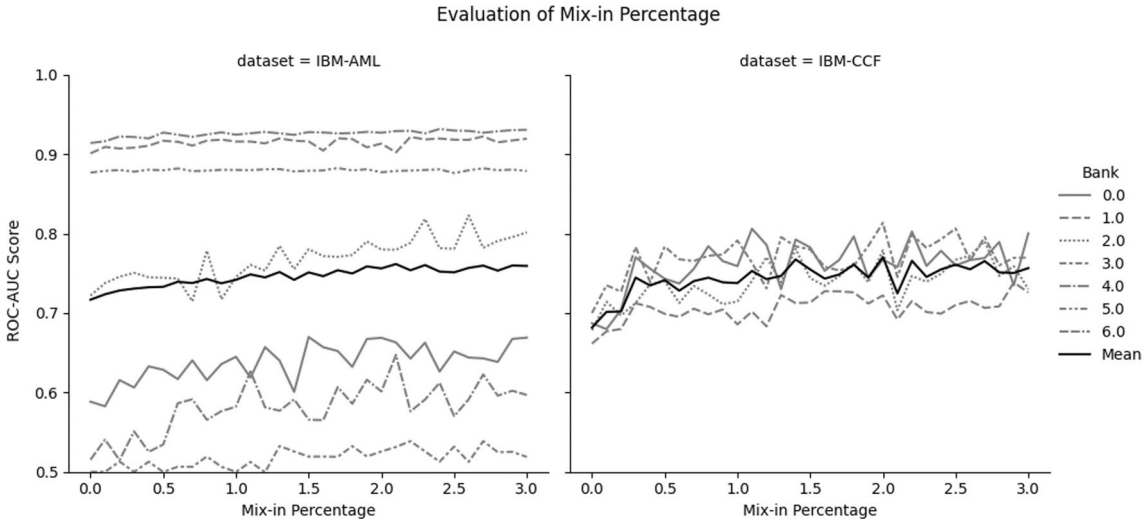


**Fig. 8** Effect of synthetic data mix-in percentage on performance

**Table 8** Comparison between different synthetic data generation models

| Dataset | Method | ROC AUC score | Dataset | Method | ROC AUC score |
|---|---|---|---|---|---|
| IBM-AML | Without shared data | 0.7168 | IBM-CCF | Without shared data | 0.6817 |
| | Full | 0.7371 | | Full | 0.7042 |
| | fullOS_10 | 0.6435 | | fullOS_10 | 0.6618 |
| | fullOS_20 | 0.6199 | | fullOS_20 | 0.6360 |
| | sep | 0.7209 | | sep | 0.6817 |
| | **sepPre** | **0.7473** | | **sepPre** | **0.7323** |

exhibit a notable decline in performance. Thus, leading us to the conclusion, that for financial transaction data, oversampling the data before training the synthetic data generation model is not suitable. Moreover, the "full" setup outperforms configurations where synthetic data generators are trained separately for each class ("sep"). This subpar performance stems from the "sep" model's poor-quality synthetic data for the minority class, which fails to capture training data patterns due to limited training dataset size. However, when the minority class model is pre-trained using data from the majority class ("sepPre"), a significant performance improvement is observed, surpassing all other methods. This enhancement is primarily due to the model's capacity to generate higher-quality samples of the minority class with greater variability. Further discussions with partner institution experts emphasized the advantage of creating class data separately as it enhances privacy by preventing leaks of sensitive information like fraud rates by independently

producing the samples for each class. Consequently, we have refined DP4 to encapsulate these insights: DP2—Provide the system with the ability to identify, validate, and apply context-specific synthetic data generation techniques with mutually agreed on over-sampling in order to remove private data, given guidelines or regulations on data sharing.

## Cycle 4: Network effects of financial data sharing

In cycle four of our DSRM project, we delve into the global data layer, guided by the literature and expert insights to address cooperative challenges within the proposed synthetic financial data ecosystems. Aiming to refine our DPs to enhance the ecosystem's capability to effectively manage these challenges.

### Design of mechanisms at the global data level

Addressing the second aspect of expert feedback and informed by the literature on data ecosystems, this cycle focuses on the global data layer and its DPs to ensure that the created ecosystem is able to handle the challenges of data ecosystems described by Gelhaar and Otto (2020). Because cooperative challenges play a dominant role in the early stage of an ecosystem, the following cycle will focus on these (Autio & Thomas, 2014). In their paper, Gelhaar and Otto (2020) describe four major cooperative challenges that need to be addressed for a data ecosystem to emerge successfully. First, it is necessary to build trust between the participants. Second, it needs to be shown that all actors benefit from participating in the ecosystem. Third, it is important to identify the right number of participants. Fourth, interoperability needs to be enabled through the agreement on standards. Thus, the focus of this section is to evaluate existing DPs through this lens and analyze whether refinements or additional principles are necessary for the development of an ecosystem capable of effectively addressing these challenges. First, trust between ecosystem partners can be built in multiple ways. On the one hand, trust can be increased by adequate control mechanisms (Geisler et al., 2021), which is already reflected in DP3—Provide the system with a back-testing mechanism in order to ensure newly generated synthetic data matches in composition and fraud detection training performance with real data given that data quality cannot be independently verified which ensures sufficient data quality in the synthetic data ecosystem. On the other hand, Majava et al. (2016) show that intermediaries play a significant role in increasing participants' trust in an ecosystem. In the financial services ecosystem, this role is typically held by public regulators. To incentivize them to participate in the ecosystem and allow them to ensure data quality and thus increase trust, we propose DP6—Provide access for external collaborators, such as regulators, to leverage the synthetic data within the ecosystem given a diverse set of synthetic data available. This gives regulators access to the ecosystem while adhering to the existing privacy measures. However, it remains unclear if access to purely synthetic data can provide enough value and thus incentivize their participation in the ecosystem. The next challenge data ecosystems face is that all actors need to benefit from participating in the ecosystem. While we already demonstrated in previous iterations that our data ecosystem is able to increase the overall performance, it remains unclear how this performance gain is distributed between institutions. To address this, further investigation is needed to check if adjustments to our design need to be made to create sufficient incentives for all institutions. Connected to this problem is identifying the right number of participants. While our previous cycles show that the ecosystem is beneficial if all institutions participate, it remains unclear if a similar effect exists, if only part of the institutions is included in the ecosystem. To incorporate this into our DPs, DP3 was extended to not only describe the monitoring of outgoing synthetic data but also cover the evaluation of performance gained by using the shared synthetic data from the data ecosystem. This results in DP3—Provide the system with a back-testing mechanism in order to ensure newly generated synthetic data matches in composition and fraud detection training performance with real data given that data quality cannot be independently verified. The last cooperative challenge that needs to be overcome is interoperability through the agreement on standards. At the moment, that is already incorporated in DP1—Provide the system with modular systems design in order to ensure independence of local data and cross-institutional proliferation of synthetic data given that the raw data is sensitive, from a data perspective where the local layer of the ecosystem is used to align the data so that it can be easily shared with the system later. Furthermore, we argue that creating DPs for the financial data ecosystem contributes to the standardization of the ecosystem from an infrastructure and ecosystem perspective and thus by creating these DPs we contribute to overcoming this challenge.

### Demonstration through the introduction of non-sharing entities and individual performance benchmarking

In this part, we further improved the prototype developed in Python, by altering the global data layer to allow for the participation of entities that do not contribute data. Additionally, we updated the system to track and report the

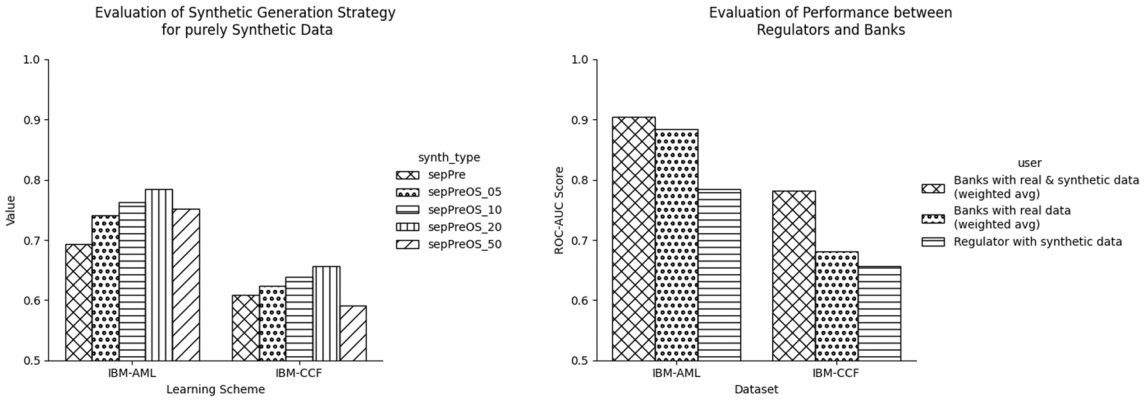


**Fig. 9** Regulator models using different resampled data and performance of the regulator model (only synthetic data) vs. the bank models

performance to each participating institution, thus allowing institutions on an individual level to see the performance gain from engaging in the data ecosystem.[9]

## Evaluation of ecosystems with non-sharing entities and individual performance benchmarking

Similar to the previous cycle, this evaluation again utilizes the two synthetic datasets (IBM-AML and IBM-CFF), due to their high data quality, size, and diversity. Moreover, the ecosystem setup and evaluation scheme are adopted from the previous cycle, utilizing the "sepPre" training scheme.

We start with evaluating DP6, which allows regulators to access purely synthetic data within the data ecosystem. To validate this DP, the synthetic data that is provided to the regulators need to be of sufficient quality for them to derive meaningful insights and effectively improve their models. However, as this cannot be easily evaluated, we use the performance of a prediction model trained on the data available to the regulator (only synthetic data) as a proxy for the quality of the data. As the architecture chosen in Cycle 3 generates separate models for different classes, more data of a specific class can easily be generated. This is especially relevant for cases where only synthetic data is used as no additional positive samples from the real data exist. Thus, the experiment conducted had two steps. In the first step, regulator models were trained on synthetic data with different amounts of minority class samples (indicated by OS_{percentage of minority class cases}). From this selection, the over-sampling ratio with the best performance was chosen and compared to the performance of the models trained at the different banks, once trained on a combination of real and synthetic data, and once trained with only real data. The results of this experiment can be seen in the following diagram (Fig. 9).

The regulator model, trained exclusively on synthetic data, exhibits performance that, while not matching that of the bank's internal models (trained on a mix of real and synthetic data), remains significant. The model approaches the performance of the bank's baseline models (trained on real data only), as illustrated in Fig. 9. This capability offers considerable advantages to collaborators who would otherwise lack access to such data. Consequently, allowing regulators to access synthetic data emerges as an effective strategy to foster collaboration and enhance trust in the ecosystem. Therefore, DP6—Provide access for external collaborators, such as regulators, to leverage the synthetic data within the ecosystem given a diverse set of synthetic data available is validated and was added to our DPs for synthetic data ecosystems.

Subsequently, the adjustment to DP3 is validated, checking if all banks profit from the synthetic data ecosystem and evaluating if the synthetic ecosystem including fewer institutions is still able to profit from the network effects of the ecosystem. To investigate this, we plot the performance of each institution against its baseline (score without any artificial data), which can be seen below (Fig. 10).

As evidenced in Fig. 10 for each single bank in both datasets, the performance increases by combining real and synthetic data. Furthermore, looking at the rightmost panel of Fig. 10, it can clearly be seen that there is a negative correlation ($-0.09$) between the performance gained by participating in the ecosystem and the size of the bank. Thus, showing that small banks over proportionally profit from participation, providing a clear incentive for them to engage in the ecosystem. However, even if absolute performance gained by bigger banks is lower, we argue that they still have a sufficient incentive to participate due to their large volume of transactions, where even small changes in the

[9] The full implementation of Cycle 4 can be found here: https://github.com/Faruman/SyntheticDataEcosystems/blob/master/Cycle3-4_EcosystemEvaluation/README.MD

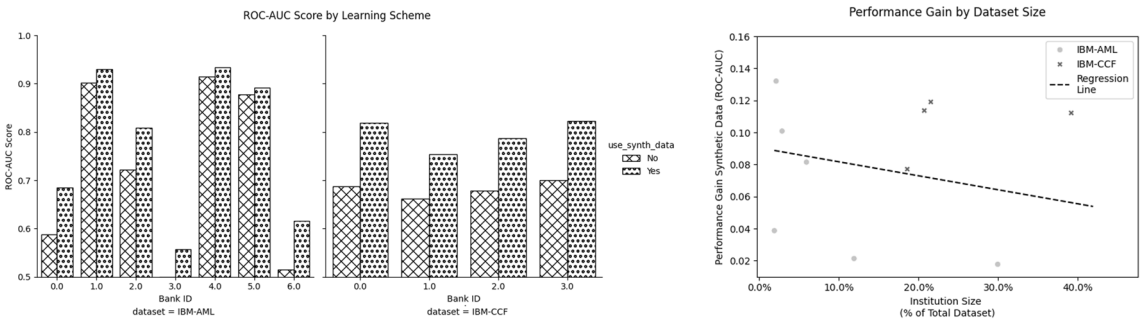


**Fig. 10** Performance gain per individual bank and performance gain by institution size

fraud detection percentage result in a high absolute sum of prevented losses. These results lead us to the conclusion that all banks contributing to the ecosystem profit from their involvement and thus the designed ecosystem is able to overcome another one of the previously outlined challenges.

Next, we investigate our synthetic data ecosystem for cases where not all institutions engage in synthetic data sharing. To achieve this, we simulated environments, where none, 50%, 75%, or 100% of all banks were part of the ecosystem. The results can be seen in Fig. 11.

Despite the significant difference between the two data sets regarding their data distribution (with IBM-CCF having an equal distribution between banks, while IBM-AML has a highly skewed one), we can clearly see that in both cases, even with only half of the banks being part of the ecosystem (IBM-CCF: 2 banks/IBM-AML: 3 banks), a significant performance gain is achieved. Thus, it seems the benefits of the synthetic data ecosystem can be realized from an early stage onwards, making it easy to overcome the hurdle of a minimum number of members needing to participate in the ecosystem, thus tackling another of the challenges outlined previously.

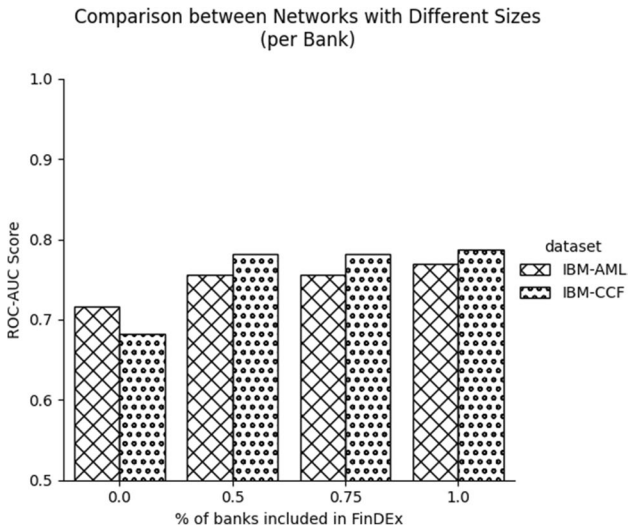


**Fig. 11** Performance (avg per bank) by percentage of participating institutions

Summarizing these results, we were able to demonstrate that the proposed data ecosystem is able to deliver excess performance for all participants in the network on an individual level and it can be seen that even for data ecosystems with only a fraction of the institutions participating in synthetic data sharing, still a significant performance gain can be achieved. Furthermore, there seem to be network effects to some extent where more partners in the ecosystem increase its overall utility. As these results validate the incentives for partners to participate in an ecosystem, we confirm our DP3—Provide the system with a back-testing mechanism in order to ensure newly generated synthetic data matches in composition and fraud detection training performance with real data given that data quality cannot be independently verified.

## Discussion

This research paper is aimed at extending the research on privacy in data ecosystems as well as machine learning of multi-organizational datasets by investigating these challenges in the field of financial fraud detection. This was done by deriving DPs for an innovative synthetic data-sharing ecosystem that allows financial institutions to exchange financial transaction data while protecting client privacy and learning effectively from this multi-institutional data. To create this artifact, we followed the process of DSRM (Peffers et al., 2007), with this paper covering four "design-implement-evaluate" cycles. Starting with the problem identification our study contributes to descriptive knowledge concerning the problem space by identifying data scarcity in combination with the inability to share data due to privacy protection as a major hurdle for financial institutions, validating the existing research on cross-organizational fraud detection collaboration within financial services (Abdul Salam et al., 2024; Kong et al., 2024). During the exploration of the solution, space synthetic data sharing

was identified as an underexplored solution to tackle data scarcity in financial fraud detection extending the literature on cross-organizational collaboration in the field (Chatterjee et al., 2024). Furthermore, the exploration of synthetic data to allow privacy-compliant data sharing as well as our experimentation on multi-organizational synthetic data during multiple "design-implement-evaluate" cycles reaches beyond financial services and addresses significant challenges in the realm of data ecosystems (Brée et al., 2024). Moreover, our research extends beyond studies that simply outline the requirements of such a data ecosystem (Immonen et al., 2014). We validate these requirements and the derived DPs through rigorous experimentation on publicly available datasets and through close collaboration with industry partners and experts, ensuring the practical applicability and robustness of our findings. Furthermore, by extending data ecosystem research into a less frequently explored domain (Cappiello et al., 2020), we are able to validate the applicability of existing knowledge and uncover new insights with potential for generalization. We achieve this by developing prescriptive knowledge and nascent theory concerning the solution space, offering a set of DPs for designing a synthetic data sharing ecosystem and providing a first instantiation in the form of a platform architecture. To provide more detailed insights into this solution space, additional key findings are encapsulated in Table 9, clustered by key areas which we deductively derived a posteriori from our study.

As shown in Table 9 under the generation dimension, we contribute to the literature on synthetic data generation in multiple ways. First, we identified the necessity for a strictly separated local layer (where real data is transformed) and a global layer (where data is shared). Second, we transfer existing algorithms to a new setup including cross-organizational data with a complex data structure and compare their performance on a prediction task (Pathare et al., 2023) identifying TVAE as the most performant algorithm for synthetic financial data generation while still showing sufficient privacy. Third, we extend the research on the generation setup by consolidating different training schemes from multiple sources (Eilertsen et al., 2021; Fan et al., 2022; Kiran & Kumar, 2024) and comparing them to each other, identifying training on data sub-clusters as the most beneficial setup.

Moving forward to training models based on synthetic data, as shown in Table 9 under the prediction dimension, we extend the literature which often looks at synthetic data generation performance separately but provides little guidance on how the generated data is best used in a data ecosystem (Dankar et al., 2022). Our research further shows that a mixture of synthetic and real data is most useful when combined; however, the exact mix-in percentage is highly organization and context-specific. Moreover, we demonstrated that using purely synthetic data can still be beneficial for players with no access to real data; however, adjustments need to be made to the composition of the data by artificially rebalancing it.

As can be seen in Table 9's ecosystem dimension, our research investigates the complexities of data ecosystems, analyzing how the incentives for participation affect performance outcomes across various sizes of institutions. This analysis also places our findings in the context of the research by Gelhaar and Otto (2020) about the initial challenges encountered within data ecosystems. By implementing design interventions that clearly articulate performance benefits and facilitate the integration of external collaborators, our research substantiates the ecosystem's capacity to overcome these early hurdles. Further, our empirical evidence suggests that even partial participation in the data ecosystem can lead to substantial performance improvements, thereby affirming the ecosystem's operational feasibility and enhancing its attractiveness to potential participants.

In the last dimension in Table 9, we demonstrate the generalizability of our derived design knowledge beyond a single use case and application area. This was done in two ways. First, through validation with experts from the field in academia and the private sector. Second, through performance evaluation in two financial services domains and three datasets which required data sharing with privacy restrictions. While performance gains might seem insignificant, small changes in fraud detection rate can have major implications on financial institutions (Levi, 1998). Thus, our research not only confirms the relevance of our DPs and system architecture but also sets the stage for their application beyond the immediate context of financial transactions, suggesting a blueprint for extending beyond financial services to other domains where data needs to be shared with privacy restrictions (Susha et al., 2019).

For practitioners, our contribution is two-fold: For managers and decision-makers, we demonstrate the value of synthetic data-sharing ecosystems that allow both large and small institutions to securely collaborate on data while ensuring privacy. This approach is particularly relevant in industries with complex, highly sensitive data, such as financial services, where data ecosystems do not emerge organically and require careful planning allowing for shared value propositions and services (Adner, 2017; Immonen et al., 2014). Furthermore, our framework addresses regulatory requirements on data privacy and our results suggest a robust foundation for scaling and sustaining privacy-focused data ecosystems. For system architects, we outline a set of DPs that guide practitioners in structuring the architecture of these ecosystems. These principles assist in selecting suitable synthetic data generation methods, implementing mechanisms for data quality assurance, and integrating data to enhance AI model performance. By focusing on these core areas, our contribution provides architects with actionable guidance toward building secure and resilient synthetic

**Table 9** Summary of insights generated in the four design-implement-evaluate cycles

| Area of insight | Focus | Activities | Reasoning |
|---|---|---|---|
| Generation | This dimension identifies the most suitable synthetic data generation algorithm and the optimal setup for generating synthetic data to enhance performance | - Performance comparison of popular synthetic data generation algorithms on real-world financial transaction data (outperformance of TVAE by 49.8%* compared to next best model)<br>- Evaluation of the optimal training setup when generating synthetic data (outperformance of pre-trained class-separated models by 2.0%* compared to next best training scheme) | The analyses lead to the selection of a modular, algorithm-agnostic approach that clearly separates between local and global layers. Further, optimal data integration strategies were identified broadening the ecosystem's applicability and enhancing its performance potential |
| Prediction | This dimension focuses on the specifics of training setups for synthetic data generation, especially in contexts with imbalanced classes | - Evaluation of the optimal mix-in percentage between real and synthetic data on the institutional level<br>- Performance evaluation of purely synthetic data using different degrees of oversampling (outperformance of data with 10% oversampling by 7.5%* compared to not oversampled baseline) | These investigations facilitate the identification of tailored training approaches that accommodate the specific needs of the ecosystem's participants, ensuring the generation of high-quality, utility-maximizing prediction models |
| Ecosystem | This dimension assesses the ecosystem's balance in terms of participation incentives across institutions of varying sizes and the effect of partial participation | - Analysis of performance gain per financial institution and the relationship between institution size and performance gain (negative correlation of − 0.09)<br>- Evaluation of the ecosystem's performance with varying numbers of participating institutions (significant performance gain with all ecosystem sizes) | The findings indicate a proportional benefit across the ecosystem, highlighting particular advantages for smaller institutions. Furthermore, even partial ecosystem participation yields substantial performance improvements. Thus, ensuring the system's viability and attractiveness |
| Generalizability | The dimension of generalizability aims to validate the created design knowledge and system architecture across diverse contexts and experts' perspectives | - Semi-structured interviews with experts from financial institutions and academic experts to validate the created DPs and system architecture<br>- Transfer of the synthetic data sharing setup to new contexts like fraud detection (improved performance by 3.6%*) and money laundering detection (improved performance by 6.6%*) using simulation-based datasets | The evaluation confirmed the relevance and applicability of the design knowledge across multiple financial services domains |

*Performance evaluation based on ROC AUC scores

data-sharing ecosystems. This framework, therefore, serves as a blueprint for future system designers working within regulated environments where data privacy and AI performance are essential.

Limitations and future research opportunities can be identified across our four key areas of insight. Regarding data generation, the current study was constrained by the available data, which prevented the consideration of advanced graph-based synthetic data generation methods such as TransGAN (X. Wang & Yang, 2024). Additionally, while privacy was tested, it was not fully guaranteed by the models used, highlighting the need for future research on the effectiveness of differentially private synthetic data generation methods such as PATEGAN (Jordon et al., 2018) in a synthetic data ecosystem. From a prediction standpoint, further investigation is required to determine how models can be aligned when data schemas—and thus the synthetic data—differ between institutions. Moreover, the design of an effective back-testing mechanism to ensure the ecosystem's predictive performance should be explored. On the ecosystem level, additional research is necessary to explore ecosystem usage incentives, building on the work by (Gelhaar et al., 2021), which was beyond the scope of this paper. Finally, while this study was limited to financial services due to resource constraints, future research should explore the applicability of the defined DPs beyond this domain, testing their general applicability.

## Conclusion

Based on the need for increased data availability to foster economic growth, this paper provides the design and evaluation of a synthetic data-sharing ecosystem for financial institutions under privacy constraints. The main contribution lies in providing guidance on how to train models based on shared data. By formulating a set of DPs, practical insights, and prototype testing, iterative design cycles were used to provide a robust framework for constructing a data ecosystem that leverages synthetic data. Each DP, from ensuring data quality and enhancing adaptability through transformation and resampling to fostering trust among ecosystem participants and facilitating regulatory access to synthetic data, extends existing research on synthetic data sharing and generation, particularly in the context of financial transaction data. For practice, our example instantiation and codebase can be used as a reference architecture for future instantiations. We not only address the identified need for an efficient, privacy-preserving financial data ecosystem but also set a foundation for future exploration in broader domains where data sharing under privacy restrictions is paramount. Thus, this contribution offers guidance for overcoming technical, trust-related, and regulatory challenges in data ecosystems, unlocking the potential for data-driven innovation and future economic development.

**Acknowledgements** The authors express their gratitude to the UnionBank of the Philippines for their valuable collaboration and the provision of key insights that contributed to this research. This research project was funded by the St. Gallen Symposium and the German Federal Ministry of Education and Research (BMBF) within the "Innovations for Tomorrow's Production, Services, and Work" Program (funding number 02K23A001) which is managed by the Project Management Agency Karlsruhe (PTKA). The authors are responsible for the content of this publication.

**Funding** Open access funding provided by University of St. Gallen.

**Data Availability** The datasets used during the current study are available in the Kaggle repositories: https://www.kaggle.com/c/ieee-fraud-detection, https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml, https://www.kaggle.com/datasets/ealtman2019/credit-card-transactions.

## Declarations

**Competing Interests** The authors declare that they have no conflict of interest.



## References

Abbas, A. E., Agahari, W., van de Ven, M., Zuiderwijk, A., & de Reuver, M. (2021). Business data sharing through data marketplaces: A systematic literature review. *Journal of Theoretical and Applied Electronic Commerce Research, 16*(7), 7. https://doi.org/10.3390/jtaer16070180

Abdul Salam, M., Fouad, K. M., Elbably, D. L., & Elsayed, S. M. (2024). Federated learning model for credit card fraud detection with data balancing techniques. *Neural Computing and Applications, 36*(11), 6231–6256. https://doi.org/10.1007/s00521-023-09410-2

Adner, R. (2017). Ecosystem as structure: An actionable construct for strategy. *Journal of Management, 43*(1), 39–58. https://doi.org/10.1177/0149206316678451

Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review, 40*, 100402. https://doi.org/10.1016/j.cosrev.2021.100402

Altman, E., Blanuša, J., von Niederhäusern, L., Egressy, B., Anghel, A., & Atasu, K. (2024). *Realistic synthetic financial transactions for anti-money laundering models* (No. arXiv:2306.16424). arXiv. https://doi.org/10.48550/arXiv.2306.16424

Altman, E. R. (2019). *Synthesizing credit card transactions* (No. arXiv:1910.03033). arXiv. https://doi.org/10.48550/arXiv.1910.03033

Asrow, K. (2021). *The role of individuals in the data ecosystem: Current debates and considerations for individual data protection and data rights in the U.S.* Federal Reserve Bank of San Francisco. https://privacysecurityacademy.com/wp-content/uploads/2021/05/The-Role-of-Individuals-in-the-Data-Ecosystem.pdf. Accessed 9 Mar 2023.

Assefa, S. (2020). Generating synthetic data in finance: Opportunities, challenges and pitfalls. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3634235

Aurna, N. F., Hossain, M. D., Taenaka, Y., & Kadobayashi, Y. (2023). Federated learning-based credit card fraud detection: Performance analysis with sampling methods and deep learning algorithms. *2023 IEEE International Conference on Cyber Security and Resilience (CSR), 2023*, 180–186. https://doi.org/10.1109/CSR57506.2023.10224978

Autio, E., spsampsps Thomas, L. D. W. (2014). Innovation ecosystems: Implications for innovation management? In M. Dodgson, D. M. Gann, spsampsps N. Phillips (Eds.), *The Oxford Handbook of Innovation Management* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199694945.013.012

Awosika, T., Shukla, R. M., & Pranggono, B. (2024). Transparency and privacy: The role of explainable AI and federated learning in financial fraud detection. *IEEE Access, 12*, 64551–64560. https://doi.org/10.1109/ACCESS.2024.3394528. IEEE Access.

Baabdullah, T., Alzahrani, A., Rawat, D. B., & Liu, C. (2024). Efficiency of federated learning and blockchain in preserving privacy and enhancing the performance of credit card fraud detection (CCFD) systems. *Future Internet, 16*(6), 6. https://doi.org/10.3390/fi16060196

Bagad, P., Mitra, S., Dhamnani, S., Sinha, A. R., Gautam, R., & Khanna, H. (2021). Data-sharing economy: Value-addition from data meets privacy. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 1105–1108. https://doi.org/10.1145/3437963.3441712

Baskerville, R., Baiyere, A., Gregor, S., Hevner, A., & Rossi, M. (2018). Design science research contributions: Finding a balance between artifact and theory. *Journal of the Association for Information Systems*, *19*(5). https://aisel.aisnet.org/jais/vol19/iss5/3. Accessed 13 June 2023.

Bauer, A., Trapp, S., Stenger, M., Leppich, R., Kounev, S., Leznik, M., Chard, K., & Foster, I. (2024). Comprehensive exploration of synthetic data generation: A survey (No. arXiv:2401.02524). arXiv. https://arxiv.org/abs/2401.02524. Accessed 15 Aug 2024.

Benchaji, I., Douzi, S., & Ouahidi, B. E. (2021). Credit card fraud detection model based on LSTM recurrent neural networks. *Journal of Advances in Information Technology, 12*(2), 113–118. https://doi.org/10.12720/jait.12.2.113-118

Bian, K., & Zheng, H. (2023). FedAvg-DWA: A novel algorithm for enhanced fraud detection in federated learning environment. *2023 4th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 13–17. https://doi.org/10.1109/ICBAIE59714.2023.10281317

Blake, M., McWaters, J., & Galaski, R. (2019). *The next generation of data-sharing in financial services* (p. 33) [White Paper]. World Economic Forum. https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/financial-services/lu-next-generation-data-sharinging-financial-services.pdf. Accessed 29 Jan 2023.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences, 99*(suppl_3), 7280–7287. https://doi.org/10.1073/pnas.082080899

Brée, T., Karger, E., & Ahlemann, F. (2024). Shaping the future of data ecosystem research—What is still missing? *IEEE Access*, *12*, 103162–103175. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3432969

Brodsky, L., & Oakes, L. (2017). *Data sharing and open banking*. McKinsey. https://www.mckinsey.com/~/media/McKinsey/Industries/Financial%20Services/Our%20Insights/Data%20sharing%20and%20open%20banking/Data-sharing-and-open-banking.pdf. Accessed 26 Feb 2024.

Bun, M., Desfontaines, D., Dwork, C., Naor, M., Nissim, K., Roth, A., Smith, A., Steinke, T., Ullman, J., & Vadhan, S. (2021). *Statistical inference is not a privacy violation*. https://differentialprivacy.org/inference-is-not-a-privacy-violation/. Accessed 14 Aug 2024.

Cappiello, C., Gal, A., Jarke, M., & Rehof, J. (2020). Data ecosystems: Sovereign data exchange among organizations (Dagstuhl Seminar 19391). *DROPS-IDN/v2/Document/*https://doi.org/10.4230/DagRep.9.9.66. *https://doi.org/10.4230/DagRep.9.9.66*

Chakravarty, S., Demirhan, H., & Baser, F. (2020). Fuzzy regression functions with a noise cluster and the impact of outliers on mainstream machine learning methods in the regression setting. *Applied Soft Computing, 96*, 106535. https://doi.org/10.1016/j.asoc.2020.106535

Chandra, L., Seidel, S., & Gregor, S. (2015). Prescriptive knowledge in IS research: Conceptualizing design principles in terms of materiality, action, and boundary conditions. *2015 48th Hawaii International Conference on System Sciences*, 4039–4048. https://doi.org/10.1109/HICSS.2015.485

Charitou, C., Dragicevic, S., & Garcez, A. d'Avila. (2021). *Synthetic data generation for fraud detection using GANs* (No. arXiv:2109.12546). arXiv. http://arxiv.org/abs/2109.12546. Accessed 11 Mar 2024.

Chatterjee, P., Das, D., & Rawat, D. B. (2024). Digital twin for credit card fraud detection: Opportunities, challenges, and fraud detection advancements. *Future Generation Computer Systems, 158*, 410–426. https://doi.org/10.1016/j.future.2024.04.057

Chen, Z., Van Khoa, L. D., Teoh, E. N., Nazir, A., Karuppiah, E. K., & Lam, K. S. (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: A review. *Knowledge and Information Systems, 57*(2), 245–285. https://doi.org/10.1007/s10115-017-1144-z

Chen, H., Grossman, M., Sen, A., & Tsao, S.-F. (2023). Establishing a FAIR, CARE, and efficient synthetic health data sharing ecosystem for canada establishing a FAIR, CARE, and efficient synthetic health data sharing ecosystem for Canada. IARIW-CIGI Conference on the Valuation of Data. https://www.researchgate.net/publication/375446378_Establishing_a_FAIR_CARE_and_Efficient_Synthetic_Health_Data_Sharing_Ecosystem_for_Canada_Establishing_a_FAIR_CARE_and_Efficient_Synthetic_Health_Data_Sharing_Ecosystem_for_Canada. Accessed 17 Dec 2024

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2018). *Generating multi-label discrete patient records using generative adversarial networks* (No. arXiv:1703.06490). arXiv. https://doi.org/10.48550/arXiv.1703.06490

Cosma, S., Cosma, S., & Pennetta, D. (2023). The rise of financial services ecosystems: Towards open banking platforms. In T. Walker, E. Nikbakht, & M. Kooli (Eds.), *The Fintech Disruption: How Financial Innovation Is Transforming the Banking Industry* (pp. 191–213). Springer International Publishing. https://doi.org/10.1007/978-3-031-23069-1_8

Dahmen, J., & Cook, D. (2019). SynSys: A synthetic data generation system for healthcare applications. *Sensors, 19*(5), 5. https://doi.org/10.3390/s19051181

Dankar, F. K., Ibrahim, M. K., & Ismail, L. (2022). A multi-dimensional evaluation of synthetic data generators. *IEEE Access, 10*,

11147–11158. https://doi.org/10.1109/ACCESS.2022.3144765. IEEE Access.

Demirgüç-Kunt, A., Klapper, L., Singer, D., & Ansar, S. (2022). *The global findex database 2021—Financial inclusion, digital payments, and resilience in the age of COVID-19*. International Bank for Reconstruction and Development / The World Bank. https://openknowledge.worldbank.org/bitstream/handle/10986/37578/9781464818974.pdf. Accessed 22 Jan 2023.

Eilertsen, G., Tsirikoglou, A., Lundström, C., & Unger, J. (2021). *Ensembles of GANs for synthetic training data generation* (No. arXiv:2104.11797). arXiv. https://doi.org/10.48550/arXiv.2104.11797

Emam, K. E., Mosquera, L., & Bass, J. (2020). Evaluating identity disclosure risk in fully synthetic health data: Model development and validation. *Journal of Medical Internet Research, 22*(11), e23139. https://doi.org/10.2196/23139

Esteban, C., Hyland, S. L., & Rätsch, G. (2017). *Real-valued (medical) time series generation with recurrent conditional GANs* (No. arXiv:1706.02633). arXiv. http://arxiv.org/abs/1706.02633. Accessed 14 Aug 2024.

European Central Bank. (2021). *Seventh report on card fraud. 2021*. https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport202110~cac4c418e8.en.html. Accessed 16 June 2023.

Fan, X., Guo, X., Chen, Q., Chen, Y., Wang, T., & Zhang, Y. (2022). Data augmentation of credit default swap transactions based on a sequence GAN. *Information Processing & Management, 59*(3), 102889. https://doi.org/10.1016/j.ipm.2022.102889

Fassnacht, M. K., Benz, C., Leimstoll, J., & Satzger, G. (2023). Is your organization ready to share? A framework of beneficial conditions for data sharing. *44th International Conference on Information Systems (ICIS 2023), Hyderabad, Indien, 10.12.2023 - 13.12.2023*. https://doi.org/10.5445/IR/1000162812

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 289–293. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). https://doi.org/10.1109/ISBI.2018.8363576

Geisler, S., Vidal, M.-E., Cappiello, C., Lóscio, B. F., Gal, A., Jarke, M., Lenzerini, M., Missier, P., Otto, B., Paja, E., Pernici, B., & Rehof, J. (2021). Knowledge-driven data ecosystems toward data transparency. *Journal of Data and Information Quality, 14*(1), 3:1-3:12. https://doi.org/10.1145/3467022

Gelhaar, J., & Otto, B. (2020). *Challenges in the emergence of data ecosystems*. PACIS 2020 Proceedings. 175. https://aisel.aisnet.org/pacis2020/175/. Accessed 17 Dec 2024.

Gelhaar, J., Groß, T., & Otto, B. (2021). A taxonomy for data ecosystems. *Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2021.739

Gelhaar, J., Henke, M., Gürpinar, T., & Otto, B. (2021). *Towards a taxonomy of incentive mechanisms for data sharing in data ecosystems*. PACIS 2021 Proceedings. 121. https://aisel.aisnet.org/pacis2021/121/. Accessed 17 Dec 2024.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative adversarial networks* (No. arXiv:1406.2661). arXiv. https://doi.org/10.48550/arXiv.1406.2661

Gray, B., & Sites, J. P. (2013). *Sustainability through partnerships*. Network for business sustainability. https://nbs.net/wp-content/uploads/2022/01/NBS-Systematic-Review-Partnerships.pdf. Accessed 18 Dec 2024.

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly, 37*(2), 337–355.

Gregor, S., Chandra Kruse, L., & Seidel, S. (2020). The anatomy of a design principle. *Journal of the Association for Information Systems, 21*, 1622–1652. https://doi.org/10.17705/1jais.00649

Gregor, S., Kruse, L. C., & Seidel, S. (2020). Research perspectives: The anatomy of a design principle. *Journal of the Association for Information Systems*, *21*(6). https://doi.org/10.17705/1jais.00649

Gröger, C. (2021). There is no AI without data. *Communications of the ACM, 64*(11), 98–108. https://doi.org/10.1145/3448247

Hashemi, S. K., Mirtaheri, S. L., & Greco, S. (2023). Fraud detection in banking data by machine learning techniques. *IEEE Access, 11*, 3034–3043. https://doi.org/10.1109/ACCESS.2022.3232287.

He, Z., Huang, J., & Zhou, J. (2023). Open banking: Credit market competition when borrowers own the data. *Journal of Financial Economics, 147*(2), 449–474. https://doi.org/10.1016/j.jfineco.2022.12.003

Heimstädt, M., Saunderson, F., & Heath, T. (2014). *Conceptualizing Open Data ecosystems: A timeline analysis of Open Data development in the UK*. [13] S. https://doi.org/10.17169/FUDOCS_DOCUMENT_000000020332

Heinz, D., Benz, C., Fassnacht, M., & Satzger, G. (2022). *Past, present and future of data ecosystems research: A systematic literature review*. PACIS 2022 Proceedings. 46. https://aisel.aisnet.org/pacis2022/46/. Accessed 18 Dec 2024.

Hevner, A. R., March, S., Park, J., & Ram, S. (2004). Design science in information systems research. *Management Information Systems Quarterly*, *28*(1). https://aisel.aisnet.org/misq/vol28/iss1/6. Accessed 13 Nov 2024.

Hevner, A., & Gregor, S. (2022). Envisioning entrepreneurship and digital innovation through a design science research lens: A matrix approach. *Information & Management, 59*(3), 103350. https://doi.org/10.1016/j.im.2020.103350

Hevner, A. R. (2007). *A three cycle view of design science research*. Scandinavian Journal of Information Systems: Vol. 19: Iss. 2, Article 4. https://aisel.aisnet.org/sjis/vol19/iss2/4/. Accessed 18 Dec 2024.

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications, 193*, 116429. https://doi.org/10.1016/j.eswa.2021.116429

Hittmeir, M., Ekelhart, A., & Mayer, R. (2019). On the utility of synthetic data: An empirical evaluation on machine learning tasks. *Proceedings of the 14th International Conference on Availability, Reliability and Security*, 1–6. https://doi.org/10.1145/3339252.3339281

Iivari, J. (2007). A paradigmatic analysis of information systems as a design science. *Scandinavian Journal of Information Systems, 19*, 39.

Immonen, A., Palviainen, M., & Ovaska, E. (2014). Requirements of an open data based business ecosystem. *IEEE Access, 2*, 88–103. https://doi.org/10.1109/ACCESS.2014.2302872

Ishfaq, H., Hoogi, A., & Rubin, D. (2023). TVAE: Triplet-based variational autoencoder using metric learning (No. arXiv:1802.04403). arXiv. https://arxiv.org/abs/1802.04403 . Accessed 16 June 2023.

Jarke, M. (2017). Data spaces: Combining goal-driven and data-driven approaches in community decision and negotiation support. In M. Schoop spsampsps D. M. Kilgour (Eds.), *Group Decision and Negotiation. A Socio-Technical Perspective* (pp. 3–14). Springer International Publishing. https://doi.org/10.1007/978-3-319-63546-0_1

Jensen, R. I. T., Ferwerda, J., Jørgensen, K. S., Jensen, E. R., Borg, M., Krogh, M. P., Jensen, J. B., & Iosifidis, A. (2023). A synthetic data set to benchmark anti-money laundering methods. *Scientific Data, 10*(1), 661. https://doi.org/10.1038/s41597-023-02569-2

Jiang, D., Zhang, G., Karami, M., Chen, X., Shao, Y., & Yu, Y. (2022). *DP$^2$-VAE: Differentially private pre-trained variational autoencoders* (No. arXiv:2208.03409). arXiv. https://doi.org/10.48550/arXiv.2208.03409

Jones, D., & Gregor, S. (2007). The anatomy of a design theory. *Journal of the Association for Information Systems, 8*(5), 1.

Jordon, J., Yoon, J., & Schaar, M. van der. (2018, September 27). PATE-GAN: Generating synthetic data with differential privacy guarantees. *International Conference on Learning Representations*. https://openreview.net/forum?id=S1zk9iRqF7. Accessed 14 Aug 2024.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). *Synthetic data—What, why and how?* (No. arXiv:2205.03257). arXiv. http://arxiv.org/abs/2205.03257. Accessed 14 Aug 2024.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *CoRR*. https://www.semanticscholar.org/paper/Auto-Encoding-Variational-Bayes-Kingma-Welling/5f5dc5b9a2ba710937e2c413b37b053cd673df02. Accessed 16 May 2024.

Kiran, A., & Kumar, S. S. (2024). A methodology and an empirical analysis to determine the most suitable synthetic data generator. *IEEE Access, 12*, 12209–12228. https://doi.org/10.1109/ACCESS.2024.3354277

Kong, Y., Li, Z., & Jiang, C. (2024). ASIA: A federated boosting tree model against sequence inference attacks in financial networks. *IEEE Transactions on Information Forensics and Security*, *19*, 6991–7004. IEEE Transactions on Information Forensics and Security. https://doi.org/10.1109/TIFS.2024.3428412

Kulatilleke, G. K. (2022). *Challenges and complexities in machine learning based credit card fraud detection* (No. arXiv:2208.10943). arXiv. http://arxiv.org/abs/2208.10943. Accessed 18 Mar 2024.

Langer, A., spsampsps Mukherjee, A. (2023). Organizing the data ecosystem. In A. Langer spsampsps A. Mukherjee (Eds.), *Developing a path to data dominance: Strategies for digital data-centric enterprises* (pp. 113–141). Springer International Publishing. https://doi.org/10.1007/978-3-031-26401-6_5

Langevin, A., Cody, T., Adams, S., & Beling, P. (2022). Generative adversarial networks for data augmentation and transfer in credit card fraud detection. *Journal of the Operational Research Society, 73*(1), 153–180. https://doi.org/10.1080/01605682.2021.1880296

Lebichot, B., Verhelst, T., Le Borgne, Y.-A., He-Guelton, L., Oble, F., & Bontempi, G. (2021). Transfer learning strategies for credit card fraud detection. *IEEE Access, 9*, 114754–114766. https://doi.org/10.1109/ACCESS.2021.3104472.

Lei, Y.-T., Ma, C.-Q., Ren, Y.-S., Chen, X.-Q., Narayan, S., & Huynh, A. N. Q. (2023). A distributed deep neural network model for credit card fraud detection. *Finance Research Letters, 58*, 104547. https://doi.org/10.1016/j.frl.2023.104547

Levi, M. (1998). Organising plastic fraud: Enterprise criminals and the side-stepping of fraud prevention. *The Howard Journal of Criminal Justice, 37*(4), 423–438. https://doi.org/10.1111/1468-2311.00110

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). *A critical review of recurrent neural networks for sequence learning* (No. arXiv:1506.00019). arXiv. https://doi.org/10.48550/arXiv.1506.00019

Liu, X., Iftikhar, N., Huo, H., Li, R., & Nielsen, P. S. (2019). Two approaches for synthesizing scalable residential energy consumption data. *Future Generation Computer Systems, 95*, 586–600. https://doi.org/10.1016/j.future.2019.01.045

Longadge, R., & Dongre, S. (2013). *Class imbalance problem in data mining review* (No. arXiv:1305.1707). arXiv. https://doi.org/10.48550/arXiv.1305.1707

Lopez-Rojas, E. A., & Axelsson, S. (2012). Money laundering detection using synthetic data. *Linköping Electronic Conference Proceedings 71*(5), 33–40. https://ep.liu.se/en/conference-article.aspx?Article_No=5&issue=71&series=ecp. Accessed 18 Dec 2024.

Lopez-Rojas, E. A., Elmir, A., & Axelsson, S. (2016). PaySim: A financial mobile money simulator for fraud detection. *European Modeling and Simulation Symposium 2016*. https://www.msc-les.org/proceedings/emss/2016/EMSS2016_249.pdf. Accessed 18 Dec 2024.

Loukides, G., Gkoulalas-Divanis, A., spsampsps Shao, J. (2010). Anonymizing transaction data to eliminate sensitive inferences. In P. G. Bringas, A. Hameurlain, spsampsps G. Quirchmayr (Eds.), *Database and Expert Systems Applications* (pp. 400–415). Springer. https://doi.org/10.1007/978-3-642-15364-8_34

Lu, Y., Wang, H., & Wei, W. (2023). *Machine learning for synthetic data generation: A review* (No. arXiv:2302.04062). arXiv. https://doi.org/10.48550/arXiv.2302.04062

Majava, J., Kinnunen, T., Foit, D., & Kess, P. (2016). An intermediary as a trust enabler in a spatial business ecosystem. *International Journal of Innovation and Learning, 20*(2), 199. https://doi.org/10.1504/IJIL.2016.077845

Major, T., & Mangano, J. (2020). *Modernising payments messaging: The ISO 20022 standard*. Reserve Bank of Australia. https://www.rba.gov.au/publications/bulletin/2020/sep/pdf/modernising-payments-messaging-the-iso-20022-standard.pdf. Accessed 18 Dec 2024.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems, 15*(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

Morley-Fletcher, E. (2022). New solutions to biomedical data sharing data sharing: Secure computation secure computationsand synthetic data synthetic data. In C. Beneduce spsampsps M. Bertolaso (Eds.), *Personalized Medicine in the Making: Philosophical Perspectives from Biology to Healthcare* (pp. 173–189). Springer International Publishing. https://doi.org/10.1007/978-3-030-74804-3_9

Mullarkey, M. T., & Hevner, A. R. (2019). An elaborated action design research process model. *European Journal of Information Systems, 28*(1), 6–20. https://doi.org/10.1080/0960085X.2018.1451811

Myalil, D., Rajan, M. A., Apte, M., & Lodha, S. (2021). Robust collaborative fraudulent transaction detection using federated learning. *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 373–378. https://doi.org/10.1109/ICMLA52953.2021.00064

Nickerson, K., Tricco, T., Kolokolova, A., Shoeleh, F., Robertson, C., Hawkin, J., spsampsps Hu, T. (2023). Banksformer: A deep generative model for synthetic transaction sequences. In M.-R. Amini, S. Canu, A. Fischer, T. Guns, P. Kralj Novak, spsampsps G. Tsoumakas (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 121–136). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-26422-1_8

O'Leary, K., O'Reilly, P., Nagle, T., Filelis-Papadopoulos, C., & Dehghani, M. (2021). The sustainable value of open banking: Insights from an open data lens. *Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2021.713.

Oliveira, M. I. S., & Lóscio, B. F. (2018). What is a data ecosystem? *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, 1–9. https://doi.org/10.1145/3209281.3209335.

Oliveira, M. I. S., & de Barros LimaFariasLóscio, G. F. B. (2019). Investigations into data ecosystems: A systematic mapping study. *Knowledge and Information Systems, 61*(2), 589–630. https://doi.org/10.1007/s10115-018-1323-6

Otto, B., Steinbuß, S., Teuscher, A., & Lohmann, S. (2019). *IDS reference architecture model 3.0* (p. 118). International Data Spaces

Association. https://internationaldataspaces.org/wp-content/uploads/IDS-Reference-Architecture-Model-3.0-2019.pdf. Accessed 17 Oct 2024.

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2023). Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys, 55*(6), 1–29. https://doi.org/10.1145/3533378

Pathare, A., Mangrulkar, R., Suvarna, K., Parekh, A., Thakur, G., & Gawade, A. (2023). Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. *International Journal of Information Management Data Insights, 3*(2), 100177. https://doi.org/10.1016/j.jjimei.2023.100177

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2016*, 399–410. https://doi.org/10.1109/DSAA.2016.49

Pazarbasioglu, C., Mora, A. G., Uttamchandani, M., Natarajan, H., Feyen, E., & Saal, M. (2020). *Digital financial services* (p. 54). World Bank Group. https://pubdocs.worldbank.org/en/230281588169110691/Digital-Financial-Services.pdf. Accessed 22 Jan 2023.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Perez, I., Wong, J., Skalski, P., Burrell, S., Mortier, R., McAuley, D., & Sutton, D. (2023). Locally differentially private embedding models in distributed fraud prevention systems. *2023 IEEE International Conference on Data Mining Workshops (ICDMW), 2023*, 475–484. https://doi.org/10.1109/ICDMW60847.2023.00068

Pranto, T. H., Hasib, K. T. A. Md., Rahman, T., Haque, A. B., Islam, A. K. M. N., & Rahman, R. M. (2022). Blockchain and machine learning for fraud detection: A privacy-preserving and adaptive incentive based approach. *IEEE Access*, *10*, 87115–87134. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3198956

Preziuso, M., Koefer, F., & Ehrenhard, M. (2023). Open banking and inclusive finance in the European Union: Perspectives from the Dutch stakeholder ecosystem. *Financial Innovation, 9*(1), 111. https://doi.org/10.1186/s40854-023-00522-1

Qiao, F., Li, Z., & Kong, Y. (2024). A privacy-aware and incremental defense method against GAN-based poisoning attack. *IEEE Transactions on Computational Social Systems*, *11*(2), 1708–1721. IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/TCSS.2023.3263241

Richhariya, P. (2012). A survey on financial fraud detection methodologies. *International Journal of Computer Applications*, *45*. https://www.ijcaonline.org/archives/volume45/number22/7080-9373/. Accessed 18 Dec 2024.

Ryman-Tubb, N. F., Krause, P., & Garn, W. (2018). How artificial intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. *Engineering Applications of Artificial Intelligence, 76*, 130–157. https://doi.org/10.1016/j.engappai.2018.07.008

Sattarov, T., Schreyer, M., & Borth, D. (2023). FinDiff: Diffusion models for financial tabular data generation. *4th ACM International Conference on AI in Finance*, 64–72. https://doi.org/10.1145/3604237.3626876

Schäfer, F., Rosen, J., Zimmermann, C., & Wortmann, F. (2023). Unleashing the potential of data ecosystems: Establishing digital trust through trust-enhancing technologies. *ECIS 2023 Research Papers*. https://aisel.aisnet.org/ecis2023_rp/325. Accessed 14 Aug 2024.

Scheider, S., Lauf, F., Möller, F., & Otto, B. (2023). A reference system architecture with data sovereignty for human-centric data ecosystems. *Business & Information Systems Engineering, 65*(5), 577–595. https://doi.org/10.1007/s12599-023-00816-9

Sein, M. K., Henfridsson, O., Purao, S., Rossi, M., & Lindgren, R. (2011). Action design research. *MIS Quarterly, 35*(1), 37–56. https://doi.org/10.2307/23043488

Sethia, A., Patel, R., & Raut, P. (2018). Data augmentation using generative models for credit card fraud detection. *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 1–6. https://doi.org/10.1109/CCAA.2018.8777628

Strelcenia, E., & Prakoonwit, S. (2023). Improving classification performance in credit card fraud detection by using new data augmentation. *AI, 4*(1), 1. https://doi.org/10.3390/ai4010008

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). *Revisiting unreasonable effectiveness of data in deep learning era*. 843–852. https://openaccess.thecvf.com/content_iccv_2017/html/Sun_Revisiting_Unreasonable_Effectiveness_ICCV_2017_paper.html. Accessed 29 Jan 2023.

Sun, C., van Soest, J., & Dumontier, M. (2023). Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *Journal of Biomedical Informatics*, *143*. Scopus. https://doi.org/10.1016/j.jbi.2023.104404

Susha, I., Grönlund, Å., & Van Tulder, R. (2019). Data driven social partnerships: Exploring an emergent trend in search of research challenges and questions. *Government Information Quarterly, 36*(1), 112–128. https://doi.org/10.1016/j.giq.2018.11.002

Tiwana, A., Konsynski, B., & Bush, A. A. (2010). Research commentary—Platform evolution: Coevolution of platform architecture, governance, and environmental dynamics. *Information Systems Research, 21*(4), 675–687. https://doi.org/10.1287/isre.1100.0323

van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., Heymann, D., & Burke, D. S. (2014). A systematic review of barriers to data sharing in public health. *BMC Public Health, 14*(1), 1144. https://doi.org/10.1186/1471-2458-14-1144

van Schalkwyk, F., Willmers, M., & McNaughton, M. (2016). Viscous open data: The roles of intermediaries in an open data ecosystem. *Information Technology for Development, 22*(sup1), 68–83. https://doi.org/10.1080/02681102.2015.1081868

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems, 25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

vom Brocke, J., Simons, A., Niehaves, B., Niehaves, B., & Reimer, K. (2009). *Reconstructing the giant: On the importance of rigour in documenting the literature search process*. https://www.semanticscholar.org/paper/European-Conference-on-Information-Systems-(-ECIS-)-Simons-Niehaves/2fc90c0163905ee89bbd72a2ba27acf3dd012526. Accessed 29 Feb 2024.

Walia, M., Tierney, B., & McKeever, S. (2020). Synthesising tabular data using wasserstein conditional GANs with gradient penalty. *Irish Conference on Artificial Intelligence and Cognitive Science*. https://ceur-ws.org/Vol-2771/AICS2020_paper_57.pdf. Accessed 18 Dec 2024.

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. *Information Systems Research, 3*(1), 36–59. https://doi.org/10.1287/isre.3.1.36

Wang, X., & Yang, Y. (2024). A data simulation method of financial fraud transactions based on TransGAN. *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, 242–246. https://doi.org/10.1145/3672758.3672798

Wang, Y., Adams, S., Beling, P., Greenspan, S., Rajagopalan, S., Velez-Rojas, M., Mankovski, S., Boker, S., & Brown, D.

(2018). Privacy preserving distributed deep learning and its application in credit card fraud detection. *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, 1070–1078. https://doi.org/10.1109/TrustCom/BigDataSE.2018.00150

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly, 26*(2), xiii–xxiii.

Weldon, J. C., Ward, T., & Brophy, E. (2021). Generation of synthetic electronic health records using a federated GAN. *ArXiv*. https://www.semanticscholar.org/reader/16f0acaec6e5d2c7421f95d81625f3c3719ff81a. Accessed 4 Sept 2023.

White, O., Madgavkar, A., Townsend, Z., Manyika, J., Olanrewaju, T., Sibanda, T., & Kaufman, S. (2021). *Financial data unbound: The value of open data for individuals and institutions* [Discussion paper]. McKinsey Global Institute. https://www.mckinsey.com/industries/financial-services/our-insights/financial-data-unbound-the-value-of-open-data-for-individuals-and-institutions#/. Accessed 24 Feb 2024.

Xing, X., Wu, H., Wang, L., Stenson, I., Yong, M., Del Ser, J., Walsh, S., & Yang, G. (2022). *Non-imaging medical data synthesis for trustworthy AI: A comprehensive survey* (No. arXiv:2209.09239). arXiv. https://arxiv.org/abs/2209.09239. Accessed 13 June 2023.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). *Modeling tabular data using conditional GAN* (No. arXiv:1907.00503). arXiv. https://doi.org/10.48550/arXiv.1907.00503

Xu, S., Marwah, M., Arlitt, M., spsampsps Ramakrishnan, N. (2021). STAN: Synthetic network traffic generation with generative neural models. In G. Wang, A. Ciptadi, spsampsps A. Ahmadzadeh (Eds.), *Deployable Machine Learning for Security Defense* (pp. 3–29). Springer International Publishing. https://doi.org/10.1007/978-3-030-87839-9_1

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2020). Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing, 416*, 244–255. https://doi.org/10.1016/j.neucom.2019.12.136

Yan, C., Yan, Y., Wan, Z., Zhang, Z., Omberg, L., Guinney, J., Mooney, S. D., & Malin, B. A. (2022). A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications*, *13*(1). Scopus. https://doi.org/10.1038/s41467-022-35295-1

Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series generative adversarial networks. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/c9efe5f26cd17ba6216bbe2a7d26d490-Abstract.html. Accessed 20 Dec 2023.

Zachariadis, M. (2020). *Data-sharing frameworks in financial services: Discussing open banking regulation for Canada* (SSRN Scholarly Paper No. 2983066). https://doi.org/10.2139/ssrn.2983066

Zhang, Z., Yan, C., Mesa, D. A., Sun, J., & Malin, B. A. (2019). Ensuring electronic medical record simulation through better training, modeling, and evaluation. *Journal of the American Medical Informatics Association : JAMIA, 27*(1), 99–108. https://doi.org/10.1093/jamia/ocz161

Zhang, Z. (2022). *Synthetic data simulation for privacy-preserving medical data sharing* [Dissertation, Vanderbilt University]. https://www.proquest.com/openview/a52c1b5ba98353ad63feac8aedc2360f/1?casa_token=EA2kv24XHBcAAAAA:_5l8NrIgKBX4sLCrWkkuay9QsZX0MsO3tYa4h5DMXmjQu48RmTiorNOGIiP6TLS9Zn1MOwcymfxo&cbl=18750&diss=y&pq-origsite=gscholar&parentSessionId=gry0yux5GXHzUSLr77QPzN7q6%2FkMyiMM8Un8M7EpXKE%3D. Accessed 6 Mar 2024.

Zhu, X., Ao, X., Qin, Z., Chang, Y., Liu, Y., He, Q., & Li, J. (2021). Intelligent financial fraud detection practices in post-pandemic era. *The Innovation, 2*(4), 100176. https://doi.org/10.1016/j.xinn.2021.100176

Zuiderwijk, A., Janssen, M., & Davis, C. (2014). Innovation with open data: Essential elements of open data ecosystems. *Information Polity, 19*(1,2), 17–33. https://doi.org/10.3233/IP-140329