big data and cognitive computing

MDPI

Article

Improving Synthetic Data Generation Through Federated Learning in Scarce and Heterogeneous Data Scenarios

Patricia A. Apellániz *, Juan Parras and Santiago Zazo

Information Processing and Telecommunications Center, ETS Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain; j.parras@upm.es (J.P.); santiago.zazo@upm.es (S.Z.)
* Correspondence: patricia.alonsod@upm.es

Abstract: Synthetic Data Generation (SDG) is a promising solution for healthcare, offering the potential to generate synthetic patient data closely resembling real-world data while preserving privacy. However, data scarcity and heterogeneity, particularly in under-resourced regions, challenge the effective implementation of SDG. This paper addresses these challenges using Federated Learning (FL) for SDG, focusing on sharing synthetic patients across nodes. By leveraging collective knowledge and diverse data distributions, we hypothesize that sharing synthetic data can significantly enhance the quality and representativeness of generated data, particularly for institutions with limited or biased datasets. This approach aligns with meta-learning concepts, like Domain Randomized Search. We compare two FL techniques, FedAvg and Synthetic Data Sharing (SDS), the latter being our proposed contribution. Both approaches are evaluated using variational autoencoders with Bayesian Gaussian mixture models across diverse medical datasets. Our results demonstrate that while both methods improve SDG, SDS consistently outperforms FedAvg, producing higher-quality, more representative synthetic data. Non-IID scenarios reveal that while FedAvg achieves improvements of 13–27% in reducing divergence compared to isolated training, SDS achieves reductions exceeding 50% in the worst-performing nodes. These findings underscore synthetic data sharing potential to reduce disparities between data-rich and data-poor institutions, fostering more equitable healthcare research and innovation.

Keywords: synthetic data generation; federated learning; medical data; data scarcity; data heterogeneity

check for updates

Academic Editor: Hui Tian

Received: 21 November 2024
Revised: 21 December 2024
Accepted: 14 January 2025
Published: 21 January 2025

Citation: Apellániz, P.A.; Parras, J.; Zazo, S. Improving Synthetic Data Generation Through Federated Learning in Scarce and Heterogeneous Data Scenarios. *Big Data Cogn. Comput.* **2025**, *9*, 18. https://doi.org/10.3390/bdcc9020018

1. Introduction

A significant challenge in healthcare research is the scarcity and often suboptimal quality of medical data, particularly in resource-limited regions. The need for infrastructure, funding, and research capacity in these areas hinders the collection of comprehensive patient datasets. Moreover, the prevalence of specific medical conditions may vary geographically, resulting not only in a lack of data for particular diseases but also in the presence of biases within the data. This disparity exacerbates the inequities between well-resourced healthcare institutions and those serving marginalized communities [1,2]. This uneven distribution of medical data intensifies the gap in healthcare research and medical innovation. Hospitals and research centers with access to extensive datasets can conduct thorough testing and develop effective treatments, while insufficient data constrain others from pursuing similar endeavors.

Addressing the scarcity of medical data requires innovative strategies. One approach involves inter-institutional data sharing, often hindered by stringent data privacy regulations. Data anonymization techniques are used to remove identifiers and standardize shared data to mitigate these restrictions [3]. However, these methods can introduce biases or distortions and may compromise data utility by removing or obscuring sensitive or unique information. Additionally, encrypted data sharing, while a widely used privacy-preserving solution, is not without its risks. Encrypted data are vulnerable to security threats, including man-in-the-middle attacks, key compromises, and the potential for re-identification when auxiliary information is available [4,5]. Such vulnerabilities highlight the inherent risks of transmitting real patient data, even when encrypted. These challenges underscore the importance of developing methods that avoid transmitting sensitive information entirely.

In recent years, Synthetic Data Generation (SDG) has emerged as a promising alternative, generating artificial patient data that mimics real data while preserving privacy. Synthetic data can augment existing datasets and facilitate research without compromising patient confidentiality [6,7]. Generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), offer powerful tools for this SDG. GANs [8,9], capable of generating diverse data, have limitations in handling mixed data types and specific feature values, particularly with imbalanced datasets. Conditional GANs, such as CTGAN [10], address these challenges using a conditional vector to specify desired labels. VAEs, on the other hand, provide a probabilistic framework for data generation, offering flexibility in handling complex data distributions. Among VAEs, a recent extension of TVAE [10] with a Bayesian Gaussian mixture model (BGM), VAE-BGM [11], demonstrates superior performance in generating high-quality synthetic tabular data. These models effectively capture the underlying structure and distribution of real-world data, enabling the creation of realistic, anonymized patient data for research and analysis. By leveraging generative artificial intelligence models, healthcare institutions can overcome the limitations of real-world data and accelerate medical research while protecting patient privacy. However, the effectiveness of these models and SDG hinges on the quality and quantity of the underlying real data. If an institution lacks sufficient data to generate reliable synthetic data, the process may require tailored approaches to minimize the impact of the lack of samples [12].

Federated Learning (FL) [13] has emerged as a promising framework for collaborative Machine Learning (ML), particularly in scenarios where data privacy is a significant concern. While traditionally employed to enhance model performance and generalization, FL's potential for improving SDG has garnered increasing attention. By enabling institutions to train models locally on their private datasets and aggregate the learned parameters, FL facilitates decentralized SDG. This approach not only protects data privacy but also leverages the diversity of data across institutions to augment model performance and generalization. FL can be particularly advantageous for institutions with limited data, as they can benefit from the collective intelligence of a more extensive network of institutions. Traditional FL techniques, such as FedSGD and FedAvg [14], are particularly effective when institutions have similar data distributions (i.e., independent and identically distributed (IID) data). However, data heterogeneity is prevalent in real-world medical contexts due to population and institutional-specific practices. These disparities can lead to biased models if not adequately addressed [15,16]. Consequently, developing techniques that can effectively handle non-IID data is imperative for achieving the full potential of FL in medical research.

Researchers have explored various data-level techniques to mitigate the challenges posed by data heterogeneity in FL [17,18]. These techniques include private data processing (e.g., data collection, filtering, cleaning, and augmentation) and leveraging external data through knowledge distillation or unsupervised representation learning. For instance, Federated Distillation (FD) methods, such as Federated Augmentation (FAug) [19], provide innovative solutions that enhance the quality of SDG. FD enables more flexible knowledge transfer between clients and the server, surpassing the limitations of only sharing model parameters. FAug, in particular, tackles data heterogeneity by generating synthetic data to augment local datasets, allowing them to resemble IID distributions. Another notable method, Astraea [20], collects local data distributions and performs data augmentation based on global distributions to alleviate imbalance. By rearranging the training of clients based on KL divergence, Astraea ensures that local models are trained on more representative data. However, it is essential to emphasize that all these methods, along with other emerging techniques such as [21,22], are primarily designed for supervised learning contexts. These approaches focus on improving model training and parameter optimization by leveraging labeled data, offering significant insights into addressing data heterogeneity in supervised tasks using FL. However, they do not directly tackle the challenges associated with SDG. Therefore, FL remains a promising framework for SDG in heterogeneous environments. However, tailored strategies are required to address the distinct challenges of training a deep generative model that takes advantage of the data in different data centers.

Building upon the existing literature, this paper proposes to address data heterogeneity in FL specifically for SDG. While this method could be compared to data augmentation, we move beyond its capabilities by focusing on generating entirely new synthetic patient data rather than transforming existing samples. Traditional data augmentation typically takes an existing dataset and applies modifications—such as rotation, scaling, or cropping in images—to create additional, varied instances within the same data collection. This transformation approach increases the dataset's diversity but does not introduce fundamentally new information, as it only reuses the original samples. In contrast, SDG, as applied in this study, involves creating entirely new, artificial patient records that replicate the underlying patterns of real data without directly mirroring specific records. Instead of relying solely on local data augmentation or knowledge distillation, we explore the potential of sharing locally generated synthetic patients among participating institutions. By leveraging the collective knowledge and diverse data distributions across the federation, we hypothesize that Synthetic Data Sharing (SDS) can enhance the quality and representativeness of generated data for all institutions, particularly those with limited or biased datasets. SDS offers several advantages: (1) it can help institutions with insufficient data benefit from the more diverse and representative synthetic data generated by others; (2) it can improve the ability of models to generalize to unseen real-world scenarios by exposing them to a broader range of synthetic patient data; and (3) it can reduce the computational burden by reusing models with minimal retraining once synthetic patients have been generated. This method parallels Domain Randomized Search (DRS), a meta-learning approach where models are trained across tasks to generalize across diverse domains [23]. Similar to DRS, where data from multiple tasks are aggregated to improve generalization, our method uses synthetic patient data from different nodes to address the issue of data heterogeneity in FL. By sharing synthetic patients, institutions with low-quality data can benefit from the more diverse data generated at other nodes, potentially improving the overall model performance. This approach could be seen as meta-learning within the FL framework, where aggregated synthetic data helps balance the disparities between nodes.

Our research contributes to the field of FL by proposing and evaluating an SDG model within a heterogeneous data environment.

- Specifically, we leverage the VAE-BGM model [11], known for its superior performance in generating high-quality synthetic tabular data. By integrating this model into an FL framework, we aim to surpass the performance of isolated SDG methods.

- To simulate realistic scenarios, we construct an FL environment comprising multiple nodes with varying data quantities and qualities. This setup enables us to assess the effectiveness of our approach under diverse conditions.
- We compare the traditional FedAvg technique with the proposed SDS method of sharing generated data across nodes under IID and non-IID data scenarios. To validate the comparative performance of these approaches, we employ statistical validation using Jensen–Shannon divergence (D_{JS}), as described in [24], and clinical utility validation, as recommended by recent research. These evaluations demonstrate the efficacy of SDS in addressing data heterogeneity and improving SDG within FL environments.

This paper is structured as follows. Section 2 outlines our methodology for the proposed SDG model within the FL framework. Section 3 presents the experimental setup, details the datasets used, and analyzes our results. Finally, Section 4 summarizes our findings, discusses the implications of our research, and proposes future research directions.

2. Materials and Methods

2.1. VAE-BGM Model

A novel approach to synthetic tabular data generation is introduced in [11], which integrates a BGM within the framework of a VAE. This approach addresses the limitations observed in existing models like CTGAN and TVAE [10]. While these earlier models demonstrate strong performance for certain data types, the VAE-BGM model offers superior results, particularly in capturing the complexity of real-world tabular data.

The model's core innovation uses a Gaussian mixture model (GMM) to model the VAE's latent space. More specifically, the model leverages a BGM, a type of GMM that offers greater flexibility. Unlike traditional GMMs, which require a pre-specified number of components, the BGM allows the model to automatically determine the appropriate number of components. This flexibility is essential for accurately capturing the complexity of real-world data, as it allows the model to adapt dynamically to the underlying data distribution. By integrating the BGM, the model avoids the restrictive assumption of a purely Gaussian latent space, which is common in models like TVAE. Instead, the BGM enables the model to handle more complex, non-Gaussian latent structures. This is achieved through a Dirichlet process that adjusts the number of Gaussian components in the mixture, allowing the model to adapt to the specific data characteristics without requiring manual specification. As a result, the VAE-BGM model provides a more nuanced and accurate latent representation, making it particularly effective for handling complex tabular datasets where simple Gaussian assumptions are insufficient. In addition to improving the latent space representation, the model excels in handling mixed data types, including continuous and discrete features. By permitting various differentiable distributions for individual features, the model ensures that the specific characteristics of different data types are preserved during the data generation process. This makes the VAE-BGM particularly suitable for applications in healthcare, where datasets often contain diverse information, ranging from binary indicators to continuous measurements.

Another key advantage of this approach is its ability to generate synthetic data that better reflects the marginal and joint distributions of the original data. Traditional VAEs are constrained by the Kullback–Leibler divergence (D_{KL}) term in the loss function, which enforces a Gaussian prior in the latent space and limits the model's ability to capture more complex data structures. Integrating a GMM into the already learned latent space overcomes this limitation, allowing for a more accurate sampling process that reflects the true diversity of the data. This enhancement leads to the generation of synthetic data that resembles the real data more closely and retains crucial feature correlations, improving its utility for downstream ML tasks.

The architecture of the proposed model follows the typical VAE design, consisting of an encoder and a decoder. The encoder learns a latent representation z of the input data x. This latent representation is assumed to be Gaussian. The encoder aims to learn a variational distribution $q_\phi(z|x)$ that is as close as possible to the true posterior distribution $p(z|x)$. This is achieved by maximizing the Evidence Lower Bound (ELBO), which is a lower bound on the marginal log-likelihood of the data represented as defined in [11]:

$$\log p_\theta(x_i) \geq -D_{KL}(q_\phi(z|x_i)||p(z)) + \mathbb{E}_{q_\phi(z|x_i)}[\log p_\theta(x_i|z)] = \mathcal{L}(x_i, \theta, \phi), \tag{1}$$

where D_{KL} represents the KL divergence. The derivation of the ELBO is critical for understanding the VAE framework. A detailed step-by-step derivation is provided in Appendix A, where it is demonstrated that $\mathcal{L}(x_i, \theta, \phi)$ in Equation (1) coincides with the ELBO expression in Equation (A5).

On the other hand, the decoder learns the conditional distribution $p_\theta(x|z)$ to generate realistic data points from the latent space. To improve the flexibility of the latent space representation, the BGM is applied to the learned latent space z, refining it into z_{GM}. The BGM models the latent space as originating from a mixture of K Gaussian distributions, each characterized by a mean vector μ_k, covariance matrix Σ_k, and mixing coefficient π_k. This allows for a more complex and multi-modal representation of the latent space, enabling the model to capture intricate data distributions. The probability density of a point in the latent space is defined as follows:

$$p(z_{GM}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z \mid \mu_k, \Sigma_k), \tag{2}$$

where $\mathcal{N}(z|\mu_k, \Sigma_k)$ is the probability density function of a multivariate Gaussian for each of the K components. The expectation–maximization algorithm estimates these parameters, enabling the model to capture richer latent structures than a single Gaussian. The BGM allows for a more flexible representation of the latent space, enabling the model to capture complex data distributions accurately. Figure 1 illustrates the schematic process of the VAE-BGM model.

Given its ability to handle complex data distributions and mixed data types and generate high-quality synthetic data, the VAE-BGM model presents a compelling synthetic tabular data generation approach. For these reasons, we have chosen to adopt this model as the generative model for our research.

2.2. FL Integration in SDG

Traditionally, ML models are trained in a centralized way, where all data are gathered in a single location. However, such centralization raises significant security concerns, especially in domains such as healthcare, where personal data are involved. FL offers a decentralized approach to ML, enabling the training of a shared model across multiple institutions (nodes) without the need to centralize their data. This paradigm is rooted in three core principles:

- Distributed Data: Training data are partitioned across various clients, preserving data locality and privacy.
- Privacy Preservation: FL mitigates privacy concerns by training models locally on each node and sharing only model updates rather than raw data.
- Model Aggregation: Model updates from all nodes are aggregated to create a global model that captures the knowledge from distributed data sources.

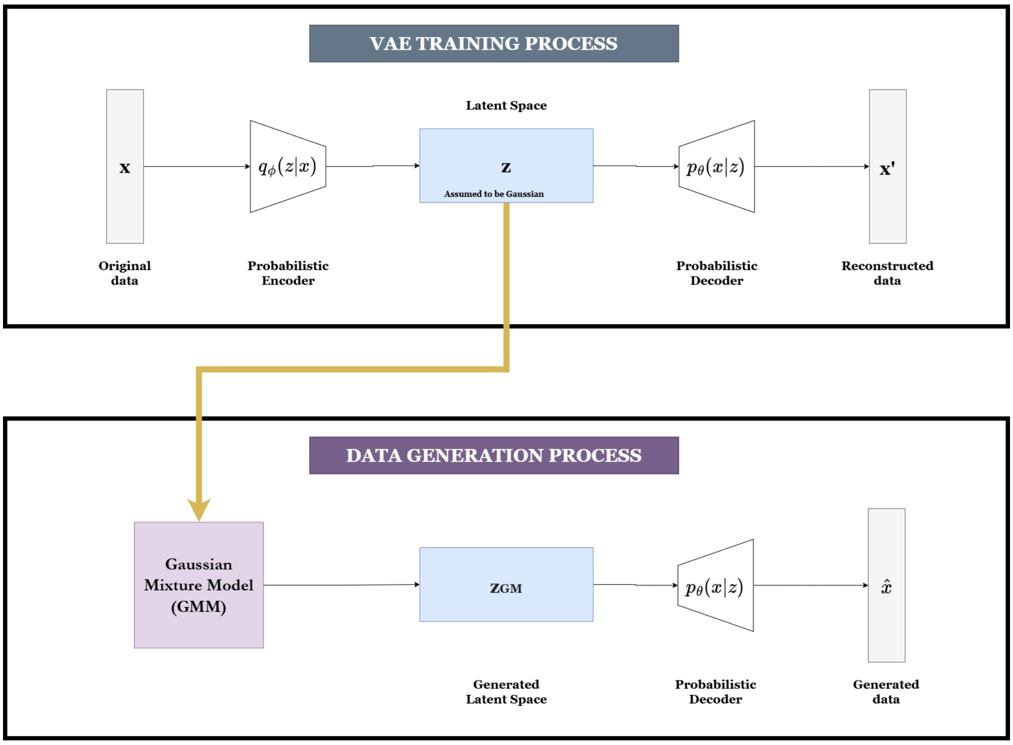


Figure 1. Schematic representation of the VAE-BGM model. The encoder extracts a latent representation of the input data. This representation is then modeled using a BGM, creating a new space z_{GM}. By sampling from this space, the model generates new distribution parameters, which are subsequently used by the already-trained decoder network to sample novel data points that closely resemble the original dataset.

This study simulates an FL environment comprising multiple nodes to generate synthetic data. In the context of SDG, using FL leverages decentralized data to create synthetic datasets that maintain the statistical properties of real-world data while protecting the privacy of the individuals involved. SDG has been widely explored in isolated settings, but challenges remain when considering data scarcity or heterogeneity across different institutions or geographical regions. This is where FL enters as a potential solution: by leveraging decentralized datasets in a privacy-preserving way, FL allows institutions to collaborate and generate synthetic data without transferring real patient records. In an FL context, SDG can be conducted across multiple distributed nodes with local, sensitive datasets. Rather than centralizing data, FL allows each node to train a generative model (VAE-BGM in this study) locally. The model parameters (not the data itself) are then shared with a central server, where they are aggregated to form a global generative model. This global model can then generate synthetic data that encapsulate the diverse statistical properties of data from all participating nodes.

2.3. Information Aggregation Techniques

In the proposed FL framework, we explore two techniques to train the VAE-BGM models across scarce and heterogeneous data environments: FedAvg and SDS. This section will explain both algorithms, detailing their mechanisms and how they address the challenges of non-IID data distribution across nodes. Comparing these two methods aims to clarify their respective advantages in improving SDG under the constraints of FL.

2.3.1. Federated Averaging

FedAvg, introduced by [14], is a foundational method in FL and serves as a baseline in this study to evaluate the performance of advanced approaches such as FedSDS.

The process begins with initializing a shared model architecture, such as the VAE-BGM, that is consistent across all nodes to ensure compatibility during aggregation. Each node trains this model locally on its dataset, producing updated parameters W_i. These local updates are then sent to a central server, where they are aggregated using a weighted averaging approach. The contribution of each node is proportional to its data size M_i, and the global model is updated as $W = \sum_{i=1}^{L} \frac{M_i}{N} W_i$, where $N = \sum_{i=1}^{L} M_i$ represents the total number of samples across all nodes. The updated global model is then distributed back to the nodes for further refinement in iterative rounds until a stopping criterion, such as convergence or a predefined number of iterations, is met. This iterative process allows FedAvg to combine distributed knowledge effectively while preserving data privacy. Figure 2 illustrates the overall FedAvg process, highlighting the interaction between local model training and global aggregation.

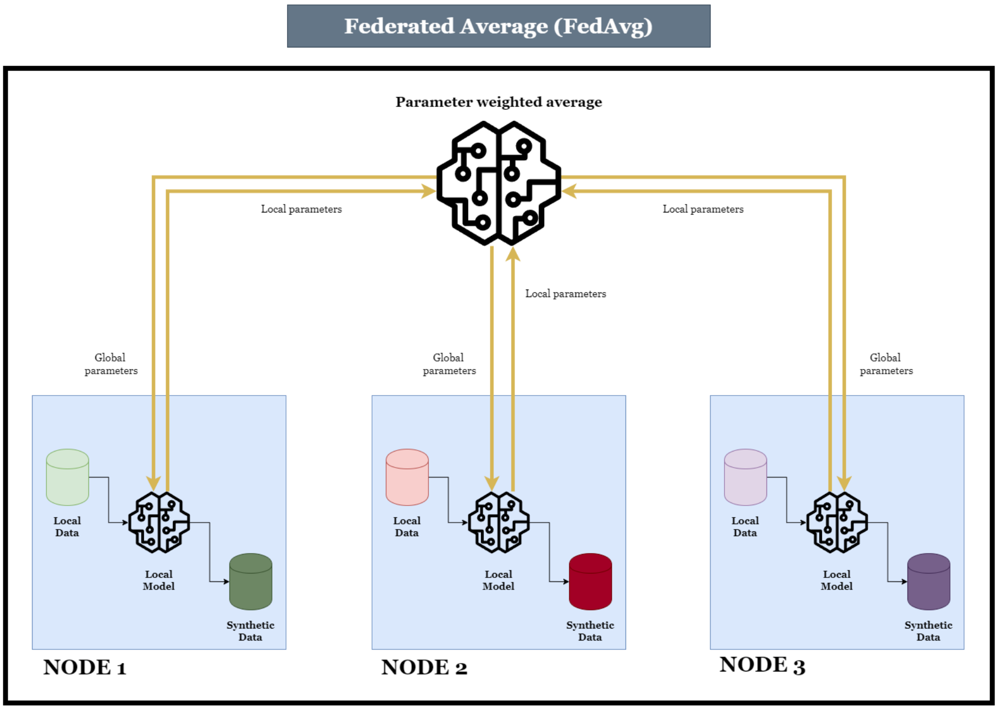


Figure 2. FedAvg process in FL. Each node trains a local model based on its data and shares the model parameters with a central server. The server averages the parameters over several rounds to create a global model, which is then distributed back to the nodes.

2.3.2. Synthetic Data Sharing

FedAvg has proven effective in many FL settings but can face challenges when applied to non-IID data [16,25]. Non-IID data are associated with scenarios where the data stored across different nodes are highly heterogeneous, leading to biased models or poor convergence. Differences in data distributions across nodes can result in variations in the local generative models. This can complicate the aggregation process, making it challenging to generate synthetic data that accurately represent the combined dataset.

Our proposal, SDS, is a technique that can address this issue: sharing synthetic patients generated locally at each node. We leverage the SDG model we intend to train at each node to generate data and enhance model performance when data are non-IID. This approach draws upon the meta-learning paradigm, specifically DRS, as introduced by [23], which approximates Model-Agnostic Meta-Learning (MAML). Meta-learning, often described as "learning to learn" [26], focuses on training algorithms that can generalize efficiently across tasks. By learning from related tasks, meta-learning models can rapidly adapt to new tasks with minimal data, addressing scenarios where large datasets are unavailable. MAML [27], one of the most significant meta-learning methods, focuses on finding a set of initial parameters, θ_{MAML}, that can be fine-tuned with minimal data for new tasks. MAML optimizes task-specific and meta-parameters through bi-level optimization, allowing rapid adaptation to new, unseen tasks. However, this bi-level optimization presents significant computational complexity, making it less suitable in environments constrained by resources or data. In contrast, DRS approximates MAML's generalization goal while reducing computational demands. Instead of performing a bi-level optimization, DRS aggregates data across tasks and trains the model directly on this aggregated dataset, resulting in a more resource-efficient solution. DRS achieves the same goal as MAML generalization across tasks by simplifying the learning process through direct training on aggregated data. Let a task instance $\mathcal{T}$ represent a tuple containing a dataset $\mathcal{D}$ and its corresponding loss function $\mathcal{L}$. Solving this task entails finding the optimal task-specific parameters ω^* that minimize the loss $\mathcal{L}$ for the particular dataset $\mathcal{D}$. Thus, the parameters θ_{DRS} for each SDG model are optimized according to the following equation:

$$\theta_{DRS} = \arg\min_{\omega} \mathcal{L}(\mathcal{D}_i^{(real)}; \omega) + \sum_{j \in -i} \mathcal{L}(\mathcal{D}_j^{(synth)}; \omega), \quad (3)$$

where $\mathcal{D}_i^{(real)}$ denotes the real, local dataset of node i, and $\mathcal{D}_j^{(synth)}$ signifies the synthetic datasets shared by the other nodes $j \in -i$, with $-i$ denoting all nodes except node i. This method optimizes the model by minimizing the loss across the aggregated data, including the real, local data from node i and the synthetic data from the other nodes. This approach mitigates the negative impact of non-IID data by leveraging synthetic data from multiple nodes to create a more representative and diverse training set. The ability of SDS to aggregate synthetic patient data from different nodes aligns with the meta-learning principles, allowing the model to generalize effectively across varying data distributions and enhancing model convergence in FL environments. A point worth emphasizing is the suitability of DRS over MAML in scenarios with a limited number of tasks (nodes). FL environments typically involve fewer data providers than more generalized ML setups. This small number of nodes can lead to situations where DRS, by aggregating synthetic data from these nodes, outperforms MAML. The reasoning behind this is grounded in the computational efficiency of DRS: unlike MAML, which requires bi-level optimization over multiple tasks, DRS aggregates data across tasks in a single round of optimization, making it less computationally intensive [23].

Thus, SDS provides the FL model with a richer and more diverse dataset, improving the quality and representativeness of the generated synthetic data, particularly in nodes where the data are scarce or biased. Unlike traditional parameter aggregation methods, SDS directly introduces additional information from other nodes, potentially improving convergence and mitigating the negative effects of data heterogeneity.

1. Local SDG: Each node initializes its local VAE-BGM model and trains it until synthetic data are generated based on the learned latent representation of the model. This aligns with the DRS strategy of generating data across domains (nodes) to capture domain-specific features.
2. Synthetic Data Sharing and Aggregation: Similar to DRS, synthetic data from each node are shared with other nodes. This aggregated data forms a more diverse and representative training dataset, mitigating the effects of data heterogeneity.

3. Model Training with Augmented Data: Each node trains its local VAE-BGM model using the augmented dataset, including real and synthetic data. This process, akin to DRS's task-based aggregation, leverages the diversity in the shared synthetic data to improve model performance. The training continues until the model converges.

This approach improves convergence and mitigates the negative impact of non-IID data by leveraging the diversity of synthetic data from multiple nodes. Sharing synthetic data instead of generative models further optimizes the system, making SDS a highly efficient option for complex FL scenarios. Figure 3 depicts the explained process.

FedAvg and SDS can operate without a central server in this revised approach, allowing direct communication between nodes. However, a notable distinction lies in the number of communication rounds necessary for model convergence. FedAvg typically requires multiple parameter updates and aggregation rounds to achieve optimal accuracy. This iterative process involves continuous communication between nodes, which can be a limitation in bandwidth-constrained environments. In contrast, SDS could theoretically be executed in a single communication round. Although this study does not apply a single-round strategy, the potential for such an approach exists. By sharing synthetic patient data only once, significant improvements in model performance could be obtained, particularly in scenarios where communication resources are limited. This single-round communication would drastically reduce overhead compared to FedAvg, which depends on numerous rounds of sharing and aggregating model parameters. Additionally, in SDS, sharing not just the synthetic data but also the generative model itself, including the decoder and the BGM-derived parameters, enhances communication efficiency compared to FedAvg. Thus, SDS offers a more scalable and bandwidth-efficient solution for real-world FL applications, especially when communication restrictions are a significant concern.

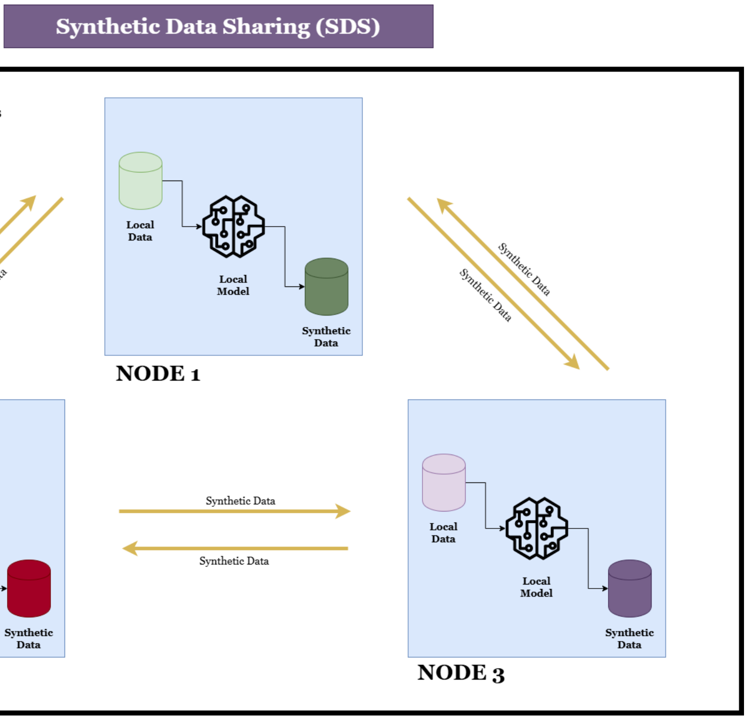


Figure 3. SDS process in FL. The SDS approach generates synthetic data locally at each node and then shares this data across nodes. Nodes incorporate the aggregated synthetic data from other nodes into their local training.

3. Results

3.1. Data

This section details the two medical datasets used in our experiments [28]. The motivation for selecting these classification datasets is their substantial number of samples and several types of varying features, making them well suited for testing our proposed approach. They offer the complexity needed to evaluate the robustness of SDG due to the intricate relationships between features and the heterogeneous nature of their data types.

- Diabetes_H (Diabetes Health Indicators) https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset (accessed on 20 January 2025): This dataset comprises survey responses collected by the Centers for Disease Control and Prevention from the Behavioral Risk Factor Surveillance System (BRFSS) 2015. BRFSS is an annual health-related telephone survey designed to gather information on health conditions and risk factors. The target variable includes three classes: 0 for no diabetes or diabetes during pregnancy, 1 for prediabetes, and 2 for diabetes.
- Heart (Heart Disease Indicators) https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease (accessed on 20 January 2025): Similar to Diabetes_H, this dataset originates from the cleaned BRFSS 2015 survey and focuses on a binary classification of heart disease presence.

Table 1 provides a more detailed description of the composition of these datasets. The data complexity stems from the many features with diverse data types, leading to intricate relationships and dependencies. This feature diversity presents a significant challenge for generating high-quality synthetic data that accurately reflect the underlying structure of the real data. Additionally, the class proportions reveal that these datasets are highly imbalanced, adding further complexity to the generation process. The imbalance introduces challenges in ensuring that the synthetic data adequately represent minority classes, which is crucial for accurate classification.

Table 1. Medical databases used in experiments. Datasets vary in number of samples, features, data types, and class proportions.

Dataset	Number of Samples	Number of Features	Data Types	Class Proportion
Diabetes_H	253,680	22	Integer, categorical, and binary	84.24%/1.83%/13.93%
Heart	253,680	22	Integer, categorical, and binary	90.58%/9.42%

3.2. Experimental Design

3.2.1. Data Distribution and Nodes Setup

We employ a federated environment consisting of three nodes to mimic real-world situations where data availability and quality vary significantly between institutions or locations.

- Node 1 represents an institution with limited data and resources, having only 100 training samples.
- Node 2 represents an institution with moderate resources, using 1000 training samples.
- Node 3 represents a well-resourced institution with access to a large dataset of 10,000 training samples.

However, all nodes share the same number of validation samples, each tested on 9500. This validation set ensures that performance is measured consistently across all nodes. The datasets selected allow us to create these divisions without losing the integrity of the data.

Two distinct scenarios are conducted to assess the performance under different data distributions: IID and non-IID.

- IID Scenario: In this case, data are split randomly across the three nodes, ensuring each node receives a statistically similar distribution of features. This ensures that the feature distributions are balanced and equal across all nodes.
- Non-IID Scenario: To represent more realistic, complex data conditions, this scenario reflects non-IID data distributions, where feature distributions vary across nodes. In an FL context, data might naturally exhibit variability across locations due to differences in population, local health factors, or socioeconomic conditions. For this scenario, we simulate distributional variation by selecting a key feature in each dataset with a substantial impact on the target variable—in this case, the *Body Mass Index* (BMI) column, due to its established correlation with diabetes [29] and heart disease [30], as well as its sensitivity to regional socioeconomic factors like healthcare access and diet quality [31]. BMI distribution was stratified to create differing distributions across nodes: one node received a balanced distribution of BMI values (50% above and 50% below the median), while the other two nodes were provided with skewed distributions, with 90% of samples having BMI values either above or below the median, respectively. Maintaining consistent features across nodes (i.e., the same columns) while introducing distributional differences allows us to model the non-IID scenario while ensuring compatibility with FedAvg realistically. By preserving identical feature columns across nodes, FedAvg remains applicable, as it requires the compatibility of model parameters based on shared input feature sets across nodes. In cases where nodes differ in feature sets, applying FedAvg would be infeasible, as model parameter aggregation depends entirely on the alignment of input features across nodes. This design choice lets us compare SDS and FedAvg, underscoring our approach's robustness and practical relevance in a realistic FL scenario.
  Figure 4 illustrates the Kernel Density Estimation (KDE) of the BMI distributions across the three nodes for both IID and non-IID scenarios. In the IID scenario, shown in Figure 4a,b, the distributions of BMI are similar across all nodes, as expected due to random data splitting. However, the distributions differ significantly across nodes in the non-IID scenario, depicted in Figure 4c,d. Specifically, Node 3 retains a distribution similar to the overall population, while Nodes 1 and 2 distributions are shifted, reflecting the intentional skew in their data. This variation highlights the challenges posed by non-IID settings in federated learning and underscores the importance of techniques like SDS to address such disparities.

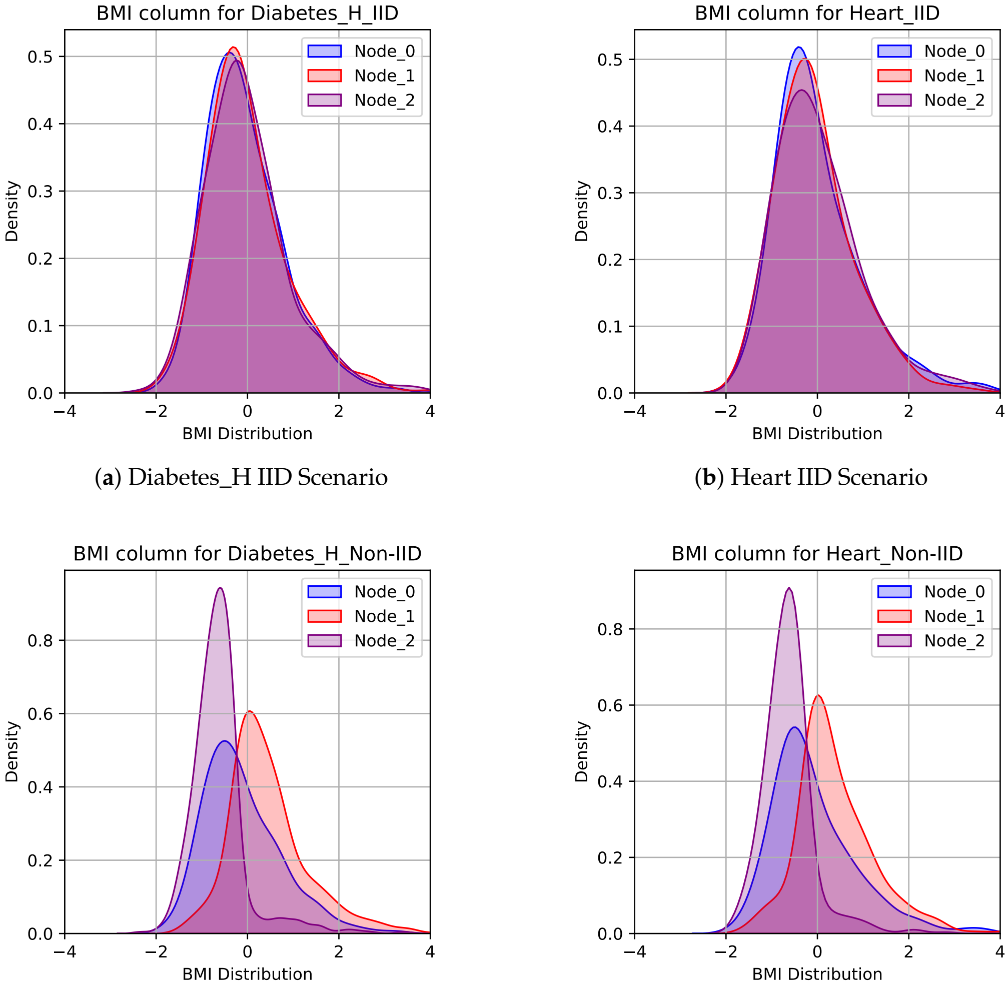
3.2.2. Data Generation in an FL Environment

Each node trains its local VAE-BGM model to generate data based on the locally available training samples. The critical aspect of this experiment is comparing two different FL techniques, FedAvg and SDS, against isolated, non-federated training. Comparing these two FL techniques will provide insights into how FL can enhance SDG in data heterogeneity and imbalance environments.

- FedAvg: This method aggregates the model weights from each node and updates the local models using a weighted average of these aggregated weights. Each node trains its local model for 200 epochs, after which the weights are shared, aggregated, and redistributed for the next training round. This process is repeated for five rounds.
- SDS: Instead of sharing model parameters, each node shares synthetic data generated locally. The synthetic data are shared across the network. These additional synthetic data improve the quality and quantity of the local datasets at each node, boosting training efficacy. The process steps are the following:

1. In the first round, each node trains its local VAE-BGM model with its original dataset, generating synthetic samples.
2. From the second round onward, the synthetic data generated by each node are shared with the other nodes, augmenting their local datasets. However, the total number of samples used for training at each node is limited to 10,000. If the sum of local and synthetic samples exceeds 10,000, the node will use only as many synthetic samples as needed to maintain this maximum training size.

Both methods aim to improve the performance of the nodes with less data or poorer data quality. In the FedAvg case, we expect that sharing model weights will help align the local models across nodes. In contrast, in SDS, the increase in data volume and diversity from shared synthetic patients is expected to mitigate the issues related to data scarcity and bias.



(**a**) Diabetes_H IID Scenario

(**b**) Heart IID Scenario

(**c**) Diabetes_H Non-IID Scenario

(**d**) Heart Non-IID Scenario

Figure 4. KDE plots for BMI distributions across nodes under IID and non-IID scenarios for the Diabetes_H and Heart datasets.

3.3. Network Architecture

The proposed network leverages a VAE and a BGM model for SDG at each node [11]. The VAE learns a latent data representation with an encoder featuring a hidden ReLU layer of 256 neurons and a hyperbolic tangent output layer. The latent space dimensionality is fixed at 20, capturing key data features. The decoder mirrors the encoder structure with tailored activation functions for covariate distributions. Dropout at 20% helps prevent overfitting. The VAE is trained for 200 epochs in each federated round with a batch size of 1024. The BGM further models the latent space as a mixture of Gaussian distributions, using a Dirichlet process prior with a maximum of 20 components. Each Gaussian component has its covariance matrix, enhancing the model's ability to represent complex data relationships.

Each node trains its local VAE-BGM model in the FL setup over five federated rounds. After each round, depending on the technique, either model weights (FedAvg) or synthetic data (SDS) are shared and aggregated across nodes. FedAvg aggregates model weights, while SDS shares synthetic data to enrich local datasets. The performance is averaged over three runs with different random seeds to account for the sensitivity of VAEs to initialization.

3.4. Evaluation Metrics

The primary goal of this study is to generate synthetic data that is indistinguishable from real data, both in terms of statistical properties and practical utility in clinical settings. We adopt a dual validation approach focused on statistical similarity and clinical utility to comprehensively evaluate the generated data, following state-of-the-art guidelines [32].

For statistical validation, although numerous statistical tests and metrics can be used to compare real and synthetic data distributions, we chose to rely on D_{JS} estimation, a well-established method for measuring the similarity between two probability distributions. The D_{JS} is a symmetrized and bounded version of the D_{KL} and is particularly effective for large datasets. With $D_{KL}(p(x)\|q(x)) = \int p(x)\log(p(x)/q(x))dx$, the D_{JS} between two distributions $p(x)$ and $q(x)$ is defined as

$$D_{JS}(p(x)\|q(x)) = \frac{1}{2}D_{KL}(p(x)\|m(x)) + \frac{1}{2}D_{KL}(q(x)\|m(x)), \tag{4}$$

where $m(x) = \frac{1}{2}(p(x) + q(x))$ is the average distribution. This divergence captures the dissimilarity between each distribution and the average distribution $m(x)$. The D_{JS} is bounded between 0 and 1, with lower values indicating greater similarity between the distributions. To estimate D_{JS}, we followed the methodology outlined by [24], which employs a discriminator network to differentiate between real and synthetic data. This probabilistic classifier is trained with two datasets: M samples from the real data, labeled as class 1, and M synthetic samples, labeled as class 0. The classifier learns a decision boundary to distinguish between real and synthetic data, and the probabilities predicted by the network are then used to compute the D_{JS}. Specifically, the network outputs probabilities $\mathcal{P}(y=1|x)$, representing the likelihood that a given sample x comes from real data. Using this information, the D_{JS} is approximated as follows:

$$\begin{aligned} D_{JS}(p(x)||q(x)) \approx \frac{1}{2L}\sum_{i=1}^{L}\log\left(\frac{2\mathcal{P}(y=1|x_i)}{\mathcal{P}(y=1|x_i)+\mathcal{P}(y=0|x_i)}\right) \\ +\frac{1}{2L}\sum_{i=1}^{L}\log\left(\frac{2\mathcal{P}(y=1|\tilde{x}_i)}{\mathcal{P}(y=1|\tilde{x}_i)+\mathcal{P}(y=0|\tilde{x}_i)}\right), \end{aligned} \tag{5}$$

where $x_i \sim p(x)$, $\tilde{x}_i \sim q(x)$ and L are the number of samples used for evaluation. This study employs $M = 7500$ samples for training and $L = 1000$ samples for evaluation, consistent with prior research [24].

To assess the effectiveness of different scenarios—specifically, isolated learning, FedAvg, and SDS—on the D_{JS} values, we introduce the Mean Reciprocal Rank (MRR). The MRR gauges the relative effectiveness of each technique by considering the rank position of the first relevant D_{JS} value within an ordered list of D_{JS} values derived from each FL technique and the isolated learning scenario. The Reciprocal Rank (RR) for each method is calculated as the inverse of the position of the first relevant D_{JS} result. For example, if the first relevant result appears in the top position, its RR is 1; if it appears in the second position, the RR is 0.5, and so forth. The MRR is then computed as the average of the RRs across all situations:

$$MRR = \frac{1}{Q} \sum_{i=1}^{Q} \frac{1}{rank_i}, \tag{6}$$

where Q denotes the total number of situations under evaluation (isolated, FedAvg, and SDS), and $rank_i$ represents the position of the first relevant D_{JS} for the i-th scenario. Higher MRR values imply that relevant D_{JS} values appear earlier in the list, indicating better performance and suggesting an advantage of one FL technique over the isolated approach. This metric thus provides insight into which FL techniques potentially enhance model performance compared to isolated scenarios.

In addition to statistical validation, evaluating whether the synthetic data can be effectively applied to real-world clinical tasks is crucial. Clinical utility validation ensures that the synthetic data are statistically similar and useful for practical applications, such as training ML models for medical decision-making. For clinical utility validation, we use a Random Forest (RF) classifier to assess the accuracy of predicting the target feature in each dataset. For this RF, we employed the default hyperparameters provided by the scikit-learn library [33]. This choice is justified by the balanced class distributions and large sample sizes in the datasets, which ensure that the default settings are sufficient to provide robust and reliable performance. Random Forests are well known for their robustness to hyperparameter choices, particularly in scenarios with ample data and balanced classes. Furthermore, this study focuses not on maximizing the classifier's performance but on evaluating the comparability of synthetic and real data in downstream tasks. Using default hyperparameters avoids introducing variability associated with tuning and provides a consistent baseline for comparing the utility of synthetic data.

Two experimental scenarios are considered:

- Training on real data and validating on real data: This is the upper-bound performance we aim to match with the synthetic data.
- Training on synthetic data and validating on real data: This tests the ability of a classifier trained on synthetic data to generalize to real-world data, indicating the practical utility of the synthetic samples.

The goal is for the classification accuracy obtained in the second scenario to closely match that of the first scenario. If the accuracy gap is minimal, synthetic data can be effectively used in clinical applications.

This dual validation approach comprehensively assesses the quality and applicability of the generated synthetic data by combining statistical similarity and clinical utility validation. Statistical validation ensures that the synthetic data closely mimics the real data distribution. In contrast, clinical validation confirms that the synthetic data retain the necessary information to perform well in real-world tasks. These metrics offer a holistic evaluation of the synthetic data's fidelity and utility.

We conducted hypothesis testing to validate our proposed approaches' effectiveness further. For statistical validation, the null hypothesis assumed that isolated training would yield lower D_{JS} values (indicating better performance) than FL techniques. A significance level of 0.01 was employed. Rejection of the null hypothesis, based on p-values below this threshold, indicated that FL techniques significantly outperformed isolated training. Regarding clinical utility validation, the accuracy achieved using real data for training and validation was considered the upper bound. The null hypothesis assumed this upper bound would be higher than the accuracy obtained using synthetic data for training and real data for validation. A significance level of 0.01 was again applied. Rejection of the null hypothesis implied that the performance of classification models trained on synthetic data was comparable to or even exceeded that of the models trained on real data.

3.5. Experiments

For each dataset, we provide a detailed comparison of the D_{JS} and accuracy scores, which reflect the statistical and clinical utility validations. These results are displayed in tables, allowing for a clear performance comparison across the different FL techniques and the isolated case. The analysis was conducted for IID and non-IID data scenarios. Additional results are included in Appendix C, where we compare key features of real and synthetic data distributions, and in Appendix **??**, which presents a study on the influence of varying sample sizes across nodes. The code to replicate the results can be found in https://github.com/Patricia-A-Apellaniz/fed_vae (accessed on 20 January 2025).

3.5.1. IID Scenarios

The results for the Diabetes_H dataset in Table 2 demonstrate the effectiveness of SDS in generating high-quality synthetic data, particularly in nodes with limited data (Nodes 1 and 2). SDS consistently outperforms isolated training and FedAvg regarding statistical similarity, as evidenced by the lower D_{JS} values. Although FedAvg enhances performance compared to independent training on Node 2, it remains less effective than SDS. The D_{JS} results for Node 3, which has the most significant data, are relatively similar across all techniques. This is expected, as the abundance of data may limit the potential gains from FL. Regarding clinical utility, all nodes achieve comparable and high accuracy in the real–real scenario, suggesting that the classification task is relatively straightforward, even with limited data. In the synthetic–real scenario, the isolated case and the FL techniques maintain comparable accuracy to the real–real scenario in Nodes 1 and 3, indicating their ability to generate clinically useful synthetic data. However, in Node 2, both FL techniques exhibit a slight decrease in accuracy, suggesting that FL may not provide significant benefits in this particular case.

Table 2. Diabetes_H results in IID scenario: Comparison of D_{JS} and accuracy for isolated training and two FL techniques. Lower D_{JS} indicates better similarity between real and synthetic data, while synthetic–real accuracy closer to real–real reflects better clinical utility. Results are expressed as *mean (std)*. * indicates p-value < 0.01. In particular, for D_{JS}, * denotes statistically significant improvement over isolated case. For accuracy, * signifies performance of models trained on synthetic data was comparable to or exceeded that of models trained on real data. **Bold** values indicate best significative performance, and ▼ denotes decrease relative to upper bounds.

Node	Technique	Similarity Validation	Clinical Utility Validation	
		Estimated D_{JS}	Accuracy (Real–Real)	Accuracy (Synthetic–Real)
Node 1	Isolated	0.718 (0.040)	0.840 (0.001)	0.833 (0.004) *
	FedAvg	0.637 (0.026)		0.835 (0.002) *
	SDS	**0.411 (0.002)** *		0.842 (0.001) *
Node 2	Isolated	0.583 (0.020)	0.846 (0.001)	0.843 (0.002) *
	FedAvg	0.444 (0.014) *		0.840 (0.001) ▼
	SDS	**0.412 (0.004)** *		0.841 (0.002) ▼
Node 3	Isolated	0.080 (0.062)	0.842 (0.001)	0.841 (0.003) *
	FedAvg	0.042 (0.008)		0.844 (0.001) *
	SDS	0.094 (0.040)		0.842 (0.001) *

For the Heart dataset in Table 3, SDS consistently outperforms isolated training and FedAvg in terms of statistical similarity, as indicated by the lowest D_{JS} values across all nodes. While FedAvg demonstrates improved performance over isolated training in Node 1, SDS significantly reduces D_{JS} values. Regarding clinical utility, SDS achieves the highest accuracy in the synthetic–real scenario for Node 2, surpassing the performance of models trained on real data. This highlights the potential of SDS to generate synthetic data that can improve model performance. In Node 3, both FedAvg and SDS achieve comparable accuracy levels in the real–real scenario, further validating their ability to produce valid synthetic data. In contrast, for Node 1, which has the most limited data, isolated training and FedAvg show a slight accuracy decline relative to the real–real scenario, underscoring SDS's superiority in generating high-quality synthetic data.

Table 3. Heart results in IID scenario: Comparison of D_{JS} and accuracy for isolated training and two FL techniques. Lower D_{JS} indicates better similarity between real and synthetic data, while synthetic–real accuracy closer to real–real reflects better clinical utility. Results are expressed as *mean (std)*. * indicates p-value < 0.01. In particular, for D_{JS}, * denotes statistically significant improvement over isolated case. For accuracy, * signifies that performance of models trained on synthetic data was comparable to or exceeded that of models trained on real data. **Bold** values indicate best significative performance, and ▼ denotes decrease relative to upper bounds.

Node	Technique	Similarity Validation	Clinical Utility Validation	
		Estimated D_{JS}	Accuracy (Real–Real)	Accuracy (Synthetic–Real)
Node 1	Isolated	0.772 (0.023)	0.907 (0.001)	0.896 (0.000) ▼
	FedAvg	0.647 (0.030) *		0.904 (0.001) ▼
	SDS	**0.431 (0.007)** *		0.905 (0.000) *
Node 2	Isolated	0.576 (0.006)	0.908 (0.000)	0.909 (0.001) *
	FedAvg	0.530 (0.022)		0.910 (0.001) *
	SDS	**0.486 (0.013)** *		**0.912 (0.001)** *
Node 3	Isolated	0.371 (0.036)	0.914 (0.001)	0.915 (0.000) *
	FedAvg	0.330 (0.047)		0.914 (0.000) *
	SDS	**0.127 (0.029)** *		0.914 (0.000) *

3.5.2. Non-IID Scenarios

The results for the non-IID scenario on the Diabetes_H dataset in Table 4 highlight the effectiveness of SDS in mitigating the impact of data heterogeneity. SDS significantly outperforms isolated training and FedAvg regarding statistical similarity in Nodes 1, 2, and 3. Regarding clinical utility, while all nodes achieve high accuracy in the real–real scenario, Node 1 exhibits a significant drop in performance when using synthetic data, particularly for isolated training. SDS, however, demonstrates a notable improvement over isolated training and FedAvg in this case, even though it does not reach the upper bound value. Nodes 2 and 3, which have more balanced data distributions, maintain comparable performance between real and synthetic data, indicating the effectiveness of FL techniques in these scenarios.

The results presented in Table 5 for the Heart dataset under non-IID conditions offer insightful observations about the performance of the SDS method compared to isolated training and FedAvg. Regarding statistical similarity, SDS consistently outperforms isolated training and FedAvg across Nodes 1 and 2. In particular, Node 1, characterized by a limited number of samples and lower data quality, benefits significantly from SDS. This suggests that SDS effectively leverages information from other nodes to generate synthetic data resembling the real data distribution, even if it substantially reduces heterogeneity. However, FedAvg fails to substantially improve isolated training in this node, further corroborating our previous findings that FedAvg may not be the optimal choice for non-IID scenarios. A similar trend is observed in Node 2, where SDS reduces D_{JS}. Conversely, Node 3, which benefits from a larger and more balanced dataset, does not exhibit significant improvements when using SDS, as expected. Interestingly, FedAvg leads to a higher D_{JS} value in Node 3 compared to the isolated case, suggesting that the exchange of model weights between nodes with disparate data distributions can negatively impact the quality of synthetic data generated in well-performing nodes. However, when examining the clinical utility, measured by the accuracy of models trained on synthetic data and tested on real data, we observe a slight decrease in performance for Nodes 1 and 2. In Node 1, accuracy drops slightly from 0.919 (real-real) to 0.915 (SDS), indicating a minor underperformance when using synthetic data generated by SDS. A similar slight decrease is observed in Node 2, with accuracy moving from 0.899 (real–real) to 0.894 (SDS). The slight underperformance of SDS in clinical utility metrics for Nodes 1 and 2 can be attributed to several factors. In non-IID settings, data distributions vary significantly across nodes. While FL techniques improve statistical similarity, they may not fully capture the unique, node-specific relationships and patterns critical for classification tasks. This is especially true for nodes with limited data, where rare or local patterns are essential for accurate predictions. In particular, SDS aims to create a more generalized synthetic dataset by incorporating information from multiple nodes. This can lead to the smoothing of important nuances or the dilution of specific features that are pivotal for the model's performance in individual nodes. Furthermore, sharing synthetic data across nodes may introduce biases if the synthetic data do not adequately represent the underlying distributions of the target nodes. This can affect the model's ability to generalize and perform well on real data from those nodes. Nevertheless, the overall accuracy remains high, indicating that the synthetic data generated by these techniques are still suitable for classification tasks.

Table 4. Diabetes_H results in non-IID scenario: Comparison of D_{JS} and accuracy for isolated training and two FL techniques. Lower D_{JS} indicates better similarity between real and synthetic data, while synthetic–real accuracy closer to real–real reflects better clinical utility. Results are expressed as *mean (std)*. * indicates p-value < 0.01. In particular, for D_{JS}, * denotes statistically significant improvement over isolated case. For accuracy, * signifies that performance of models trained on synthetic data was comparable to or exceeded that of models trained on real data. **Bold** values indicate best significative performance, and ▼ denotes decrease relative to upper bounds.

Node	Technique	Similarity Validation	Clinical Utility Validation	
		Estimated D_{JS}	Accuracy (Real–Real)	Accuracy (Synthetic–Real)
Node 1	Isolated	0.823 (0.011)	0.846 (0.000)	0.815 (0.001) ▼
	FedAvg	0.715 (0.005) *		0.826 (0.002) ▼
	SDS	**0.381 (0.002)** *		0.837 (0.001) ▼
Node 2	Isolated	0.542 (0.026)	0.808 (0.001)	0.810 (0.003) *
	FedAvg	0.394 (0.017)*		0.806 (0.002) *
	SDS	**0.376 (0.011)** *		0.803 (0.005) *
Node 3	Isolated	0.489 (0.006)	0.911 (0.001)	0.910 (0.000) *
	FedAvg	0.479 (0.020)		0.911 (0.001) *
	SDS	**0.349 (0.036)** *		0.910 (0.000) *

Table 5. Heart results in non-IID scenario: Comparison of D_{JS} and accuracy for isolated training and two FL techniques. Lower D_{JS} indicates better similarity between real and synthetic data, while synthetic–real accuracy closer to real–real reflects better clinical utility. Results are expressed as *mean (std)*. * indicates p-value < 0.01. In particular, for D_{JS}, * denotes statistically significant improvement over isolated case. For accuracy, * signifies that performance of models trained on synthetic data was comparable to or exceeded that of models trained on real data. **Bold** values indicate best significative performance, and ▼ denotes decrease relative to upper bounds.

Node	Technique	Similarity Validation	Clinical Utility Validation	
		Estimated D_{JS}	Accuracy (Real–Real)	Accuracy (Synthetic–Real)
Node 1	Isolated	0.803 (0.002)	0.919 (0.000)	0.918 (0.000) *
	FedAvg	0.586 (0.007) *		0.914 (0.001) ▼
	SDS	**0.361 (0.006)** *		0.915 (0.000) ▼
Node 2	Isolated	0.479 (0.054)	0.899 (0.000)	0.893 (0.001) ▼
	FedAvg	0.441 (0.023)		0.895 (0.000) ▼
	SDS	0.359 (0.020)		0.894 (0.001) ▼
Node 3	Isolated	0.336 (0.021)	0.917 (0.001)	0.915 (0.000) *
	FedAvg	0.476 (0.024) ▼		0.914 (0.001) *
	SDS	0.280 (0.084)		0.916 (0.000) *

3.5.3. Discussion

The results presented in Table 6 further highlight the effectiveness of the SDS in both IID and non-IID scenarios. In both settings, SDS consistently outperforms both isolated training and FedAvg regarding MRR, indicating its superior ability to generate synthetic data that enhance model performance. In the IID setting, where data are evenly distributed across nodes, SDS effectively leverages the collective knowledge of the federation to generate high-quality synthetic data. This leads to significant improvements in model performance, particularly in scenarios with limited data, as observed in Nodes 1 and 2 of the Diabetes_H dataset.

Table 6. MRR values across different scenarios (IID and Non-IID) and datasets for isolated FedAvg and SDS approaches. Higher MRR values indicate superior performance. **Bold** values denote best performances.

Scenario	Dataset	Isolated	FedAvg	SDS
IID	Diabetes_H	0.389	0.667	**0.778**
	Heart	0.333	0.667	**0.833**
Non-IID	Diabetes_H	0.333	0.500	**1.000**
	Heart	0.389	0.444	**1.000**

The robustness of SDS becomes even more apparent in non-IID settings, where data are unevenly distributed across nodes. In such scenarios, SDS demonstrates remarkable robustness, mitigating the negative impact of data heterogeneity. By sharing synthetic data generated from diverse data distributions, SDS helps improve models' generalization ability and enhance their performance in unseen data. In contrast, FedAvg, while providing moderate improvements over isolated training, struggles to fully address the imbalance in data distribution, particularly in nodes with scarce or biased datasets. This is particularly evident in the non-IID scenarios, where FedAvg's performance can be negatively impacted by the imbalance in data distribution. Overall, these results underscore the potential of SDS as a powerful tool for generating high-quality synthetic data in FL environments, particularly in data scarcity and heterogeneity. By effectively leveraging the collective knowledge of the federation, SDS can improve model performance and enhance the generalization ability of models trained on synthetic data.

Beyond improving model performance, we emphasize the security and privacy-preserving benefits of synthetic data generation within the FL framework. As outlined in Appendix B, we conducted a rigorous evaluation of privacy preservation by analyzing minimum distances between real and synthetic samples. In this analysis, we compared the distributions of pairwise distances between real–real samples and synthetic–real samples across Node 3 under both IID and non-IID scenarios. The results, visualized through histograms and KDE plots, demonstrate that while the distance distributions are statistically similar, they are not identical, and minimum distances are consistently non-zero. This finding is significant because our VAE-BGM architecture generates synthetic data by sampling from the latent space rather than directly replicating real samples. Consequently, synthetic data avoid exact duplication of real data points, effectively mitigating privacy risks. This outcome underscores a key advantage of SDG: unlike raw or encrypted data, synthetic data inherently reduce the risk of privacy breaches during sharing, such as those associated with man-in-the-middle attacks or key compromises. These properties make SDG a robust solution for data privacy preservation, particularly when compared to sharing raw datasets, encrypted data, or even trained model parameters, which may still leak sensitive information through model inversion or membership inference attacks.

In summary, SDS not only enhances performance in FL environments under both IID and non-IID conditions but also offers significant security advantages by reducing the risks associated with sharing real or encrypted data. By leveraging synthetic data, institutions can collaborate effectively while ensuring data privacy, thereby fostering trust and enabling more widespread adoption of FL frameworks in healthcare and other sensitive domains.

4. Conclusions

This research underscores the effectiveness of FL for SDG in healthcare, particularly in addressing the challenges posed by heterogeneous and scarce data distributions. By employing VAE-BGM models across diverse medical datasets, this study demonstrates that SDS consistently outperforms traditional approaches like FedAvg and isolated training in both IID and non-IID scenarios. A key strength of SDS lies in its ability to expose nodes to diverse synthetic samples, effectively approximating a more IID-like environment even in non-IID settings. This results in significant advantages for generating high-quality synthetic data, as reflected in lower D_{JS} values, and supports robust model performance across nodes. Clinical utility validation confirms the practicality of synthetic data generated using SDS, achieving comparable accuracy to real data in downstream tasks. In non-IID environments, SDS proves particularly robust, addressing the challenges of unevenly distributed data among institutions by leveraging the diversity of synthetic samples to enhance representativeness and mitigate the negative effects of heterogeneity. In contrast, FedAvg demonstrates limited improvements in these scenarios, often failing to match SDS's effectiveness, particularly in nodes with constrained data availability or skewed distributions.

These findings highlight the potential of sharing synthetic data within FL frameworks. By fostering data diversity and reducing the disparities between data-rich and data-poor nodes, SDS enables improved model generalization and supports collaborative research without exposing sensitive patient information. This approach not only bridges gaps in data accessibility and quality but also sets a foundation for advancing medical research and innovation in under-resourced regions. Future work should continue exploring the role of synthetic data in FL, focusing on increasingly heterogeneous and imbalanced data distributions to further validate and refine the methodology.

Future research should explore optimizing FL architectures for even more complex data types and more extensive networks of institutions, further refining the integration of SDG with FL to maximize efficiency and scalability. In particular, exploring SDG in low-sample settings as in [12] could be highly beneficial. This approach, which integrates meta-learning (like DRS) and transfer learning techniques to SDG, could be adapted to augment FL environments, where leveraging knowledge from previously trained models or similar tasks could significantly enhance the quality of synthetic data in low-sample nodes. Furthermore, while the VAE-BGM model has already been compared with state-of-the-art tabular generative models (CTGAN and TVAE), future research should investigate additional architectures tailored for tabular data, further validating its performance. Promising approaches such as TabDDPM [34], a diffusion-based generative model designed for tabular data, could be evaluated to enhance the quality of synthetic data generation and improve the overall effectiveness of SDS. Expanding our evaluation to include datasets from other domains, such as financial markets [35] or sustainable energy [36], would demonstrate the broader applicability of SDS. Additionally, focusing on increasingly heterogeneous and imbalanced data distributions can be used to further validate and refine the methodology. While this study varied the distribution of a single feature (BMI) across nodes, a more realistic setup could involve modifying multiple features to emulate extreme heterogeneity better. However, such modifications may limit direct comparisons with techniques like FedAvg, which rely on consistent feature sets across nodes. Exploring these scenarios independently of FedAvg could provide deeper insights into SDS's performance under extreme variability. Exploring other data types, such as imaging or sequential data, could provide new opportunities to extend the methodology to domains requiring diverse data modalities. Incorporating these data types would further validate the flexibility and robustness of SDS in addressing challenges across a wide range of applications. Lastly, addressing privacy risks must be a future line of research on this topic. Investigating techniques to mitigate privacy risks associated with FL, such as differential privacy [37] or homomorphic encryption [38], can help protect sensitive patient data while enabling collaborative training. By pursuing these research directions, we can continue advancing the FL field for SDG in healthcare and develop more robust and effective methods for generating high-quality synthetic data.

Author Contributions: P.A.A., J.P. and S.Z. conceived the study and designed the general architecture. P.A.A. and J.P. developed the code, conducted the experiments, and analyzed the results. P.A.A. wrote the original manuscript. All authors have read and agreed to the the published version of the manuscript.

Funding: This work was supported by the GenoMed4All and SYNTHEMA projects from the European Union's Horizon 2020 Research and Innovation Program under Grant 101017549 and Grant 101095530. However, the views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used in this research are publicly available and can be found in the following repository: https://github.com/Patricia-A-Apellaniz/fed_vae (accessed on 20 January 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. VAE-BGM Methodology

We include detailed derivations and explanations in this appendix to provide further clarity on the VAE-BGM model [11]. The full implementation of the model can be accessed via the following GitHub repository: https://github.com/Patricia-A-Apellaniz/vae-bgm_data_generator (accessed on 20 January 2025). Below, we elaborate on the VAE framework's underlying principles and the ELBO's derivation.

The VAE is a generative model first introduced in 2013 [39]. It employs deep neural networks for Bayesian inference, aiming to represent a dataset $x_i{}_{i=1}^{N}$ of N independent and identically distributed (i.i.d.) samples. These samples are generated via a two-step stochastic process:

1. A latent variable z_i is drawn from a prior distribution $p(z)$. A simple isotropic Gaussian prior is often assumed for simplicity and generality.
2. The observed variable x_i is then sampled from a conditional distribution $p_\theta(x|z)$ governed by model parameters θ. This process is referred to as the generative model.

Figure A1 illustrates this process, where the latent variable z forms the basis for generating observable data x.

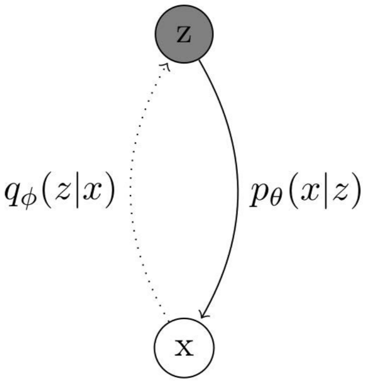


Figure A1. Bayesian VAE model. The shaded circle refers to the latent variable z, and the white circle refers to the observable x. Probabilities $p_\theta(x|z)$ and $q_\phi(z|x)$ denote, respectively, the generative model and the variational approximation to the posterior, since the true posterior $p_\theta(z|x)$ is unknown.

The true posterior $p_\theta(z|x)$ is generally intractable because the marginal likelihood $p_\theta(x)$ is difficult to compute directly. Variational inference addresses this by introducing an approximation $q_\phi(z|x)$, parameterized by ϕ, to the true posterior.

Appendix A.1. ELBO Derivation

Formulating the corresponding optimization problem is necessary to establish the optimization objective. Assuming that x_i are independent and identically distributed (i.i.d.), the marginal likelihood of a dataset comprising $\{x_i\}_{i=1}^{N}$ can be expressed as

$$\log p_\theta(x_1, x_2, ..., x_N) = \sum_{i=1}^{N} \log p_\theta(x_i), \tag{A1}$$

where the marginal likelihood for a single data point x is given by the following:

$$p_\theta(x) = \int p_\theta(x,z)dz = \int p_\theta(x,z)\frac{q_\phi(z|x)}{q_\phi(z|x)}dz = \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right]. \tag{A2}$$

Using Jensen's inequality, we can obtain the following:

$$\log p_\theta(x) = \log \left[\mathbb{E}_{q_\phi(z|x)} \left[\frac{p_\theta(x,z)}{q_\phi(z|x)} \right] \right] \geq \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)} \right]. \tag{A3}$$

Rearranging Equation (A3), we can express it as follows:

$$\begin{aligned} \mathbb{E}_{q_\phi(z|x)} \left[\log \left(\frac{p_\theta(x,z)}{q_\phi(z|x)} \right) \right] \\ = \int q_\phi(z|x) \log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} dz \\ = \int q_\phi(z|x) \log \frac{p(z)}{q_\phi(z|x)} dz + \int q_\phi(z|x) \log p_\theta(x|z) dz \\ = -\int q_\phi(z|x) \log \frac{q_\phi(z|x)}{p(z)} dz + \int q_\phi(z|x) \log p_\theta(x|z) dz \\ = -D_{\mathbb{KL}}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \\ = \mathcal{L}(x,\theta,\phi), \end{aligned} \tag{A4}$$

where $D_{\mathbb{KL}}(p||q)$ is the $D_{\mathbb{KL}}$ between distributions p and q, and $\mathcal{L}(x,\theta,\phi)$ is the ELBO, which is defined as follows:

$$\log p_\theta(x) \geq -D_{\mathbb{KL}}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] = \mathcal{L}(x,\theta,\phi). \tag{A5}$$

Thus, the ELBO is a lower bound for the marginal log-likelihood of the relevant set of points. Maximizing the ELBO maximizes the log-likelihood of the data. This would be the optimization problem to solve. This last equation, labeled (A5), coincides with the ELBO formulation previously introduced as Equation (1).

Appendix B. Privacy Concerns

To ensure that the proposed SDG process is privacy-preserving, we conducted an empirical evaluation by analyzing the similarity between real and synthetic data. This analysis aims to confirm that the generated synthetic data maintain sufficient statistical similarity to the real data for utility while avoiding direct replication of real samples, thereby protecting sensitive information. Such privacy assurance is particularly critical in federated environments where synthetic data are shared across nodes.

We designed the study as follows:

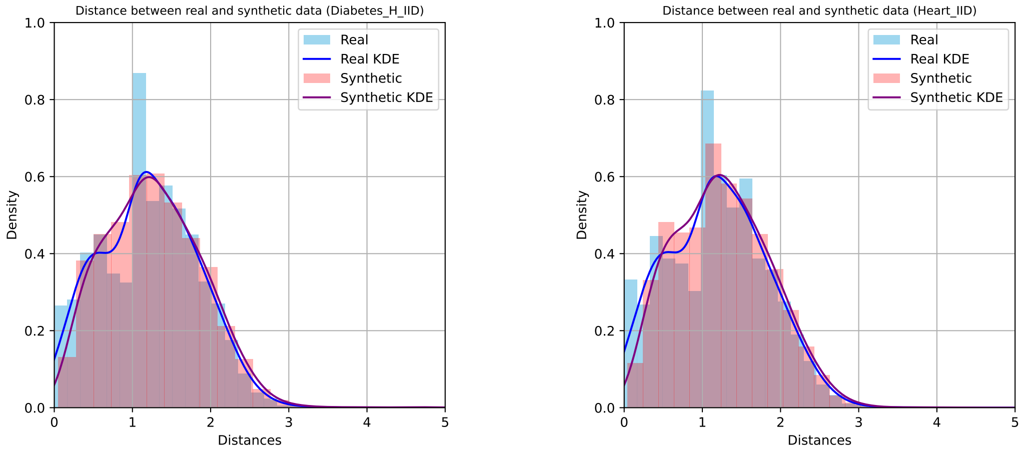
1. Real data were input to our proposed VAE-BGM generative framework to produce synthetic data.
2. These synthetic data were shared with a specific node within the federated framework.
3. To quantify the similarity and verify privacy preservation, we calculated the minimum pairwise distances between the following:
   - Real samples and other real samples.
   - Synthetic samples and real samples.
4. The minimum distances were visualized using histograms and KDE plots for both comparisons.
5. To statistically validate the differences between these minimum distances, we applied one-sided Wilcoxon and Kolmogorov–Smirnov (KS) tests to compare the two distance distributions.

The results of the privacy evaluation are presented in Figure A2. This evaluation was specifically performed on Node 3 in both IID and non-IID scenarios for the two datasets used in the study (Heart and Diabetes_H), resulting in four distinct plots. These analyses comprehensively evaluate privacy preservation under different data distributions and experimental settings. The histograms and KDE curves of the minimum distances between real–real samples (blue) and synthetic–real samples (pink) exhibit high similarity. However, they are not identical, which is a desirable outcome. Specifically:

- Non-zero minimum distances: Since the VAE-BGM framework generates synthetic data by sampling from the latent space rather than directly reconstructing the real samples, the minimum distances are never zero. This ensures that no synthetic sample exactly replicates any real data point, mitigating privacy risks.
- Larger distances for synthetic–real comparisons: The *p*-values obtained through the Wilcoxon and KS tests (both lower than 10^{-3}) confirm that the minimum distances between synthetic and real samples are statistically larger than those observed between real samples themselves. This outcome aligns with our expectations, as synthetic data are generated from a latent space and are not direct copies of the real data.
- Similarity, not equality: While the distributions of the distances are similar, the histograms and KDEs demonstrate slight deviations, reflecting the stochastic nature of the latent space sampling process. This confirms that the synthetic data preserve the statistical properties of the real data without compromising privacy.

These results demonstrate that the synthetic data provide sufficient privacy protection. Specifically, even in the event of an attack such as man-in-the-middle during the federated data sharing process, the exposure of synthetic data would present a far lower risk than raw data transmission, as the synthetic data are inherently different from the real samples. The minimum distances provide further assurance, as the largest values consistently arise in the synthetic–real comparisons, reinforcing that the synthetic data do not overlap with real data points.

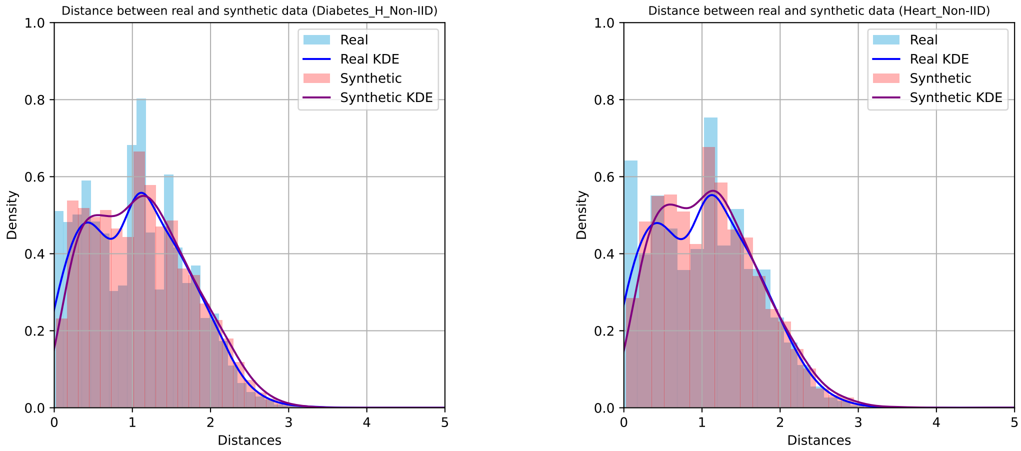
This study validates the privacy-preserving nature of our SDG process. The proposed framework effectively mitigates privacy risks such as re-identification by ensuring that the generated synthetic data do not replicate real data points while retaining statistical similarity. Furthermore, the statistical tests confirm that the synthetic data maintain an appropriate level of separation from real data, making them robust to adversarial attacks within a federated learning environment. These findings underscore the utility of the VAE-BGM architecture in generating high-quality, privacy-preserving synthetic data suitable for FL environments.



(**a**) Diabetes_H IID Scenario

(**b**) Heart IID Scenario

Figure A2. *Cont.*



(**c**) Diabetes_H Non-IID Scenario

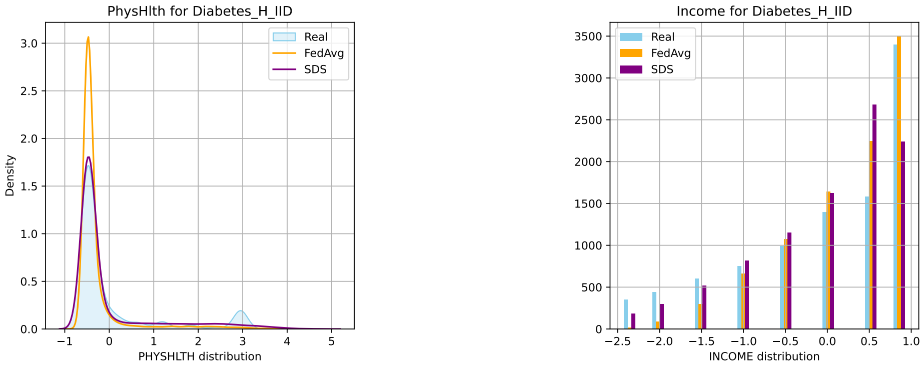
(**d**) Heart Non-IID Scenario

Figure A2. Comparison of minimum pairwise distances between real–real samples and synthetic–real samples. The histograms and KDE plots show similar distributions, ensuring statistical resemblance while maintaining privacy.

Appendix C. Comparison of Features Distributions

To evaluate the performance of SDG techniques in capturing the underlying distributions of critical features, we focused on the worst-performing node (Node 0) in both IID and non-IID scenarios for the Diabetes_H dataset. We identified the most important features for classification tasks using an RF classifier trained on real data. The distributions of these key features were then compared between real data and synthetic data generated using FedAvg and SDS, aiming to assess the quality of the synthetic data generated by each technique.

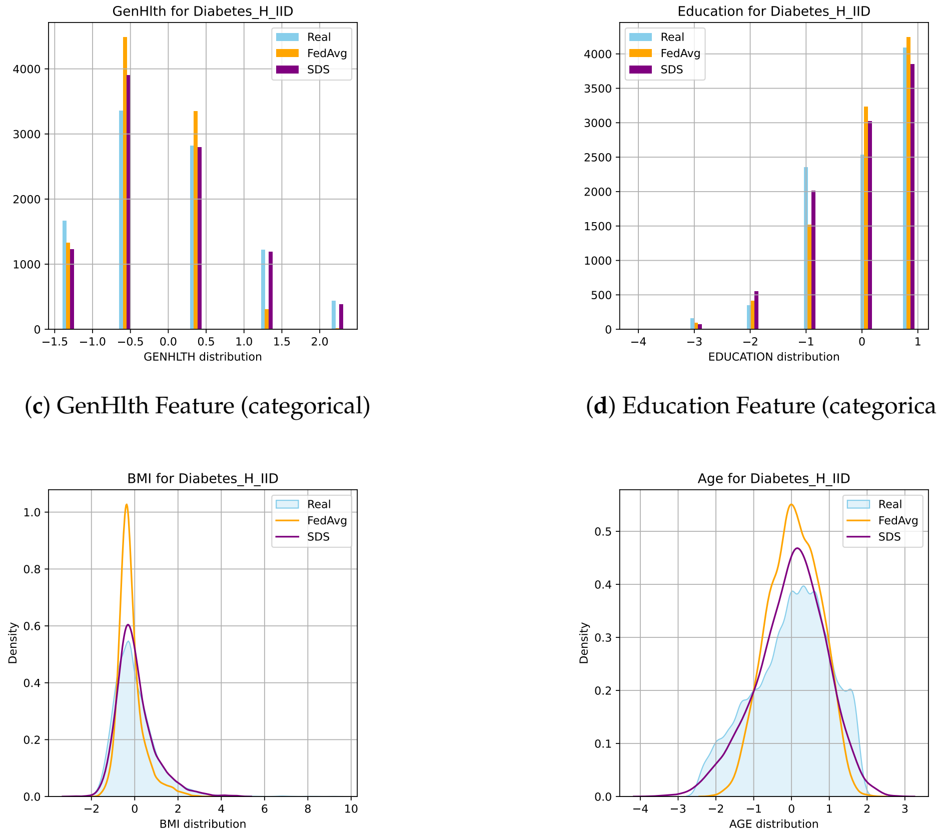
Figures A3 and A4 show the KDEs and histograms of the selected features for both IID and non-IID scenarios. These comparisons provide insights into how effectively each SDG technique captures the distributions of the real data. We can confirm that SDS captures critical continuous features more accurately. SDS better approximates the real data distribution for features such as Age and BMI compared to FedAvg. This suggests that SDS can model complex continuous variables more effectively. In addition, SDS improves the handling of categorical features. SDS demonstrates a superior ability to generate realistic distributions for categorical features such as Education. In contrast, FedAvg overemphasizes the most predominant category, resulting in less representative synthetic data distributions. Finally, consistency across IID and non-IID scenarios since the ability of SDS to generate accurate feature distributions remains evident in both of them.



(**a**) PhysHlth Feature (continuous)

(**b**) Income Feature (categorical)

Figure A3. *Cont.*



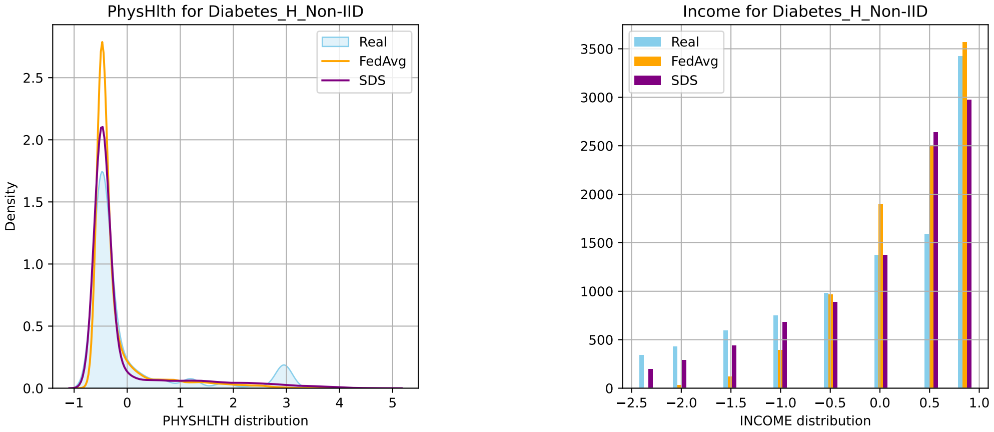
(**c**) GenHlth Feature (categorical)

(**d**) Education Feature (categorical)

(**e**) BMI Feaure (continuous)

(**f**) Age Feature (continuous)

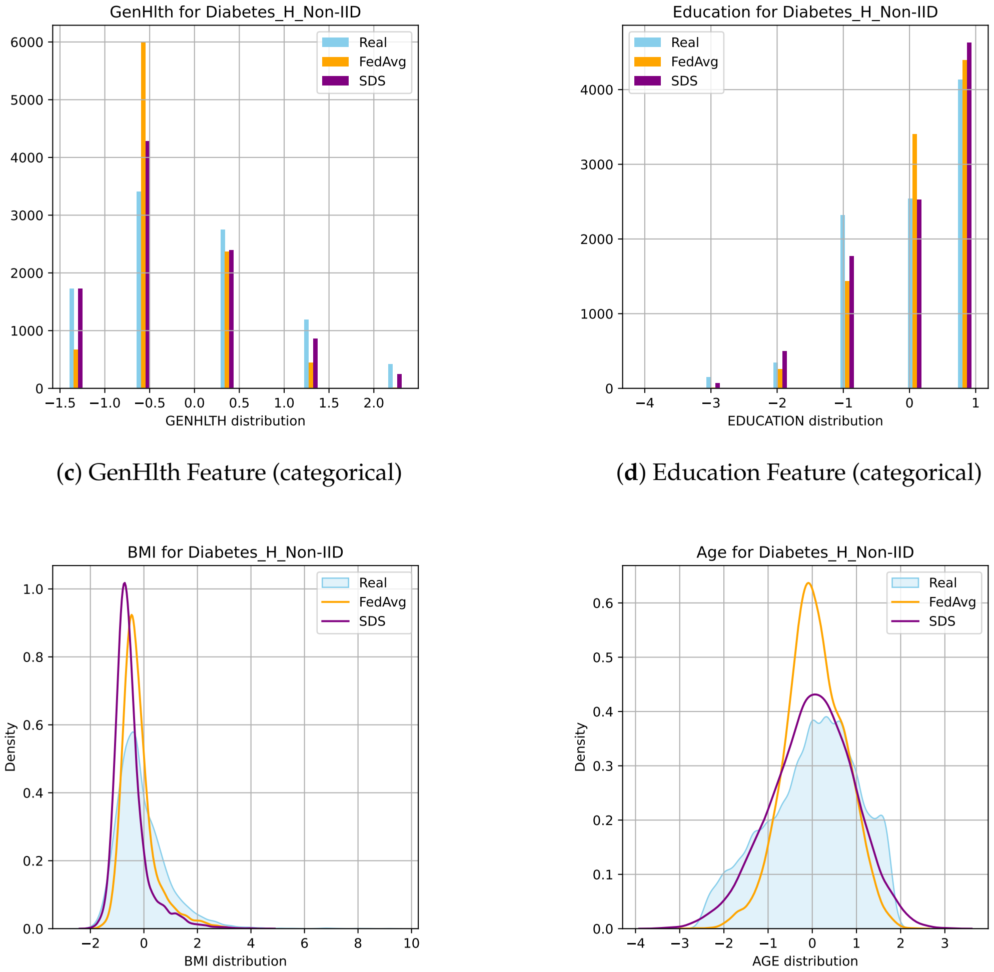
Figure A3. Distribution plots of selected features from the real data, synthetic data generated by FedAvg, and synthetic data generated by SDS in the IID scenario for the Diabetes_H dataset. All displayed feature data are normalized.



(**a**) PhysHlth Feature (continuous)

(**b**) Income Feature (categorical)

Figure A4. *Cont.*



(**c**) GenHlth Feature (categorical) (**d**) Education Feature (categorical)

(**e**) BMI Feature (continuous) (**f**) Age Feature (continuous)

Figure A4. Distribution plots of selected features from the real data, synthetic data generated by FedAvg, and synthetic data generated by SDS in the non-IID scenario for the Diabetes_H dataset. All displayed feature data are normalized.

Appendix D. Impact of Sample Distribution Across Nodes

This section examines the impact of varying sample sizes across nodes on the proposed FL framework's performance in both IID and non-IID scenarios. Tables A1 and A2 summarize the results for different combinations of sample sizes, evaluating statistical similarity using D_{JS} and clinical utility through accuracy metrics for real–real and synthetic–real setups.

- Clinical Utility: Across all sample size combinations, models trained on synthetic data generally achieve slightly lower accuracy than those trained on real data (Real-Real accuracy). However, the gap is small, often within a few decimal points, indicating that synthetic data remains highly effective for downstream classification tasks. In both IID and non-IID scenarios, in most cases, SDS outperforms FedAvg in clinical utility validation. This trend is particularly evident in nodes with smaller sample sizes, where SDS produces synthetic data that generalize more effectively to real-world applications.
- Similarity Validation: SDS consistently achieves lower D_{JS} values than both FedAvg and isolated training, particularly in nodes with limited data. This indicates that SDS generates synthetic data that more closely mirror the real data distribution, effectively mitigating the effects of data scarcity and heterogeneity. In contrast, FedAvg often

struggles to reduce D_{JS}, particularly in non-IID scenarios where the data distributions across nodes are highly skewed.

- Node-Level Performance: In Node 1, SDS demonstrates substantial improvements over FedAvg and isolated training in both D_{JS} and accuracy metrics, highlighting its ability to compensate for limited local data through shared synthetic samples. While SDS continues to outperform FedAvg in most cases in Node 2, the differences are less pronounced, reflecting the node's relatively balanced data availability. Finally, Node 3 does not benefit from any technique since the large dataset allows effective model training without significant reliance on external synthetic data.

Table A1. Diabetes_H results in IID scenario with different sample sizes: Comparison of D_{JS} and accuracy for isolated training and two FL techniques. Lower D_{JS} indicates better similarity between real and synthetic data, while synthetic–real accuracy closer to real–real reflects better clinical utility. Results are expressed as *mean (std)*. * indicates p-value < 0.01. In particular, for D_{JS}, * denotes statistically significant improvement over isolated case. For accuracy, * signifies that performance of models trained on synthetic data was comparable to or exceeded that of models trained on real data. **Bold** values indicate best significative performance, and ▼ denotes decrease relative to upper bounds.

Sample Size Combination	Node	Technique	Similarity Validation	Clinical Utility Validation	
			Estimated D_{JS}	Accuracy (Real–Real)	Accuracy (Synthetic–Real)
(50, 100, 500)	Node 1	Isolated	0.872 (0.002)	0.843 (0.002)	0.823 (0.001) ▼
		FedAvg	0.843 (0.005) *		0.833 (0.002) ▼
		SDS	**0.753 (0.016)** *		0.827 (0.003) ▼
	Node 2	Isolated	0.789 (0.008)	0.850 (0.000)	0.841 (0.001) ▼
		FedAvg	0.771 (0.011)		0.834 (0.001) ▼
		SDS	**0.692 (0.004)** *		0.839 (0.001) ▼
	Node 3	Isolated	0.560 (0.045)	0.853 (0.001)	0.843 (0.002) ▼
		FedAvg	0.631 (0.017)		0.839 (0.002) ▼
		SDS	0.623 (0.030)		0.839 (0.002) ▼
(50, 100, 1000)	Node 1	Isolated	0.887 (0.01)	0.843 (0.001)	0.824 (0.004)▼
		FedAvg	0.816 (0.011) *		0.821 (0.003) ▼
		SDS	**0.690 (0.031)** *		0.838 (0.000) ▼
	Node 2	Isolated	0.756 (0.016)	0.847 (0.0)	0.839 (0.001) ▼
		FedAvg	0.753 (0.012)		0.841 (0.002) *
		SDS	**0.625 (0.003)** *		0.837 (0.001) ▼
	Node 3	Isolated	0.547 (0.027)	0.852 (0.000)	0.845 (0.001) ▼
		FedAvg	0.520 (0.022)		0.838 (0.001) ▼
		SDS	0.483 (0.010)		0.848 (0.001)*
(50, 500, 1000)	Node 1	Isolated	0.862 (0.004)	0.843 (0.002)	0.824 (0.002) ▼
		FedAvg	0.853 (0.005)		0.823 (0.002) ▼
		SDS	**0.593 (0.006)** *		0.840 (0.003) *
	Node 2	Isolated	0.605 (0.022)	0.849 (0.001)	0.845 (0.002) *
		FedAvg	0.614 (0.013)		0.846 (0.002) *
		SDS	0.611 (0.005)		0.845 (0.002)*
	Node 3	Isolated	0.462 (0.057)	0.853 (0.000)	0.848 (0.002) *
		FedAvg	0.403 (0.031)		0.831 (0.003) ▼
		SDS	0.572 (0.018)		0.832 (0.001) ▼
(100, 500, 1000)	Node 1	Isolated	0.775 (0.009)	0.842 (0.001)	0.832 (0.002) ▼
		FedAvg	0.692 (0.009)*		0.827 (0.004) ▼
		SDS	**0.541 (0.043)***		0.838 (0.003) *
	Node 2	Isolated	0.587 (0.005)	0.850 (0.000)	0.850 (0.001) *
		FedAvg	0.599 (0.011)		0.845 (0.000) ▼
		SDS	0.543 (0.008)*		0.845 (0.001) ▼

Table A1. *Cont.*

Sample Size Combination	Node	Technique	Similarity Validation	Clinical Utility Validation	
			Estimated D_{JS}	Accuracy (Real–Real)	Accuracy (Synthetic–Real)
(100, 500, 1000)	Node 3	Isolated	0.502 (0.002)	0.852 (0.001)	0.842 (0.003) ▼
		FedAvg	0.549 (0.008) ▼		0.839 (0.003)▼
		SDS	**0.593 (0.029)**		0.838 (0.003) ▼

Table A2. Diabetes_H results in non-IID scenario with different sample sizes: Comparison of D_{JS} and accuracy for isolated training and two FL techniques. Lower D_{JS} indicates better similarity between real and synthetic data, while synthetic–real accuracy closer to real–real reflects better clinical utility. Results are expressed as *mean (std)*. * indicates p-value < 0.01. In particular, for D_{JS}, * denotes statistically significant improvement over isolated case. For accuracy, * signifies that performance of models trained on synthetic data was comparable to or exceeded that of models trained on real data. **Bold** values indicate best significative performance, and ▼ denotes decrease relative to upper bounds.

Sample Size Combination	Node	Technique	Similarity Validation	Clinical Utility Validation	
			Estimated D_{JS}	Accuracy (Real–Real)	Accuracy (Synthetic–Real)
(50, 100, 500)	Node 1	Isolated	0.882 (0.009)	0.845 (0.001)	0.821 (0.001) ▼
		FedAvg	0.850 (0.006) *		0.825 (0.000) ▼
		SDS	**0.764 (0.017)** *		0.836 (0.002) ▼
	Node 2	Isolated	0.780 (0.005)	0.788 (0.002)	0.767 (0.002)) ▼
		FedAvg	0.774 (0.021)		0.761 (0.003) ▼
		SDS	0.767 (0.000)		0.778 (0.003) ▼
	Node 3	Isolated	0.595 (0.023)	0.902 (0.000)	0.902 (0.000) *
		FedAvg	0.628 (0.020)		0.903 (0.001) *
		SDS	0.576 (0.005)		0.901 (0.000) *
(50, 100, 1000)	Node 1	Isolated	0.871 (0.004)	0.846 (0.002)	0.812 (0.001) ▼
		FedAvg	0.878 (0.005)		0.824 (0.000) ▼
		SDS	**0.627 (0.028)** *		0.842 (0.001) *
	Node 2	Isolated	0.734 (0.005)	0.788 (0.002)	0.787 (0.000) *
		FedAvg	0.724 (0.025)		0.782 (0.002) ▼
		SDS	**0.714 (0.004)** *		0.786 (0.002) *
	Node 3	Isolated	0.527 (0.048)	0.898 (0.001)	0.900 (0.000) *
		FedAvg	0.542 (0.045)		**0.902 (0.001)** *
		SDS	0.526 (0.063)		**0.902 (0.001)** *
(50, 500, 1000)	Node 1	Isolated	0.886 (0.008)	0.846 (0.001)	0.801 (0.002)▼
		FedAvg	0.830 (0.013) *		0.823 (0.000) ▼
		SDS	**0.556 (0.005)** *		0.838 (0.001) ▼
	Node 2	Isolated	0.549 (0.039)	0.793 (0.002)	0.784 (0.002) ▼
		FedAvg	0.558 (0.022)		0.782 (0.002) ▼
		SDS	0.514 (0.016)		0.783 (0.000) ▼
	Node 3	Isolated	0.523 (0.031)	0.905 (0.0)	0.904 (0.002) *
		FedAvg	0.519 (0.011)		0.900 (0.001) ▼
		SDS	0.422 (0.037)		0.906 (0.000) *
(100, 500, 1000)	Node 1	Isolated	0.793 (0.012)	0.846 (0.000)	0.829 (0.002) ▼
		FedAvg	0.829 (0.003)		0.827 (0.002) ▼
		SDS	**0.694 (0.009)***		0.831 (0.001) ▼
	Node 2	Isolated	0.575 (0.046)	0.810 (0.000)	0.793 (0.002) ▼
		FedAvg	0.619 (0.002)		0.806 (0.006) *
		SDS	0.639 (0.005)		0.798 (0.002) ▼
	Node 3	Isolated	0.561 (0.007)	0.914 (0.000)	0.913 (0.001) *
		FedAvg	**0.478 (0.009)** *		0.914 (0.001) *
		SDS	0.594 (0.026)		0.912 (0.001) *

References

1. Gkiouleka, A.; Wong, G.; Sowden, S.; Bambra, C.; Siersbaek, R.; Manji, S.; Moseley, A.; Harmston, R.; Kuhn, I.; Ford, J. Reducing health inequalities through general practice. *Lancet Public Health* **2023**, *8*, e463–e472. [CrossRef]
2. Chelak, K.; Chakole, S. The role of social determinants of health in promoting health equality: A narrative review. *Cureus* **2023**, *15*, e33425. [CrossRef] [PubMed]
3. Puri, V.; Sachdeva, S.; Kaur, P. Privacy preserving publication of relational and transaction data: Survey on the anonymization of patient data. *Comput. Sci. Rev.* **2019**, *32*, 45–61. [CrossRef]
4. Jayabalan, M.; Rana, M.E. Anonymizing healthcare records: A study of privacy preserving data publishing techniques. *Adv. Sci. Lett.* **2018**, *24*, 1694–1697. [CrossRef]
5. Salem, O.; Alsubhi, K.; Shaafi, A.; Gheryani, M.; Mehaoua, A.; Boutaba, R. Man-in-the-Middle attack mitigation in internet of medical things. *IEEE Trans. Ind. Inform.* **2021**, *18*, 2053–2062. [CrossRef]
6. Abowd, J.M.; Vilhuber, L. How protective are synthetic data? In Proceedings of the International Conference on Privacy in Statistical Databases, Istanbul, Turkey, 24–26 September 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 239–246.
7. Kotal, A.; Piplai, A.; Chukkapalli, S.S.L.; Joshi, A. Privetab: Secure and privacy-preserving sharing of tabular data. In Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics, Baltimore, MD, USA, 24–27 April 2022; pp. 35–45.
8. Choi, E.; Biswal, S.; Malin, B.; Duke, J.; Stewart, W.F.; Sun, J. Generating multi-label discrete patient records using generative adversarial networks. In Proceedings of the Machine Learning for Healthcare Conference. PMLR, Boston, MA, USA, 18–19 August 2017; pp. 286–305.
9. Park, N.; Mohammadi, M.; Gorde, K.; Jajodia, S.; Park, H.; Kim, Y. Data synthesis based on generative adversarial networks. *arXiv* **2018**, arXiv:1806.03384. [CrossRef]
10. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling tabular data using conditional gan. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 7335–7345.
11. Apellaniz, P.A.; Parras, J.; Zazo, S. An Improved Tabular Data Generator with VAE-GMM Integration. In Proceedings of the 2024 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 26–30 August 2024; pp. 1886–1890.
12. Apellániz, P.A.; Jiménez, A.; Galende, B.A.; Parras, J.; Zazo, S. Artificial Inductive Bias for Synthetic Tabular Data Generation in Data-Scarce Scenarios. *arXiv* **2024**, arXiv:2407.03080.
13. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol. (TIST)* **2019**, *10*, 1–19. [CrossRef]
14. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial intelligence and statistics, PMLR, Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
15. Ye, M.; Fang, X.; Du, B.; Yuen, P.C.; Tao, D. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Comput. Surv.* **2023**, *56*, 1–44. [CrossRef]
16. Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; Chandra, V. Federated learning with non-iid data. *arXiv* **2018**, arXiv:1806.00582. [CrossRef]
17. Efthymiadis, F.; Karras, A.; Karras, C.; Sioutas, S. Advanced Optimization Techniques for Federated Learning on Non-IID Data. *Future Internet* **2024**, *16*, 370. [CrossRef]
18. Wang, H.; Kaplan, Z.; Niu, D.; Li, B. Optimizing federated learning on non-iid data with reinforcement learning. In Proceedings of the IEEE INFOCOM 2020-IEEE Conference on Computer Communications, Toronto, ON, Canada, 6–9 July 2020; pp. 1698–1707.
19. Jeong, E.; Oh, S.; Kim, H.; Park, J.; Bennis, M.; Kim, S.L. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv* **2018**, arXiv:1811.11479.
20. Duan, M.; Liu, D.; Chen, X.; Tan, Y.; Ren, J.; Qiao, L.; Liang, L. Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications. In Proceedings of the 2019 IEEE 37th international conference on computer design (ICCD), Abu Dhabi, United Arab Emirates, 17–20 November 2019; pp. 246–254.
21. Nguyen, A.T.; Torr, P.; Lim, S.N. Fedsr: A simple and effective domain generalization method for federated learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 38831–38843.
22. Fallah, A.; Mokhtari, A.; Ozdaglar, A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3557–3568.
23. Gao, K.; Sener, O. Modeling and optimization trade-off in meta-learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 11154–11165.
24. Apellániz, P.A.; Jiménez, A.; Galende, B.A.; Parras, J.; Zazo, S. Synthetic Tabular Data Validation: A Divergence-Based Approach. *IEEE Access* **2024**, *12*, 103895–103907. [CrossRef]
25. Li, X.; Huang, K.; Yang, W.; Wang, S.; Zhang, Z. On the convergence of fedavg on non-iid data. *arXiv* **2019**, arXiv:1907.02189.
26. Thrun, S.; Pratt, L. Learning to learn: Introduction and overview. In *Learning to Learn*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 3–17.

27. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
28. Woźnica, K.; Wilczyński, P.; Biecek, P. SeFNet: Bridging Tabular Datasets with Semantic Feature Nets. *arXiv* **2023**, arXiv:2306.11636.
29. Boles, A.; Kandimalla, R.; Reddy, P.H. Dynamics of diabetes and obesity: Epidemiological perspective. *Biochim. Biophys. Acta-(BBA)-Mol. Basis Dis.* **2017**, *1863*, 1026–1036. [CrossRef]
30. Alexander, J.K. Obesity and coronary heart disease. *Am. J. Med. Sci.* **2001**, *321*, 215–224. [CrossRef] [PubMed]
31. Seiglie, J.A.; Marcus, M.E.; Ebert, C.; Prodromidis, N.; Geldsetzer, P.; Theilmann, M.; Agoudavi, K.; Andall-Brereton, G.; Aryal, K.K.; Bicaba, B.W.; et al. Diabetes prevalence and its relationship with education, wealth, and BMI in 29 low-and middle-income countries. *Diabetes Care* **2020**, *43*, 767–775. [CrossRef] [PubMed]
32. Hernadez, M.; Epelde, G.; Alberdi, A.; Cilla, R.; Rankin, D. Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. *Methods Inf. Med.* **2023**, *62*, e19–e38. [CrossRef] [PubMed]
33. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
34. Kotelnikov, A.; Baranchuk, D.; Rubachev, I.; Babenko, A. Tabddpm: Modelling tabular data with diffusion models. In Proceedings of the International Conference on Machine Learning. PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 17564–17579.
35. Naritomi, Y.; Adachi, T. Data augmentation of high frequency financial data using generative adversarial network. In Proceedings of the 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Melbourne, Australia, 14–17 December 2020; pp. 641–648.
36. Thonglek, K.; Ichikawa, K.; Takahashi, K.; Nakasan, C.; Yuasa, K.; Babasaki, T.; Iida, H. Enhancing the prediction accuracy of solar power generation using a generative adversarial network. In Proceedings of the 2021 IEEE Green Energy and Smart Systems Conference (IGESSC), Long Beach, CA, USA, 1–2 November 2021; pp. 1–6.
37. Dwork, C. Differential privacy. In Proceedings of the International Colloquium on Automata, Languages, and Programming, Venice, Italy, 10–14 July 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–12.
38. Paillier, P. Public-key cryptosystems based on composite degree residuosity classes. In Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques, Zagreb, Croatia, 10–14 May 2020; Springer: Berlin/Heidelberg, Germany, 1999; pp. 223–238.
39. Kingma, D.P. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.