Chapter 4

Synthetic Data Generation

4.1 Introduction

The availability of high-quality, diverse, and representative datasets is a cornerstone of medical research and ML applications in healthcare. However, the sensitive nature of medical data, strict privacy regulations, ethical considerations, and the scarcity of well-annotated datasets pose significant barriers to accessing and sharing such data. SDG has emerged as a promising solution to address these challenges. By generating realistic and privacy-preserving synthetic datasets, SDG enables researchers to develop and validate ML models without compromising patient confidentiality.

Despite its potential, SDG in the medical domain remains nascent and faces several challenges. Current GMs, such as VAEs, GANs, and their variations, often fail to preserve the complex joint distributions of high-dimensional medical data. Instead, existing evaluation standards focus on marginal distributions (e.g., column-wise similarity), neglecting the joint dependencies critical for downstream tasks such as predictive modeling or causal inference.

This chapter proposes a novel GM, VAE-BGM (Variational Autoencoder with Bayesian Gaussian Mixture), tailored to generate high-dimensional medical data. VAE-BGM is designed to capture the intricate joint distributions of medical datasets while ensuring scalability and privacy preservation. Beyond proposing a model, this chapter also addresses the broader challenge of evaluation in SDG. We introduce a standardized validation methodology that leverages divergence approximators to quantify the similarity between synthetic and real datasets, explicitly accounting for joint distributions. Additionally, considering the scarcity of data in the medical domain, this chapter explores the application of transfer learning and meta-learning techniques to enhance SDG for small or limited datasets. By leveraging knowledge from related domains or tasks, these approaches enable the generation of synthetic data that faithfully represents the characteristics of limited real-world datasets.

The following contributions and associated publications support the methodologies and frameworks presented in this chapter:

- **VAE-BGM:** A. Apellániz P., Parras J., and Zazo S., *An Improved Tabular Data Generator with VAE-GMM Integration,'* in the 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 2024, pp. 1886-1890, doi: 10.23919/EUSIPCO63174.2024.10715230
- **Validation methodology:** A. Apellániz P., Jiménez A., Arroyo Galende B., Parras J., and Zazo S., *Synthetic Tabular Data Validation: A Divergence-Based Approach,'* in IEEE Access, vol. 12, pp. 103895-103907, 2024, doi: 10.1109/ACCESS.2024.3434582.
- **Enhanced SDG methodology for scarce-data scenarios:**
  - **General-data purpose:** A. Apellániz P., Jiménez A., Arroyo Galende B., Parras J., and Zazo S., *Artificial Inductive Bias for Synthetic Tabular Data Generation in Data-Scarce Scenarios,'* under review in Pattern Recognition journal. Preprint available: 2407.03080.
  - **Medical data:** A. Apellániz P.,Arroyo Galende B., Jiménez A., Parras J., and Zazo S., *Advancing Cancer Research with Synthetic Data Generation in Low-Data Scenarios,'* under review in the IEEE Journal of Biomedical and Health Informatics.

This chapter aims to advance the state of SDG in healthcare by addressing key methodological and evaluative challenges. The proposed solutions contribute to developing robust, scalable, and clinically relevant frameworks for SDG, ultimately supporting ML applications in healthcare. This work aligns with the broader vision of enabling inclusive and equitable healthcare research and innovation by bridging gaps in generative modeling, evaluation, and data scarcity.

4.2 Proposed Synthetic Data Generation Model: VAE-BGM

4.2.1 Methodology of the Proposed Model (VAE-BGM)

The foundation of the proposed VAE-BGM model is the VAE, whose detailed explanation can be found in Appendix A.

Gaussian Mixture model

The GMM offers a probabilistic framework for clustering data points based on Gaussian distributions. Formally, given the same dataset $X = \{x_i\}_{i=0}^{N}$ that contains N samples with Cov features each, the GMM assumes that the data originate from a mixture of K Gaussian distributions. Each component within this mixture is characterized by its mean vector μ_k, covariance matrix $\sum_k$, and mixing coefficient π_k. These parameters collectively define the shape of each Gaussian component, location, and contribution to the mixture. The likelihood of observing a data point x as a weighted sum of the densities of individual Gaussian components is:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k), \tag{4.1}$$

where $\mathcal{N}(x|\mu_k, \Sigma_k)$ represents the probability density function of a multivariate Gaussian distribution with mean μ_k and covariance matrix $\sum_k$. Estimating the model parameters, μ_k, $\sum_k$, and π_k, is typically achieved by iterative optimization algorithms. The commonly employed Expectation-Maximization (EM) algorithm provides a widely used approach for this task.

GMMs are known to be universal approximators for continuous densities [245], and we will rely on that property. However, GMMs face the challenge of needing K as a parameter, i.e., it has to be predetermined before training. To overcome this issue, K can be estimated using a Dirichlet process [246] using Variational Inference tools. This approach, called Bayesian Gaussian Mixture (BGM), has the advantage of automatically adjusting K from the data.

Our Contribution

This work introduces a novel approach to data generation by integrating a BGM into a standard VAE architecture. This integration leverages the strengths of both models: the ability of the VAE to learn a latent representation of the data z and the flexibility of the BGM in modeling complex distributions, potentially non-Gaussian, within this latent space. Figure 4.1 depicts our model.

State-of-the-art models, such as TVAE [109], implicitly assume that the latent space z aligns with the prior Gaussian distribution. Therefore, to generate new samples of z, sampling from the prior distribution, which is isotropic Gaussian, suffices. The reason is the presence of the $D_{\mathbb{KL}}$ term in the loss function Equation (A.6). However, this assumption may not hold because of complex dependencies and correlations in real-world data. These factors are represented by the other term in the loss function, the marginal log-likelihood distributions.

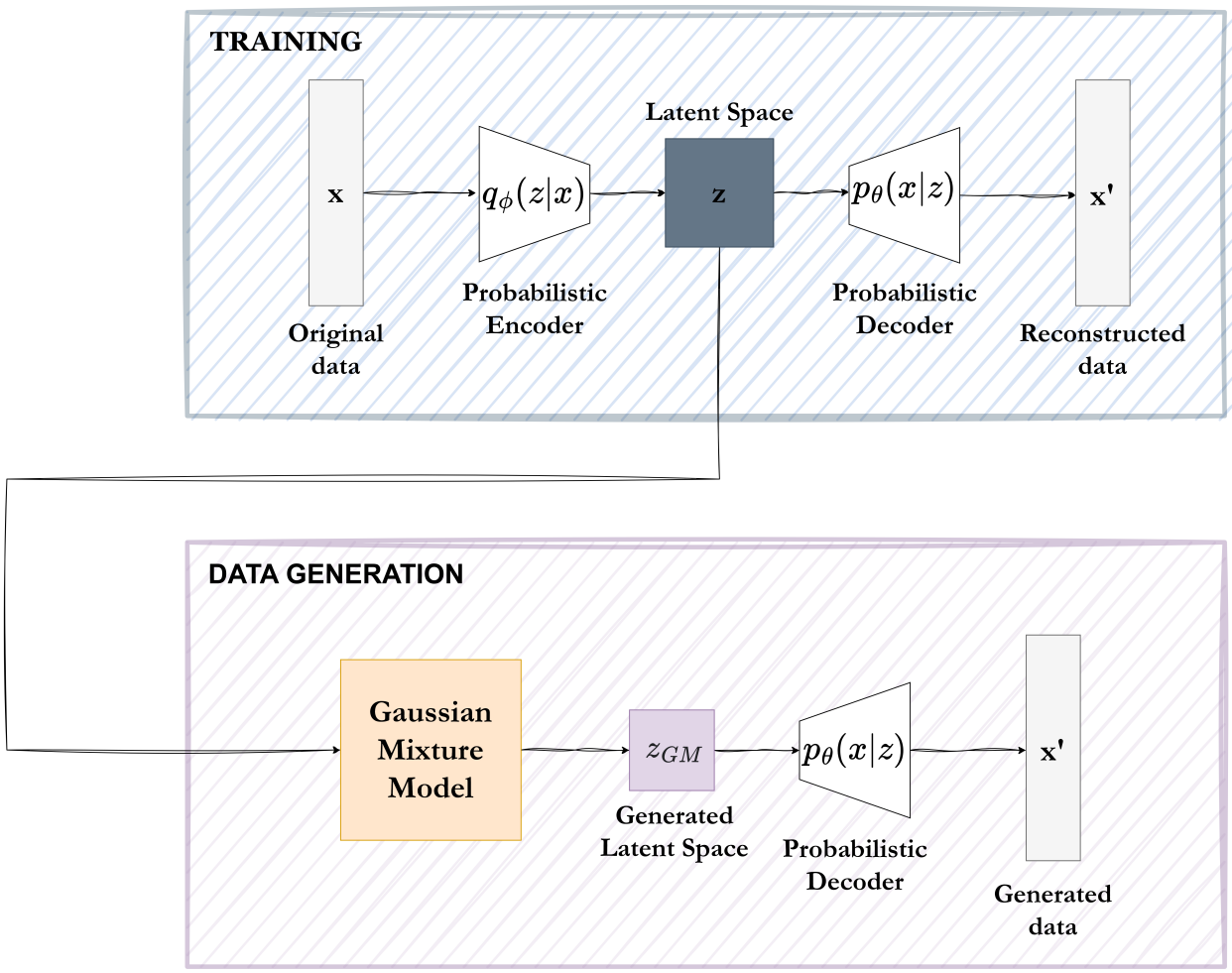


Figure 4.1: Proposed VAE-BGM model architecture. It is built on a standard VAE. After training, the latent space z is modeled using a GMM. This creates a new space z_{GM}, which serves as the basis for generating new distribution parameters and ultimately sampling new data points.

This term can push the latent space away from a Gaussian distribution. In other words, the VAE loss in Equation (A.6) has two components: the log-likelihood of the reconstructed data, which allows the reconstruction of samples from the latent space with high similarity, and the $D_{\mathbb{KL}}$, which acts as a regularizer in the latent space. The $D_{\mathbb{KL}}$ limits the complexity of the latent space so that it can be thought of as the allowed latent representation complexity. As it is typical to use an isotropic Gaussian as prior, due to the KL being analytical, there is an equilibrium between having a good reconstruction with a low-complexity latent space representation. In practice, the actual latent space learned need not be an isotropic Gaussian (as we will show experimentally), and thus, sampling z from the prior may not provide the best results in generating new samples.

Therefore, our key contribution lies in employing a BGM for the sampling process of new tabular data. Unlike TVAE, our model does not assume that the actual latent space follows the isotropic Gaussian prior: instead, we additionally model the already learned latent space

z as a mixture of K Gaussian distributions, each characterized by its mean, covariance matrix, and mixing coefficient. Notably, if z is truly an isotropic Gaussian, the BGM could still effectively approximate it, as it corresponds to the case where $K = 1$.

4.2.2 Overview of State-of-the-Art Comparative Models

The work in [109] introduces two powerful GMs for tabular data: CTGAN (Conditional Tabular GAN) and TVAE (Tabular Variational Autoencoder). These models address key challenges in tabular data generation, including the coexistence of mixed data types, non-Gaussian distributions, and imbalanced categorical features. Both aim to generate realistic synthetic data while preserving the statistical properties of the original dataset.

Given a real dataset $X_r = \{x_i\}_{i=0}^{N}$ containing N samples and Cov features, indexed by $f = 1, 2, ..., Cov$, we assume that the dataset consists of N_c continuous $(C_1, ..., C_{N_c})$ and N_d discrete features $(D_1, ..., D_{N_d})$. Each sample x_i is a combination of continuous and discrete features $x_i = \{c_{i,1}, c_{i,2}, ..., c_{i,N_c,d_{i,1},d_{i,2},...,d_{i,N_d}}\}$, where $Cov = N_c + N_d$. The following subsections describe the working mechanisms of CTGAN and TVAE.

CTGAN

CTGAN introduces multiple innovations to tackle the complexities of tabular data generation. One of its key features is mode-specific normalization, designed to handle continuous variables that exhibit multimodal and non-Gaussian distributions. To achieve this, CTGAN applies a Variational GMM to estimate the modes of each continuous feature. The probability of a given continuous value $c_{i,f}$ belonging to a mode ψ is computed as:

$$\mathcal{P}_\psi = \mu_\psi \mathcal{N}(c_{i,f}; \eta_\psi; \phi_\psi), \tag{4.2}$$

where μ_ψ is the weight, η_ψ is the mean, and ϕ_ψ is the standard deviation of mode ψ. The value $c_{i,f}$ is then normalized using the most probable mode ψ:

$$\hat{c}_{i,f} = \frac{c_{i,f} - \eta_\psi}{4\phi_\psi}, \tag{4.3}$$

while the mode is represented as a one-hot vector:

$$\beta_{i,f} = [\text{one-hot vector indicating mode } \psi]. \tag{4.4}$$

This mode-specific normalization ensures NNs can effectively process continuous data with complex distributions.

CTGAN uses a conditional generator that learns the conditional distribution $P_{Gen}(x_i|D_f = \psi)$, where D_f is a selected categorical column with value ψ. This approach helps address the issue of imbalanced categorical variables by ensuring that synthetic data accurately represent minority categories. To enforce this conditioning, CTGAN incorporates a cross-entropy loss in the objective of the generator:

$$\mathcal{L}_{\text{CTGAN}_{Gen}} = -\frac{1}{m}\sum_{j=1}^{m}\sum_{k=1}^{|D_j|} 1[d_{i,j} = k] \log P_{Gen}(\hat{d}_{i,j} = k), \tag{4.5}$$

where m is the batch size, D_j is the categorical domain of feature j, $1[\cdot]$ is an indicator function and $P_{Gen}(\hat{d}_{i,j})$ represents the predicted categorical distribution for feature j. This formulation ensures that generated samples align with the categorical feature distributions of the real data.

The discriminator evaluates the quality of the generated data using the Wasserstein GAN loss with gradient penalty:

$$\mathcal{L}_{\text{CTGAN}_{Disc}} = \mathbb{E}[\text{Disc}(x_g)] - \mathbb{E}[\text{Disc}(x_r)] + \lambda \mathbb{E}[(||\nabla_{\tilde{x}_r} \text{Disc}(\tilde{x}_r)||_2 - 1)^2], \tag{4.6}$$

where $\tilde{x}_r$ is a linear interpolation between real samples x_r and synthetic samples x_g.

TVAE

TVAE, on the other hand, adapts the VAE architecture to tabular data, explicitly modeling the probabilistic distributions of both continuous and discrete variables. The model consists of an encoder $q_\phi(z|x_r)$ to map rows x_i to a latent space z and the decoder $p_\theta(x_r|z)$ to reconstruct the rows. The decoder represents continuous variables α_j as Gaussian distributions:

$$\alpha_j \sim \mathcal{N}(\mu_j, \sigma_j), \tag{4.7}$$

where μ_j is the mean and σ_j is the variance. Categorical variables β_j are modeled using Softmax distributions:

$$\beta_j \sim \text{Softmax(h)}, d_j \sim \text{Softmax(h)}, \tag{4.8}$$

where h represents the logits (pre-activation values) for the categorical outputs.

The joint distribution for all columns is defined as:

$$p_\theta(x_r|z) = \prod_{j=1}^{N_c} P(\alpha_j) \prod_{j=1}^{N_c} P(\beta_j) \prod_{j=1}^{N_d} P(d_j), \tag{4.9}$$

where N_c and N_d are the numbers of continuous and discrete columns, respectively. The encoder outputs the mean μ and variance σ^2 of the latent variables:

$$q_\phi(z|x_r) \sim \mathcal{N}(\mu, \text{diag}(\sigma^2)). \tag{4.10}$$

The model optimizes the ELBO to maximize the likelihood of the data:

$$\mathcal{L}_{\text{TVAE}} = \mathbb{E}_{q_\phi(z|x_r)}[\log p_\theta(x_r|z)] - D_{KL}(q_\phi(z|x_r)||p(z)). \tag{4.11}$$

Suitability for Comparative Evaluation

CTGAN and TVAE stand out as robust GMs for tabular data, both offering unique capabilities that make them highly suitable for comparative evaluation. CTGAN excels in handling the challenges of mixed data types by employing innovations such as mode-specific normalization and a conditional generator, enabling it to effectively model non-Gaussian and multimodal distributions. Its ability to handle highly imbalanced categorical columns is particularly

noteworthy, as it ensures that all categories are well-represented in the synthetic data. This feature makes CTGAN a versatile tool across various applications, from medical research to finance, where data often exhibits such complexities.

TVAE complements CTGAN by taking a probabilistic approach to data generation. By modeling the joint distribution of continuous and discrete variables explicitly within a latent space, TVAE achieves competitive performance and has the potential to become a state-of-the-art solution. Its reliance on VAEs allows it to generate realistic synthetic data while maintaining scalability. The straightforward architecture of TVAE makes it easier to implement and integrate with other systems, offering an attractive alternative for practitioners seeking a balance between performance and simplicity.

However, these models are not without limitations. CTGAN, while capable of handling both categorical and continuous features, can encounter convergence challenges when generating continuous data. This issue may arise in cases where the continuous columns exhibit extreme distributions or sparse coverage in the training dataset, leading to instability during training. Future work could explore improved optimization strategies or adaptive architectures to mitigate this issue.

Similarly, the reliance of TVAE on a Gaussian assumption for its latent space and sampling process, though effective for many datasets, might not generalize well to complex real-world scenarios with non-Gaussian characteristics. This limitation could reduce similarity in the generated data when the underlying data distribution significantly deviates from a Gaussian one. Exploring alternative sampling techniques, such as those based on copulas or non-parametric methods, could enhance the applicability of TVAE to a broader range of datasets.

In summary, CTGAN and TVAE represent significant advancements in tabular data generation. Their strengths in addressing the core challenges of tabular data make them excellent candidates for benchmarking against other methods. However, addressing their respective limitations would further solidify their positions as state-of-the-art tools for SDG.

4.2.3 Experiments and Results

Evaluation Metrics

Due to the lack of standardized metrics, as explained in Section 2.3.4, evaluating data generation models requires a multifaceted approach. We used a combination of resemblance and utility-based evaluations.

We assessed resemblance through multiple metrics to ensure a comprehensive assessment. First, we used an RF discriminator to measure the similarity of the joint distribution between real and synthetic data. In an ideal scenario of high model performance, the RF accuracy should be close to a random guessing rate (around 0.5), indicating that it can not distinguish between real and synthetic data. Furthermore, we used the technique proposed in [247] to validate the marginal distributions and the correlation between pairs of columns. Additionally, we incorporated the MMD to measure the divergence between two probability distributions without requiring parametric assumptions. Specifically, we used the Radial Basis Function (RBF) kernel, which is well-suited for capturing non-linear relationships by mapping data into

a high-dimensional feature space, allowing a more fine-grained comparison of distributions. Recall that given two datasets, MMD is computed as:

$$MMD^2(p,q) = \mathbb{E}_p[\hat{k}(x_r, x_r)] - 2\mathbb{E}_{p,q}[\hat{k}(x_r, x_g)] + \mathbb{E}_{\text{ıı}}[\hat{k}(x_g, x_g)], \tag{4.12}$$

where x_r and x_g represent real and synthetic data, respectively, and $\hat{k}(x, y)$ is the RBF kernel function:

$$\hat{k}(x,y) = \exp\left(-\frac{|x-y|^2}{2\sigma^2}\right). \tag{4.13}$$

Utility evaluation, tailored to the type of task, involved training CoxPH for SA and the RF discriminator for classification tasks. This evaluation considered training and testing on real data (benchmark) and training on synthetic data with real data testing. This assessment ensured that the generated data exhibited high resemblance (capturing feature distributions and correlations) and utility (supporting robust statistical inference).

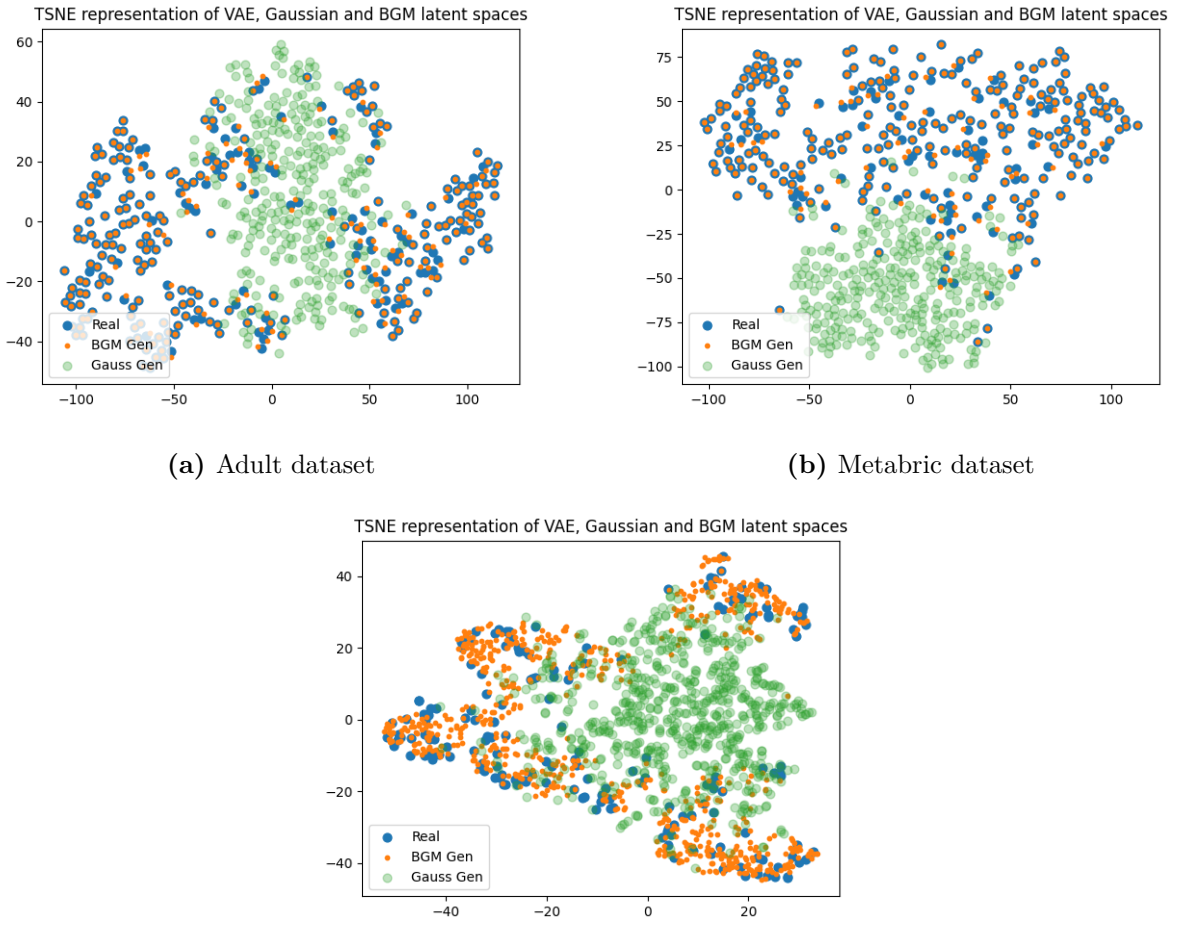
To benchmark the performance of our model, we compared it with CTGAN and TVAE, the current state-of-the-art models for SDG. TVAE and our model were implemented similarly for a coherent comparison, but we added the BGM sampling to ours. All models were evaluated using the same methodology for a fair and consistent comparison.

Implementation Details

Our experimental setup involved two models. The VAE architecture comprised a simple encoder with a hidden layer of ReLU and a hyperbolic tangent output layer. The decoder mirrored this structure with a ReLU hidden layer, a 20% dropout layer, and an output layer with activation functions adjusted to the covariate distributions. We fixed the latent space dimensionality (number of Gaussian distributions) to 5 and the number of hidden neurons to 50. During the training of $1,000$ epochs and a batch size of 500, an early stop was configured based on the validation loss. The GMM employed a Dirichlet process prior with a fixed maximum number of components set to the dimensionality of the latent space (in this case, 5) and individual general covariance matrices for each component. Data were divided into 80% training and 20% validation sets for training and evaluation. Due to the sensitivity of VAEs to initial conditions, 15 training runs with different seeds were performed. The final results were averaged from the three seeds with the best performance.

Results

The code and data to replicate the results are publicly available in https://github.com/Patricia-A-Apellaniz/vae-bgm_data_generator. The datasets used for these experiments are Adult, Metabric, and STD, which are described along with their characteristics in Appendix B. We obtained the Adult dataset from [247] for this study and used 10,000 samples. Note that the Metabric dataset used in this study is the version preprocessed by [59], where the dimensionality is reduced from 21 features to 9, plus the two survival features, resulting in a more compact dataset for these experiments.



(a) Adult dataset

(b) Metabric dataset

(c) STD dataset

Figure 4.2: Latent space comparison in VAE-BGM experiments. Three hundred samples from each VAE, TVAE, and the latent space of the proposed model are shown. The BGM-modeled latent space (orange) closely aligns with z (blue) compared to the latent space of TVAE (green). This demonstrates the importance of BGM for capturing the latent space distribution and potentially leading to higher-quality generated samples.

Figure 4.2 verifies our hypothesis that the actual latent space learned by the VAE does not need to follow the prior distribution, so our choice of BGM is justified. The figure compares the latent spaces obtained using three different approaches: (1) the original latent space from VAE, obtained by passing the dataset samples through the encoder, (2) latent space sampled using a BGM (ours), and (3) latent space sampled using the prior isotropic Gaussian (TVAE). We can see that the BGM-modeled latent space aligns closely with the original VAE latent space in all three cases. In contrast, the TVAE latent space exhibits less dispersion, suggesting that the Gaussian assumption does not accurately capture the underlying distribution of the real-world data. This visually confirms our hypothesis that the latent space in real-world datasets often deviates from a Gaussian distribution, which justifies the use of BGM for a more accurate representation in the sampling process.

The resemblance results of our model compared to CTGAN and the VAE-based Gaussian sampling model can be seen in Table 4.1, Table 4.2 and Table 4.3. Table 4.1 reports the accuracy and its CIs obtained by the RF for each model. For VAE-based models, the table shows the average of the three best seeds, and the CIs are calculated based on the ones obtained for each seed. Our model consistently outperformed the other two models by a significant margin in all datasets, indicating superior generative capabilities and robustness. In addition, Table 4.2 presents MMD results, which further validate the resemblance between real and synthetic data distributions. The results demonstrate that our model achieves the lowest MMD scores across all datasets, significantly outperforming CTGAN and TVAE. This suggests that VAE-BGM not only preserves marginal distributions but also maintains the joint feature relationships more effectively. The CIs, estimated using bootstrapping with 10 resamples, reinforce the stability of these findings. Furthermore, the results of the column analysis in Table 4.3 confirm the superiority of our model in terms of similarity to ground-truth values.

Dataset	CTGAN	TVAE	Our model
Adult	0.75 (0.74, 0.76)	0.78 (0.74, 0.81)	**0.68 (0.64, 71)**
Metabric	**0.73 (0.70, 0.76)**	0.77 (0.74, 0.80)	**0.67 (0.62, 0.71)**
STD	0.94 (0.91, 0.96)	0.77 (0.70, 0.82)	**0.64 (0.57, 0.70)**

Table 4.1: Resemblance evaluation using RF in VAE-BGM experiments. Accuracy and 99% CI. Lower is better. The best values are in **bold**.

Dataset	CTGAN	TVAE	Our model
Adult	**0.0004 (0.0002, 0.0005)**	0.0570 (0.0349, 0.0895)	**0.0002 (0.0000, 0.0008)**
Metabric	0.0103 (0.0077, 0.0126)	0.0159 (0.0084, 0.0318)	**0.0022 (0.0005, 0.0043)**
STD	0.0611 (0.0441, 0.0903)	0.0314 (0.0188, 0.0470)	0.**0042 (0.0005, 0.0075)**

Table 4.2: Resemblance evaluation using MMD in VAE-BGM experiments. MMD values with 99% confidence intervals, computed using bootstrapping with 10 resamples. Lower values indicate better resemblance between real and synthetic data. The best values are highlighted in **bold**.

Table 4.4 presents the validation results for each ML task based on the dataset. For the Adult dataset, the accuracy is reported, as it is a classification dataset, while the C-index metric is used for the other two datasets. Similarly to the previous evaluation, we present the average value for the three best-performing random seeds in the case of TVAE and our proposed method. The 'Real-Real' value represents the benchmark. The benchmark results are the upper bound in performance, where training and testing are performed using real data. Then, for each GM, we also show the results for these ML tasks training with synthetic data and testing with real data to assess whether the generated data are useful for the intended purpose of the dataset. The results in this validation stage indicate that, in general, the performance

metrics obtained using data generated by each model (TVAE and ours) are comparable to the results obtained using real data. This suggests that the generated data exhibit sufficient utility and quality for practical ML applications.

Dataset	CTGAN	TVAE	Our model
Adult	0.87	0.87	**0.93**
Metabric	0.89	0.88	**0.92**
STD	0.86	0.87	**0.95**

Table 4.3: Resemblance evaluation using column analysis [247] in VAE-BGM experiments. Higher is better: a score of 1 means that the patterns captured for real and synthetic data are the same. The best values are in **bold**.

Dataset	Benchmark	CTGAN	TVAE	Our approach
Adult	0.80 (0.79, 0.82)	0.79 (0.77, 0.80)	0.80 (0.77, 0.82)	0.79 (0.76, 0.81)
Metabric	0.58 (0.53, 0.63)	0.57 (0.52, 0.62)	0.60 (0.54, 0.65)	0.60 (0.54, 0.65)
STD	0.64 (0.57, 0.70)	0.54 (0.47, 0.61)	0.54 (0.46, 0.60)	0.65 (0.55, 0.75)

Table 4.4: Utility results comparing CTGAN, TVAE, and VAE-BGM. Accuracy results for Adult. C-index results for Metabric and STD. Note that all the methods tested provide useful data for the tasks tested, as the results fall in the CI of the benchmark.

4.2.4 Conclusions

This work proposes a novel approach to generate synthetic tabular data by integrating a BGM into a VAE architecture. Our experiments demonstrate that the proposed model outperforms the state-of-the-art models CTGAN and TVAE on several validation criteria. This superior performance can be attributed to two key strengths. First, our model effectively captures the diverse distributions in the data at the individual feature level (marginal) and across features (joint). Unlike many models that struggle with mixed data types, our approach handles various data types effectively. Second, incorporating the BGM enables high-quality sampling by approximating the latent space of the VAE. This is crucial, as the assumption of TVAE of a Gaussian latent space can lead to worse performance with real-world data that often deviates from this distribution. Looking ahead, several promising lines for future research exist. One direction involves exploring the privacy implications of data generation, particularly in sensitive domains like healthcare. Analyzing privacy concerns and developing secure methods to share synthetic data in medical settings is a valuable area of investigation. Furthermore, we propose investigating novel FL strategies based on synthetic data sharing rather than trained model information (addressed in Section 5.3). This approach could improve collaboration and knowledge sharing among entities while protecting sensitive information.

4.3 Synthetic Data Validation through Divergence Estimation

The evaluation of SDG remains a critical challenge in the field, as there is currently no standardized validation framework. Existing validation methods often focus exclusively on assessing marginal distributions, neglecting the joint distributions of the synthetic data. However, in real-world applications, particularly in healthcare, the relationships and dependencies between multiple features are crucial for ensuring the utility and fidelity of synthetic data.

This section proposes a novel validation approach that evaluates the joint distributions of the generated synthetic data relative to the real data. By incorporating the joint distributions, we aim to provide a more comprehensive measure of the quality of the synthetic data, ensuring that the complex relationships between features are accurately preserved. This proposed methodology addresses the limitations of marginal-focused validation techniques and establishes a more robust evaluation framework for SDG.

4.3.1 Methodology of the Proposed Validation Approach

Divergence definition

Considering two probability distributions, $p(x)$ and $q(x)$, for a random variable x, in probability distributions, divergence quantifies the dissimilarity between $p(x)$ and $q(x)$. It measures how different these distributions are regarding the probabilities assigned to each possible value of x. Divergences play a crucial role in various fields, including ML, as they provide a way to compare and analyze the behavior of different probability distributions.

The $D_{\mathbb{KL}}$ is a common measure in information theory to quantify the discrepancy between $p(x)$ and $q(x)$. Denoted $D_{\mathbb{KL}}\big(p(x)\|q(x)\big)$, the $D_{\mathbb{KL}}$ divergence captures the asymmetry of this difference by measuring the expected penalty incurred by assuming $q(x)$ to be an approximation of the true distribution $p(x)$. Due to its extensive applicability in various fields, the $D_{\mathbb{KL}}$ divergence remains a dominant metric for comparing probability distributions. The $D_{\mathbb{KL}}$ divergence between $p(x)$ and $q(x)$ is given by

$$D_{\mathbb{KL}}\big(p(x)\|q(x)\big) = \int p(x) \log \frac{p(x)}{q(x)} dx. \tag{4.14}$$

Unfortunately, this expression is usually intractable. Thus, we must approximate it using Monte Carlo (MC) simulation, assuming we can draw L samples from $p(x)$:

$$\int p(x) \log \frac{p(x)}{q(x)} dx \approx \frac{1}{L} \sum_{i=1}^{L} \log \frac{p\left(x_i\right)}{q\left(x_i\right)}, \quad x_i \sim p(x). \tag{4.15}$$

To create a symmetric and bounded measure of divergence between $p(x)$ and $q(x)$, the Jensen-Shannon divergence is based on $D_{\mathbb{KL}}$. $D_{\mathbb{JS}}$ between two probability distributions denoted $D_{\mathbb{JS}}\big(p(x)\|q\big(x)\big)\big)$, is a true distance defined as the $D_{\mathbb{KL}}$ average between both distributions

and a common reference:

$$D_{\mathbb{JS}}\Big(p(x)\|q(x)\Big) = \frac{1}{2}D_{\mathbb{KL}}\Big(p(x)\|m(x)\Big) + \frac{1}{2}D_{\mathbb{KL}}\Big(q(x)\|m(x)\Big). \tag{4.16}$$

The common reference distribution, $m(x)$, is typically chosen as the midpoint between the two original distributions, such as the average: $m(x) = \Big(p(x) + q(x)\Big)/2$. Notably, the $D_{\mathbb{JS}}$ is inherently bounded. When employing a base-2 logarithm, the maximum divergence value is 1.

Similarly to $D_{\mathbb{KL}}$, the $D_{\mathbb{JS}}$ can also be estimated using MC simulation:

$$\begin{aligned} & D_{\mathbb{JS}}\Big(p(x)\|q(x)\Big) \\ \approx & \frac{1}{2L}\sum_{i=1}^{L}\log\frac{p\left(x_i\right)}{m\left(x_i\right)} + \frac{1}{2L}\sum_{i=1}^{L}\log\frac{q\left(\tilde{x}_i\right)}{m\left(\tilde{x}_i\right)} \\ = & \frac{1}{2L}\sum_{i=1}^{L}\log\left(\frac{2p\left(x_i\right)}{p\left(x_i\right)+q\left(x_i\right)}\right) + \frac{1}{2L}\sum_{i=1}^{L}\log\left(\frac{2q\left(\tilde{x}_i\right)}{p\left(\tilde{x}_i\right)+q\left(\tilde{x}_i\right)}\right), \end{aligned} \tag{4.17}$$

assuming that we can generate L samples from both distributions

$$\begin{cases} x_i \sim p(x) \\ \tilde{x}_i \sim q(x) \end{cases}. \tag{4.18}$$

Estimating density ratio using probabilistic classification

Density ratio estimation focuses on estimating the ratio $r^*(x)$ of two probability densities, $p(x)$ and $q(x)$, based solely on samples drawn from these distributions. This field has seen significant theoretical development, particularly regarding complex ratio estimators and their convergence properties [248]. The ratio is therefore defined as

$$r^*(x) = \frac{p(x)}{q(x)}, \tag{4.19}$$

where $*$ refers to the exact value. This approach avoids directly estimating individual densities since errors in the denominator (i.e., $q(x)$) are dramatically amplified during the ratio calculation. The classical probabilistic classification approach remains popular among the various methods of estimating the density ratio due to its relative simplicity. This subsection reviews the method described by [249] for estimating the $D_{\mathbb{KL}}$. We also demonstrate its application to the $D_{\mathbb{JS}}$ divergence estimation.

Estimating Equation (4.14) and Equation (4.16) is based on determining the unknown density ratio r^*. This section establishes a connection between the density ratio of two distributions, $p(x)$ and $q(x)$, and a probabilistic classifier that optimally distinguishes samples drawn from these distributions. Consider a scenario in which we possess samples from both distributions, $p(x)$ and $q(x)$, with each sample labeled according to its origin. Using a classifier to estimate the class-membership probabilities for each sample, we can derive an estimator for the density

ratio. Let $X_p = x^{(1)}, x^{(2)}, ..., x^{(M_p)}$ and $X_q = \tilde{x}^{(1)}, \tilde{x}^{(2)}, ..., \tilde{x}^{(M_q)}$ represent sets of real and synthetic data samples, respectively, where M_p and M_q denote the corresponding sample sizes.

We form a combined dataset, denoted as $D = (x_n, y_n)_{n=1}^{N}$, where $M_T = M_p + M_q$ represents the total number of samples. The label y_n associated with each sample x_n indicates its source distribution: $y_n = 1$ for samples from $p(x)$ and $y_n = 0$ for samples from $q(x)$. Consequently, we have $p(x) = \mathcal{P}(x \mid y = 1)$ and $q(x) = \mathcal{P}(x \mid y = 0)$, where $\mathcal{P}$ denotes the probability.

By applying the theorem of Bayes theorem, we can express the density ratio as:

$$\begin{aligned} & r^*(x) \\ & \quad = \frac{p(x)}{q(x)} \\ & \quad = \frac{\mathcal{P}(x \mid y = 1)}{\mathcal{P}(x \mid y = 0)} \\ & \quad = \left(\frac{\mathcal{P}(y = 1 \mid x)\mathcal{P}(x)}{\mathcal{P}(y = 1)}\right)\left(\frac{\mathcal{P}(y = 0 \mid x)\mathcal{P}(x)}{\mathcal{P}(y = 0)}\right)^{-1} \\ & \quad = \frac{\mathcal{P}(y = 0)}{\mathcal{P}(y = 1)}\frac{\mathcal{P}(y = 1 \mid x)}{\mathcal{P}(y = 0 \mid x)}. \end{aligned} \tag{4.20}$$

Equal prior probabilities are assumed for the source distributions, specifically $\mathcal{P}(y = 0) = \mathcal{P}(y = 1)$. This assumption is strategically made to prevent the estimation process from being skewed towards any particular class. By setting the prior probabilities to be equal, the ratio of marginal probabilities effectively cancels out, leaving:

$$r^*(x) = \frac{\mathcal{P}(y = 1 \mid x)}{\mathcal{P}(y = 0 \mid x)}. \tag{4.21}$$

This expression reveals that r^* can be estimated solely based on the posterior probability, $\mathcal{P}(y = 1 \mid x)$, which represents the probability that a sample originates from $p(x)$ given its features.

To simplify even more, the logit function, denoted by σ^{-1}, offers a convenient way to transform the posterior probability into the ratio of the original probabilities. The logit function is defined as the inverse of the logistic function σ:

$$\sigma^{-1}(z) = \ln\left(\frac{z}{1 - z}\right), \tag{4.22}$$

where z represents any probability. The logit function has the property that the log odds of two probabilities equal the log of their ratio. In our case, applying the logit function to both the numerator and denominator of the expression for Equation (4.21) results in the log ratio of the original probabilities:

$$\ln\left(\frac{\mathcal{P}(y = 1 \mid x)}{\mathcal{P}(y = 0 \mid x)}\right) \quad = \sigma^{-1}\left(\mathcal{P}(y = 1 \mid x)\right) - \sigma^{-1}\left(\mathcal{P}(y = 0 \mid x)\right) \tag{4.23}$$

Since the prior probabilities are assumed to be equal, the second term cancels out, leaving us with:

$$\ln\left(r^*(x)\right) = \sigma^{-1}(\mathcal{P}(y = 1 \mid x)) \tag{4.24}$$

Taking the exponent of both sides recovers the original expression for $r^*(x)$ but expressed in terms of the odds of the posterior probability:

$$r^*(x) = \exp\left[\sigma^{-1}(\mathcal{P}(y = 1 \mid x))\right]. \tag{4.25}$$

Therefore, by applying the logit function, we can simplify the calculation of the density ratio. The transformation converts the posterior probability into a log-odds representation, ultimately leading to the desired log ratio. This establishes a crucial link between the density ratio and the probabilistic classifier. By training a classifier to effectively distinguish between samples from $p(x)$ and $q(x)$, we can obtain estimates of posterior probabilities and subsequently derive an estimate of the density ratio using the logit transformation.

Implementation

This research uses an NN classifier, denoted D_θ with parameters θ, to approximate the posterior probability $P(y = 1 \mid x)$. This network acts as a discriminator to classify input samples originating from the $p(x)$ or $q(x)$ distributions. The network architecture is designed to capture the discriminative features that distinguish these distributions effectively. This study aims to understand how input data variations influence the ratio estimation between distributions. To isolate the impact of input data, we deliberately chose not to fine-tune the model across different experiments, maintaining a consistent architecture throughout. This approach allows us to examine the robustness of the estimator across various data types without the confounding effects of model optimization. The model parameters were selected to balance interpretability and provide meaningful insights, ensuring that our results illustrate the practical capabilities of the estimator rather than achieving optimal performance in each scenario. This strategy emphasizes the versatility and applicability of the estimator in real-world contexts, offering valuable insights into the challenges of synthetic data validation.

Specifically, D_θ employs a three-layer architecture with a decreasing number of neurons per layer: 256 in the first hidden layer, followed by 64 and 32 neurons in the subsequent hidden layers, respectively. The Leaky Rectified Linear Unit activation function is used throughout the hidden layers to introduce nonlinearity into the model. In addition, dropout, batch normalization, and early stopping techniques are incorporated to prevent overfitting during the training process.

The estimator for r^* can be constructed as a function of the output of the classifier:

$$r_\theta(x) = \exp\left[\sigma^{-1}\left(D_\theta(x)\right)\right] \approx \exp\left[\sigma^{-1}(\mathcal{P}(y = 1 \mid x))\right] = r^*(x). \tag{4.26}$$

The optimal class probability estimator is learned by minimizing a suitable loss function, such

as the binary cross-entropy loss:

$$\mathcal{L}(\theta) = -\mathbb{E}_{p(x)}\Big[\log \mathcal{D}_\theta(x)\Big] - \mathbb{E}_{q(x)}\Big[\log(1-\mathcal{D}_\theta(x))\Big]. \tag{4.27}$$

Once trained, the class probability estimator can construct MC estimates of $D_{\mathbb{KL}}$ and $D_{\mathbb{J}}$. The $D_{\mathbb{KL}}$ between $p(x)$ and $q(x)$ can be estimated as

$$\begin{aligned} D_{\mathbb{KL}}\Big(p(x)\|q(x)\Big) & \\ &= \mathbb{E}_{p(x)}\Big[\log r^*(x)\Big] \\ &\approx \frac{1}{L}\sum_{i=1}^{L}\log r^*(x_i) \approx \frac{1}{L}\sum_{i=1}^{L}\log r_\theta(x_i) \\ &= \frac{1}{L}\sum_{i=1}^{L}\sigma^{-1}\Big(\mathcal{D}_\theta\left(x_i\right)\Big), \end{aligned} \tag{4.28}$$

where $x_i \sim p(x)$, L is the number of samples used for the estimate.

Similarly, the $D_{\mathbb{JS}}$ divergence can be estimated as:

$$\begin{aligned} & D_{\mathbb{JS}}\Big(p(x)\|q(x)\Big) \\ \approx & \frac{1}{2L}\sum_{i=1}^{L}\log\Big(2\mathcal{D}_\theta(x_i)\Big) + \frac{1}{2L}\sum_{i=1}^{L}\log\Big(2-2\mathcal{D}_\theta(\tilde{x}_i)\Big) \\ = & \frac{1}{2L}\sum_{i=1}^{L}\log\Big(\mathcal{D}_\theta(x_i)\Big) + \frac{1}{2L}\sum_{i=1}^{L}\log\Big(1-\mathcal{D}_\theta(\tilde{x}_i)\Big), \end{aligned} \tag{4.29}$$

where $\tilde{x}_i \sim q(x)$ denotes the number of samples used from the second distribution during divergence estimation.

Interestingly, the discriminator loss function, $\mathcal{L}(\theta)$, converges to a lower bound of $D_{\mathbb{JS}}$ up to a constant, $-2\cdot D_{\mathbb{JS}}\Big(p(x)\|q(x)\Big) + \log 4$ [249]. This relation can be established by analyzing the upper bound of the loss function:

$$\begin{aligned} & \sup_\theta \mathbb{E}_{p(x)}\Big[\log D_\theta(x)\Big] + \mathbb{E}_{q(x)}\Big[\log(1-D_\theta(x))\Big] \\ & = \mathbb{E}_{p(x)}[\log \mathcal{P}(y=1 \mid x)] + \mathbb{E}_{q(x)}[\log \mathcal{P}(y=0 \mid x)] \\ & = \mathbb{E}_{p(x)}\Big[\log \frac{p(x)}{p(x)+q(x)}\Big] + \mathbb{E}_{q(x)}\Big[\log \frac{q(x)}{p(x)+q(x)}\Big] \\ & = \mathbb{E}_{p(x)}\Big[\log \frac{1}{2}\frac{p(x)}{m(x)}\Big] + \mathbb{E}_{q(x)}\Big[\log \frac{1}{2}\frac{q(x)}{m(x)}\Big] \\ & = \mathbb{E}_{p(x)}\Big[\log \frac{p(x)}{m(x)}\Big] + \mathbb{E}_{q(x)}\Big[\log \frac{q(x)}{m(x)}\Big] - 2\log 2 \\ & = 2\cdot D_{\mathbb{JS}}\Big(p(x)\|q(x)\Big) - \log 4. \end{aligned} \tag{4.30}$$

Then, we can establish:

$$2\cdot D_{\mathbb{JS}}\Big(p(x)\|q(x)\Big) - \log 4 \geq \sup_\theta \mathbb{E}_{p(x)}\Big[\log D_\theta(x)\Big] + \mathbb{E}_{q(x)}\Big[\log(1-D_\theta(x))\Big]. \tag{4.31}$$

Following this analysis, we can arrive at the desired inequality. However, we can manipulate the inequality by negating both sides for a more convenient form.

$$\begin{aligned} & -2 \cdot D_{\mathbb{JS}}\Big(p(x) \| q(x)\Big) + \log 4 \\ & \leq -\sup_{\theta} \mathbb{E}_{p(x)}\Big[\log D_\theta(x)\Big] + \mathbb{E}_{q(x)}\Big[\log(1 - D_\theta(x))\Big] \\ & = \inf_{\theta} -\mathbb{E}_{p(x)}\Big[\log D_\theta(x)\Big] - \mathbb{E}_{q(x)}\Big[\log(1 - D_\theta(x))\Big] \\ & = \inf_{\theta} \mathcal{L}(\theta). \end{aligned} \tag{4.32}$$

The architecture of the divergence estimator proposed between two distributions is depicted in Figure 4.3. The discriminator network receives two sets of samples: M samples from the first distribution $p(x)$ labeled class 1 and M samples from the second distribution $q(x)$ labeled class 0. During training, the discriminator aims to distinguish between these two sets. Subsequently, L samples from each distribution estimate the divergence between the underlying probability distributions.

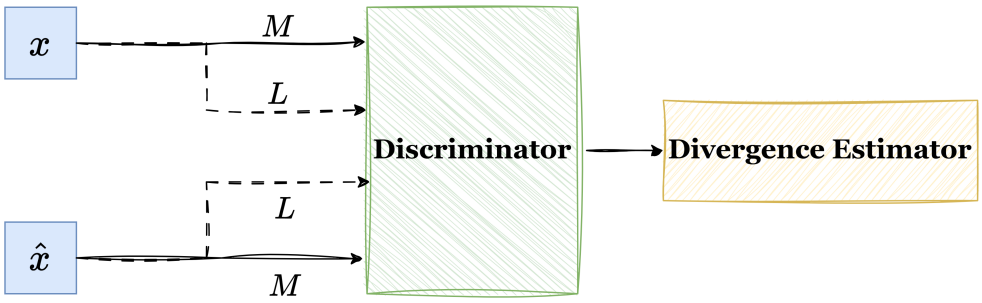


Figure 4.3: Architecture of the NN-based divergence estimator to assess the dissimilarity between samples from two datasets. The discriminator takes two sets of samples as input: M samples from each set to train and L samples from each to infer the divergence.

This study evaluates the dissimilarity between real and synthetic data generated by a GM. Minimizing the divergence between the real and synthetic data distributions is crucial in this context. Ideally, the divergence should approach zero, indicating that the generated data become virtually indistinguishable from the real data. This interchangeability allows synthetic data in various applications where real data might be scarce or sensitive. From the perspective of the discriminator, achieving minimal divergence implies that it cannot reliably differentiate between real and synthetic samples, indicating successful data generation. For GMs, the distributions of interest become:

- $p_R(x) = p(x)$, representing the real data distribution.
- $p_G(x) = q(x)$, representing the synthetic data distribution generated by the model.

Training data consist of N samples drawn from the real data distribution, denoted $x_r \sim p_R(x)$. These data are used to train the GM, as illustrated in Figure 4.4. The GM learns to

approximate the real data distribution as $p_G(x)$, allowing us to sample from this distribution and obtain synthetic data points $x_g \sim p_G(x)$. Following the architecture described in Figure 4.3, these synthetic samples are used along with the real data samples to train the discriminator and estimate the divergence between the real and synthetic data distributions.

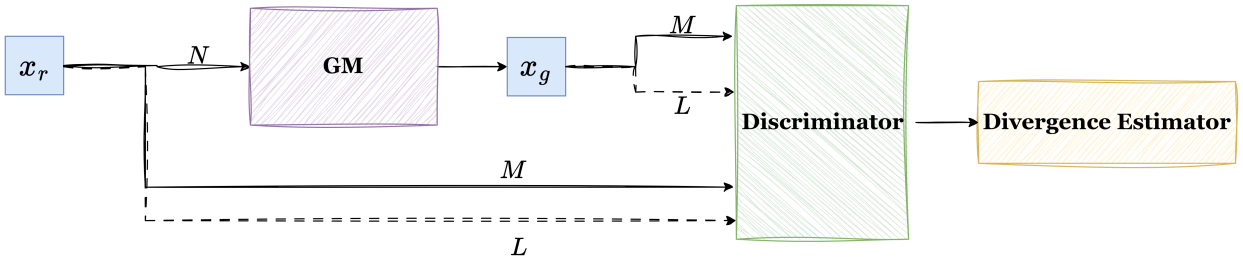


Figure 4.4: GM for divergence estimation between real and synthetic data. The GM learns an approximation of N samples from real data x_r, denoted x_g. Subsequently, M samples are drawn from each distribution to train the divergence estimation discriminator. Finally, L samples from each distribution estimate the divergence between real and synthetic data.

4.3.2 Experiments and Results

Experimental Setting

We aim to validate the effectiveness of our proposed divergence estimator for data generation. We achieve this through an exhaustive experimental evaluation in four different experiments, described in Table 4.5 designed to assess the performance of the estimator under various conditions. The initial set of experiments focuses on controlled scenarios that involve comparisons of simple theoretical data distributions. Subsequently, the complexity is gradually increased by incorporating generative scenarios that reproduce real-world applications. This progression in complexity tests the ability of the estimator to handle increasingly challenging conditions and demonstrates its scalability and reliability. The increasing complexity is essential to ensure that our estimator remains accurate and efficient, even as the distributional assumptions become less defined and more reflective of the irregularities typically found in real data.

We comprehensively analyze the influence of sample size on the performance of the estimator. This analysis covers the impact of the training and validation set sizes for the discriminator network and the number of samples used during the generative process. To gain a deeper understanding, we evaluated different configurations for each parameter, as the table shows. For every experiment, all possible combinations of N, M, and L are executed five times with unique random seeds. This rigorous approach employing multiple random seeds mitigates the effects of inherent randomness within training processes. Consequently, it provides a statistically robust evaluation and a more reliable representation of the observed trends. Ultimately, this comprehensive analysis facilitates identifying the configuration that yields the most accurate assessment of the divergence between data distributions.

Experiment	Compared Distributions	Generative Model	N	M	L
Experiment 1	Multivariate Gaussian Distributions	-	-	[20, 200, 2000]	[20, 200, 2000]
Experiment 2	Gaussian Mixture Distributions	-	-	[20, 200, 2000]	[20, 200, 2000]
Experiment 3	Gaussian Mixture and Synthetic Data	Gaussian Mixture Model	[10, 20, 30, ... , 150]	2,000	2,000
Experiment 4	Real and Synthetic Data	CTGAN[109], VAE	10,000	7,500	1,000

Table 4.5: Generation methodology experiments configuration summary. Experiments carried out to assess the effectiveness of the proposed approach. Experiments 1 and 2 investigate the impact of distribution complexity by varying the number of M and L. Experiments 3 and 4 focus on the application of generative processes. A GM is introduced to create synthetic data based on N samples, enabling a comparison between real and synthetic scenarios. Different values of N, M, and L could have been used; however, we believe that the chosen values are sufficiently illustrative of our aims. Moreover, we sought a trade-off between simplicity in interpretation and the ability to produce exemplary results.

The inherent variability in data nature and complexity across different experiments needs diverse divergence estimation methods. Table 4.6 details the specific techniques employed for each experiment, allowing for a comparative analysis whenever possible. The level of complexity of the experiment guides the selection of the validation technique.

- **Analytical Divergence:** This method, calculated only for $D_{\mathbb{KL}}$ for simplicity, represents the true divergence value due to the specific distributions used in the experiment.
- **MC Estimated Divergence:** This is a widely used estimation approach, but it can be computationally expensive.
- **Discriminator Estimated Divergence:** This method uses our proposed discriminator network to learn the density ratio between the two distributions and estimate the divergence.

Experiment	Validation Technique
Experiment 1	Analytical, MC and Discriminator Estimations
Experiment 2	MC and Discriminator Estimations
Experiment 3	MC and Discriminator Estimations
Experiment 4	Discriminator Estimation

Table 4.6: Validation procedure for each experiment in the generation methodology.

In the experiments conducted, the parameter values of the different models used were explicitly chosen to yield illustrative results to analyze the performance of the proposed estimator in validating SDG. We recognize that alternative parameters could be employed or fine-tuning could be undertaken to identify optimal settings; however, we intend to observe the effect of the input data on the models as such and to demonstrate their potential. This approach allows us to focus on understanding how variations in input data influence the ability of the estimator to validate synthetic datasets, thereby showcasing the practical applicability and robustness of the estimator across different data scenarios.

Results

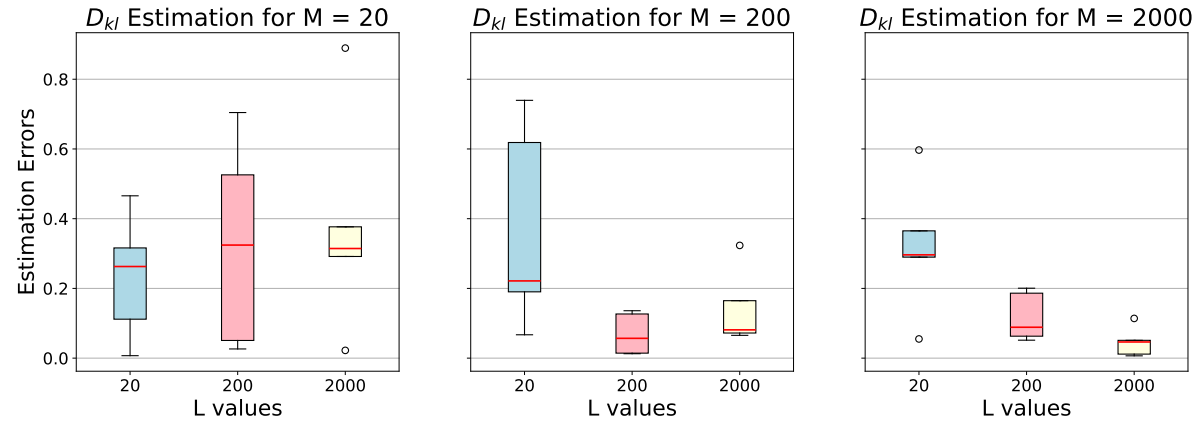
For complete transparency and reproducibility, the data and code used in this study are publicly available in https://github.com/Patricia-A-Apellaniz/divergence_estimator. Additional results can be found in Appendix D.1.

→ ***Experiment 1***

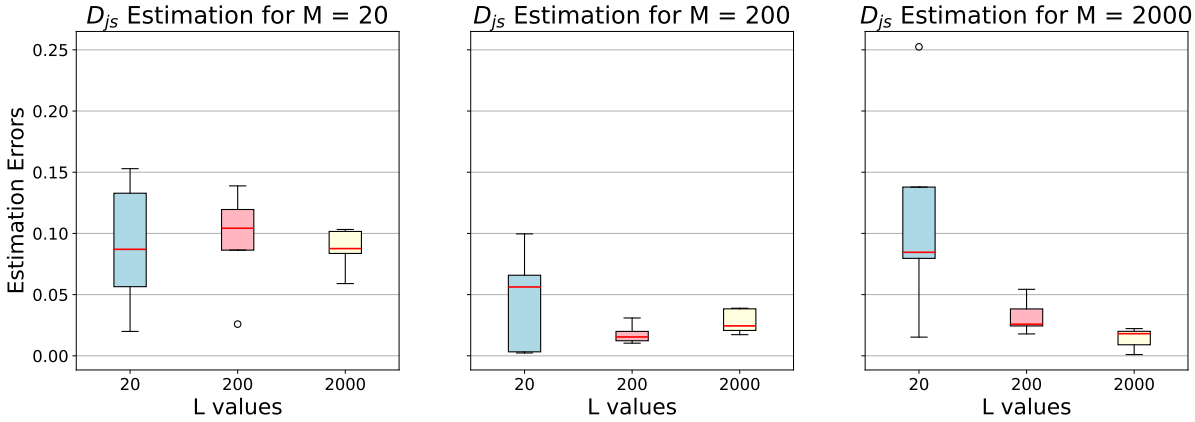
M	L	Analytical $D_{\mathbb{KL}}$	MC Estimated $D_{\mathbb{KL}}$	Discriminator Estimated $D_{\mathbb{KL}}$	MC Estimated $D_{\mathbb{JS}}$	Discriminator Estimated $D_{\mathbb{JS}}$
20	20	1.035	1.168 ± 0.203	0.802 ± 0.152	0.320 ± 0.105	0.213 ± 0.072
	200		1.026 ± 0.221	0.709 ± 0.248	0.301 ± 0.055	0.200 ± 0.038
	2000		1.024 ± 0.193	0.665 ± 0.084	0.288 ± 0.047	0.209 ± 0.028
200	20		1.221 ± 0.422	0.991 ± 0.520	0.320 ± 0.055	0.289 ± 0.053
	200		1.114 ± 0.257	1.099 ± 0.227	0.313 ± 0.055	0.294 ± 0.049
	2000		1.026 ± 0.191	0.955 ± 0.101	0.289 ± 0.038	0.267 ± 0.038
2000	20		0.841 ± 0.190	0.833 ± 0.229	0.219 ± 0.113	0.218 ± 0.111
	200		1.004 ± 0.190	0.978 ± 0.185	0.300 ± 0.047	0.290 ± 0.049
	2000		1.055 ± 0.212	1.060 ± 0.200	0.299 ± 0.051	0.293 ± 0.050

Table 4.7: Impact of training and validating samples on $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ estimation for Experiment 1. Analytical $D_{\mathbb{KL}}$ along with MC $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ estimations, as well as proposed discriminator estimations for both divergences. Results are displayed for various combinations of training samples M and validation samples L. There is a clear correlation between the number of samples used and the estimation error.

Table 4.7 presents the results obtained for $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ between random multivariate Gaussian distributions. A significant advantage of this study is the availability of a closed-form solution for D_{KL}, which serves as the ground truth value. We use the MC simulation estimate for the D_{JS} as the ground truth. To ensure the reliability of this experiment, we opted for multivariate Gaussian distributions with a dimensionality of 10. This allows us to compute the true divergence values and assess the accuracy of estimating the ratio of the proposed discriminator. The results reveal a critical relation between sample size and estimation accuracy, particularly for discriminator-based approaches. When the number of samples is limited ($M = 20$, $L = 20, 200$), the estimated divergence from the discriminator deviates significantly from both the analytical value and the MC estimate. This suggests a potential overfitting due to insufficient data, leading to underestimating the true divergence. However, as the number of samples increases, the estimated divergence from the discriminator network progressively converges towards the ground truth value, followed by improved CI precision. Figure 4.5 confirms this trend. Both subfigures illustrate how increasing the number of training samples M and validation samples L reduces the error associated with the estimated divergence ratio. Further support for the correlation between sample size and estimation accuracy comes from the discriminator loss curves depicted in Figure 4.6. As detailed in



(a) $D_{\mathbb{KL}}$



(b) $D_{\mathbb{JS}}$

Figure 4.5: Estimation error representation for $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ in Experiment 1. Results are shown for different combinations of training sample sizes M and validation sample sizes L. As expected, a decrease and precision in the error is observed with increasing values of M and L.

[249], the loss function converges to a constant value approximating the Jensen-Shannon divergence. The figure visually confirms this concept, as the loss curves flatten with increasing training M and validation L samples. In contrast, a low number of samples (e.g., $M = 20$, $L = 20$) results in larger fluctuations in the loss function, potentially indicating discriminator overfitting. These validations demonstrate the effectiveness of our proposed method: with sufficient training data, the discriminator can accurately learn the density ratio and provide reliable estimates of the divergences, particularly for the $D_{\mathbb{JS}}$.

Leveraging Experiment 1, we further emphasize the superior robustness of our proposed approach in estimating the D_{JS}, regardless of variations in distribution separation. Figure 4.7 illustrates the relation between estimated divergences and ground truth values as the separation between two 4-dimensional multivariate Gaussian distributions increases. As observed,

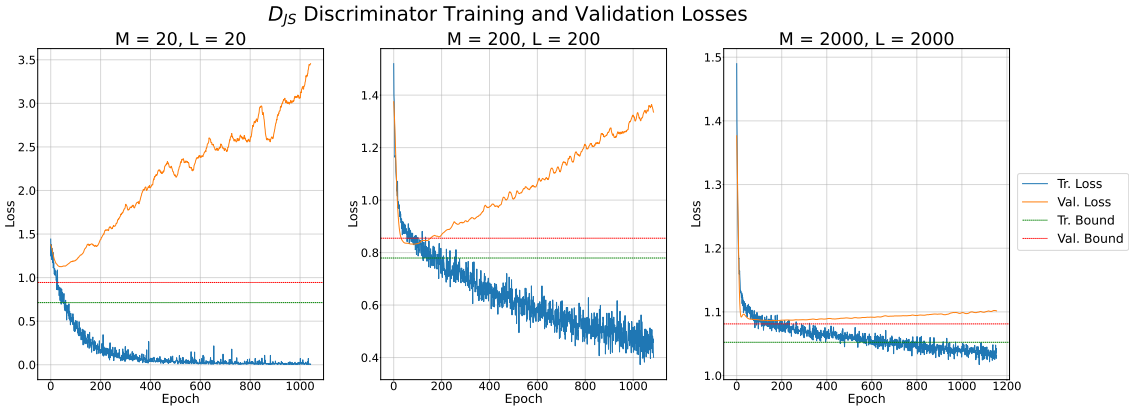


Figure 4.6: Discriminator loss curves for Experiment 1. The loss curves show a clear overfitting due to low sample sizes. Green and red dashed lines represent theoretical convergence values.

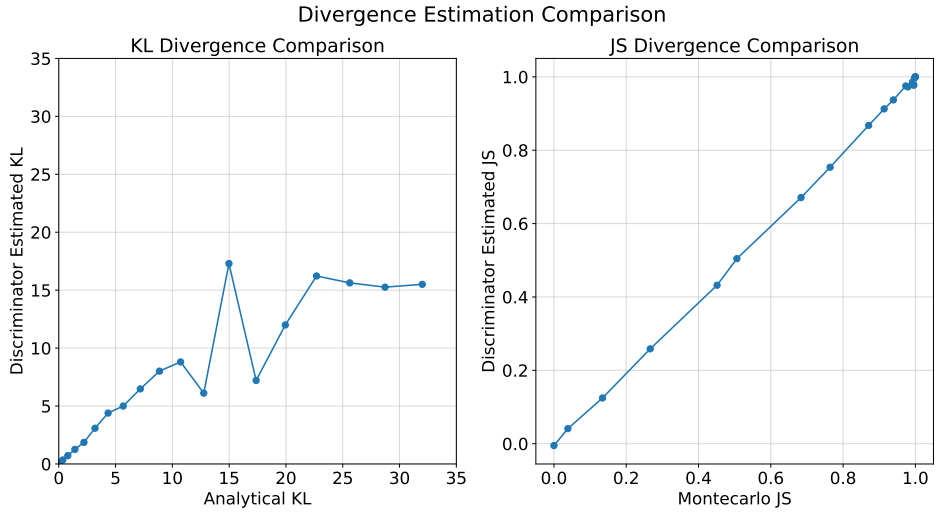


Figure 4.7: Relation between ground truth and estimated divergences as distribution separation increases in Experiment 1.

the relation for the D_{JS} divergence remains almost linear until it approaches 1. While the relation appears less linear for values near 1, it is essential to note that the primary application of this discriminator lies in comparing similar distributions. Therefore, the focus should be on the performance of low D_{JS} divergences. Meanwhile, the D_{KL} exhibits a nonlinear relation starting from the first comparisons, with the estimated D_{KL} deviating from the ground truth probably because of bias introduced by the discriminator.

→ *Experiment 2*

This experiment investigates the performance of the proposed divergence estimation method in Gaussian mixture distributions. These distributions offer a higher complexity level than those used in Experiment 1. Each Gaussian mixture distribution comprises two independent

isotropic Gaussian components with distinct mixing probabilities (weights assigned to each component within the mixture). Unlike the prior case, where analytical ground truth for divergence was obtainable, we rely solely on MC approximations as ground truth here and in the following experiment. Similar results were obtained for this experiment compared to the first.

M	L	MC Estimated $\boldsymbol{D}_{\mathbb{KL}}$	Discriminator Estimated $\boldsymbol{D}_{\mathbb{KL}}$
200	200	2.669 ± 0.118	2.686 ± 0.294
	2000	2.887 ± 0.045	3.037 ± 0.517
2000	200	2.710 ± 0.100	2.628 ± 0.357
	2000	2.850 ± 0.071	2.913 ± 0.129

(a) $D_{\mathbb{KL}}$

M	L	MC Estimated $\boldsymbol{D}_{\mathbb{JS}}$	Discriminator Estimated $\boldsymbol{D}_{\mathbb{JS}}$
200	200	0.392 ± 0.013	0.374 ± 0.020
	2000	0.428 ± 0.009	0.412 ± 0.009
2000	200	0.415 ± 0.020	0.412 ± 0.019
	2000	0.423 ± 0.014	0.420 ± 0.013

(b) $D_{\mathbb{JS}}$

Table 4.8: Impact of training and validating samples on divergence estimation for Experiment 2. MC and proposed method divergence estimation for various combinations of training and validation samples M and L. The results highlight a clear correlation between the sample number and the estimation error.

Table 4.8 focuses on a practical range for training sample sizes M and validation sample sizes L (both set to $200, 2000$) and present the estimated divergences obtained using our proposed method compared to MC estimates. The results demonstrate that, for both divergences, the proposed method achieves results comparable to the MC estimates. These findings further support the previously observed correlation between sample size and estimation accuracy. Additionally, we observe a narrowing of CIs with increasing sample sizes for both divergence measures. This trend indicates that the estimation becomes progressively more precise as the available data increases. Overall, the results of this experiment underscore the effectiveness of the proposed method in estimating divergences for more intricate distributions, such as Gaussian mixtures. The remaining analysis can be found in Appendix D.1.1.

$\rightarrow$ ***Experiment 3***

The following experiments introduce the concept of a generative process, where synthetic data are generated based on N samples from a particular data distribution. We investigate the influence of the quality of the GM in terms of the number of samples used to generate. This experiment uses a Gaussian mixture with two components as the real distribution. We employ a GMM as the GM. The trained GMM then generates synthetic data that serve as

an approximation of the real data. Finally, we estimate the divergence using our approach with sufficient samples to achieve a small CI and compare it to the MC simulation estimation. We analyze the behavior of the GMM under different training configurations, varying the number of training samples N. Since previous experiments demonstrated that a high number of. samples for both M and L achieve better divergence estimations, we have fixed $M = 2000$ and $L = 2000$ for this experiment. We compare the estimated divergence errors for this combination of training samples M and validation L used by the divergence estimator.

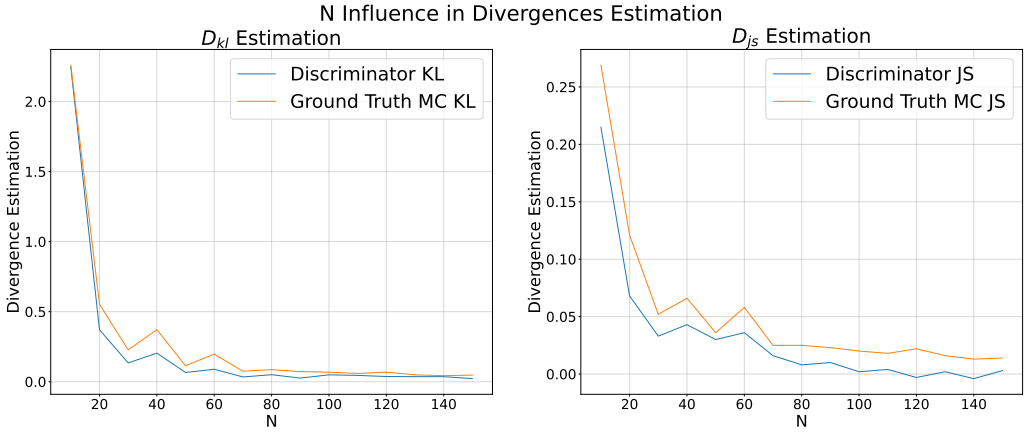


Figure 4.8: Estimated $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ compared to ground truth values for varying sample sizes of N in Experiment 3. Note the independent y-axes for each divergence measure due to their inherent scale differences, with $D_{\mathbb{JS}}$ exhibiting lower values than $D_{\mathbb{KL}}$.

Figure 4.8 shows the estimated divergences as N increases. The results demonstrate a clear correlation between the size of the training data for the GMM, N, and the estimated divergence errors. When the GMM is poorly trained due to a limited number of samples ($N = 10$), it cannot effectively capture the underlying data patterns. This leads to generated data that significantly deviate from the real distribution. Consequently, the real divergence values increase significantly, and a large estimation error occurs, particularly for unbounded $D_{\mathbb{KL}}$. $D_{\mathbb{JS}}$, which is inherently bounded between 0 and 1, exhibits less extreme error values. As the number of training samples for the GMM increases ($N = 200$ and $N = 2000$), the generated data become more representative of the real distribution. This results in lower divergence values and reduced estimation errors for $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$. This finding supports the validity of our proposed validation technique for generative processes involving Gaussian mixture distributions.

$\rightarrow$ ***Experiment 4***

This final experiment evaluates the performance of GMs in a more realistic setting by incorporating a real-world dataset and its synthetic counterpart. Unlike previous experiments that used purely synthetic distributions, this scenario assesses how well a generative model can replicate real-world data. The chosen dataset is Adult. As described in Appendix B, it contains information on $32,561$ individuals used to predict their annual income exceeding \$50,000. A subset of 10,000 samples obtained from [247] was used for this study. The dataset consists of 14 features, including categorical, binary, and integer values. Based on previous

experiments analyzing sample size influence, we set the hyperparameters at $N = 10,000$, $M = 7,500$, and $L = 1,000$.

To assess the GMs, we compared two state-of-the-art approaches: CTGAN [109] and VAE-BGM, the latter being our proposed method detailed in Section 4.2. VAE-BGM consistently outperformed other models in previous validation experiments. To further strengthen our evaluation, we computed MMD with an RBF kernel in addition to divergence measures. This complementary analysis helps determine whether the estimated divergences align with another widely used distributional comparison metric, reinforcing the robustness of our evaluation framework.

CTGAN $D_{\mathbb{KL}}$	**VAE** $D_{\mathbb{KL}}$	**CTGAN** $D_{\mathbb{JS}}$	**VAE** $D_{\mathbb{JS}}$	**CTGAN MMD**	**VAE MMD**
0.342 ± 0.016	0.185 ± 0.030	0.138 ± 0.002	0.099 ± 0.002	0.0013 ± 0.000	0.0001 ± 0.000

Table 4.9: Comparison of GMs for real-world data. Similarity comparison between real data and their synthetic counterparts generated by each model using estimated divergences, $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$, and computed MMD. Lower values in both metrics indicate a greater similarity between real and synthetic data.

Section 4.3.2 presents the results, confirming that VAE-BGM achieves the lowest values across all metrics. The reduced divergence values indicate a closer resemblance between the real and synthetic distributions, while the MMD results further validate these findings. Given that MMD assesses probability distribution similarity in a reproducing kernel Hilbert space, its consistently lower values for VAE-BGM strongly support its superior generative performance. The agreement between MMD and divergence estimates further reinforces our confidence in the evaluation methodology, suggesting that VAE-BGM generates more realistic synthetic data compared to CTGAN.

To explain why the VAE may outperform the CTGAN in our experiments, we refer to Section 4.2.3, which compares the VAE generator with CTGAN and TVAE using metrics similar to the ones used in this work. The results support our observations. We suggest that VAEs perform better due to their stability in managing complex data distributions, unlike GANs, which struggle with convergence. This stability in VAEs contributes to their enhanced ability to generate more realistic synthetic data, highlighting significant differences in model architecture and distribution handling. The inclusion of MMD as an additional validation metric further substantiates these findings by providing an alternative perspective on distributional resemblance.

This experiment demonstrates the practical applicability of the proposed evaluation framework, showing that divergence-based estimates are consistent with MMD in measuring similarity between real and synthetic data. These findings highlight the effectiveness of VAE-BGM in generating high-fidelity synthetic data, which is crucial for real-world applications where ensuring data quality and representativeness is essential.

4.3.3 Conclusions

This research proposes a novel and practical approach for validating synthetic tabular data generated by various models. The core of this method lies in using a divergence estimator based on a probabilistic classifier to capture the discrepancies between the real and synthetic data distributions. This approach overcomes the limitations associated with traditional marginal divergence comparisons by considering the joint distribution. While marginal comparisons assess the similarity of individual features between real and synthetic data, they can be misleading. Even if the marginal distributions of each feature appear similar, the joint distribution, which captures the relations between features, may differ significantly. This can lead to unrealistic synthetic data, where individual features appear plausible, but their co-occurrences do not represent the real data. By considering the joint distribution, our proposed method provides a more comprehensive assessment of the quality of synthetic tabular data.

The efficacy of the proposed method is comprehensively evaluated through a series of experiments with progressively increasing complexity. The initial phase establishes a solid foundation by analyzing the performance in controlled scenarios with well-defined theoretical distributions. The results demonstrate that the accuracy of divergence estimates is highly dependent on the amount of training data available for both the GM and the divergence estimator network. Subsequent experiments explore more intricate scenarios involving Gaussian mixture distributions and real-world datasets. The findings consistently support the effectiveness of the proposed method in approximating the true divergences. Our emphasis in this work has been to show the advantages of using divergences as synthetic data validation for tabular data.

This research offers significant contributions beyond the specific tabular data validation domain. The proposed methodology facilitates better validation practices for various fields dependent on GMs. Its key strength is capturing complex data distribution relations, leading to more robust and reliable validation processes. In addition to the positive results, the study also highlights the importance of the quality of the GM. When it is inadequately trained due to insufficient data, it can significantly impact the accuracy of the divergence estimation. This emphasizes the need to consider GM training procedures carefully to ensure reliable validation results.

Several promising avenues exist for future research. One direction involves exploring the potential to extend the proposed approach to more complex data structures, such as images or time series data. Additionally, investigating the integration of this validation technique within GM training pipelines could enable the development of self-improving GMs that can automatically adjust their parameters to generate data that closely resemble the target distribution. Moreover, addressing the impact of changing the assumption of prior probabilities for the source distributions to be compared, where the number of real and synthetic samples may vary, presents a significant challenge. Techniques to mitigate the effects of unbalanced classes in the regression framework could be explored. These techniques may include adaptive sampling strategies, class weighting, or methods specifically tailored to handle imbalanced data distributions. Implementing such approaches could enhance the robustness and applicability of the validation technique across diverse datasets and real-world applications. Finally,

developing a robust methodology to estimate divergences when limited samples are available presents a crucial challenge (addressed in Section 4.4). Addressing this challenge would further enhance the versatility and practicality of the proposed validation technique in real-world scenarios with restricted data availability. Furthermore, it would be beneficial to investigate alternative density ratio estimation techniques, such as those presented in [250], [251]. These methods may offer advantages regarding sample size requirements, estimation error, or other relevant estimator properties.

4.4 A Novel Synthetic Data Generation Methodology to Address Data Scarcity

One of the significant challenges in SDG is ensuring the quality and representativeness of the generated data when working with limited real data. Data scarcity is a prevalent issue, particularly in healthcare applications, where acquiring large, high-quality datasets is often constrained by privacy regulations, collection costs, and the rarity of certain medical conditions. Traditional generative models, such as GANs or VAEs, may struggle to produce reliable synthetic data in low-data regimes, leading to poor generalization, overfitting, or under-representative samples.

This section proposes a novel methodology tailored to address data scarcity in SDG. Our approach focuses on enhancing the generative process when training data are limited, ensuring that the synthetic data maintains sufficient statistical fidelity and practical utility. By leveraging strategies such as transfer learning and meta-learning techniques, as well as the flexibility of VAEs, our proposed method aims to overcome the challenges posed by low-data scenarios and generate synthetic data that accurately reflect the underlying real distributions.

4.4.1 Generation Methodology

Assuming that we have a tabular dataset composed of N entries $\{x_r^i\}_{i=1}^N$, where N represents the number of samples available and each entry x_r^i has a dimensionality of Cov features. In other words, Cov represents the number of attributes associated with each data point. A Deep Generative Model (DGM) can be defined as a high-dimensional probability distribution p_θ, where θ represents the learnable parameters of the model. The objective of the DGM is to learn a representation, p_θ, that closely approximates the true underlying data distribution, denoted by $p(x_r)$. Once trained, the DGM can generate new synthetic generated samples x_g by drawing from its learned distribution:

$$x_g \sim p_\theta. \tag{4.33}$$

Ideally, a well-trained DGM should produce synthetic data x_g that are statistically indistinguishable from real data x_r.

In the prevalent big data setting, characterized by many training samples ($N \gg Cov$), DGMs with sufficient complexity can effectively capture the underlying data distribution $p(x_r)$. This is evidenced by the impressive results achieved in recent research, where high-dimensional synthetic samples are generated using vast amounts of training data [109], [252]–[254]. However, for scenarios with limited training data, which is common in tabular domains, DGMs struggle. Consequently, the synthetic samples generated x_g deviate significantly from the true data distribution $p(x_r)$, leading to high $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ between real and synthetic data.

We propose an approach that uses artificially generated inductive biases to address this challenge. Figure 4.9 illustrates the overall architecture. In the standard big data setting, a DGM p_θ is directly trained using real data x_r generating high-quality synthetic data $x_g \sim p_\theta$. However, when the number of real samples N is limited, the quality of the generated data x_g deteriorates. To mitigate this issue, we introduce an artificial inductive bias generator. This

module inputs the initial synthetic data x_g and outputs an initial set of weights θ_0. These weights are then used as the inductive bias to train a second DGM $p_{\hat{\theta}}$ using real data x_r. This second DGM generates a new set of synthetic samples, $\hat{x}_g$. Notably, the only distinction between p_θ and $p_{\hat{\theta}}$ lies in the initial weights: $p_{\hat{\theta}}$ leverages the inductive bias encoded in θ_0 to potentially achieve faster convergence to a distribution that better resembles $p(x_r)$. At the same time, p_θ begins training with random weights. As our simulations will demonstrate, this seemingly minor difference translates into significant improvements in the quality of the generated synthetic data.

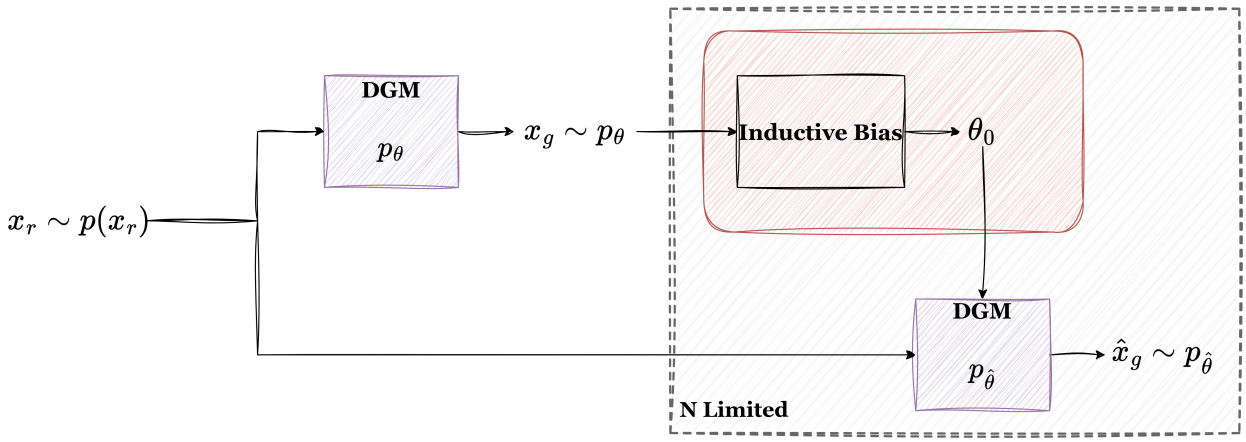


Figure 4.9: Block diagram for the proposed architecture for the SDG methodology. In a standard big data setting, the first GM p_θ generates good enough samples. However, in cases where large amounts of data are not available (N is limited), we propose to use the data generated by the first DGM as input to an artificial inductive bias generator, which in return provides a set of initial weights θ_0 for a DGM that contains the artificially generated inductive bias. This initial weight is then used to train a second DGM $p_{\hat{\theta}}$ using real data x_r, generating a second set of synthetic data $\hat{x}_g$, which is of higher quality than the synthetic data x_g.

The proposed approach (depicted in Figure 4.9) hinges on two key concepts: the importance of inductive biases and the feasibility of their artificial generation. The importance of inductive biases in supervised learning is well established. The no-free-lunch theorems state that a universally optimal learner does not exist. Consequently, specific learning biases can produce substantial performance gains for particular problem domains (see [255] and the references therein). CNNs exemplify this principle. Their inherent inductive bias, the fact that the image information possesses spatial correlation, makes them the preferred architecture for image processing tasks. Similarly, as highlighted in [256], using inductive biases is a cornerstone of the success of DL. In scenarios with limited training data, regularizers are commonly employed as inductive biases to prevent overfitting. This underscores the dual role of inductive biases: not only do they contribute to the effectiveness of Deep Learning, but they are also crucial in avoiding overfitting. However, effective use of inductive biases is often contingent on having specific knowledge about the problem. In the aforementioned example of CNNs, we inherently understand spatial correlation in images. However, in tabular data, this domain-specific knowledge is often scarce. Recent efforts have focused on designing large models trained

on artificially generated data as inductive biases to address this challenge. The underlying hope is that the actual problem to be solved exhibits similarities to those encountered during training of the large model (e.g., [257] and [258]).

Therefore, our proposed approach departs from existing methods for incorporating inductive biases in STDG. Unlike the brute-force approach employed in [258], we use data generated by a potentially low-quality DGM p_θ. This strategy aims to obtain an initial set of weights θ_0, which act as an inductive bias. Ideally, these weights should guide the model to a region of the parameter space that facilitates convergence towards a high-quality solution. In particular, we assess our ideas using a state-of-the-art VAE architecture for the DGM, although we hypothesize that similar results could be achieved with other DGM architectures. Due to its demonstrated superiority against other leading models, we will use the VAE-BGM (Section 4.2). VAEs are known to be sensitive to the initial random conditions (seeds) used during training. This dependence on seeds requires training with multiple seeds and selecting the one(s) that exhibit the best performance based on a chosen metric, such as minimum validation loss. The remaining runs, often discarded, may still contain valuable problem-specific information despite not achieving optimal solutions using traditional metrics. Our key idea lies in exploiting these potentially informative data from discarded VAE runs to create an artificial inductive bias for the final DGM trained with real data.

The following subsections explore two distinct paradigms for generating the initial set of weights, θ_0: transfer learning and meta-learning. Transfer learning techniques encompass pre-training and model averaging, while meta-learning techniques include Model-Agnostic Meta-Learning (MAML) and Domain Randomized Search (DRS). Pre-training offers a versatile approach applicable to any DGM architecture, regardless of inherent characteristics. In contrast, model averaging and meta-learning techniques are particularly well suited for VAEs trained with multiple seeds due to their intrinsic variability in learned representations. Consequently, we will evaluate the two methods within the chosen VAE architecture. Additionally, to assess the efficacy of pre-training across different DGM architectures, we will compare its performance on the CTGAN.

Transfer learning

Transfer learning is an ML paradigm that leverages knowledge acquired from a context domain (also called the source domain) to enhance learning performance in a new target domain [259]. This approach aims to improve the learning process in the target domain by capitalizing on the knowledge gained from solving related tasks in the context domain. This technique has demonstrated its efficacy in fields where data scarcity is a common challenge, such as the medical field [260].

Formally, based on the definition in [259], we can define a domain $\mathcal{D}$ by a feature space $\mathcal{X}$ and a marginal probability distribution $p(x_r)$. Two domains are considered distinct if their feature spaces $\mathcal{X}_1, \mathcal{X}_2$ or marginal probability distributions $p(x_1), p(x_2)$ differ, i.e., if $\mathcal{X}_1 \neq \mathcal{X}_2$ or $p(x_1) \neq p(x_2)$. The core objective of transfer learning is to leverage the knowledge learned in a context domain $\mathcal{D}_{context}$ to improve learning in a target domain $\mathcal{D}_{target}$. This is typically achieved when the context and target domains differ, i.e., $\mathcal{D}_{context} \neq \mathcal{D}_{target}$.

Our work focuses on a scenario where the context domain $\mathcal{D}_{context}$ consists of data x_g generated by a DGM. On the other hand, the target domain $\mathcal{D}_{target}$ consists of x_r. Our approach leverages the representational power learned by the DGM p_θ on x_g to provide a strong starting point for learning in the target domain with real data x_r. This knowledge transfer is achieved by initializing the model weights for the target domain with the weights learned from the DGM model trained on the generated data.

Transfer learning can be categorized into homogeneous and heterogeneous settings based on the feature spaces of the domains [261]. Homogeneous transfer learning applies when the context and target domains share the same feature space $\mathcal{X}_1 = \mathcal{X}_2$, while heterogeneous transfer learning deals with scenarios where feature spaces differ $\mathcal{X}_1 \neq \mathcal{X}_2$. This work focuses on homogeneous transfer learning, where the context domain is an augmented version of the target domain. The key difference between the domains in our case lies in the number of samples, leading to situations where the empirical distributions of the data differ, i.e., $p_\theta(x_g) \neq p(x_r)$.

Within homogeneous transfer learning, various methodologies exist to improve target task performance by capitalizing on knowledge from a related source domain. These techniques encompass instance-based [97], relational knowledge transfer [262], feature-based [263], and, as employed in this work, parameter-based [264] transfer through shared model parameters or hyperparameter distributions. This study leverages a two-stage parameter-based transfer learning approach. The first stage involves pre-training or model averaging, followed by fine-tuning in the second stage. Subsequent sections will delve deeper into both pre-training and model averaging techniques. Upon completion of one of these initial phases, fine-tuning serves to refine the model parameters, ultimately achieving optimal adaptation for the target domain.

$\rightarrow$ ***Pre-training***

Pre-training is a frequently adopted strategy for introducing an inductive bias into a model. By leveraging a pre-trained model on a context domain, the target model gains generalizable features that enhance its performance on a target domain. However, while pre-training is a standard in computer vision and natural language processing, achieving similar success with tabular data remains challenging. This disparity arises from the inherent heterogeneity of the features of the tables, which creates substantial feature space shifts between pre-training and downstream datasets, hindering effective knowledge transfer. Despite these challenges, recent efforts like [265] and [266] explore tabular transfer learning with promising results. Although these studies demonstrate potential, achieving comprehensive parameter transfer in tabular data requires further research to establish best practices and unlock the full potential of pre-training in this domain.

In this work, pre-training involves the following steps. First, we train a separate DGM $p_{\theta_{pt}}$ using synthetic data x_g as training data. Since x_g is sampled from the initial DGM, $x_g \sim p_\theta$, we can generate a vast amount of synthetic data. This abundance circumvents the limitations associated with training in small datasets, such as overfitting. Then, the optimal weights θ^*_{pt} from DGM $p_{\theta_{pt}}$ are used as initial weights θ_0 to train the GM $p_{\hat{\theta}}$ (see Figure 4.9).

In essence, our approach aligns with the well-established concept of data augmentation. We generate synthetic data x_g, which may not perfectly capture the intricacies of the original data

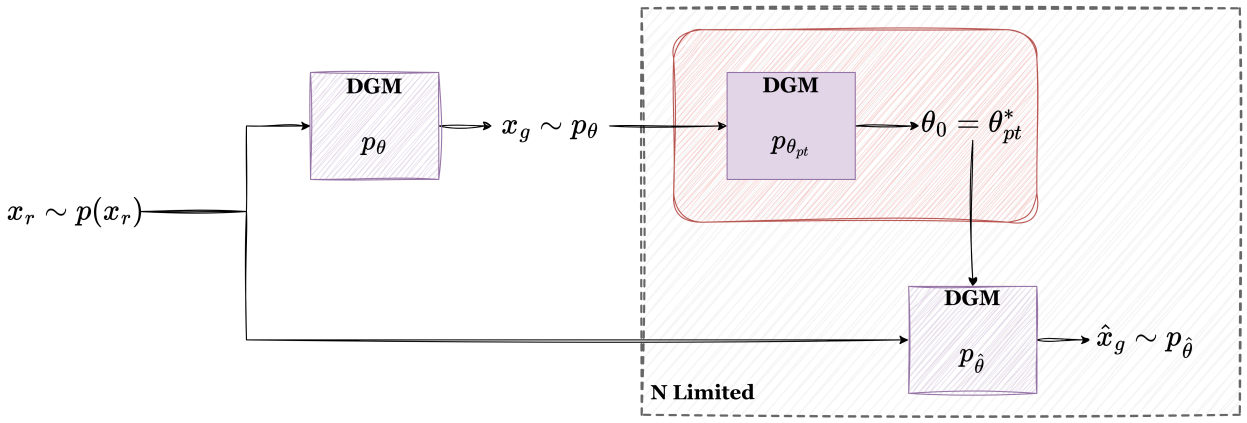


Figure 4.10: Block diagram for the pre-training case. The inductive bias is introduced by training a DGM $p_{\theta_{pt}}$ on a large collection of x_g samples. The weights learned from this training process with abundant samples serve as the initial parameters $\hat{\theta}$ for the fine-tunning process using the real data x_r to obtain $p_{\hat{\theta}}$.

x_r. However, we used these synthetic data to train another DGM $p_{\theta_{pt}}$. Although $p_{\theta_{pt}}$ might generate lower-quality synthetic samples, our objective is to exploit the information encoded within this DGM to establish an initial set of weights for the DGM that will eventually be trained on x_r. In other words, we exploit the knowledge of the GM $p_{\theta_{pt}}$, the context domain, to obtain a better GM $p_{\hat{\theta}}$, which is our target domain. Figure 4.10 visually represents this pre-training procedure.

$\rightarrow$ ***Model Averaging***

The concept of model averaging emerged in the 1960s, primarily within the field of economics [267], [268]. Traditional empirical research often selects a single 'best' model after searching a wide space of possibilities. However, this approach can underestimate the real uncertainty, leading to overly confident conclusions. Model averaging offers a compelling alternative. By combining multiple models, the resulting ensemble can outperform any individual model. This approach aligns with the core principles of statistical modeling: maximizing information use and balancing flexibility with overfitting. In essence, model averaging extends the concept of model selection by leveraging insights from all the models considered.

While pre-training can be incorporated with any DGM, our approach focuses on models where the training process is sensitive to initial conditions, such as VAEs. In such cases, it is common to train the DGM p_θ with multiple initial conditions (seeds) and potentially discard 'bad' seeds based on a specific metric. We propose using these discarded seeds to create an artificial inductive bias. The simplest implementation involves averaging the model parameters. In this case, our context domains are the different results of each seed, and the target domain is obtained by averaging across the context domains. If we train S different seeds for p_θ, resulting in S models with parameters θ_s, we propose using the average of these weights as the inductive bias:

$$\theta_0 = \frac{1}{S}\sum_{s=1}^{S}\theta_s \tag{4.34}$$

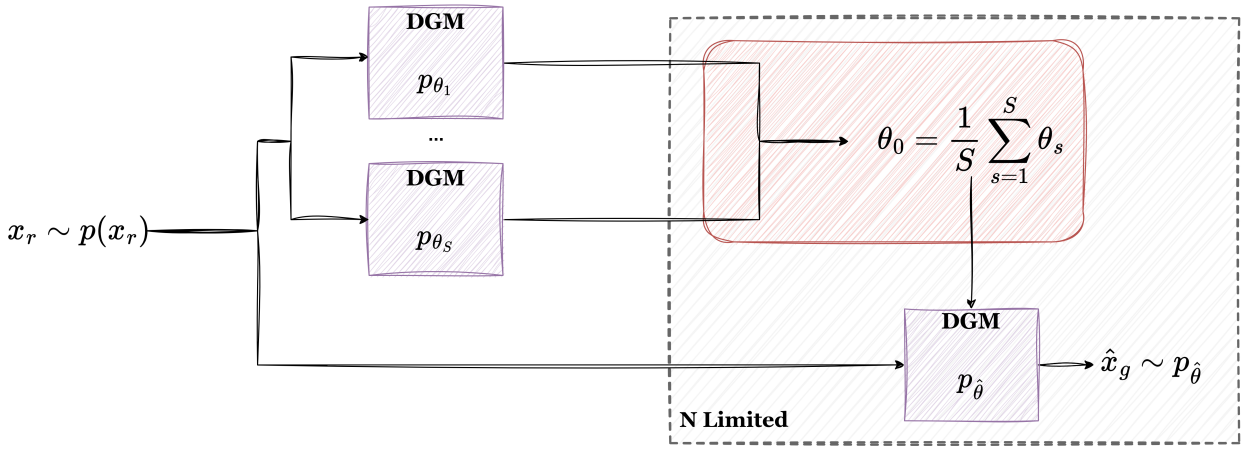


Figure 4.11: Block diagram for the model averaging case. The inductive bias is introduced by training a DGM p_θ on x_r using S different seeds. The average of the weights learned from these training processes serve as the initial parameters $\hat{\theta}$ for the fine-tunning process using the real data x_r to obtain $p_{\hat{\theta}}$.

This straightforward approach is computationally efficient, requiring only calculating the average across the precomputed weights. It assumes that the average model may capture a robust inductive bias, leading to improved performance. Figure 4.11 summarizes this process.

Meta-learning

Traditional ML models often rely on large volumes of data to achieve optimal performance in specific tasks. In contrast, meta-learning introduces a distinct paradigm by training algorithms that can 'learn to learn' [269], enabling them to adapt to new tasks with minimal data rapidly. This departure from the conventional requirement of extensive datasets for each new task allows meta-learning algorithms to leverage knowledge by addressing numerous related tasks. Through reflective analysis of past experiences, these models dynamically adjust their learning strategies when confronted with novel situations, making them more efficient learners and requiring less data to perform well on tasks with similar characteristics.

In this work, we exploit the multi-seed training configuration of certain DGMs. We construct a meta-learning framework by treating each S different seeds obtained after training the DGM as a distinct task.

$\rightarrow$ *MAML*

MAML is a prevalent approach within the field of meta-learning [270]. It identifies the initial set of weights denoted by θ_{MAML} by leveraging various tasks, enabling rapid and data-efficient adaptation to new tasks. This efficiency comes from fine-tuning θ_{MAML} with minimal data for each new task. However, the successful application of MAML requires access to diverse tasks for effective learning.

We can frame the problem by starting with a common single-task learning scenario and transforming it into a meta-learning framework. Consider a task $\mathcal{T}$ that consists of an input

x sampled from a probability distribution $\mathcal{D}$. For simplicity, we define a task instance $\mathcal{T}$ as a tuple comprising a dataset $\mathcal{D}$ and its corresponding loss function $\mathcal{L}$. To solve the task $\mathcal{T}$, we need to obtain an optimal model parameterized by a task-specific parameter ω^*, which minimizes a loss function $\mathcal{L}$ on the data of the task as follows:

$$\omega^* = \arg\min_{\omega} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathcal{L}(\mathcal{D};\omega)\Big]. \tag{4.35}$$

In single-task learning, hyperparameter optimization is achieved by splitting the dataset $\mathcal{D}$ into two disjoint subsets $\mathcal{D} = \mathcal{D}^{(t)} \cup \mathcal{D}^{(v)}$, which are the training and validation sets, respectively. The meta-learning setting aims to develop a general-purpose learning algorithm that excels across a distribution of tasks represented by $p(\mathcal{T})$ [271]. The objective is to use training tasks to train a meta-learning model θ_{MAML} that can be fine-tuned to obtain ω to perform well on unseen tasks sampled from the same task environment $p(\mathcal{T})$. Meta-learning methods utilize meta-parameters to model the common latent structure of the task distribution $p(\mathcal{T})$. Therefore, we consider meta-learning an extension of hyperparameter optimization, where the hyperparameter of interest – often called a meta-parameter – is shared across many tasks.

In this work, the distribution of tasks is defined by the set of S training seeds obtained after training the DGM. Given a set of S training seeds following $p(\mathcal{T})$, each task $\mathcal{T} \sim p(\mathcal{T})$ is therefore formalized as $\mathcal{T} = \{\mathcal{D}, \mathcal{L}\}$. Each dataset $\mathcal{D}$ consists of synthetic data points x_g^s drawn from the model for the different training seeds. The loss function $\mathcal{L}$ corresponds to the DGM loss function. The specific $\mathcal{L}$ form depends on the chosen DGM. If the chosen DGM is a VAE, the loss function $\mathcal{L}$ would be the negative of the ELBO [10]. In contrast, if a GAN is used, the loss function $\mathcal{L}$ would be the minimax loss function arising from the interplay between the generator and discriminator networks [32]. It is important to note that both VAEs and GANs use two neural networks within their architecture, different from the single network architectures commonly found in state-of-the-art applications [272], [273].

Solving this problem using the MAML approach requires access to B tasks sampled from $p(\mathcal{T})$. We denote this set of tasks $\mathcal{T}_b$ used for training as $\mathcal{D}_b = \{(\mathcal{D}_b^{(t)}, \mathcal{D}_b^{(v)})\}_{b=1}^{B}$, where each task b has dedicated meta-training and meta-validation data, respectively. The goal of meta-training is to find the optimal ω_b^* for a given task b given θ_{MAML}. This θ_{MAML} essentially captures the ability to learn effectively from new data. In this context, the task-related parameter ω_b denotes the parameters of the two networks comprising the VAE, i.e., the task-specific parameters of the encoder and decoder. After meta-training, the learned ω_b^* is used to guide the training of a base model θ_{MAML}. This procedure is called meta-testing. This essentially means that the model leverages the knowledge gained from previous tasks to improve the efficiency of learning on new tasks. This can be viewed as a bi-level optimization problem [274]:

$$\begin{aligned} &\min_{\theta_{MAML}} \mathop{\mathbb{E}}_{\mathcal{T}_b\sim p(\mathcal{T})} \left[\mathop{\mathbb{E}}_{x_{g_b}^{(v)}\sim\mathcal{D}_b^{(v)}} \Big[\mathcal{L}_b(\mathcal{D}_b^{(v)}; \omega_b^*(\theta_{MAML}))\Big] \right] \\ &\text{s.t: } \omega_b^*(\theta_{MAML}) = \arg\min_{\omega_b} \mathop{\mathbb{E}}_{x_{g_b}^{(t)}\sim\mathcal{D}_b^{(t)}} \Big[\mathcal{L}_b(\mathcal{D}_b^{(t)}; \omega_b(\theta_{MAML}))\Big]. \end{aligned} \tag{4.36}$$

This equation minimizes the expected loss across all tasks on the meta-validation sets, subject

to the constraint that the task-specific parameter ω is optimized on the corresponding meta-training data for each task.

Since in our work, we are upgrading the parameters using gradient descent, we can reformulate Equation (4.36) as follows:

$$\omega_b \leftarrow \theta - \alpha \nabla_{\omega_b} \mathcal{L}_b(\mathcal{D}_b^{(t)}; \omega_b) \tag{4.37}$$

$$\theta_{MAML} \leftarrow \theta_{MAML} - \gamma \nabla_{\theta_{MAML}} \sum_{b=1}^{B} \mathcal{L}_b(\mathcal{D}_b^{(v)}; \theta_{MAML}). \tag{4.38}$$

Here, α and γ represent the learning rates for the inner and outer loops, respectively. The inner loop updates the task-specific parameters ω for each task b using the gradient of the loss function $\mathcal{L}_b$ in the meta-training data. The outer loop updates the meta-parameters θ_{MAML} based on the accumulated meta-validation loss across all tasks.

Figure 4.12 illustrates the integration of the MAML procedure within the framework of our proposed methodology. In this context, the task space denoted by $p(\mathcal{T})$ corresponds to the various seeds S obtained during the training process. Essentially, the task space encompasses the different probability distributions p_{θ_s} associated with each training seed. Ultimately, the meta-training steps lead to identifying the desired parameters, denoted by θ_{MAML}. Note that θ_{MAML} represents a set of parameters that adapt fast to new data; in our case, the DGM initial parameters are chosen so that they adapt fast to generate real data.

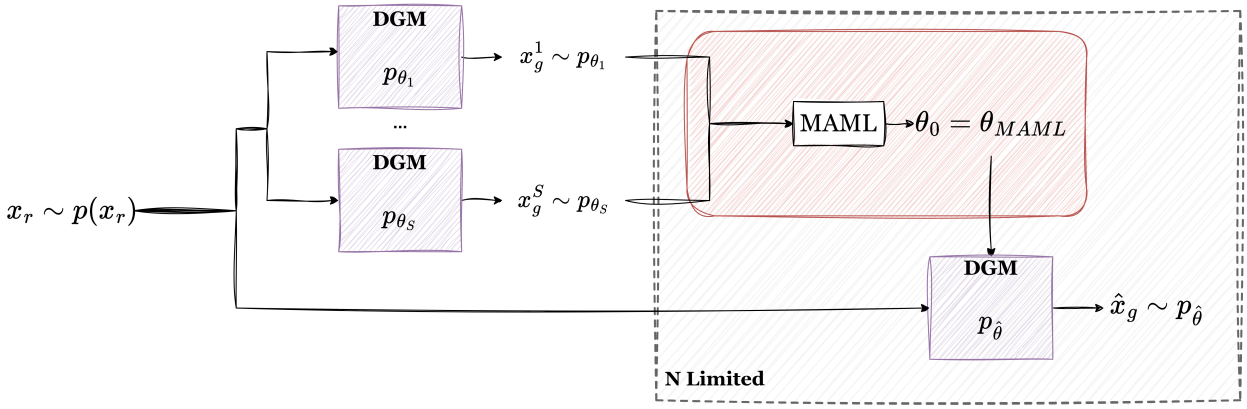


Figure 4.12: Block diagram for the MAML case. The inductive bias is introduced by training a DGM p_θ on x_r using S different seeds. The synthetic dataset x_g^s generated by each seed serves as a task for MAML. The starting point to fine-tune using the real data x_r and obtain $p_{\hat{\theta}}$ is the MAML solution obtained, θ_{MAML}.

$\rightarrow$ *DRS*

Although MAML offers the potential to leverage the underlying structure of learning problems through a powerful optimization framework, it introduces a significant computational cost. Therefore, while we should seek a trade-off between accuracy and computational efficiency, there is no management approach. An understanding of the domain-specific characteristics inherent to the meta-problem itself is needed.

DRS presents an alternative meta-learning approach that circumvents the computational burden of bilevel optimization problems. Unlike MAML, DRS trains a model on the combined data from all tasks. This eliminates the need for the complex optimization procedures present in MAML, leading to a more computationally efficient solution. However, it is important to acknowledge that DRS offers an approximation to the ideal solution [275].

Formally, DRS focuses on the meta-information, denoted by θ_{meta}, as the initialization of an iterative optimizer used in a new meta-testing task, $\mathcal{T}_S$. In this context of meta-learning initialization, a straightforward alternative involves solving the following pseudo-meta problem:

$$\theta_{DRS} = \arg\min_{\omega} \mathop{\mathbb{E}}_{\mathcal{T}_S \sim p(\mathcal{T})} \mathcal{L}(\mathcal{D}^*; \omega). \tag{4.39}$$

In this context, $\mathcal{D}^*$ represents the aggregated synthetic data collection, x_g, obtained across all training seeds S. We refer to this approach as Domain-Randomized Search due to its alignment with the domain randomization method presented in [276] and its core principle of directly searching over a distribution of domains (tasks).

Figure 4.13 shows the application of the DRS procedure within the framework of our proposed methodology. we aim to identify θ_0.

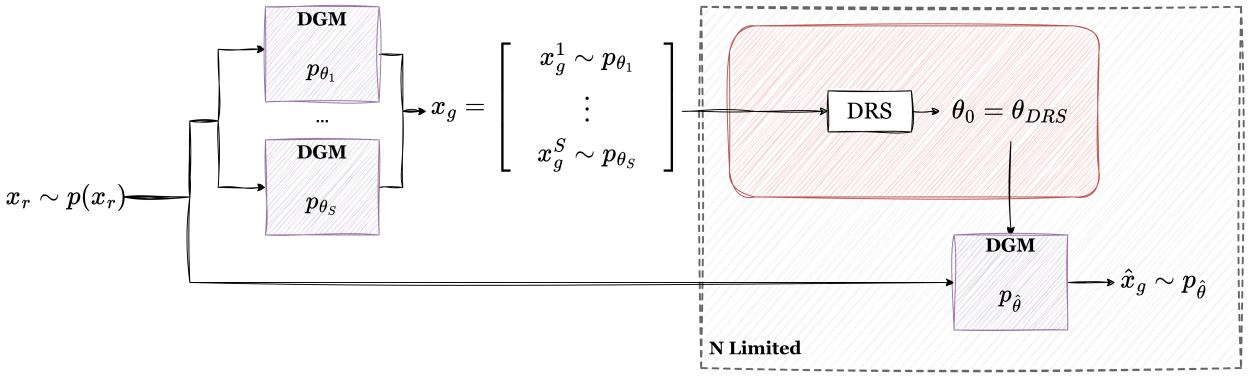


Figure 4.13: Block diagram for the DRS case. The inductive bias is introduced by training a DGM p_θ on x_r using S different seeds. The synthetic dataset x_g contains data generated by each seed and serves as input to DRS. The starting point to fine-tune using the real data x_r and obtain $p_{\hat{\theta}}$ is the DRS solution obtained, θ_{DRS}.

Both MAML and DRS offer complementary approaches with a trade-off between modeling complexity and optimization cost [275]. DRS delivers an approximate solution with lower computational demands, while MAML offers higher precision at the expense of greater computational resources. DRS is also advantageous when dealing with a limited number of learning tasks. In our case, where data generated by each seed ($s = 1, 2, ..., S$) is considered a task, and S typically takes values around 10, DRS is expected to provide better solutions than MAML, aligning with the findings of [275]. Finally, note that DRS is similar to the pre-training approach. While both techniques aim to improve model performance, they utilize data differently. Pre-training leverages data from the best VAE seed, whereas DRS capitalizes on data from all VAE seeds. This distinction reflects the core principle of DRS: exploring a

wider range of possibilities by searching across a distribution of domains (tasks) represented by the various seed variations.

4.4.2 Generation Methodology Applied to General-Purpose Datasets

Validation Metrics

To rigorously evaluate the effectiveness of our proposed method in capturing the real data distribution, we strictly adhere to the validation approach described in Section 4.3. This approach leverages a probabilistic classifier (discriminator) to estimate the ratio of probability densities between the real and synthetic distributions, subsequently calculating $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$. Traditional validation methods often focus on individual data points and the marginal distribution of each separate feature. In contrast, this approach considers the entire data distribution, including complex relationships between features. Additionally, divergences are robust to noise and offer clear interpretations, making them ideal for evaluating the effectiveness of the DGM. This provides a comprehensive approach to measuring the discrepancy between two probability distributions, making them suitable for assessing the similarity between real data $p(x_r)$ and the distribution of the synthetic data generated by the DGM p_θ.

The discriminator network plays a crucial role in the validation process. This neural network architecture is trained to distinguish between real and synthetic data samples. The network receives two sets of samples as input. The first consists of M samples from the real data distribution $p(x_r)$ labeled as class 1. The second consists of M samples labeled as class 0 from the synthetic data distribution generated by the DGM p_θ or $p_{\hat{\theta}}$, depending on the N number of samples in the dataset. During training, the discriminator aims to learn a decision boundary that effectively separates these two sets of samples. This process forces the discriminator to capture the underlying differences between the real and synthetic distributions. Once the discriminator network is trained, it is used to estimate $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ between the real and synthetic probability distributions. This estimation involves using L samples from each distribution and feeding them to the trained discriminator. The output probabilities of the discriminator for these samples are then used to compute the divergence metrics.

Figure 4.14 illustrates the overall scheme of the approach. The figure emphasizes the separate but related components: the inductive bias generator for synthetic data creation and the validation process with the discriminator network. It also highlights the number of samples used from each distribution for each process step (N to generate samples, M to train the discriminator, and L to estimate divergence). By adhering to this rigorous validation approach, we ensure that our proposed methodology for generating synthetic data from small datasets is thoroughly evaluated and its effectiveness is quantitatively demonstrated using established metrics like $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$. Note that M and L need to be large enough to prevent inaccurate divergence estimations, so we consider this in our experiments.

Beyond divergence estimation for validation, we further assess distributional similarity using MMD with an RBF kernel. Unlike divergence metrics that rely on probability density estimations, MMD operates in an RKHS to directly compare distributions through kernel-based representations. The RBF kernel is particularly suited for this task, as it effectively captures complex data structures while ensuring smooth similarity estimations. By incorporating

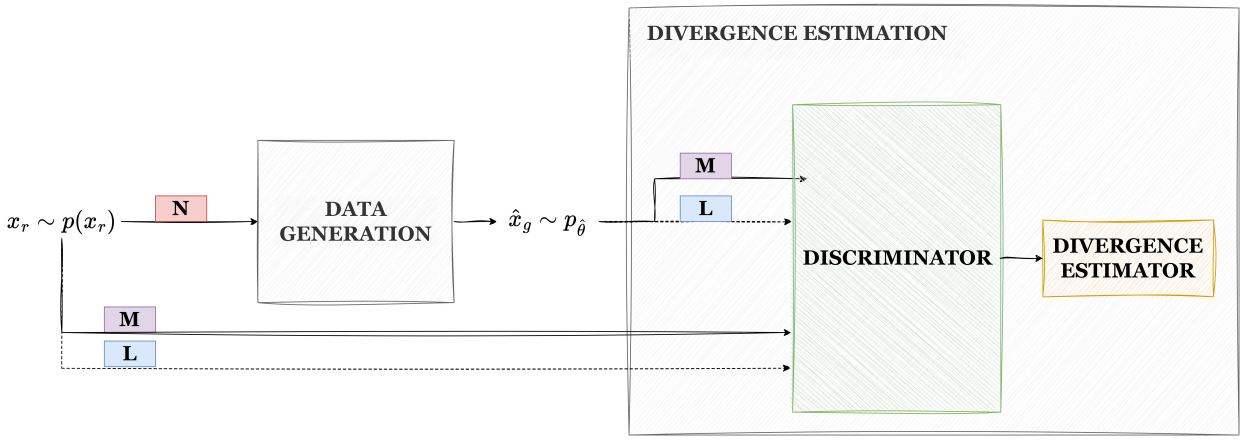


Figure 4.14: General Scheme of the proposed SDG methodology approach. Overall scheme of the approach and its validation process. This last one consists of a discriminator and a divergence estimator. The number of samples used from each distribution for each step is also highlighted: N to generate samples, M to train the discriminator, and L to estimate the divergences.

MMD alongside $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$, we provide an additional independent measure of resemblance between real and synthetic data, reinforcing the robustness of our validation framework.

Implementation Details

In terms of the experimental design to evaluate the proposed method, and as described in the methodology section, we used the VAE-BGM architecture to train ten different seeds. Subsequently, we applied methodologies based on transfer learning and meta-learning. For pre-training, we also included results using another state-of-the-art model, CTGAN. It is important to note that while we maintained the default parameters for CTGAN, we adjusted the dimensionality of the latent space in the VAE depending on the dataset. This allows the VAE to capture the specific characteristics of each dataset more effectively. We used a latent space dimension of 10 for the Adult and Intrusion datasets, 20 for the News dataset, and 15 for the King dataset. We maintained a consistent hidden size of 256 neurons for all VAE models. Regarding the configuration of parameters M and L, we defined two different validation configurations for each methodology: a reliable case and a more realistic case. The parameter N remains unchanged, reflecting the actual number of samples available to train the DGM, which is beyond our control. However, we can vary the number of generated samples for validation, i.e., M and L. Specifically, the results presented for each dataset are as follows.

- **'Big data':** First, we present the optimistic scenario with a sufficient N of $10,000$ samples, where no methodology is needed to calculate the inductive bias. This provides results of the divergences that serve as an 'upper bound' or reference for the best possible outcome. For M and L, we maintain high values of the validation samples, 7500 and 1000, respectively.
- **'Low data':** Next, we show results for a realistic scenario with few samples ($N = 300$)

without applying our methodology. This allows us to quantify the gain using the method and determine its benefits. We use two configurations for parameters M and L: $[M = 7500, L = 1000]$ and $[M = 100, L = 100]$. The second configuration is more realistic for few-data scenarios. When limited training data are available, there is also a limitation on the amount of data that can be effectively used for validation. The first configuration, with much larger values for M and L, serves as a rigorous evaluation of the impact of our methodology. However, it is acknowledged that using small values for M and L can lead to unreliable metric estimations.

- **'Pre-train':** In this case, we apply the proposed methodology using the pre-training technique. Results are presented for both CTGAN and VAE. The parameter configurations chosen are: $[N = 300, M = 7500, L = 1000]$ and $[N = 300, M = 100, L = 100]$.
- **'AVG,' 'MAML,' 'DRS':** These scenarios apply the model averaging ('AVG') and the meta-learning techniques ('MAML,' 'DRS'). We solely utilize the VAE architecture for multiple training runs. The following parameter configurations will be presented: $[N = 300, M = 7500, L = 1000]$ and $[N = 300, M = 100, L = 100]$.

This setup aims to thoroughly evaluate the performance of the proposed methodologies and robustness in various data availability scenarios and parameter configurations.

Experiments and Results

In this section, we present the results obtained from the experiments, focusing on scenarios that promote the validation of reliable synthetic data characterized by higher values of M and L. The experiments were carried out on four public datasets (Adult, News, King and Intrusio) obtained from the SDV environment [247], which also implements various data generation models, including the CTGAN implementation we use. Refer to Appendix B for a more detailed data description. The experiment design prioritized datasets with a sufficient number of samples. This allows us to create multiple data splits for various training and validation parameters¡ configurations. This approach comprehensively evaluates the proposed method under different parameter settings. The results for scenarios with $M = 100$ and $L = 100$, considered unreliable due to their low information content, are presented in Appendix D.2.1 for comparison purposes. For each database, we present a table summarizing the scenarios defined previously defined scenarios and their respective $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ values. The results for each metric are displayed in the following format: mean (standard deviation) (lower is better). The code to replicate our results and the data used can be found in https://github.com/Patricia-A-Apellaniz/low_sample_data_generator.

The Adult subtable in Table 4.10 shows the validation results regarding divergence obtained for the Adult dataset. We focus primarily on the $D_{\mathbb{JS}}$ divergence due to its interpretability as a bounded metric (ranging from 0 to 1). The table shows the upper and lower bounds used to assess the efficacy of the proposed methodology in the reliable case of higher validation samples ($M = 7500$ and $L = 1000$). These bounds are 0.079 (upper) and 0.331 (lower), highlighting a significant gap and room for improvement in the base VAE model (without any techniques applied). A consistent decrease in divergence is observed by examining the

Scenario	N	VAE $D_{\mathbb{JS}}$	CTGAN $D_{\mathbb{JS}}$	VAE $D_{\mathbb{KL}}$	CTGAN $D_{\mathbb{KL}}$	VAE MMD	CTGAN MMD
Big data	10000	0.079 (0.001)	0.150 (0.002)	0.153 (0.019)	0.420 (0.025)	0.0007 (0.0002)	0.0047 (0.0003)
Low data	300	0.331 (0.004)	0.563 (0.002)	0.697 (0.018)	1.653 (0.015)	0.0032 (0.0004)	0.0148 (0.0007)
Pre-train	300	**0.171 (0.004)**	0.563 (0.002)	**0.427 (0.021)**	1.753 (0.040)	**0.0013 (0.0003)**	0.0174 (0.0007)
AVG	300	**0.157 (0.004)**	N/A	**0.380 (0.043)**	N/A	**0.0019 (0.0002)**	N/A
MAML	300	**0.300 (0.002)**	N/A	0.686 (0.037)	N/A	**0.0007 (0.0001)**	N/A
DRS	300	**0.189 (0.006)**	N/A	**0.427 (0.043)**	N/A	0.0036 (0.0003)	N/A

(a) Adult dataset

Scenario	N	VAE $D_{\mathbb{JS}}$	CTGAN $D_{\mathbb{JS}}$	VAE $D_{\mathbb{KL}}$	CTGAN $D_{\mathbb{KL}}$	VAE MMD	CTGAN MMD
Big data	10000	0.253 (0.009)	0.463 (0.003)	0.647 (0.045)	1.506 (0.031)	0.0006 (0.0001)	0.0014 (0.0002)
Low data	300	0.840 (0.003)	0.962 (0.002)	4.582 (0.136)	8.994 (0.909)	0.0024 (0.0002)	0.0122 (0.0001)
Pre-train	300	**0.746 (0.003)**	**0.937 (0.003)**	**3.516 (0.082)**	8.603 (0.463)	0.0024 (0.0002)	0.0137 (0.0003)
AVG	300	**0.609 (0.003)**	N/A	**2.596 (0.060)**	N/A	0.0026 (0.0001)	N/A
MAML	300	0.851 (0.001)	N/A	5.176 (0.242)	N/A	0.0028 (0.0002)	N/A
DRS	300	**0.645 (0.006)**	N/A	**2.449 (0.057)**	N/A	0.0026 (0.0002)	N/A

(b) News dataset

Table 4.10: Resemblance metrics results across scenarios I. 'Big data' represents the ideal case where many samples ($N = 10,000$) are available to generate reliable synthetic data. 'Low data' represents a more realistic scenario in which a limited number of samples ($N = 300$) are available, posing a challenge for SDG. The next rows compare the divergences and MMD obtained by each methodology ('Pre-train', 'AVG', 'MAML', and 'DRS') applied to the 'Low data' scenarios. **Bold** indicates improvements. Results are presented as *mean (standard deviation)*, with lower values being preferable.

$D_{\mathbb{JS}}$ divergence results for applying different proposed techniques. The worst improvement is obtained for 'MAML' (0.300) and the best for AVG (0.157). This implies that improvement is always present and, in the best cases, significantly high in $D_{\mathbb{JS}}$ for VAE. A similar pattern is observed for $D_{\mathbb{KL}}$ divergence in VAE: better divergence results are obtained for transfer learning cases, but improvements are always achieved. However, for the CTGAN model (where only 'Pre-train' results are available), we see no significant improvement in either $D_{\mathbb{KL}}$ or $D_{\mathbb{JS}}$.

The News subtable in Table 4.10 summarizes the results obtained for the News dataset. The values demonstrate the effectiveness of the proposed techniques in improving divergence metrics compared to the established lower bounds. Notably, the VAE model improves divergence metrics in most methodologies, except the 'MAML' technique. We hypothesize that the 'MAML' technique might require a larger number of tasks to achieve comparable

performance. Consistent with the findings for the previous dataset, model averaging emerges as the methodology that generally achieves the best results. When considering CTGAN, pre-training improves $D_{\mathbb{JS}}$ metric. While $D_{\mathbb{KL}}$ results for pre-trained CTGAN do not show significant worsening compared to the lower bound, they remain within the established CIs.

Scenario	N	VAE $D_{\mathbb{JS}}$	CTGAN $D_{\mathbb{JS}}$	VAE $D_{\mathbb{KL}}$	CTGAN $D_{\mathbb{KL}}$	VAE MMD	CTGAN MMD
Big data	10000	0.862 (0.002)	0.777 (0.003)	4.768 (0.072)	3.124 (0.115)	0.0225 (0.0016)	0.0006 (0.0001)
Low data	300	0.927 (0.002)	0.940 (0.003)	13.763 (0.696)	7.470 (0.392)	0.0029 (0.0003)	0.0109 (0.0009)
Pre-train	300	**0.862 (0.002)**	0.945 (0.002)	**5.286 (0.327)**	9.533 (0.453)	**0.0020 (0.0003)**	0.0137 (0.0010)
AVG	300	**0.740 (0.002)**	N/A	**3.489 (0.209)**	N/A	**0.0010 (0.0003)**	N/A
MAML	300	**0.910 (0.002)**	N/A	**6.436 (0.496)**	N/A	0.0057 (0.0006)	N/A
DRS	300	**0.809 (0.003)**	N/A	**4.321 (0.215)**	N/A	0.0028 (0.0003)	N/A

(a) King dataset

Scenario	N	VAE $D_{\mathbb{JS}}$	CTGAN $D_{\mathbb{JS}}$	VAE $D_{\mathbb{KL}}$	CTGAN $D_{\mathbb{KL}}$	VAE MMD	CTGAN MMD
Big data	10000	0.760 (0.013)	0.531 (0.033)	2.744 (0.084)	2.623 (0.537)	0.0001 (0.0000)	0.0057 (0.0006)
Low data	300	0.920 (0.003)	0.961 (0.002)	6.216 (0.154)	8.841 (0.710)	0.0644 (0.0011)	0.0793 (0.0020)
Pre-train	300	**0.793 (0.004)**	0.959 (0.001)	**3.831 (0.151)**	8.443 (0.630)	**0.0600 (0.0017)**	**0.0682 (0.0022)**
AVG	300	**0.867 (0.007)**	N/A	5.798 (0.295)	N/A	**0.0617 (0.0013)**	N/A
MAML	300	**0.913 (0.003)**	N/A	6.359 (0.054)	N/A	0.0639 (0.0017)	N/A
DRS	300	**0.835 (0.009)**	N/A	**4.587 (0.166)**	N/A	0.0612 (0.0026)	N/A

(b) Intrusion dataset

Table 4.11: Resemblance metrics results across scenarios II. 'Big data' represents the ideal case where many samples ($N = 10,000$) are available to generate reliable synthetic data. 'Low data' represents a more realistic scenario in which a limited number of samples ($N = 300$) are available, posing a challenge for SDG. The next rows compare the divergences and MMD obtained by each methodology ('Pre-train', 'AVG', 'MAML', and 'DRS') applied to the 'Low data' scenarios. **Bold** indicates improvements. Results are presented as *mean (standard deviation)*, with lower values being preferable.

King results in the first subtable in Table 4.11 further reinforce the efficacy of the methodologies proposed in the King dataset. The VAE model consistently improves on the lower bounds established for $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ across all techniques. For CTGAN, the results for $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ are consistent with the lower bounds. This suggests that while CTGAN does not produce significant reductions in divergence metrics, it appears to maintain the quality of the data distribution compared to the lower bounds. However, it should be noted that CTGAN exhibits consistently lower gains across datasets compared to the VAE. This may be attributed to inherent GAN framework instabilities.

The Intrusion dataset results are summarized in the second subtable in Table 4.11. These values align with the findings for the News dataset, demonstrating consistent improvements in divergence metrics for reliable cases ($M = 7500$ and $N = 1000$) across different methodologies. Similarly to the News dataset, Intrusion presents a high number of features, resulting in higher dimensionality. This increased dimensionality poses a greater challenge in generating synthetic data resembling real-world data distribution. Consequently, the divergence results for Intrusion either maintain or exhibit smaller improvements compared to lower-dimensional datasets. Despite the challenges posed by the high dimensionality of Intrusion, the proposed methodologies still demonstrate their effectiveness in improving divergence metrics, particularly for reliable cases.

The MMD results presented in Table 4.10 and Table 4.11 further support the findings obtained through divergence metrics. Across all datasets and scenarios, the VAE-based models consistently achieve lower MMD values than CTGAN, reinforcing the observation that VAE-BGM generates synthetic data that more closely resembles the real distribution. For the Adult, King and Intrusion datasets, the MMD values follow a similar trend to the divergence results, where some techniques demonstrate significant improvements over the base VAE model in 'Low data' scenarios. In particular, 'AVG' and 'Pre-train' achieve the lowest MMD values, indicating that these approaches effectively enhance generative performance. The CTGAN model, on the other hand, shows significantly higher MMD values, suggesting a weaker ability to generalize under data scarcity. The King dataset presents an interesting contrast, where the baseline CTGAN model exhibits a lower MMD value than VAE in the 'Big data' setting. However, as seen in the divergence metrics, CTGAN struggles under low-data conditions, with its MMD value increasing substantially. Finally, we observe no significant improvements using the different methodologies for the News dataset, which poses additional challenges due to its high dimensionality (58 features). While the VAE-based models still outperform CTGAN, MMD values remain the same as those observed for lower-dimensional datasets. This suggests that the difficulty in accurately capturing inter-feature dependencies increases with dimensionality, aligning with the trends seen in divergence metrics. Overall, the MMD results corroborate the divergence-based evaluations, reinforcing the robustness of the proposed validation framework. The consistent alignment between MMD and divergence scores indicates that the improvements in generative performance achieved by VAE-BGM are well-reflected across multiple validation approaches.

Finally, Table 4.12 summarizes the results of applying various methodologies to generate synthetic tabular data using the VAE model. Across different methods, the results consistently show positive gains, except for 'MAML'. These gains are computed as the difference between the metric in the lower bound and any of the proposed methods. This indicates that the proposed techniques generate synthetic data closer to the real-world data distribution than when not using them. This improvement is particularly evident in the model averaging methodology, which consistently outperforms other techniques in terms of divergence reduction. The performance of MAML might be limited in this scenario due to its reliance on a large number of seeds. Since the results presented do not specify the number of seeds used, MAML might not have had enough to learn effectively [275]. The consistent improvement in divergence metrics highlights the robustness and generalizability of the proposed techniques. These findings suggest that our approach is an effective methodology for generating high-quality

synthetic tabular data that can be used for various applications.

Dataset	Pre-train Gain		AVG Gain		MAML Gain		DRS Gain	
	$D_{\mathbb{JS}}$	$D_{\mathbb{KL}}$	$D_{\mathbb{JS}}$	$D_{\mathbb{KL}}$	$D_{\mathbb{JS}}$	$D_{\mathbb{KL}}$	$D_{\mathbb{JS}}$	$D_{\mathbb{KL}}$
Adult	**0.159 (0.482)**	**0.271 (0.388)**	**0.173 (0.525)**	**0.317 (0.455)**	**0.030 (0.092)**	**0.011 (0.016)**	**0.142 (0.430)**	**0.271 (0.388)**
News	**0.093 (0.111)**	**1.065 (0.233)**	**0.230 (0.274)**	**1.985 (0.433)**	-0.011 (-0.013)	-0.594 (-0.130)	**0.194 (0.231)**	**2.133 (0.465)**
King	**0.064 (0.070)**	**8.477 (0.616)**	**0.187 (0.202)**	**10.274 (0.746)**	**0.017 (0.018)**	**7.327 (0.532)**	**0.118 (0.128)**	**9.442 (0.686)**
Intrusion	**0.127 (0.138)**	**2.385 (0.384)**	**0.053 (0.057)**	**0.419 (0.067)**	**0.006 (0.007)**	-0.143 (-0.023)	**0.085 (0.092)**	**1.629 (0.262)**
Average	0.111 (0.200)	3.050 (0.405)	0.161 (0.265)	3.249 (0.425)	0.011 (0.026)	1.650 (0.099)	0.135 (0.220)	3.369 (0.450)

Table 4.12: Gains using the proposed methodology for the VAE. Gains are represented in the following format: *absolute gain (relative gain)*. The methodology achieves relative gains of up to 50% in $D_{\mathbb{JS}}$ divergence, which is bounded, and up to 75% in $D_{\mathbb{KL}}$ divergence. Bold values indicate positive gain. Higher is better.

It is important to acknowledge the varying computational demands of the compared methods. Model averaging is the most efficient approach, as inductive bias generation only involves calculating a mean, resulting in minimal computational overhead. In contrast, MAML exhibits the highest computational load due to its intricate optimization procedure. Pre-training and DRS fall between these two extremes, both requiring the training of a DGM to establish the inductive bias. Considering these findings alongside the results presented in Table 4.12, we recommend against using MAML. It offers minimal performance gains with significant computational costs. The other methods, on the other hand, provide a more favorable trade-off between computational efficiency and performance. Furthermore, as detailed in Appendix D.2.1, reliably quantifying the benefits of our methodology in a realistic, limited-data setting is challenging. This implies that validation with a large number of samples is necessary to definitively assess which of our proposed inductive bias techniques yields superior results for a specific dataset. However, the experimental results strongly suggest potential gains that warrant further exploration.

Conclusions

This research proposed a novel approach to generate synthetic tabular data using DGMs in the context of limited datasets. Our approach leverages four distinct techniques to artificially introduce an inductive bias that guides the DGM toward generating more realistic and informative synthetic data samples. These techniques encompass two transfer learning approaches, MAML and DRS. To facilitate the application of model averaging, MAML, and DRS, we employ the VAE-BGM proposed in this thesis and train multiple instances with different random seeds. This allows us to leverage the ensemble properties of the VAE models for techniques like model averaging and further enables the application of meta-learning algorithms like MAML and DRS. We also used the CTGAN [109] to assess pre-training to compare other well-known model architectures for STDG. We used divergence metrics, in particular $D_{\mathbb{JS}}$ and $D_{\mathbb{KL}}$ divergences, and MMD to compare the real and synthetic data distributions generated. The experimental results consistently demonstrate the effectiveness of our proposed approach in generating high-quality synthetic tabular data, particularly when

using transfer learning techniques. These techniques significantly improve the resemblance metrics, indicating a closer resemblance between the synthetic and real data distributions. Our approach offers several advantages over existing methods. Firstly, it effectively addresses the challenge of generating realistic synthetic data from small datasets, a common limitation in many real-world applications. Secondly, the use of transfer learning and meta-learning techniques enhances the inductive bias of the DGM, leading to more meaningful and informative synthetic data samples. However, it is also important to acknowledge the trade-offs associated with our methodology. Training VAEs with these techniques requires training multiple VAE models with different random seeds. This can lead to a significant increase in computational cost compared to simpler DGM training methods. While resemblance metrics provide a valuable measure of distributional similarity, their ability to reliably assess the improvement in synthetic data quality for specific downstream tasks can be limited, especially with small datasets, as detailed in Appendix D.2.1. Our experimental results show that our methodology may provide significant gains in $D_{\mathbb{JS}}$ divergence of up to 50%.

In conclusion, our proposed approach provides a promising solution for generating high-quality synthetic tabular data from small datasets, particularly when VAEs apply transfer learning techniques. We believe that this work has the potential to significantly contribute to the field of SDG and ML applications that rely on small datasets. However, there are several research lines to be addressed. While the current study focuses on VAEs and GANs, investigating the applicability of our framework to other DGM architectures could provide valuable insights. In addition, our current approach does not explicitly incorporate domain knowledge. Future research could explore mechanisms to integrate domain-specific information from an expert into the inductive bias generation process, potentially leading to even more realistic and informative synthetic data. Lastly, although resemblance metrics offer a valuable measure of distributional similarity, exploring additional evaluation techniques that assess the quality and usefulness of synthetic data for specific downstream tasks would provide a more comprehensive understanding of the effectiveness of our methodology, valid also for the case when little amount of data is available for validation, which is a current limitation of this work.

4.4.3 Generation Methodology Applied to Medical Datasets

Validation Metrics

As with general-purpose datasets, divergence metrics are employed to validate the quality of synthetic data. These calculations rely on a probabilistic discriminator network to estimate the density ratio between real and synthetic data distributions. This approach effectively captures discrepancies, addressing the lack of standardized validation methods in STDG. Traditional similarity validation techniques often assess variables independently [277], failing to account for inter-variable correlations and complex non-linear relationships. To overcome this limitation, our methodology incorporates three key metrics: $D_{\mathbb{KL}}$, $D_{\mathbb{JS}}$, and MMD.

$D_{\mathbb{KL}}$ measures the difference between two probability distributions, quantifying the amount of information lost when one distribution is used to approximate the other. $D_{\mathbb{JS}}$, a symmetrized and bounded version of $D_{\mathbb{KL}}$, offers advantages in interpretability and robustness. Specifically, $D_{\mathbb{JS}}$ provides a value between 0 and 1, making it easy to interpret, and it effectively considers

complex correlations between variables, providing a more comprehensive similarity assessment.

Additionally, as in other SDG contributions, we incorporate MMD as a resemblance metric to further assess distributional similarity. MMD enables a non-parametric comparison between real and synthetic data distributions using a kernel-based approach. Specifically, we employ again the RBF kernel, which maps data into a high-dimensional space, allowing for a more robust evaluation of structural similarities. Lower MMD values indicate a closer match between real and synthetic datasets, reinforcing the effectiveness of the generative models. By integrating MMD alongside divergence metrics, we provide a more comprehensive validation framework for assessing synthetic data quality in medical applications.

However, in medical contexts, where practical application is critical, similarity validation alone is insufficient. Clinical utility validation is essential to assess whether synthetic data can perform specific tasks effectively. For this, we validate synthetic data for classification and SA tasks.

- **Classification Tasks:** An MLP classifier is used for its flexibility and robustness in handling diverse classification problems. The accuracy score is the evaluation metric, measuring the proportion of correctly classified instances.
- **SA Tasks:** SAVAE is used for superior performance over classic models like CoxPH [45] and other DL models. Evaluation metrics include the C-index, which quantifies the concordance between predicted and observed survival times, and the IBS, which assesses the precision of survival predictions over time. Additionally, KM curves visualize and compare survival probabilities, providing an intuitive method to evaluate similarities between real and synthetic data distributions.

Clinical utility validation is conducted across three cases to compare synthetic data performance:

1. **Real case:** Metrics obtained by training and validating on real data serve as the upper-bound benchmark.
2. **Synthetic case:** Metrics calculated by training on synthetic data and validating on real data.
3. **Synthetic Fine-tuned case:** Metrics obtained by training synthetic data, fine-tuning on a separate real dataset, and validating the same real data used in the other cases.

This approach ensures consistent validation using the same real dataset, enabling direct comparison across cases and assessing the impact of synthetic data on utility tasks.

Clinical utility validation is crucial, but we can also have a better vision of how well the synthetic data are generated by changing the main task of the dataset. Therefore, to extend the scope of validation, additional tests are conducted by altering the main tasks of the datasets:

- **Classification Datasets:** Target labels are modified to assess the adaptability of synthetic data to new classification tasks.
- **SA Datasets:** Variables other than time, selected for their medical relevance, are used

as targets to explore alternative predictive capabilities.

This comprehensive approach evaluates whether synthetic data can effectively support tasks beyond its original purpose (e.g., instead of predicting A, we want to predict B). It demonstrates the flexibility of synthetic datasets in addressing various clinical questions, enhancing their value and applicability in medical research. This method provides a broader clinical utility validation and helps accumulate evidence supporting the effectiveness of STDG.

The dual validation strategy—combining similarity and clinical utility—ensures that synthetic data approximates real data distributions and proves practical for specific medical applications. Similarity validation assures the similarity of generated data, while clinical utility validation demonstrates task-specific effectiveness. The inclusion of MMD as a complementary resemblance metric strengthens this validation, providing an additional perspective on distributional similarity. This is crucial in medical settings where synthetic data may need to support diverse tasks beyond initial expectations. By incorporating both validation approaches, this methodology establishes the reliability of synthetic data for real-world use and its potential to adapt to evolving research needs.

Implementation Details

Parameters were configured to balance complexity, computational efficiency, and data quality to implement the VAE-based model for STDG. The latent space dimensions were tailored to dataset types: 20 for classification datasets with more features and 10 for SA datasets with fewer features. Each hidden layer comprised 256 neurons, and the model was trained with ten different seeds, a batch size of 256, and up to 10,000 epochs, using early stopping to prevent overfitting. Other parameters for the VAE and the BGM models adhered to defaults outlined in the original methodology.

For clinical utility validation, the MLP classifier consisted of three dense layers (256, 64, and 32 neurons) with leaky ReLU activation, batch normalization, and dropout for regularization. The SAVAE model for SA was implemented with a latent dimension of 10 and hidden layers of 256 neurons, trained for 5,000 epochs using early stopping and a batch size of 256.

All models were cross-validated with five splits, ensuring robust and unbiased performance estimates. This is especially important for small datasets. This approach mitigates overfitting and allows comprehensive evaluation across data splits.

Three key experimental scenarios were designed to evaluate the methodology under varying data availability conditions:

1. **'Big data' Scenario:** This optimistic scenario featured sufficient samples to train the DGM, with $N = 10,000$ for classification datasets and 80% of the total data for SA datasets. No inductive bias techniques were applied, establishing an upper-bound benchmark for divergences.
2. **'Low data' Scenario:** This realistic scenario featured a limited sample size ($N = 100$) to assess baseline performance. The goal was to quantify potential gains from applying the proposed methodology under data-scarce conditions.

3. **'Pre-train,' 'AVG,' and 'DRS' Scenarios:** In the 'Low Data' scenario, advanced methods like pre-training ('Pre-train'), model averaging ('AVG') and DRS were applied to evaluate their effectiveness in generating high-quality synthetic data with limited samples.

To calculate divergences, $7,500$ samples were used for training, and $1,000$ for estimating the density ratio, consistent with the general-purpose data case. While feasible for classification datasets with abundant data, this approach is less applicable to SA datasets, which typically have fewer samples. These experiments provided insights into the performance of the methodology and divergence estimation under constrained data conditions.

Finally, p-values were used to rigorously assess differences in validation metrics, providing statistical insights into model performance across varying scenarios. Appendix D.3.3 presents the p-values obtained for each dataset, comparing the 'Low-data' scenario to the different technique-enhanced scenarios. Given the large number of p-values generated, the adjustment of Holm [232] was applied to control the family-wise error rate and ensure meaningful comparisons. The adjusted p-values are also provided in Appendix D.3.3, allowing for a clearer interpretation of the statistical significance of the results.

The experimental setup assessed the robustness and effectiveness of the methodology, enabling SDG for diverse medical applications. Code and datasets are openly available in https://github.com/Patricia-A-Apellaniz/medical_low_sample_generator to promote replication and further research.

Experiments and Results

In this study, we used four datasets—Heart, Metabric, GBSG, and NWTco—to evaluate the effectiveness of the proposed methodology for SDG. Detailed explanations of these datasets are provided in Appendix B. These datasets span diverse tasks, including classification and SA, and reflect the variability and complexity of medical data, ranging from abundant samples in the Heart dataset (253,680 samples) to smaller, clinically relevant cancer-related SA datasets (e.g., Metabric with 1,904 samples). The Heart dataset, characterized by its large size and binary/discrete features, is a benchmark for assessing performance under optimal conditions. In contrast, the Metabric, GBSG, and NWTco datasets, sourced from the Pycox package, focus on SA tasks, capturing real-world challenges like data scarcity and heterogeneity. These datasets allow for a robust evaluation of the ability of the methodology to generate high-quality synthetic data that maintain data similarity and practical utility across a range of medical applications. Additional experiments using other classification and SA datasets, detailed in Appendix D.3.1, further validate the applicability of the methodology. Through this evaluation, our study highlights the potential of STDG techniques to address the complexities of medical data, contributing valuable insights for advancing SDG in healthcare.

→ Classification datasets

The classification experiments focus on the Heart dataset. Validation consists of two key components: similarity validation (using $D_{\mathbb{KL}}$, $D_{\mathbb{JS}}$ and MMD) and clinical utility validation (accuracy results across training and testing configurations). Configurations include training and validating with real data (Real Acc.), training with synthetic data and validating with

real data (Synth. Acc.), and training with synthetic data, fine-tuning with real data, and validating with real data (Synth. Fine-Tuned Acc.).

Scenario	SIMILARITY VALIDATION			CLINICAL UTILITY VALIDATION		
	$D_{\mathbb{JS}}$	$D_{\mathbb{KL}}$	MMD	Real Acc	Synth Acc	Synth Fine-Tuned Acc
Big data	0.096 (0.057)	0.171 (0.056)	0.0003 (0.0001)	0.631 (0.018)	0.615 (0.021)	0.629 (0.021)
Low data	0.852 (0.002)	6.642 (0.463)	0.0161 (0.0004)	0.601 (0.046)	0.640 (0.062)	0.636 (0.022)
Pre-train	**0.788 (0.004)**	**4.575 (0.242)**	**0.0073 (0.0002)**	N/A	0.687 (0.011)	0.692 (0.011)
AVG	**0.748 (0.008)**	**4.300 (0.135)**	**0.0057 (0.0002)**	N/A	0.696 (0.024)	0.665 (0.021)
DRS	**0.767 (0.017)**	**3.691 (0.122)**	**0.0063 (0.0002)**	N/A	0.649 (0.045)	0.661 (0.035)

Table 4.13: Validation results for the Heart dataset across different scenarios. The 'Big data' scenario represents an optimal condition where many samples ($N = 10,000$) are available, facilitating the generation of reliable synthetic data. The 'Low data' scenario reflects a more practical situation with a limited sample size ($N = 100$), posing a significant challenge for STDG. The similarity validation section presents the divergences ($D_{\mathbb{JS}}$, $D_{\mathbb{KL}}$, and MMD) obtained using various techniques (pre-training, model averaging, and DRS) applied to the 'Low data' scenario. In the similarity validation, values in **bold** indicate improvements achieved by the applied techniques. Lower values are preferable for divergence metrics. The clinical utility validation section provides accuracy metrics comparing the performance of models trained on real, synthetic, and synthetic data with fine-tuning on real data. Higher values indicate better accuracy. In the clinical utility validation, **bold** denotes that the adjusted p-value is below the significance threshold of 0.01. All results are expressed as *mean (standard deviation)*.

Table 4.13 summarizes the results for the Heart dataset. The 'Big data' scenario, with $N = 10,000$ samples, achieves the lowest $D_{\mathbb{JS}}$ divergence (0.096 ± 0.057), representing the optimal condition. In contrast, the 'Low data' scenario, with $N = 100$ samples, exhibits a significantly higher $D_{\mathbb{JS}}$ divergence (0.852 ± 0.002), reflecting the challenges of limited data. Advanced techniques, such as model averaging and DRS, reduce divergences effectively, though they still fall short of the 'Big data' scenario, highlighting the persistent difficulty of STDG under data scarcity. A similar pattern is observed for $D_{\mathbb{KL}}$, reinforcing that sample size significantly influences the quality of generated synthetic data.

Beyond divergence metrics, MMD results provide an additional similarity validation perspective. Lower MMD values indicate a higher resemblance between real and synthetic distributions, complementing the divergence findings. As expected, the 'Big data' scenario achieves the lowest MMD (0.0003 ± 0.0001), confirming that synthetic data closely match real data when ample samples are available. The 'Low data' scenario, in contrast, exhibits a significantly higher MMD (0.0161 ± 0.0004), aligning with the increased divergence values. 'Pre-train', 'AVG' and 'DRS' improve MD scores under low-data conditions, with 'AVG' achieving the lowest MMD (0.0057 ± 0.0002), further validating its effectiveness in enhancing synthetic data quality. These findings underscores that similarity validation results heavily depend on the number of samples used to generate synthetic data.

In clinical utility validation, three scenarios are analyzed: Real Acc. as the benchmark, Synth. Acc., and Synth. Fine-Tuned Acc. Results from the 'Big data' scenario serve as the upper bound, while the 'Low data' scenario highlights the challenges of the methodology. In the benchmark scenario, accuracy metrics such as Real Acc. (0.631 ± 0.018) are slightly higher than in the 'Low data' scenario (0.601 ± 0.046). In this dataset, we observe in the benchmark scenario (Real Acc.) that when only a few real data samples are used, the accuracy obtained (0.601 ± 0.046) overlaps with the 'Big data' scenario (0.631 ± 0.018). This pattern persists in the other two cases where the methodology is not used. The transfer-learning techniques do not show significant advantages or disadvantages in the Synthetic and Synthetic Fine-tuned cases compared to the benchmark accuracies obtained in the Real Acc. case.

The methodology yields notable improvements in divergences under low-data scenarios, indicating better alignment between synthetic and real data distributions. However, clinical utility validation without the methodology often achieves results comparable to the benchmark scenarios, suggesting that the accurate generation of critical variables for classification can compensate for discrepancies in the overall distribution. This behavior might be explained by the accurate generation of key variables, which contain sufficient information for correct classification even if other variables are not well-represented. Consequently, despite high divergence values indicating discrepancies in the synthetic joint distribution, utility metrics for synthetic data can still closely approximate those achieved with real data. In Appendix D.3.2, we provide further insights into this behavior, discussing how the generation of these critical variables can suffice for clinical utility validation, even when the joint distribution of the synthetic data is suboptimal. This underscores the importance of combining similarity and utility validation to assess synthetic data quality comprehensively.

→ SA results

In the previous section, classification data demonstrated that the methodology improves SDG, particularly regarding divergences. Additionally, we reiterated that divergences are a robust and reliable metric to validate SDG. Building on this, SA experiments focus on datasets with limited sample sizes to test the robustness of the methodology in real-world, data-scarce conditions. These cancer-related datasets, chosen for their scarcity and heterogeneity, present a challenging but realistic testbed for evaluating synthetic data quality.

Due to the limited sample sizes in SA datasets, accurate calculation of divergences is not feasible, as inadequate data affects the training of DGMs. Thus, similarity validation is omitted for SA datasets, and the focus shifts exclusively to clinical utility validation. Clinical utility metrics, including C-index and IBS, assess the practical applicability of the methodology under previously defined scenarios.

Table 4.14 presents the C-index and IBS results for the three different cases (Real, Synthetic, and Synthetic Fine-Tuned) across each scenario. The results highlight two key observations: (1) there is no significant loss in any metric when comparing the 'Big data' scenario to the 'Low data' scenario, aligning with the classification data results in clinical utility validation, and (2) there is no difference in performance metrics when utilizing the methodologies, consistent with previous experiments. This confirms that clinical utility validation alone is insufficient to evaluate the quality of synthetic data. We hypothesize that using the methodology would

yield better divergence metrics, which is in line with our previous experiment and the findings in the general-purpose data case; however, it is essential to note that we cannot reliably assess divergences with a low number of samples.

Scenario	Real CI	Synth CI	Synth Fine-Tuned CI	Real IBS	Synth IBS	Synth Fine-Tuned IBS
Big data	0.633 (0.035)	0.622 (0.035)	0.626 (0.035)	0.183 (0.028)	0.196 (0.028)	0.185 (0.028)
Low data	0.595 (0.037)	0.589 (0.040)	0.587 (0.041)	0.194 (0.029)	0.199 (0.029)	0.200 (0.029)
Pre-train	N/A	0.614 (0.037)	0.617 (0.038)	N/A	0.184 (0.028)	0.186 (0.028)
AVG	N/A	0.613 (0.035)	0.615 (0.036)	N/A	0.182 (0.028)	0.183 (0.028)
DRS	N/A	0.614 (0.036)	0.604 (0.040)	N/A	0.183 (0.028)	0.184 (0.028)

(a) Metabric dataset

Scenario	Real CI	Synth CI	Synth Fine-Tuned CI	Real IBS	Synth IBS	Synth Fine-Tuned IBS
Big data	0.683 (0.032)	0.690 (0.031)	0.685 (0.031)	0.193 (0.026)	0.184 (0.026)	0.192 (0.026)
Low data	0.638 (0.041)	0.598 (0.053)	0.595 (0.052)	0.208 (0.028)	0.232 (0.034)	0.228 (0.033)
Pre-train	N/A	0.656 (0.033)	0.657 (0.033)	N/A	0.213 (0.028)	0.211 (0.029)
AVG	N/A	0.647 (0.043)	0.641 (0.051)	N/A	0.214 (0.028)	0.217 (0.029)
DRS	N/A	0.654 (0.036)	0.672 (0.032)	N/A	0.198 (0.027)	0.196 (0.027)

(b) GBSG dataset

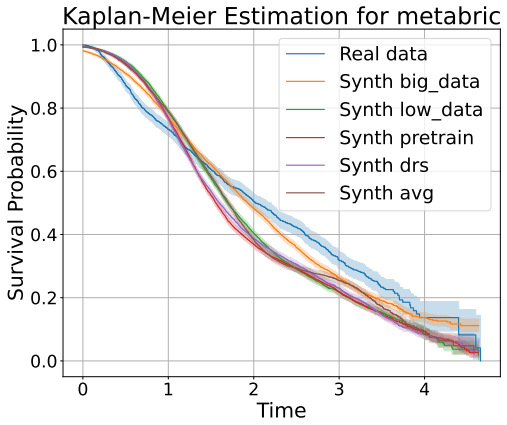
Scenario	Real CI	Synth CI	Synth Fine-Tuned CI	Real IBS	Synth IBS	Synth Fine-Tuned IBS
Big data	0.694 (0.023)	0.682 (0.028)	0.683 (0.025)	0.112 (0.016)	0.111 (0.016)	0.111 (0.015)
Low data	0.587 (0.044)	0.549 (0.028)	0.542 (0.028)	0.139 (0.023)	0.141 (0.017)	0.136 (0.017)
Pre-train	N/A	0.590 (0.025)	0.596 (0.025)	N/A	0.152 (0.019)	0.146 (0.018)
AVG	N/A	0.591 (0.026)	0.593 (0.028)	N/A	0.138 (0.019)	0.142 (0.017)
DRS	N/A	**0.611 (0.033)**	**0.594 (0.037)**	N/A	0.133 (0.018)	0.131 (0.018)

(c) NWTco dataset

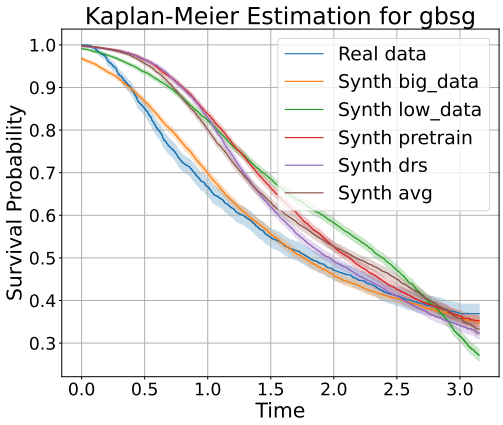
Table 4.14: Validation results for the SA datasets across different scenarios. The 'Big data' scenario represents an ideal condition with a larger sample size ($N = 1,524$, $N = 1,786$, and $N = 3,223$, 80% of the data for Metabric, GBSG and NWTco, respectively), enabling reliable SDG. The 'Low data' scenario reflects a more realistic constraint with a smaller sample size ($N = 100$), posing challenges for SDG. The table presents SA metrics (C-index and IBS) comparing models trained on real data, synthetic data, and synthetic data fine-tuned on real data. Higher C-index values indicate better predictive performance, while lower IBS values are preferable. **Bold** highlights significant improvements using the methodology, while * indicates a significant disadvantage. Results are reported as *mean (standard deviation)*.

KM estimations were performed to supplement clinical utility validation. With CIs, these survival curves visually compare survival probabilities for real and synthetic data across scenarios, offering additional insights into synthetic data quality. Figure 4.15 highlights

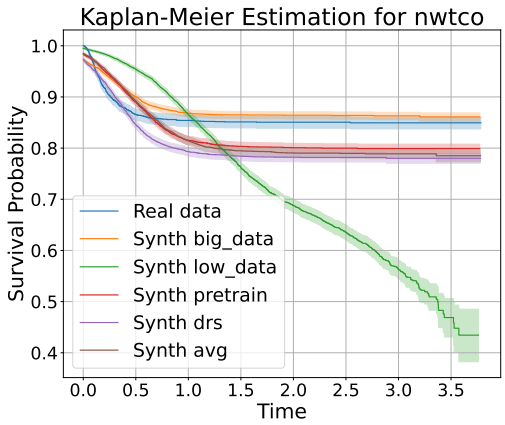
notable trends, particularly in the NWTco dataset. KM curves generated with 'Big data' closely align with those from real data, serving as the benchmark. In the 'Low data' scenario, curves deviate significantly from the benchmark without the methodology, with broader CIs. Methodology-enhanced scenarios, such as DRS, produce KM curves that converge toward the upper bound, narrowing the gap between synthetic and real data. The GBSG dataset exhibits similar trends, particularly in later survival time regions, where methodology-applied curves diverge less from the benchmark than the 'Low data' curves. These results confirm that the methodology effectively improves synthetic data quality under limited conditions.



(a) Metabric dataset



(b) GBSG dataset



(c) NWTco dataset

Figure 4.15: KM estimations with CIs using real and synthetic data across different scenarios for each dataset. The survival functions for the upper bounds are represented in blue and orange, illustrating the survival probabilities of the real data and synthetic data generated from many samples. The survival functions for the synthetic data generated using the proposed methodology, shown in red, purple, and brown, demonstrate convergence towards the upper bounds. In contrast, the survival function for the lower bound, depicted in green, shows significant deviation from the upper bounds.

→ Different data utility results

This section evaluates the robustness of synthetic data generated using STDG by modifying original tasks for classification and SA datasets. Target labels are altered in classification datasets, while categorical variables replace survival time in SA datasets, transforming them into classification problems. This approach assesses the adaptability and reliability of synthetic data across diverse clinical applications.

One of the primary goals of STDG is to create datasets that can be repurposed for tasks beyond their original intent. Traditional clinical utility validation focuses on relationships between covariates and target variables. However, testing whether synthetic data retains utility when target variables change is critical, as it mirrors real-world scenarios where datasets are used for different studies.

Heart					
Feature	**Low data**	**DRS**	**Feature**	**Low data**	**DRS**
HighBP	0.664 (0.005)	**0.695 (0.005)**	**Veggies**	0.519 (0.167)	0.609 (0.069)
HighChol	0.603 (0.008)	0.609 (0.004)	**HvyAlcoholConsump**	0.456 (0.294)	0.934 (0.038)
CholCheck	0.665 (0.146)	0.493 (0.199)	**AnyHealthcare**	0.629 (0.048)	**0.755 (0.012)**
Smoker	0.599 (0.006)	0.604 (0.008)	**NoDocbcCost**	0.728 (0.047)	0.639 (0.011)
Stroke	0.768 (0.280)	0.664 (0.010)	**DiffWalk**	0.799 (0.019)	0.795 (0.003)
PhysActivity	0.671 (0.017)	0.675 (0.004)	**Sex**	0.539 (0.016)	0.528 (0.018)
Fruits	0.588 (0.022)	0.585 (0.028)	**HeartDiseaseorAttack**	0.679 (0.123)	0.671 (0.055)

(a) Classification dataset

Metabric			GBSG			NWTco		
Feature	**Low data**	**DRS**	**Feature**	**Low data**	**DRS**	**Feature**	**Low data**	**DRS**
x4	0.676 (0.012)	0.652 (0.018)	**x0**	0.595 (0.012)	0.607 (0.021)	**in.subcohort**	0.678 (0.150)	0.578 (0.216)
x5	0.578 (0.015)	0.579 (0.020)	**x1**	0.326 (0.071)	0.322 (0.064)	**instit_2**	0.605 (0.034)	**0.726 (0.046)**
x6	0.748 (0.003)	0.740 (0.023)	**x2**	0.873 (0.007)	**0.890 (0.004)**	**histol_2**	0.714 (0.153)	0.695 (0.037)
x7	0.795 (0.013)	0.724 (0.017)*	**event**	0.805 (0.006)	0.809 (0.001)	**study_4**	0.782 (0.008)	0.763 (0.012)
event	0.696 (0.004)	0.701 (0.006)				**event**	0.687 (0.029)	**0.788 (0.005)**

(b) SA datasets

Table 4.15: Clinical utility validation results for the datasets. Accuracy comparison between the $N = 100$ samples without methodology case ('Low data') and the DRS technique applied to the lower bound case ('DRS') for each feature used as a classification label. Higher values indicate better performance. **Bold** values indicate a significant advantage in using the methodology, while * values indicate a significant disadvantage in using the methodology. All results are expressed as *mean (standard deviation)*.

Table 4.15 summarizes the results for each dataset. Accuracy values were compared across two scenarios: (1) training with $N = 100$ samples of synthetic data generated without the methodology ('Low data') and (2) training with $N = 100$ samples of synthetic data generated using the DRS technique and validating with real data ('DRS'). The first subtable shows the Heart dataset results, where different features were used as target labels. These features were binary or categorical, with fewer than ten classes, ensuring suitability for classification

tasks. The results show that the DRS technique improves accuracy compared to the 'Low data' scenario. Notable improvements include using *'HighBP'* and *'AnyHealthcare'* as target labels, where accuracy increased significantly. For example, *'AnyHealthcare'* improved from 0.629 ± 0.048 to 0.755 ± 0.012 with the DRS technique. These findings suggest that the methodology effectively models complex relationships between features, enhancing data utility for classification tasks. Even in scenarios where no significant improvement is observed, the methodology does not negatively affect performance. In the SA datasets from the second subtable, categorical variables were used as target labels instead of survival time to assess the utility of the data in classification tasks. The results align with those from the classification datasets, with the DRS method consistently improving accuracy for specific features. For example, in the NWTco dataset, the feature *'instit_2'* increased accuracy from 0.605 ± 0.034 to 0.726 ± 0.046. Similarly, the feature *'x2'* from GBSG dataset improved from 0.873 ± 0.007 to 0.890 ± 0.004. While some features, such as *'x7'* in the Metabric dataset, experienced slight decreases, these cases are rare, and the overall trend indicates improved or maintained performance with the methodology.

By validating our approach through these transformed tasks, we demonstrate that using the generation methodology either yields benefits in specific scenarios or does not negatively affect the outcomes. This evidence supports that incorporating an inductive bias for STDG in low-sample datasets enhances the modeling of complex relationships between variables. Consequently, this approach improves the generation process, providing better utility and robustness for diverse clinical applications.

Conclusions

In DL, large amounts of data are ideal, but data are often scarce and heterogeneous in the medical field. This research applied the methodology proposed for STDG in medical environments, focusing on cancer-related SA datasets with limited samples. As a preliminary step, classification datasets with more samples were used to evaluate the performance of the methodology by comparing results from subsets of data to those obtained using the entire dataset.

The validation process combined similarity validation using divergences, particularly the interpretable and bounded $D_{\mathbb{JS}}$ divergence, and clinical utility validation. These assessments demonstrated that transfer learning and meta-learning techniques in STDG environments enhance model performance, especially under data-scarce conditions. Initially, the methodology was tested on classification datasets, where its application improved the generation process. Divergences proved to be robust metrics for comparing real and synthetic datasets, while clinical utility validation showed limited sensitivity, as metrics like accuracy remained stable across scenarios. Clinical utility validation metrics (C-index and IBS) showed minimal variation compared to upper-bound results for cancer-related SA datasets characterized by few samples. These findings confirmed that the synthetic data were sufficient SA tasks, though divergences were more reliable than utility metrics for validating STDG. When datasets were repurposed for alternative tasks, the methodology occasionally generated better synthetic data, supporting its ability to model complex relationships and produce data usable across different clinical applications.

Calculating the $D_{\mathbb{JS}}$ divergence can sometimes be challenging due to limited sample sizes. However, when it is possible to compute it reliably, the observed results suggest that the methodology effectively aligns the synthetic and real data distributions. Further evidence of this alignment comes from the repurposing of synthetic datasets for alternative tasks, where they have demonstrated utility and yielded robust results. Additionally, a more comprehensive analysis of clinical utility across multiple applications confirms that the generated synthetic data maintain practical relevance and usability.

This research demonstrated that the methodology significantly improves STDG for medical applications with limited data. It often makes validation metrics statistically comparable to those in the 'Big data' scenario and outperforms the naive 'Low data' approach. $D_{\mathbb{JS}}$ divergence emerged as a reliable tool for comparing real and synthetic datasets, while clinical utility validation requires further refinement for low-sample scenarios. Despite its limitations, the methodology enables SDG for broader medical analyses.

Future work should focus on developing more robust validation metrics by combining alternative divergence measures with clinical utility assessments for a comprehensive evaluation framework. Collaboration with clinical experts is essential to refine the STDG process and ensure that generated data meet medical standards. Establishing public repositories of high-quality synthetic datasets would facilitate further research while maintaining patient privacy and data security, promoting innovation in medical research.

4.5 Chapter Conclusions

This chapter has explored the methodologies, challenges, and validation strategies in SDG, particularly for tabular data, a format that plays a pivotal role in fields like healthcare. The collective findings underscore both the progress made and the opportunities for further advancements in this area.

SDG models, such as VAEs and GANs, have demonstrated significant potential in addressing challenges posed by real-world datasets, including mixed data types, imbalances, and limited sample sizes. Innovative techniques, such as integrating Bayesian Gaussian Mixtures into VAE architectures, enhance the ability to model complex data distributions at both the marginal and joint levels, setting new benchmarks for synthetic data quality. Similarly, incorporating transfer learning and meta-learning into GMs highlights the potential for improving data realism, particularly in data-scarce scenarios.

Validation remains a critical focus, as traditional metrics often fail to capture the nuances of joint feature relationships. This chapter presented a probabilistic divergence-based framework as a robust alternative, offering a more comprehensive assessment of synthetic data quality. However, the effectiveness of any validation approach depends on the quality of the GM and the adequacy of training data, emphasizing the need for methodological rigor in GM training and validation design.

The application of these methodologies in medical settings revealed the practical potential of SDG to replicate complex relationships in sensitive domains like healthcare. Resemblance metrics, such as $D_{\mathbb{JS}}$ and MMD, emerged as reliable tools for validating synthetic data quality, while clinical utility metrics showed promise but required further refinement for scenarios involving small datasets. Moreover, the adaptability of synthetic datasets for alternative clinical tasks underscores the robustness and versatility of the proposed approaches.

Despite these advancements, the chapter also highlights areas for future research. Key directions include developing comprehensive validation frameworks integrating distributional similarity with downstream task utility, incorporating domain-specific knowledge to enhance data realism, and improving computational efficiency to make advanced techniques more accessible. Additionally, establishing public repositories of high-quality synthetic datasets would enable broader adoption and collaboration while safeguarding privacy.

In summary, this chapter demonstrates that while significant strides have been made in SDG, particularly for tabular data, there remains substantial scope for innovation. By addressing the challenges of data scarcity, complex distributions, and validation limitations, SDG can play a transformational role in advancing machine learning and fostering collaboration across diverse domains. The methodologies and findings presented here provide a strong foundation for future research, emphasizing the potential of synthetic data to drive impactful applications while ensuring privacy and security.

Chapter 5

Federated Learning

5.1 Introduction

This chapter presents the contributions of this thesis to advancing FL methodologies, particularly in non-IID settings where techniques such as FedAvg [182] fail to perform satisfactorily. We propose Federated Synthetic Data Sharing (FedSDS) as a novel FL technique that overcomes the limitations of parameter-sharing approaches by leveraging locally generated synthetic data to collaborate. FedSDS is designed to address the dual challenges of data scarcity and heterogeneity, which are particularly prevalent in healthcare research scenarios.

Data heterogeneity is a significant challenge in real-world medical contexts due to population and institutional-specific practices, which lead to disparities in patient demographics, clinical protocols, and data collection methods. These variations can result in biased models if not adequately addressed [278], [279], limiting the effectiveness and generalizability of FL solutions. Consequently, developing techniques capable of handling non-IID data is imperative for realizing the full potential of FL in medical research. Traditional FL techniques, such as FedAvg, aggregate model parameters iteratively across institutions to produce a global model. However, FedAvg assumes that data across participating clients are IID, which is rarely the practice case. In non-IID settings, such as SA, the inherent diversity in patient populations and healthcare settings often results in significant variations in data distributions. As a result, the reliance of FedAvg on parameter-sharing can lead to biased global models, poor generalization, and suboptimal convergence. Furthermore, there are non-IID scenarios where traditional parameter-sharing FL techniques cannot be applied effectively due to these challenges.

FedSDS represents a departure from conventional FL paradigms by introducing a data-sharing mechanism that operates through SDG at each participating node. Using advanced generative models such as the VAE-BGM model proposed in this thesis (Section 4.2), nodes generate synthetic datasets that encapsulate the statistical properties of their local data. These datasets are then shared among institutions, enabling collaborative model training without compromising patient privacy or data security. Essential to FedSDS is an innovative data aggregation technique based on similarity to local data distributions. This ensures that

the shared synthetic data contributes meaningfully to the training process, even in highly heterogeneous environments.

This chapter explores two key applications of FedSDS:

1. **Federated VAE (FedVAE) for Synthetic Data Generation:** By deploying the VAE-BGM model in a federated environment, we demonstrate how FedSDS facilitates effective SDG across multiple nodes. The framework addresses data scarcity and heterogeneity by augmenting limited datasets with high-quality synthetic samples. This approach corresponds to the work presented in the paper:
    - **A. Apellániz P.,** Parras J., and Zazo S., *'Improving Synthetic Data Generation through Federated Learning in Scarce and Heterogeneous Data Scenarios,'* in Big Data and Cognitive Computing, vol. 9(2), 18, 2025, doi: 10.3390/bdcc9020018.
2. **Federated SAVAE (FedSAVAE) for Survival Analysis:** By leveraging the SAVAE model within the FedSDS framework, we extend the capabilities of FL to Survival Analysis (SA). This integration enables robust survival modeling across geographically and demographically diverse institutions, overcoming the challenges of non-IID data distributions. This contribution is documented in the paper:
    - A. Apellániz P., Parras J., and Zazo S., *'Enhancing Survival Analysis Through Federated Learning in Non-IID and Scarce Data Scenarios,'* under review in Computers in Biology and Medicine journal.

The remainder of this chapter is structured as follows. First, we introduce the FedSDS framework, outlining its methodology, the synthetic data aggregation technique, and key advantages over existing FL approaches. We then delve into the specific applications of FedSDS to SDG (FedVAE) and SA (FedSAVAE), presenting detailed methodologies, experimental results, and discussions. By integrating SDG and SA into the FL paradigm, this chapter highlights the revolutionary potential of FedSDS in advancing collaborative research while preserving data privacy and addressing critical challenges in real-world medical datasets.

5.2 Federated Synthetic Data Sharing (FedSDS)

5.2.1 Definition of the Proposed Aggregation Strategy

FL has demonstrated its utility in enabling collaborative model training across decentralized institutions while preserving data privacy. However, traditional FL methods, such as FedAvg, face significant challenges in non-IID settings, where data distributions vary across nodes. This variability often leads to imbalanced node contributions, biased global models, and suboptimal convergence [279], [280]. Differences in data distributions across nodes result in variations in local models, complicating the aggregation process and making it difficult to perform the specific task accurately.

To address these challenges, we propose FedSDS, an innovative FL strategy that leverages SDG instead of parameter-sharing methods to mitigate non-IID data limitations. Our approach aligns with the meta-learning paradigm, particularly DRS [275], which approximates MAML [270]. Unlike MAML, which requires bi-level optimization over multiple tasks, DRS aggregates synthetic data across nodes in a single optimization round, making it more computationally efficient in typical FL environments, where the number of nodes (tasks) is limited. This ability to aggregate diverse synthetic data enables FedSDS to improve generalization, ensuring better model convergence in non-IID FL settings.

The core of FedSDS lies in the VAE-BGM model (Section 4.2), which refines the latent space of a standard VAE by integrating a BGM model. The BGM models the latent representation as a mixture of multiple Gaussian components. Recall that, unlike traditional Gaussian priors, the BGM dynamically adjusts the number of components using a Dirichlet process, enabling the model to capture complex, multi-modal data distributions effectively. This flexibility allows the VAE-BGM to handle mixed data types, including continuous, binary, and categorical variables, ensuring that synthetic datasets accurately reflect the underlying structure of real-world data.

Additionally, in cases where the available data are scarce, FedSDS can leverage the data-scarce generation methodology proposed in Section 4.4, which includes techniques such as pre-training, model averaging, and meta-learning strategies to enhance synthetic data quality under limited-data conditions. This approach ensures that synthetic data are representative, even in resource-constrained settings.

Unlike traditional FL approaches, FedSDS does not rely on model parameter aggregation. Instead, synthetic data are generated locally at one or multiple nodes, depending on data quality and quantity. These generated samples are then shared with selected nodes that require additional data, ensuring fairer and more balanced data distributions across the federation.

The framework consists of the following key steps:

1. **Local Synthetic Data Generation:** Each selected node independently trains a VAE-BGM model using its local data. The generated synthetic data aim to preserve critical patterns and correlations while preventing privacy leakage. In cases of data scarcity, the generation methodology in Section 4.4 can be applied to improve data

representativeness.

2. **Data Sharing and Aggregation:** Once synthetic data are generated, they are shared with other nodes needing additional samples. To address potential biases and ensure effective aggregation, FedSDS employs two distinct strategies:
   - **Random Aggregation (*naive* case):** Synthetic data from other nodes is randomly combined with the local dataset. This straightforward approach provides a baseline for comparison but may introduce biases if the synthetic data significantly differs from the local data.
   - **Similarity-Based Aggregation (*biased* case):** Synthetic samples are filtered based on their proximity to the local data in the latent space, ensuring that only the most relevant samples are integrated. This approach preserves the inherent characteristics of the local dataset while leveraging the diversity of the shared synthetic data.
3. **Training with Augmented Data:** Each node trains its local model using the augmented dataset that combines real and synthetic data. This process improves the generalization ability of the local model by leveraging the additional diversity introduced through data sharing.

A key advantage of FedSDS is its robustness in non-IID scenarios. By sharing synthetic data instead of model parameters, FedSDS accommodates nodes with varying data distributions, including those with missing covariates or demographic biases. Synthetic data from one node can compensate for unavailable features in another, enabling collaborative training even in heterogeneous environments. Additionally, FedSDS significantly reduces communication overhead compared to traditional FL methods. FedAvg typically necessitates multiple iterative rounds of model parameter updates between nodes, which can introduce significant communication overhead. In contrast, FedSDS has the potential to operate in a single communication round, drastically reducing bandwidth consumption. The specific implementation of this strategy varies depending on the study. In FedVAE, multiple rounds of synthetic data sharing are performed to iteratively improve the quality of the generated data iteratively, incorporating better representations in each round. In contrast, FedSAVAE operates with a single communication round, as synthetic data are shared once and directly integrated into the learning process. The following sections will further detail the rationale behind these choices and their implications for model performance. Moreover, FedSDS can share not just synthetic data but also the generative model itself, including the decoder and the BGM-derived parameters, further enhancing communication efficiency. These features make FedSDS a more scalable and bandwidth-efficient FL strategy, particularly in real-world scenarios with communication restrictions.

Figure 5.1 illustrates the FedSDS architecture, where nodes generate and share synthetic data based on local data availability and quality. This method ensures the effective participation of all nodes while preserving data privacy.

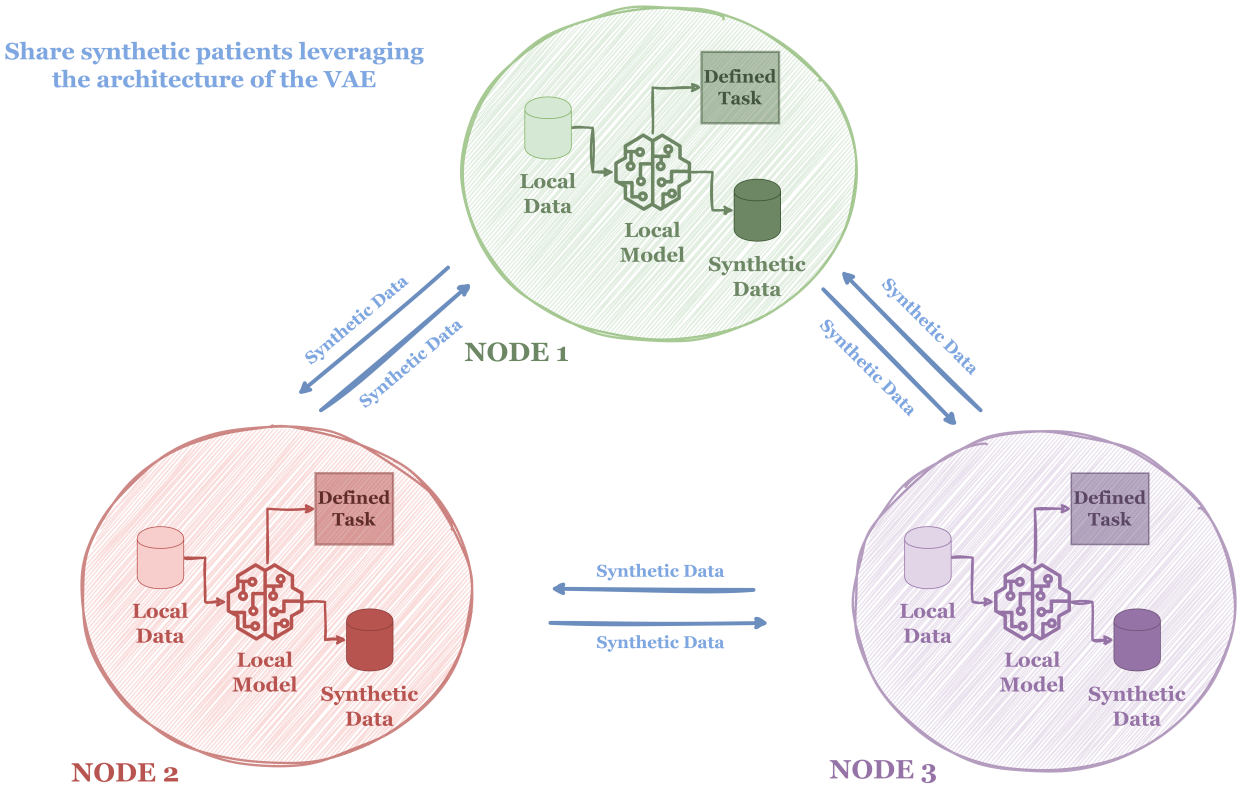


Figure 5.1: Architecture of FedSDS leveraging synthetic data sharing. Each node generates synthetic data locally using VAE-BGM and shares it with others for collaborative training. Each node uses these data to perform the defined task.

Biased Aggregation Strategy

The *biased* aggregation strategy represents a key technical innovation in the FedSDS framework, addressing the challenge of ensuring that synthetic data shared between nodes aligns closely with local data distributions. This alignment is crucial in non-IID settings, where data heterogeneity can lead to suboptimal updates if irrelevant synthetic samples are integrated. For this type of aggregation to be feasible, the model performing the specific task must incorporate the latent space representation characteristic of encoder-decoder architectures, which is essential for the *biased* aggregation strategy.

To achieve this alignment, the following steps are implemented:

1. **Latent Representation Generation:** Each synthetic sample generated at a node is passed through the encoder of the local model at the receiving node (e.g., the encoders of VAE-BGM or SAVAE). This process maps the synthetic data into a latent space, resulting in a vector representation $z_{synthetic}$ for each sample. Similarly, the local dataset at each receiving node is encoded into its latent space representation z_{local}.

2. **Proximity Calculation:** To evaluate the relevance of each synthetic sample, the distance between $z_{synthetic}$ and z_{local} is calculated. This metric quantifies the similarity between the synthetic sample and the local data. The distance is calculated using the

Euclidean norm as defined below:

$$d(z^i_{synthetic}, z^j_{local}) = \|z^i_{synthetic} - z^j_{local}\|, \tag{5.1}$$

where $z^i_{synthetic}$ and z^j_{local} represent the latent space vectors of the i-th synthetic sample and the j-th local sample, respectively. The norm $||\cdot||$ denotes the Euclidean norm. For each synthetic sample $z^i_{synthetic}$, the minimum distance to all local samples z^j_{local} is calculated.

3. **Relevance Filtering:** Based on the calculated distances, only the most relevant synthetic samples with the smallest distances to z_{local} are selected for integration into the local dataset. The synthetic samples are ranked based on their computed d_{min} distances. To filter the most relevant samples, the synthetic points are sorted in ascending order of distance, and the samples with the smallest distances are selected. This ensures that the aggregated synthetic data complements the local data characteristics rather than introducing noise or misaligned distributions.
4. **Dataset Augmentation:** The selected synthetic samples are then integrated into the local dataset, enhancing its size and diversity while maintaining consistency with its original distribution.

By aligning the shared synthetic data with local distributions, the *biased* aggregation strategy minimizes the risk of introducing distributional noise. This alignment accelerates model convergence and enhances predictive performance. Additionally, the use of the encoder-decoder architecture of the SAVAE model allows nodes to dynamically filter synthetic data based on latent space proximity, making this approach adaptable to varying levels of heterogeneity. Moreover, unlike the *naive* aggregation, which can be computationally wasteful, the *biased* approach optimizes the selection process, ensuring that only the most relevant samples are used. By implementing this strategy, FedSDS ensures that the synthetic data exchange preserves the unique data characteristics in each node while benefiting from the diversity introduced by external samples.

Figure 5.2 illustrates the process of *biased* aggregation in detail, showcasing the steps from latent space representation to selecting the most relevant synthetic samples.

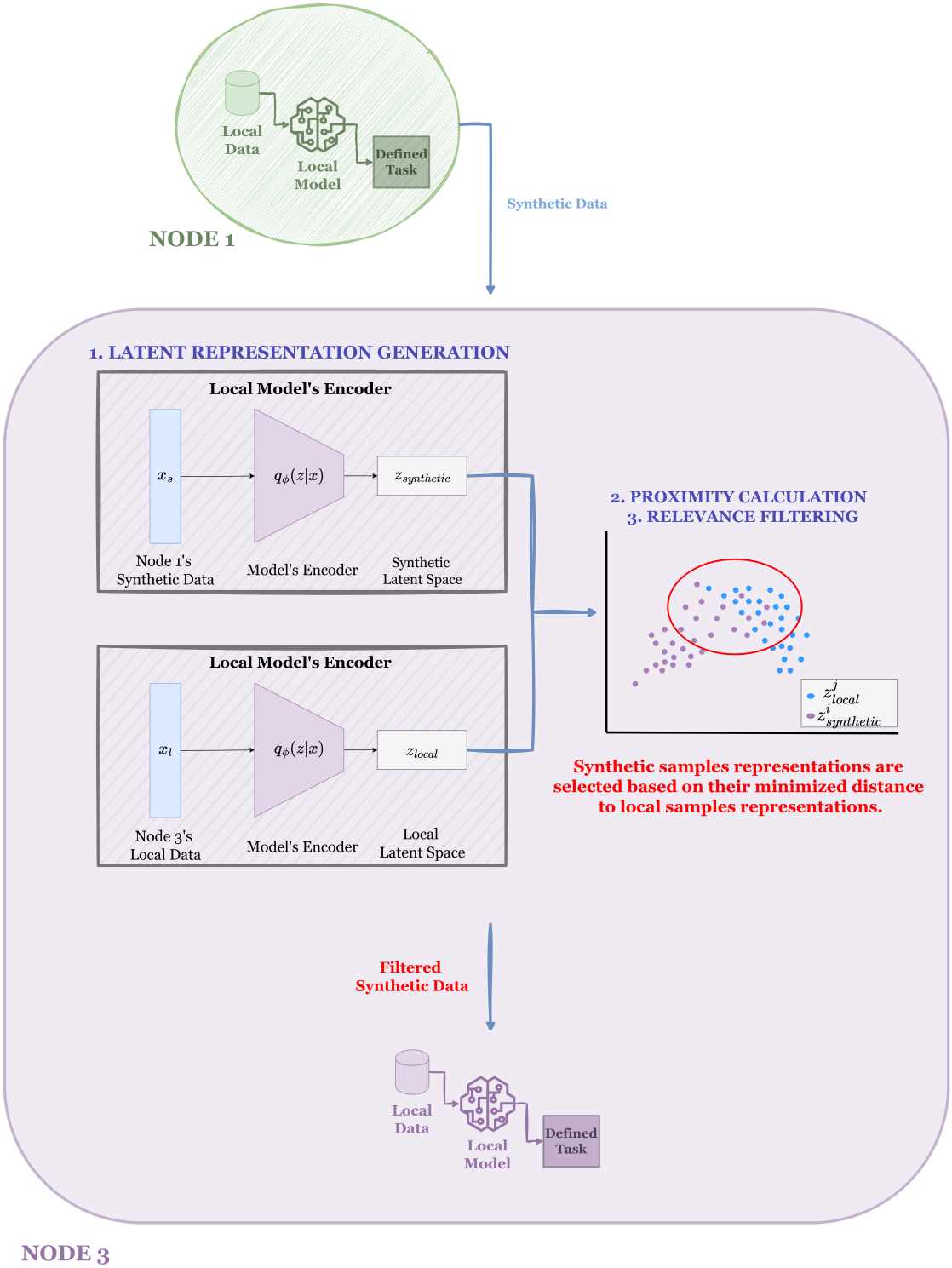


Figure 5.2: Schematic representation of the *biased* aggregation process. Node 1 generates and shares synthetic data with Node 3, where latent representations $z_{synthetic}$ and z_{local} are generated using the encoder of the model. Relevant synthetic samples are selected based on the minimal distance to z_{local} and integrated into the local dataset. *Note: The local model must incorporate an encoder-decoder architecture to enable this aggregation type, as it relies on latent space representations for filtering and integration.*

5.3 Federated Synthetic Data Generation (FedVAE)

As previously discussed, healthcare research faces persistent challenges related to data scarcity, privacy constraints, and the heterogeneity of medical datasets. Stringent privacy regulations often limit traditional approaches to data sharing, while the uneven distribution of data across institutions exacerbates inequities and hinders collaborative advancements. SDG and FL have been identified as promising solutions to address these issues.

Building on these foundations, this study leverages the VAE-BGM model within a federated framework to enable the decentralized generation and sharing of synthetic data. By allowing each institution to generate synthetic datasets locally, FedVAE facilitates collaborative learning while preserving privacy and enhancing the quality and diversity of the data available to each node. This approach is particularly advantageous in non-IID settings, where traditional FL techniques often struggle to address disparities in data distributions across institutions.

FedVAE is designed to tackle the dual challenges of data scarcity and heterogeneity by combining the strengths of FL and VAE-BGM. The framework not only supports high-quality SDG in decentralized environments but also ensures that the generated data captures the critical statistical properties of real-world datasets, making it a powerful tool for advancing medical research and improving equity in data access.

5.3.1 Federated Learning integration in Synthetic Data Generation

In the proposed FL framework, we explore two techniques to train the VAE-BGM models across scarce and heterogeneous data environments: FedAvg and FedSDS. Comparing these two methods aims to clarify their respective advantages in improving SDG under the constraints of FL.

FedAvg has proven effective in many FL settings but can face challenges when applied to non-IID data. Our proposal, FedSDS, can address this issue by sharing synthetic patients generated locally at each node. FedSDS 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.

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 FedSDS a highly efficient option for complex FL scenarios. Figure 5.3 builds upon the framework outlined in Figure 5.1, specifically adapting it to SDG as the defined task and utilizing VAE-BGM as the generative model to produce and share synthetic tabular data across nodes.

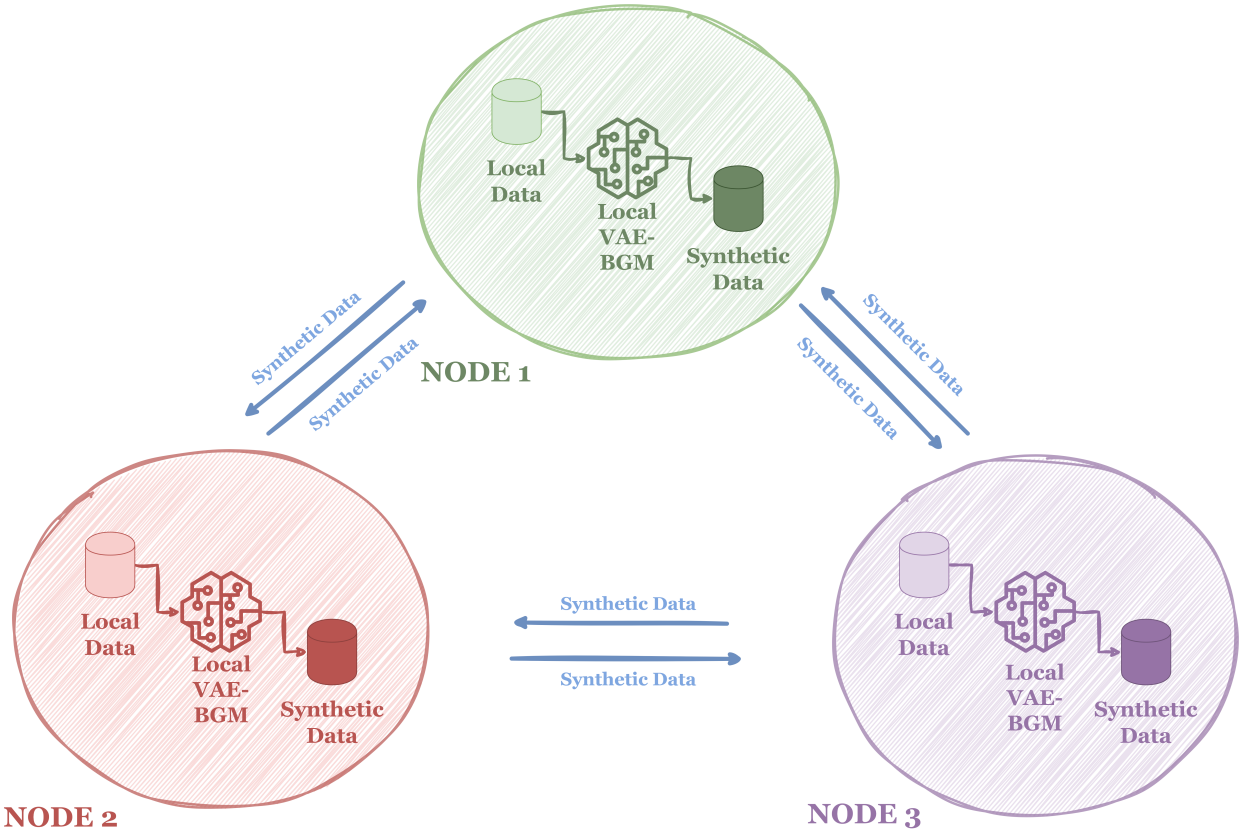


Figure 5.3: FedSDS process in FL for SDG. The FedSDS approach generates synthetic data locally using VAE-BGM at each node and then shares this data across nodes. Nodes incorporate the aggregated synthetic data from other nodes into their local training.

5.3.2 Experiments and Results

Validation Design

*→ **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 0** represents an institution with limited data and resources, having only 100 training samples.
- **Node 1** represents an institution with moderate resources, using $1,000$ training samples.
- **Node 2** 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 $9,500$. 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 distribu-

tions: IID and non-IID.

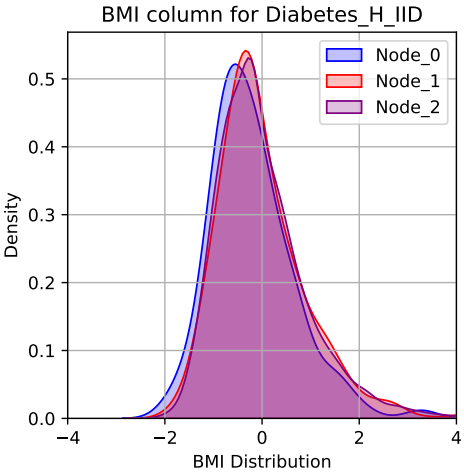
- **IID Scenario:** In this case, the data are split randomly across the three nodes, ensuring each node receives a statistically similar distribution of features. This ensures that the feature distributions are balanced and equal across all nodes.
- **Non-IID Scenario:** To represent more realistic, 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 [281] and heart disease [282], as well as its sensitivity to regional socioeconomic factors like healthcare access and diet quality [283]. 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, Node 0 and Node 1, were provided with skewed distributions, with 90% of samples having BMI values either below or above 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 FedSDS and FedAvg, underscoring the robustness of our approach and practical relevance in a realistic FL scenario.

Figure 5.4 illustrates the KDE of the BMI distributions across the three nodes for both IID and non-IID scenarios. In the IID scenario, 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. Specifically, Node 2 retains a distribution similar to the overall population, while Nodes 0 and 1 distributions are shifted, reflecting the intentional skew in their data. This variation highlights the challenges posed by non-IID settings in FL and underscores the importance of techniques like FedSDS to address such disparities.

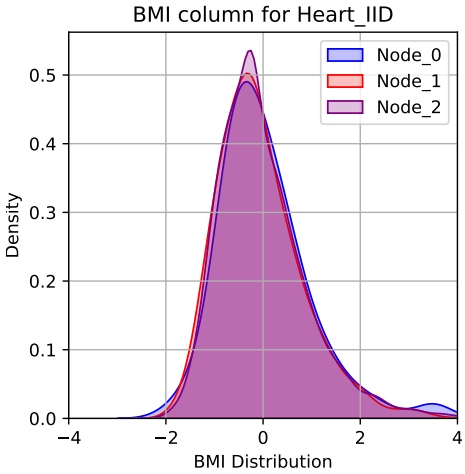
→ Data generation in an FL environment

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 FedSDS, 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.

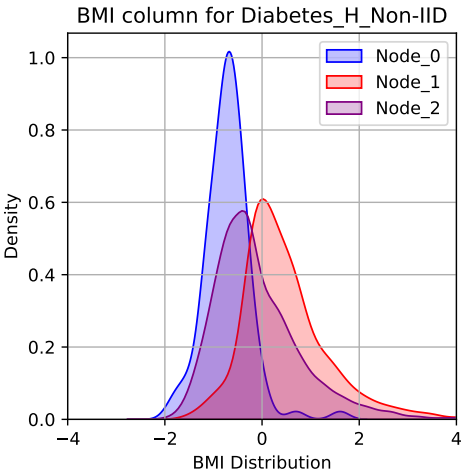
- **FedAvg:** This method aggregates the model weights from each node and updates the local models using a weighted average of these aggregated weights. Each node trains its local model for 200 epochs, after which the weights are shared, aggregated, and redistributed for the next training round. This process is repeated for five rounds.



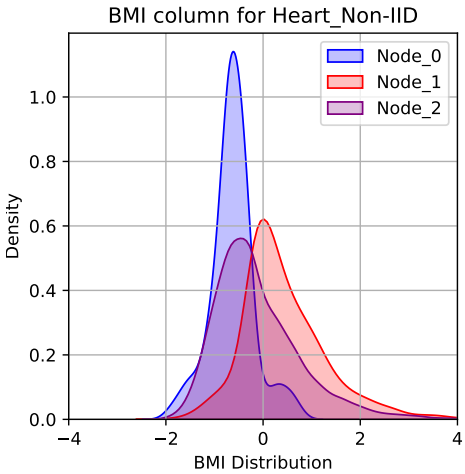
(a) Diabetes_H dataset IID scenario



(b) Heart dataset IID scenario



(c) Diabetes_H dataset non-IID scenario



(d) Heart dataset non-IID scenario

Figure 5.4: KDE plots for BMI distributions across nodes under IID and non-IID scenarios for the Diabetes_H and Heart datasets.

- **FedSDS:** Instead of sharing model parameters, each node shares synthetic data generated locally. The synthetic data are exchanged across the network, enhancing local datasets at other nodes. The two aggregation strategies (*naive* and *biased*) are applied to assess their impact on performance. The process is structured as follows:

  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, nodes share their generated synthetic data with others. The shared data are then integrated using either *naive* or *biased* aggregation, depending on the strategy applied.

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 FedSDS, the increase in data volume and diversity from shared synthetic patients is expected to mitigate the issues related to data scarcity and bias.

Validation Metrics

The primary goal of this study is to generate synthetic data that is indistinguishable from real data as in Section 4.2, 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 [8].

We estimate the $D_{\mathbb{JS}}$ for statistical validation to measure the similarity between the real and synthetic data distributions. Following the methodology outlined in Section 4.3, a probabilistic discriminator network is trained to distinguish between real and synthetic data. Using the output probabilities of this classifier, we approximate the $D_{\mathbb{JS}}$, where lower values indicate greater similarity between the distributions. This approach ensures a robust statistical assessment of the quality of synthetic data.

To assess the effectiveness of different scenarios—specifically, isolated learning, FedAvg, and FedSDS (*naive* and *biased* aggregation strategies) on the $D_{\mathbb{JS}}$ values, we calculate the MRR. In this case, the MRR considers the rank position of the first relevant $D_{\mathbb{JS}}$ value within an ordered list of $D_{\mathbb{JS}}$ values derived from each FL technique and the isolated learning scenario. The RR for each method is calculated as the inverse of the position of the first relevant $D_{\mathbb{JS}}$ 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 on. Recall that the MRR is then computed as the average of the RRs across all situations:

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}, \tag{5.2}$$

where Q denotes the total number of situations under evaluation (isolated, FedAvg, FedSDS *naive*, and FedSDS *biased*), and $rank_i$ represents the position of the first relevant $D_{\mathbb{JS}}$ for the i-th scenario. Higher MRR values imply that relevant $D_{\mathbb{JS}}$ values appear earlier in the list, indicating better performance and 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 applied effectively to real-world clinical tasks is crucial, as confirmed previously. For the clinical utility validation, we assess the practical applicability of the synthetic data in real-world tasks. Specifically, we use an RF classifier to predict target features. Two experimental scenarios

are conducted:

- **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 match that of the first scenario closely. 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 retains the necessary information to perform well in real-world tasks. These metrics offer a holistic evaluation of the similarity and utility of the synthetic data.

We conducted hypothesis testing to validate the effectiveness of our proposed approaches further. For statistical validation, the null hypothesis assumed that isolated training would yield lower $D_{\mathbb{JS}}$ 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.

Experimental Setting

The proposed network leverages the VAE-BGM model for SDG at each node (Section 4.2). 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 ability of the model 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 (FedSDS) are shared and aggregated across nodes. FedAvg aggregates model weights, while FedSDS 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.

Results

We used two medical datasets for these experiments, Heart and Diabetes_H, whose detailed descriptions and characteristics are provided in Appendix B. These datasets were chosen due to their substantial number of samples and diverse feature types, making them ideal for evaluating the robustness of the proposed approach. Their complexity stems from intricate relationships between features and the heterogeneous nature of their data types, which are critical for testing the effectiveness of SDG methods.

For each dataset, we provide a detailed comparison of the $D_{\mathbb{JS}}$ 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 is conducted for IID and non-IID data scenarios. Additional experiments are included in Appendix E.1.2, where we compare key features of real and synthetic data distributions. The code to replicate the results can be found in https://github.com/Patricia-A-Apellaniz/fed_vae.

Note: *It is important to note that divergences are estimations, and small negative values reflect an approximation error, suggesting near-zero divergence.*

→ *IID Scenarios*

The results presented in Table 5.1 for Diabetes_H and Heart datasets confirm the advantages of using FedSDS over isolated training and FedAvg in terms of similarity validation and clinical utility validation. Across all nodes in both datasets, FedSDS consistently outperforms FedAvg and the isolated case in reducing $D_{\mathbb{JS}}$ values. This demonstrates that sharing synthetic data significantly enhances the similarity between real and synthetic distributions, leading to more reliable data generation for FL. The most notable improvement is observed in FedSDS *biased*, where the estimated divergences approach zero, indicating an almost perfect alignment between real and synthetic data distributions. In Nodes 0 and 1, which initially have fewer samples, the reductions in $D_{\mathbb{JS}}$ are particularly substantial, proving that FedSDS can effectively compensate for data scarcity. For Node 2, where the number of real samples is already high, the difference compared to the isolated case is less pronounced, as expected. However, the improvement in FedSDS *biased* still confirms that even well-resourced nodes benefit from the refined data-sharing approach. Regarding accuracy in the clinical utility validation, FedSDS methods consistently match or exceed the performance of FedAvg and isolated training across both datasets. In Nodes 0 and 1, which are more resource-constrained, FedSDS leads to the highest accuracy values, confirming that better similarity validation translates directly into improved model performance. In some cases, FedSDS *naive* and *biased* even surpass the accuracy of models trained on real data, highlighting the potential of synthetic data augmentation to improve generalization and predictive performance. The results demonstrate that FedSDS significantly outperforms traditional FL approaches like FedAvg, particularly when using the *biased* technique.

Node	Technique	SIMILARITY VALIDATION	CLINICAL UTILITY VALIDATION	
		Estimated $D_{\mathbb{JS}}$	Accuracy (Real-Real)	Accuracy (Synthetic-Real)
Node 0	Isolated	0.696 (0.026)	0.844 (0.001)	0.836 (0.001)▼
	FedAvg	0.521 (0.016)*		0.837 (0.000)▼
	FedSDS *naive*	0.269 (0.095)*		0.841 (0.001)*
	FedSDS *biased*	**-0.001 (0.014)***		0.838 (0.001)▼
Node 1	Isolated	0.302 (0.001)	0.842 (0.000)	0.843 (0.003)*
	FedAvg	0.177 (0.007)*		0.845 (0.002)*
	FedSDS *naive*	0.119 (0.017)*		**0.846 (0.001)***
	FedSDS *biased*	**-0.035 (0.076)***		0.845 (0.001)*
Node 2	Isolated	0.107 (0.021)	0.842 (0.001)	0.843 (0.001)*
	FedAvg	0.195 (0.036)		0.846 (0.002)*
	FedSDS *naive*	0.208 (0.061)		0.847 (0.002)*
	FedSDS *biased*	**-0.014 (0.002)***		0.838 (0.002)*

(a) Diabetes_H dataset

Node	Technique	SIMILARITY VALIDATION	CLINICAL UTILITY VALIDATION	
		Estimated $D_{\mathbb{JS}}$	Accuracy (Real-Real)	Accuracy (Synthetic-Real)
Node 0	Isolated	0.850 (0.004)	0.909 (0.000)	0.898 (0.000)▼
	FedAvg	0.493 (0.017)*		0.902 (0.001)▼
	FedSDS *naive*	0.297 (0.010)*		0.902 (0.001)▼
	FedSDS *biased*	**-0.033 (0.007)***		0.899 (0.001)▼
Node 1	Isolated	0.373 (0.018)	0.907 (0.001)	0.908 (0.001)*
	FedAvg	0.205 (0.039)*		0.910 (0.001)*
	FedSDS *naive*	0.142 (0.016)*		**0.912 (0.001)***
	FedSDS *biased*	**0.092 (0.003)***		**0.910 (0.001)***
Node 2	Isolated	0.150 (0.003)	0.906 (0.001)	0.909 (0.001)*
	FedAvg	0.147 (0.002)		**0.914 (0.001)***
	FedSDS *naive*	0.152 (0.007)		**0.912 (0.001)***
	FedSDS *biased*	**-0.017 (0.028)***		0.910 (0.000)*

(b) Heart dataset

Table 5.1: Diabetes_H and Heart results in IID scenario. Comparison of $D_{\mathbb{JS}}$ and accuracy for isolated training and three FL techniques. Lower $D_{\mathbb{JS}}$ 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 (standard deviation)*. * indicates p-value < 0.01. In particular, for $D_{\mathbb{JS}}$, * denotes statistically significant improvement over the isolated case. For accuracy, * signifies that the performance of models trained on synthetic data was comparable to or exceeded that of models trained on real data. **Bold** values indicate the best significative performance, and ▼ denotes a decrease relative to upper bounds.

→ Non-IID Scenarios

The results for the non-IID scenario on the Diabetes_H dataset in Table 5.2 highlight the effectiveness of FedSDS in mitigating the impact of data heterogeneity. In terms of similarity validation, FedSDS, particularly FedSDS *biased*, significantly outperforms FedAvg and isolated training in the worst-performing nodes (Node 0 and Node 1), leading to much lower $D_{\mathbb{JS}}$ values, indicating better alignment between real and synthetic data distributions. FedAvg also improves over the isolated case in Node 0 but remains noticeably less effective than FedSDS. Regarding clinical utility validation, in the best-performing nodes (Nodes 1 and 2), no FL technique fully reaches the accuracy obtained when training with real data. However, FedSDS consistently outperforms isolated training and remains aligned with FedAvg, with FedSDS *biased* achieving the highest mean accuracy in Node 1, where the data distribution is skewed. In Node 2, where real data availability is high, FedSDS *naive* exhibits the best mean accuracy, slightly surpassing FedAvg. Despite not fully matching real-data accuracy, both FedSDS techniques achieve remarkably close results. In Node 0, the most challenging case, all FL techniques achieve the accuracy of the real-data scenario, demonstrating the advantages of collaborative training in data-scarce environments. Notably, FedSDS *biased* achieves the highest mean accuracy, reinforcing its effectiveness in leveraging synthetic data to enhance classification performance.

The results presented in Table 5.2 for the Heart dataset under non-IID conditions follow the same trend observed in the Diabetes_H dataset. Regarding similarity validation, FedSDS significantly improves $D_{\mathbb{JS}}$ in Nodes 0 and 1 compared to isolated training and FedAvg. This improvement is particularly evident in Node 1, where FedAvg fails to provide any significant reduction in divergence compared to the isolated case, while FedSDS achieves a substantial decrease. In Node 2, only FedSDS *biased* improves the similarity validation results over the isolated and FedAvg cases, which is expected since Node 2 already possesses the largest dataset. Across all nodes, the FedSDS *biased* approach consistently yields the lowest divergences, reinforcing its effectiveness in mitigating the challenges posed by data heterogeneity. Regarding clinical utility validation, the results follow a similar pattern to those observed in the Diabetes_H dataset. In Nodes 1 and 2, none of the FL techniques fully match the accuracy obtained when training with real data. However, FedSDS consistently outperforms isolated training and FedAvg, coming very close to real-data accuracy, with differences of only one or two decimal places. Notably, in Node 0, the most challenging case, FedSDS techniques successfully achieve the same accuracy as real-data training, while FedAvg falls short. These findings confirm that FedSDS provides a superior approach in non-IID settings, particularly in resource-limited nodes, by effectively aligning synthetic data distributions with real data and preserving the clinical utility of synthetic datasets.

Node	Technique	SIMILARITY VALIDATION	CLINICAL UTILITY VALIDATION	
		Estimated $D_{\mathbb{JS}}$	Accuracy (Real-Real)	Accuracy (Synthetic-Real)
Node 0	Isolated	0.829 (0.006)	0.898 (0.000)	0.887 (0.000)▼
	FedAvg	0.463 (0.003)*		0.898 (0.001)*
	FedSDS *naive*	0.328 (0.011)*		0.897 (0.001)*
	FedSDS *biased*	**0.148 (0.039)***		0.901 (0.001)*
Node 1	Isolated	0.382 (0.019)	0.899 (0.001)	0.788 (0.001)▼
	FedAvg	0.215 (0.062)		0.789 (0.001)▼
	FedSDS *naive*	0.174 (0.001)*		0.789 (0.003)▼
	FedSDS *biased*	**0.092 (0.029)***		0.794 (0.001)▼
Node 2	Isolated	0.173 (0.007)	0.899 (0.001)	0.853 (0.002)▼
	FedAvg	0.260 (0.006)▼		0.856 (0.001)▼
	FedSDS *naive*	0.172 (0.003)		0.860 (0.001)▼
	FedSDS *biased*	**-0.030 (0.004)***		0.856 (0.001)▼

(a) Diabetes_H dataset

Node	Technique	SIMILARITY VALIDATION	CLINICAL UTILITY VALIDATION	
		Estimated $D_{\mathbb{JS}}$	Accuracy (Real-Real)	Accuracy (Synthetic-Real)
Node 0	Isolated	0.807 (0.002)	0.925 (0.001)	0.906 (0.001)▼
	FedAvg	0.526 (0.011)*		0.919 (0.001)▼
	FedSDS *naive*	0.243 (0.013)*		0.923 (0.000)*
	FedSDS *biased*	**0.043 (0.002)***		0.925 (0.001)*
Node 1	Isolated	0.371 (0.005)	0.926 (0.001)	0.886 (0.001)▼
	FedAvg	0.294 (0.030)		0.884 (0.000)▼
	FedSDS *naive*	0.179 (0.003)*		0.886 (0.002)▼
	FedSDS *biased*	**0.022 (0.005)***		0.889 (0.002)▼
Node 2	Isolated	0.169 (0.015)	0.924 (0.001)	0.918 (0.001)▼
	FedAvg	0.164 (0.012)		0.917 (0.001)▼
	FedSDS *naive*	0.170 (0.004)		0.918 (0.001)▼
	FedSDS *biased*	**-0.013 (0.019)***		0.919 (0.001)▼

(b) Heart dataset

Table 5.2: Diabetes_H and Heart results in non-IID scenario. Comparison of $D_{\mathbb{JS}}$ and accuracy for isolated training and three FL techniques. Lower $D_{\mathbb{JS}}$ 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 (standard deviation)*. * indicates p-value < 0.01. In particular, for $D_{\mathbb{JS}}$, * denotes statistically significant improvement over the isolated case. For accuracy, * signifies that the performance of models trained on synthetic data was comparable to or exceeded that of models trained on real data. **Bold** values indicate the best significative performance, and ▼ denotes a decrease relative to upper bounds.

→ *Discussion*

The results presented in Table 5.3 emphasize the exceptional robustness of FedSDS across both IID and non-IID scenarios. Regardless of the data distribution setting, FedSDS consistently outperforms isolated training and FedAvg, demonstrating its effectiveness in improving model performance. However, the FedSDS *biased* strategy stands out overwhelmingly, delivering the best results in every scenario.

Scenario	Dataset	Isolated	FedAvg	FedSDS *naive*	FedSDS *biased*
IID	Diabetes_H	0.333	0.333	0.417	**1.000**
	Heart	0.333	0.389	0.500	**1.000**
Non-IID	Diabetes_H	0.333	0.333	0.500	**1.000**
	Heart	0.278	0.389	0.417	**1.000**

Table 5.3: MRR values across different scenarios (IID and Non-IID) and datasets for isolated, FedAvg and FedSDS (*naive* and *biased*) approaches. Higher MRR values indicate superior performance. **Bold** values denote best performances.

The FedSDS *biased* approach dominates in every case, achieving perfect MRR values (1.000) across all datasets and scenarios. This demonstrates that, by strategically selecting the most relevant synthetic samples, FedSDS *biased* optimally aligns the generated synthetic data with real data distributions, maximizing its effectiveness in federated settings. Even FedSDS *naive* surpasses FedAvg and isolated training, reinforcing that synthetic data sharing is inherently more beneficial than parameter aggregation in FL environments. However, the *biased* approach far surpasses all others, proving to be the most robust and effective method for tackling data heterogeneity. The advantages of FedSDS become even more pronounced in non-IID settings, where data distribution imbalances introduce significant challenges. While FedAvg provides some improvements over isolated training, it fails to fully address data heterogeneity, particularly in more challenging nodes. In contrast, FedSDS effectively mitigates the negative effects of non-IID data by leveraging diverse synthetic data distributions, allowing models to generalize better and achieve superior performance across varying conditions.

Beyond improving model performance, we emphasize the security and privacy-preserving benefits of synthetic data generation within the FL framework. As outlined in Appendix E.1.1, 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 2 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, FedSDS 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.

5.3.3 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 FedSDS, particularly FedSDS *biased*, consistently outperforms traditional approaches like FedAvg and isolated training in both IID and non-IID scenarios. The results show significant advantages in FedSDS for generating high-quality synthetic data, as reflected in lower D_{JS} values. Clinical utility validation confirms the practicality of synthetic data generated using FedSDS, achieving comparable accuracy to real data in downstream tasks. In non-IID environments, FedSDS 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 the effectiveness of FedSDS, particularly in nodes with constrained data availability or skewed distributions. These findings highlight the potential of sharing synthetic data within FL frameworks. FedSDS enables improved model generalization and supports collaborative research without exposing sensitive patient information by fostering data diversity and reducing disparities between data-rich and data-poor nodes. 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 research should explore optimizing FL architectures for even more complex data types and more extensive networks of institutions and further refining the integration of SDG with FL to maximize efficiency and scalability. In particular, exploring SDG in low-sample settings 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. This would ensure that even nodes with minimal data could generate synthetic datasets with robust statistical similarity and practical utility, further improving the performance of FL in heterogeneous healthcare settings. In addition, addressing privacy risks must be a future line of research on this topic. Investigating techniques to mitigate privacy risks associated with FL, such as DP [284] or HE [285], 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.

5.4 Federated Survival Analysis (FedSAVAE)

By now, it has already been explained several times that SA is a critical tool in healthcare for understanding time-to-event outcomes, such as patient survival or disease progression. However, real-world SA datasets often suffer from limitations, including data scarcity, heterogeneity, and high levels of censoring, which hinder the development of robust and generalizable predictive models. These challenges are further compounded by privacy regulations and institutional barriers, which restrict the centralization of patient data for collaborative research.

To address these limitations, we propose FedSAVAE, an FL framework that integrates the advanced generative capabilities of SAVAE into a decentralized training environment. FedSAVAE leverages the latent space representation of SAVAE to perform survival modeling across multiple nodes, enabling collaboration between institutions without sharing sensitive patient data. By exchanging synthetic data tailored to the characteristics of the dataset in each node, FedSAVAE ensures that local models can benefit from the diversity and richness of external data while preserving privacy.

This framework is particularly designed to address non-IID scenarios, where data distributions vary significantly across institutions. Traditional FL techniques often struggle in such environments, leading to biased models or suboptimal performance. By incorporating synthetic data sharing and utilizing the encoder-decoder architecture of SAVAE, FedSAVAE overcomes these limitations, enabling accurate and scalable survival modeling in decentralized and privacy-sensitive settings. This approach addresses the challenges of data scarcity and heterogeneity and ensures the equitable advancement of SA methodologies across diverse and resource-constrained institutions.

5.4.1 Federated Learning integration in Survival Analysis

In line with the approach outlined for FedVAE, this section evaluates the integration of SA into an FL framework by comparing two techniques: FedAvg and FedSDS. Both methods are applied to survival modeling using the SAVAE architecture, leveraging its latent space representation for handling censored and non-linear data relationships.

The comparison aims to assess the effectiveness of these techniques in addressing the unique challenges posed by non-IID datasets, which are common in real-world SA scenarios. Figure 5.5 compares both setups.

As mentioned, a key advantage of FedSDS over FedAvg is its communication efficiency. Traditional FL methods like FedAvg require multiple iterative rounds of parameter updates and aggregations to achieve convergence, which can be resource-intensive in bandwidth-constrained environments. In this research, FedSDS operates in a single communication round, wherein nodes can share synthetic data or their generative models (including decoders and BGM parameters) with others. Sharing the generative model allows nodes to generate as much synthetic data as needed locally, enabling greater flexibility in addressing data scarcity and enhancing local model training. This single-round approach drastically reduces communication overhead while preserving model performance, making it particularly advantageous in settings with limited communication resources. FedSDS accelerates model convergence and enhances

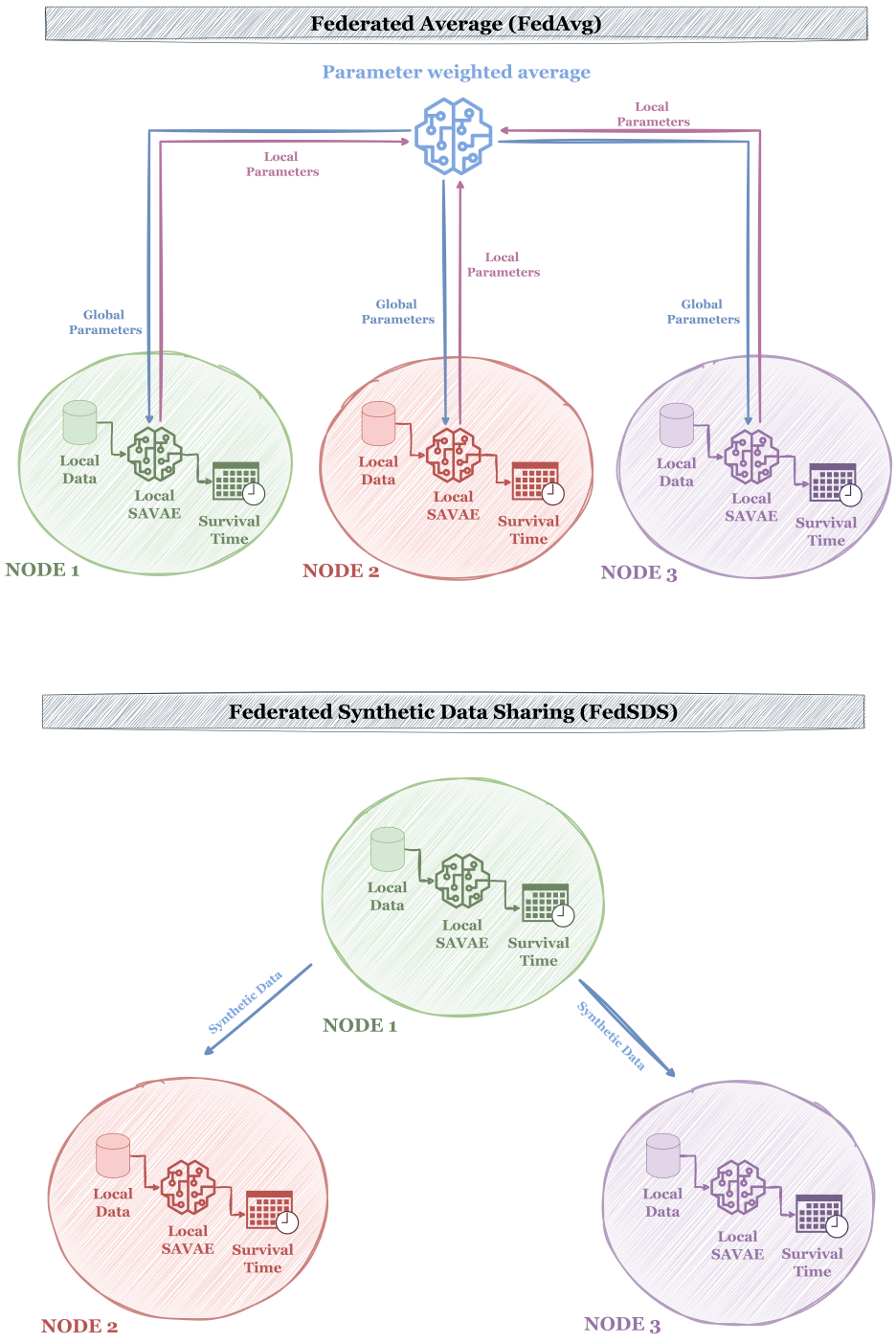


Figure 5.5: Comparison of FL techniques for SA. FedAvg (Top) aggregates local model parameters from each node into a global model through iterative communication rounds. FedSDSD (bottom) enables nodes to share locally generated synthetic data, reducing communication overhead. In FedSDS, it is assumed that Node 1 is the 'best' node, which shares synthetic samples with the other two nodes. Both approaches use SAVAE models trained locally for survival time prediction and covariate reconstruction.

scalability for practical FL deployments by eliminating the need for iterative exchanges. In addition, FedSDS is particularly effective in non-IID environments, where nodes may have different covariates or missing features. By sharing synthetic data tailored to the local characteristics of each node, FedSDS allows nodes to integrate data that complement their own, ensuring robust and unbiased updates to the global model. For example, synthetic data from one node can help another node predict unavailable covariates, enhancing the performance of its local model. FedAvg, by contrast, cannot address these disparities, often resulting in suboptimal global models in heterogeneous settings.

5.4.2 Experiments and Results

Validation Design

This study evaluates the proposed methodology using distributed data across three nodes, simulating IID and non-IID scenarios. The experimental design is structured to assess the performance of the proposed framework under diverse data distribution settings and varying degrees of heterogeneity. Given the small sample sizes of the original datasets, synthetic data was generated using the methodology proposed in Section 4.2 to ensure sufficient data to design the various nodes and scenarios. Below, we describe the distribution of data across nodes and the specific scenarios considered for the different cases.

Case	Scenario	Data Distribution	Node 1	Node 2	Node 3	Data Quality
IID	**Scenario 1**	Equal	2000	2000	2000	Homogeneous
IID	**Scenario 2**	Unequal	2000	500	50	Homogeneous
IID	**Scenario 3**	Unequal with missing data	2000	500	25 (50% missing)	Homogeneous
Non-IID	**Scenario 4**	Equal	2000	2000	2000	Node 2: 95% below median age, Node 3: 95% above median age
Non-IID	**Scenario 5**	Unequal	2000	500	50	Node 2: 95% below median age, Node 3: 95% above median age
Non-IID	**Scenario 6**	Unequal with missing data	2000	500	25 (50% missing)	Node 2: 95% below median age, Node 3: 95% above median age
Non-IID	**Scenario 7**	Two nodes, Node 2 lacks *age* covariate	3500	3500	N/A	Node 2: *age* covariate predicted from the synthetic data in Node 1

Table 5.4: Overview of scenarios defined for the experimental design in FedSAVAE. Data quantity and quality across nodes under IID and non-IID conditions are detailed. Node 1, Node 2, and Node 3 refer to training data allocated across the nodes in each scenario.

Three nodes, each differing in the quantity of data they possess, are used for the experiments. The study examines two primary cases: IID and non-IID data distributions. In each case, three distinct scenarios are defined to reflect varying data allocation strategies and the introduction of heterogeneity. Additionally, we include a special non-IID scenario designed to test the

ability of the FedSDS framework to handle missing covariates. Table 5.4 provides a structured summary of all defined scenarios, detailing data distribution and quality across nodes.

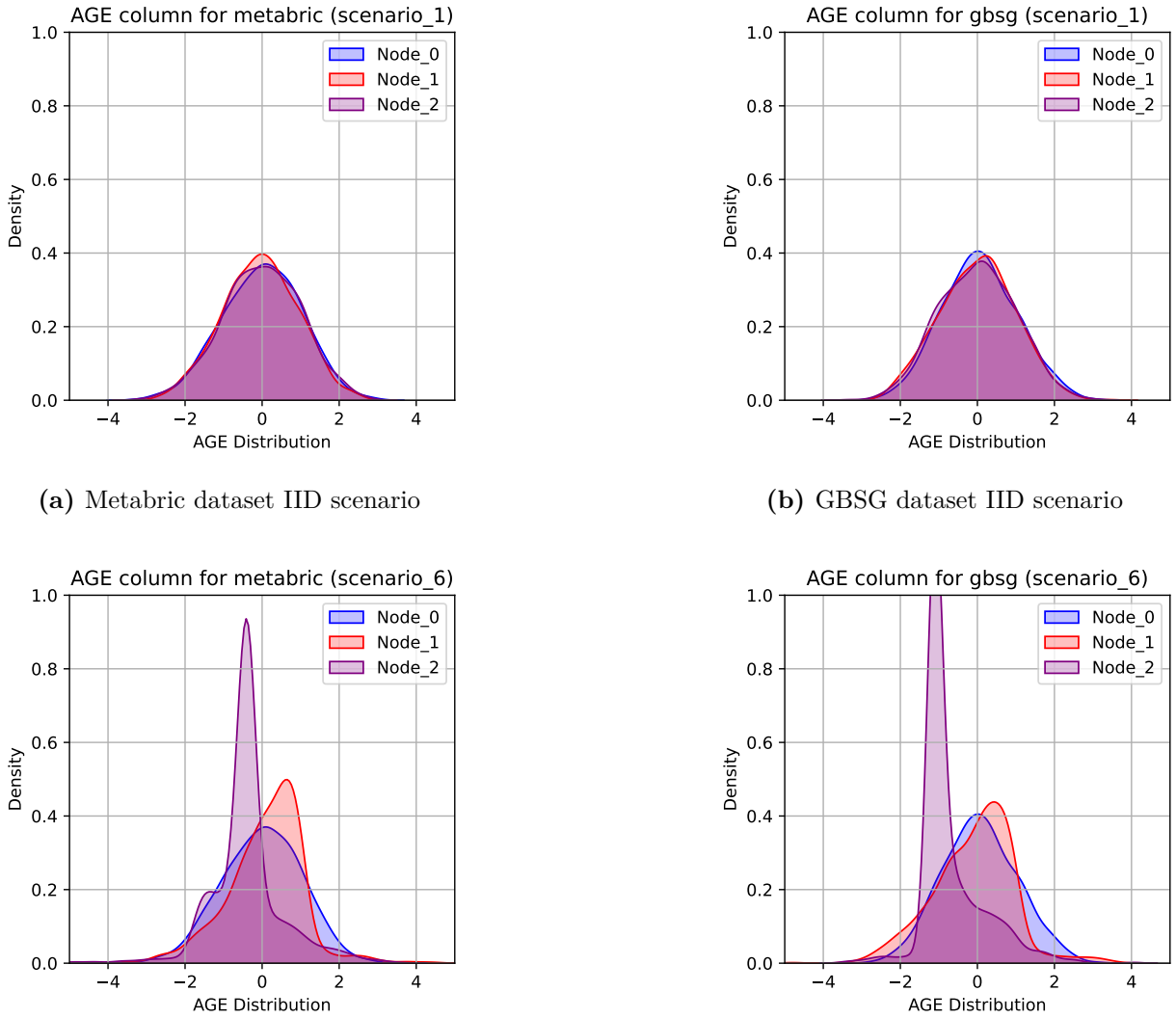
→ IID Data Distribution
In the IID setting, the data across nodes follow similar distributions regarding covariates and event times. Three distinct scenarios are defined:

- **Scenario 1. Equal data distribution across nodes:** Each of the three nodes is allocated 3,000 samples, with 2,000 samples used for training and 1,000 for validation. This scenario represents the most balanced and favorable case for FL, ensuring equal contributions from all nodes.
- **Scenario 2. Unequal data distribution across nodes:** The nodes have differing quantities of samples: Node 1 has 3,000 samples, Node 2 has 1,500 samples, and Node 3 has 1,050 samples. Validation sets of 1,000 samples are maintained for all nodes, leaving 2,000, 500, and 50 samples, respectively, for training. This scenario introduces imbalances in data quantity between nodes, simulating real-world conditions.
- **Scenario 3. Unequal data distribution with missing data:** This scenario builds on distribution of samples from Scenario 2 but introduces missing data in Node 3. Specifically, 50% of the data are missing. This scenario highlights the challenge of handling incomplete datasets in FL.

→ Non-IID Data Distribution
In the non-IID setting, the data across nodes are heterogeneous regarding covariate distributions. Three scenarios analogous to the IID case are defined but with added heterogeneity in the distribution of a key covariate, *regarding*, which is critical for SA.

- **Scenario 4. Equal data distribution with heterogeneous covariates:** Each node is allocated 2,000 training and 1,000 validation samples. However, heterogeneity is introduced in the *age* covariate:
  - **Node 1:** Uniform distribution of *age* samples.
  - **Node 2:** 95% of samples have ages below the median, with only 5% above the median.
  - **Node 3:** 95% of samples have ages above the median, with only 5% below the median.
- **Scenario 5. Unequal data distribution with heterogeneous covariates:** This scenario mirrors Scenario 2 regarding data quantity across nodes but incorporates the same heterogeneity in the age covariate as Scenario 4.
- **Scenario 6. Unequal data distribution with missing data and heterogeneous covariates:** This scenario extends Scenario 3 by introducing heterogeneity in the age covariate for Nodes 2 and 3, as defined in Scenario 4. Node 3 also retains 50% missing data, making this the most challenging non-IID scenario for both FL techniques.
- **Scenario 7. Addressing missing covariates:** This scenario cannot be evaluated

using traditional parameter-sharing methods like FedAvg and is designed specifically to test the capabilities of FedSDS. Two nodes are used, each with 3,500 training and 1,000 validation samples. However, Node 2 lacks a crucial covariate—*age*—significantly impacting SA model performance. Using FedSDS, synthetic data generated by Node 1 is employed to train a predictor for the missing *age* column in Node 2. This predicted column is then integrated into the dataset from Node 2, allowing the FedSDS framework to handle this non-IID scenario effectively and demonstrate its adaptability to situations with fully missing covariates.



(a) Metabric dataset IID scenario

(b) GBSG dataset IID scenario

(c) Metabric dataset non-IID scenario

(d) GBSG dataset non-IID scenario

Figure 5.6: KDE plots for AGE distributions across nodes under IID and non-IID scenarios for the Metabric and GBSG datasets.

Figure 5.6 illustrates the KDE plots for the age distributions in both IID (Scenario 1) and non-IID (Scenario 6) settings for the Metabric and GBSG datasets. In the IID scenario, shown

in subfigures (a) and (b), the distributions of the age variable across nodes are relatively similar, as expected, since data are split randomly across the three nodes. However, in the non-IID scenario (subfigures (c) and (d)), the distributions vary significantly between nodes, reflecting the intentional skew in the dataset. Specifically, Nodes 1 and 2 exhibit clear shifts in age distribution, aligning with the experimental design where one node contains predominantly younger individuals (below the median age), and another consists mainly of older individuals (above the median age).

Validation Metrics

As defined in Section 2.2.5 and used in Chapter 3, the C-index is a key metric in SA that evaluates the ability of a model to rank predicted risks relative to observed event times, measuring the probability that higher predicted risks correspond to shorter time-to-event. These experiments are validated using the time-dependent C-index version to account for dynamic changes in risk over time.

Additionally, hypothesis testing was conducted to compare the mean C-index values of the FL cases against the isolated cases to ensure robust evaluation. The null hypothesis assumes that isolated training would yield worse C-index values (indicating worse performance of isolated training) than FL techniques. The validity of this hypothesis was assessed using p-values, with a significance threshold set at 0.05. A p-value below this threshold led to the rejection of the null hypothesis, indicating statistically significant superiority of the FL techniques. Conversely, a p-value exceeding 0.05 suggested no significant difference between the cases. This statistical approach ensured a comprehensive evaluation of model performance by accounting for variations in the metrics across different experiments.

Given the multiple hypothesis tests conducted, the risk of Type I errors (false positives) increases with the number of tests, as noted in [229]–[231]. To mitigate this, the Holm adjustment of the p-values [232] was applied, effectively controlling FWER inflation and ensuring the reliability of the statistical conclusions.

Experimental Setting

→*Local SAVAE Model at Each Node*

The architecture of the SAVAE model implemented at each node follows the design outlined in Section 3.2. The model comprises three different DNNs: one encoder and two decoders. The encoder maps the input data to a Gaussian latent space while the decoders reconstruct covariates and time parameters. The encoder features a simple architecture consisting of a single hidden layer with 50 neurons and a rectified linear unit activation function. The output layer applies a hyperbolic tangent activation function to generate the Gaussian latent space. The encoder processes covariate vectors from the training dataset, projecting them into a latent space of a fixed dimensionality ($d_z = 5$). The latent representation generated by the encoder serves as input to both decoders. Each decoder comprises two linear layers. The first layer employs 50 neurons, a ReLU activation function, and a dropout rate of 20% to mitigate overfitting. The second layer features activation functions tailored to the specific distributions of the covariates or time parameters, ensuring output suitability. Early stopping

is applied during training to prevent overfitting, with a batch size of 250. Furthermore, due to the inherent variability in VAE initialization, the SAVAE model is trained using ten random seeds for robustness.

→Synthetic Data Generation with VAE-BGM
We employ the VAE-BGM model for SDG as described in Section 4.2. This model introduces a BGM in the VAE framework, enhancing the flexibility and expressiveness of the latent space. The VAE-BGM model architecture mirrors that of the SAVAE, featuring an encoder with a hidden ReLU layer of 50 neurons and a hyperbolic tangent output layer alongside a decoder that adapts activation functions to match the distributions of the covariates. The latent space dimensionality is set to $d = 5$, balancing feature representation and model complexity. Dropout with a 20% rate is incorporated to mitigate overfitting, and the model is trained for up to 10,000 epochs with early stopping, using a batch size of 250. A key enhancement of VAE-BGM is its treatment of the latent space as a mixture of Gaussian distributions. This is achieved using a Dirichlet process prior with a maximum of 20 components, allowing the model to adjust to the complexity of the underlying data dynamically. Each Gaussian component is parameterized with its covariance matrix, enabling the model to capture intricate dependencies and non-Gaussian structures in the data effectively. To further refine the performance of the VAE-BGM in low-data scenarios, we integrate the model averaging technique proposed in Section 4.2. This approach leverages multiple training runs (10 seeds in our implementation) with different initializations to enhance robustness.

→Federated Learning Framework
In the FL framework, the two techniques, FedAvg and FedSDS, are applied differently:

- **FedAvg**: This technique involves iterative training with five federated steps. During each step, nodes share their locally updated SAVAE model weights for aggregation, followed by a global update. This process ensures a progressively refined global model but incurs higher communication costs due to the multiple rounds of information exchange.
- **FedSDS**: In contrast, in FedSDS, each node trains the VAE-BGM locally to generate synthetic data. The generated synthetic datasets are shared only in the first round of FL, eliminating the need for iterative communication rounds. This approach significantly reduces communication overhead, making FedSDS a scalable and efficient solution for real-world FL applications.

To evaluate the performance of the SA models in each node under the isolated case and both FL techniques, the final results are averaged over the three best runs with different random seeds. This approach accounts for the sensitivity of VAEs to initialization, ensuring a robust and comprehensive evaluation of the models.

Results

This section presents the C-index performance comparisons across various scenarios for IID and non-IID distributions using two datasets: Metabric and GBSG, which are described in detail in Appendix B. Additional results using the IBS, which evaluates calibration and discrimination, are included in Appendix E.2. The code for reproducing all experiments is

available at https://github.com/Patricia-A-Apellaniz/fed_savae.

→ *IID Scenarios*

The results below compare the performance of various approaches: isolated node training, FedAvg, and the proposed FedSDS framework under both naive and biased synthetic data aggregation strategies. The results are shown for three distinct scenarios across three nodes, including centralized and isolated settings. Each cell reports the C-index values in the format (lower bound - mean - upper bound), reflecting the variability of results across multiple runs. Additionally, adjusted p-values are provided to evaluate the statistical significance of the differences between FL methods. Significant values (less than 0.05) are highlighted in bold. This analysis is replicated across all datasets used in the study to ensure a comprehensive evaluation.

The results for the Metabric and GBSG datasets, presented in Table 5.5 and Table 5.6, highlight distinct patterns across scenarios. In Metabric, Scenario 1, with equal data distribution, shows no significant improvement from FL methods. In contrast, Scenario 2 demonstrates that the disadvantaged Nodes 2 and 3 achieve substantial C-index gains exclusively with the FedSDS *biased* approach, matching centralized upper-bound results. Scenario 3 highlights significant improvements for Node 2 with FedSDS methods, particularly the *biased* strategy, and for Node 3 across all FL techniques, again favoring FedSDS *biased*. In GBSG, Scenario 1 reveals significant gains in Node 3, attributed to using FedAvg and including synthetic data. Scenario 2 sees improvements for Node 2 with FedAvg and FedSDS *biased*, while Node 3 benefits from all FL techniques, with FedSDS *biased* yielding the strongest results. Scenario 3 follows a similar trend, with Node 2 improving exclusively with FedSDS methods and Node 3 showing gains across all techniques, again led by FedSDS *biased*.

Scenario	Nodes	Isolated	FedAvg	FedSDS *naive*	FedSDS *biased*	Adjusted p-values
Centralized	**Node 1**	(0.619 - 0.647 - 0.673)	-	-	-	-
	Node 1	(0.606 - 0.637 - 0.666)	(0.611 - 0.638 - 0.666)	(0.613 - 0.641 - 0.668)	(0.609 - 0.640 - 0.671)	1.000 / 0.386 / 1.000
Scenario 1	**Node 2**	(0.621 - 0.648 - 0.676)	(0.625 - 0.652 - 0.681)	(0.617 - 0.646 - 0.676)	(0.623 - 0.651 - 0.679)	0.214 / 1.000 / 0.494
	Node 3	(0.621 - 0.649 - 0.677)	(0.620 - 0.648 - 0.677)	(0.619 - 0.647 - 0.677)	(0.620 - 0.652 - 0.681)	1.000 / 1.000 / 1.000
	Node 1	(0.613 - 0.641 - 0.669)	(0.608 - 0.635 - 0.662)	(0.611 - 0.638 - 0.666)	(0.612 - 0.640 - 0.669)	1.000 / 1.000 / 1.000
Scenario 2	**Node 2**	(0.576 - 0.608 - 0.644)	(0.589 - 0.621 - 0.658)	(0.583 - 0.628 - 0.673)	(0.611 - 0.640 - 0.673)	0.122 / 0.244 / **0.000**
	Node 3	(0.587 - 0.621 - 0.651)	(0.598 - 0.629 - 0.660)	(0.595 - 0.634 - 0.675)	(0.614 - 0.645 - 0.678)	0.185 / 0.371 / **0.001**
	Node 1	(0.608 - 0.638 - 0.666)	(0.612 - 0.640 - 0.669)	(0.609 - 0.637 - 0.663)	(0.607 - 0.636 - 0.665)	0.712 / 1.000 / 1.000
Scenario 3	**Node 2**	(0.577 - 0.610 - 0.641)	(0.589 - 0.622 - 0.659)	(0.592 - 0.631 - 0.667)	(0.614 - 0.644 - 0.673)	0.244 / **0.025** / **0.000**
	Node 3	(0.492 - 0.542 - 0.580)	(0.561 - 0.593 - 0.622)	(0.557 - 0.590 - 0.623)	(0.568 - 0.601 - 0.637)	**0.006** / **0.004** / **0.001**

Table 5.5: C-index comparison of isolated, FedAvg, and FedSDS (*naive* and *biased*) methods for the Metabric dataset in IID scenarios. Average C-index results are shown with CIs. Adjusted p-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in **bold**.

Scenario	Nodes	Isolated	FedAvg	FedSDS *naive*	FedSDS *biased*	Adjusted p-values
Centralized	**Node 1**	(0.660 - 0.688 - 0.714)	-	-	-	-
	Node 1	(0.643 - 0.673 - 0.702)	(0.648 - 0.674 - 0.699)	(0.646 - 0.673 - 0.698)	(0.646 - 0.673 - 0.700)	1.000 / 1.000 / 1.000
Scenario 1	**Node 2**	(0.648 - 0.677 - 0.706)	(0.654 - 0.681 - 0.707)	(0.649 - 0.681 - 0.711)	(0.650 - 0.684 - 0.715)	0.190 / 0.402 / 0.248
	Node 3	(0.658 - 0.686 - 0.714)	(0.666 - 0.693 - 0.721)	(0.661 - 0.691 - 0.720)	(0.664 - 0.693 - 0.719)	**0.016** / 0.181 / **0.015**
	Node 1	(0.646 - 0.672 - 0.697)	(0.647 - 0.674 - 0.701)	(0.646 - 0.673 - 0.699)	(0.645 - 0.673 - 0.701)	0.107 / 0.262 / 1.000
Scenario 2	**Node 2**	(0.620 - 0.650 - 0.680)	(0.628 - 0.658 - 0.686)	(0.624 - 0.656 - 0.686)	(0.634 - 0.664 - 0.695)	**0.030** / 0.248 / **0.008**
	Node 3	(0.563 - 0.601 - 0.649)	(0.627 - 0.666 - 0.710)	(0.633 - 0.668 - 0.696)	(0.650 - 0.683 - 0.714)	**0.000** / **0.001** / **0.000**
	Node 1	(0.646 - 0.673 - 0.698)	(0.647 - 0.674 - 0.700)	(0.643 - 0.672 - 0.699)	(0.644 - 0.671 - 0.697)	0.248 / 1.000 / 1.000
Scenario 3	**Node 2**	(0.617 - 0.649 - 0.679)	(0.625 - 0.656 - 0.685)	(0.625 - 0.662 - 0.692)	(0.633 - 0.666 - 0.701)	0.181 / **0.045** / **0.012**
	Node 3	(0.502 - 0.545 - 0.603)	(0.565 - 0.597 - 0.630)	(0.580 - 0.609 - 0.638)	(0.578 - 0.609 - 0.647)	**0.014** / **0.010** / **0.005**

Table 5.6: C-index comparison of isolated, FedAvg, and FedSDS (*naive* and *biased*) methods for the GBSG dataset in IID scenarios. Average C-index results are shown with CIs. Adjusted p-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in **bold**.

→ *Non-IID Scenarios*

The following results are depicted as in the IID scenarios. C-index values are presented comparing isolated, FedAvg, and FedSDS methods (*naive* and *biased*) for the different datasets under non-IID settings. In these scenarios, Nodes 2 and 3 exhibit biased distributions in the covariate *age* while maintaining the same number of samples as the corresponding IID scenarios. Specifically, the age distributions are deliberately skewed across nodes to introduce heterogeneity, creating a realistic challenge for FL techniques.

Table 5.7 and Table 5.8 present the results for Metabric and GBSG in non-IID scenarios, highlighting consistent patterns in the most disadvantaged nodes. No significant improvements are observed in Scenario 4 of Metabric despite the biased *age* distributions in Nodes 2 and 3. However, in Scenarios 5 and 6, both nodes benefit significantly. Node 2 improves with FedAvg and FedSDS *biased* in Scenario 5, while Node 3 achieves gains across all FL techniques, with FedSDS *biased* delivering the strongest results. A similar trend appears in Scenario 6, where Node 3 significantly improves all methods, but FedSDS *biased* yields the most substantial gain, while Node 2 improves only with this approach. For GBSG, Scenario 4 shows gains in Node 3 using FedAvg and FedSDS *biased*, but Node 2 sees improvement just using FedAvg. Scenarios 5 and 6 display significant improvements in Node 3 with all FL methods, particularly in the most challenging Scenario 6, where FedSDS *biased* again achieves the most pronounced results. These findings reinforce the effectiveness of FedSDS *biased* in addressing non-IID challenges.

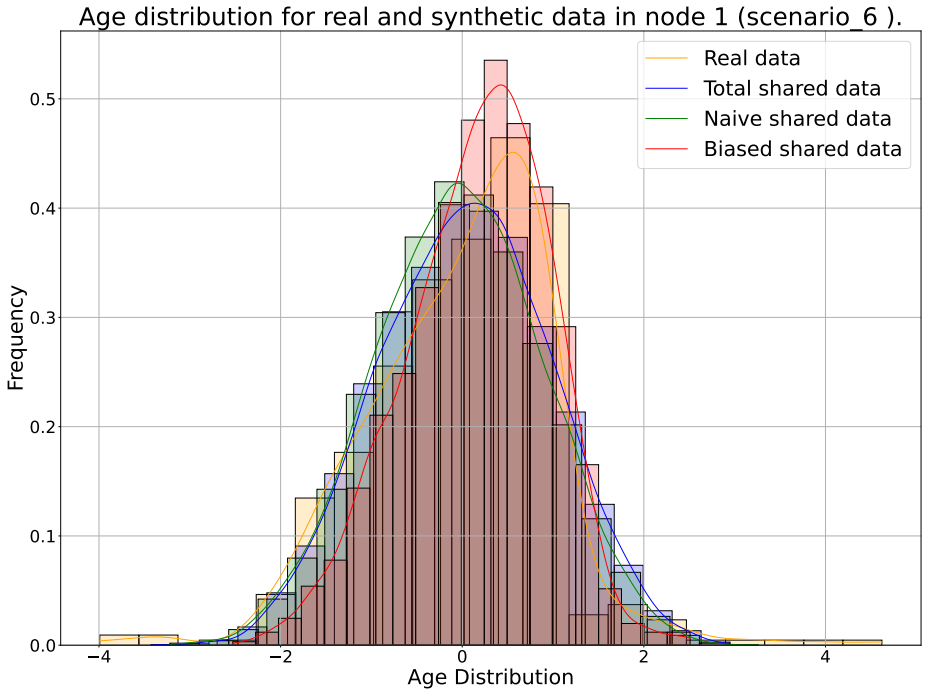
Scenario	Nodes	Isolated	FedAvg	FedSDS *naive*	FedSDS *biased*	Adjusted p-values
Centralized	**Node 1**	(0.619 - 0.647 - 0.673)	-	-	-	-
	Node 1	(0.611 - 0.638 - 0.664)	(0.611 - 0.638 - 0.665)	(0.612 - 0.642 - 0.671)	(0.612 - 0.640 - 0.668)	1.000 / 0.298 / 0.550
Scenario 4	**Node 2**	(0.586 - 0.617 - 0.649)	(0.586 - 0.616 - 0.646)	(0.579 - 0.615 - 0.653)	(0.590 - 0.619 - 0.648)	1.000 / 1.000 / 1.000
	Node 3	(0.616 - 0.646 - 0.675)	(0.622 - 0.649 - 0.676)	(0.617 - 0.646 - 0.674)	(0.622 - 0.650 - 0.679)	0.501 / 1.000 / 0.262
	Node 1	(0.612 - 0.640 - 0.670)	(0.610 - 0.638 - 0.665)	(0.610 - 0.639 - 0.669)	(0.613 - 0.641 - 0.672)	1.000 / 1.000 / 1.000
Scenario 5	**Node 2**	(0.582 - 0.616 - 0.646)	(0.600 - 0.631 - 0.661)	(0.582 - 0.625 - 0.673)	(0.602 - 0.635 - 0.667)	**0.004** / 1.000 / **0.004**
	Node 3	(0.502 - 0.545 - 0.589)	(0.566 - 0.599 - 0.634)	(0.559 - 0.590 - 0.622)	(0.563 - 0.594 - 0.632)	**0.002** / **0.005** / **0.002**
	Node 1	(0.607 - 0.638 - 0.667)	(0.611 - 0.639 - 0.667)	(0.611 - 0.640 - 0.669)	(0.609 - 0.636 - 0.665)	1.000 / 1.000 / 1.000
Scenario 6	**Node 2**	(0.589 - 0.620 - 0.653)	(0.593 - 0.625 - 0.655)	(0.576 - 0.620 - 0.660)	(0.612 - 0.643 - 0.671)	1.000 / 1.000 / **0.001**
	Node 3	(0.496 - 0.533 - 0.567)	(0.523 - 0.557 - 0.592)	(0.525 - 0.563 - 0.603)	(0.544 - 0.573 - 0.603)	**0.002** / **0.005** / **0.000**

Table 5.7: C-index comparison of isolated, FedAvg, and FedSDS (*naive* and *biased*) methods for the Metabric dataset in non-IID scenarios. Average C-index results are shown with CIs. Adjusted p-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in **bold**.

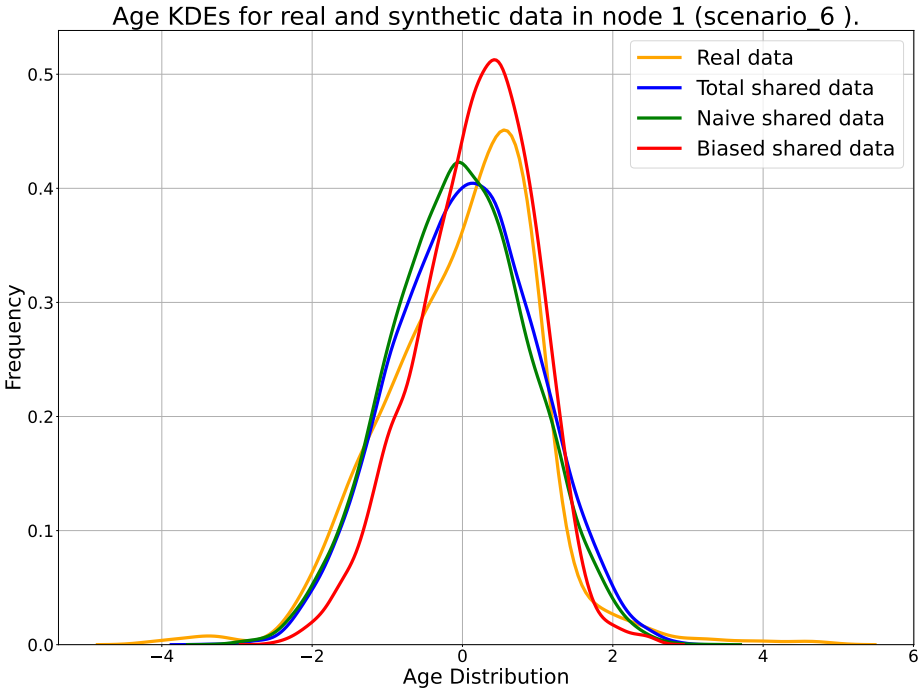
Scenario	Nodes	Isolated	FedAvg	FedSDS *naive*	FedSDS *biased*	Adjusted p-values
Centralized	**Node 1**	(0.660 - 0.688 - 0.714)	-	-	-	-
	Node 1	(0.647 - 0.674 - 0.701)	(0.645 - 0.673 - 0.701)	(0.645 - 0.673 - 0.700)	(0.644 - 0.675 - 0.703)	1.000 / 1.000 / 1.000
Scenario 4	**Node 2**	(0.666 - 0.693 - 0.720)	(0.673 - 0.699 - 0.727)	(0.659 - 0.689 - 0.719)	(0.665 - 0.691 - 0.718)	**0.013** / 1.000 / 1.000
	Node 3	(0.657 - 0.686 - 0.712)	(0.665 - 0.692 - 0.718)	(0.651 - 0.684 - 0.713)	(0.663 - 0.691 - 0.717)	**0.013** / 1.000 / **0.037**
	Node 1	(0.643 - 0.673 - 0.699)	(0.646 - 0.675 - 0.701)	(0.648 - 0.677 - 0.704)	(0.644 - 0.672 - 0.702)	0.777 / 0.117 / 1.000
Scenario 5	**Node 2**	(0.636 - 0.672 - 0.708)	(0.660 - 0.686 - 0.711)	(0.648 - 0.680 - 0.710)	(0.648 - 0.681 - 0.711)	0.180 / 0.658 / 0.538
	Node 3	(0.606 - 0.642 - 0.675)	(0.638 - 0.668 - 0.698)	(0.623 - 0.663 - 0.699)	(0.632 - 0.663 - 0.694)	**0.004** / **0.037** / **0.009**
	Node 1	(0.641 - 0.671 - 0.698)	(0.648 - 0.674 - 0.700)	(0.642 - 0.671 - 0.697)	(0.645 - 0.672 - 0.700)	0.275 / 1.000 / 1.000
Scenario 6	**Node 2**	(0.641 - 0.676 - 0.706)	(0.655 - 0.685 - 0.713)	(0.635 - 0.670 - 0.699)	(0.653 - 0.682 - 0.709)	0.204 / 1.000 / 0.426
	Node 3	(0.510 - 0.556 - 0.608)	(0.571 - 0.600 - 0.629)	(0.581 - 0.611 - 0.638)	(0.581 - 0.616 - 0.651)	**0.034** / **0.016** / **0.006**

Table 5.8: C-index comparison of isolated, FedAvg, and FedSDS (*naive* and *biased*) methods for the GBSG dataset in non-IID scenarios. Average C-index results are shown with CIs. Adjusted p-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in **bold**.

The impact of different synthetic data-sharing strategies is further examined in Figure 5.7, which presents histograms and KDEs of the normalized age distributions for Node 1 in Scenario 6 of the Metabric dataset. These visualizations allow for a direct comparison between real and synthetic data, highlighting how different sharing techniques influence the alignment of distributions. The differences between the real and synthetic data distributions in the first subfigure, which displays histograms, are not clearly distinguishable. While some variation can be observed, the overlap between real and synthetic data remains ambiguous, making it



(a) Histograms for real and synthetic data



(b) KDEs for real and synthetic data

Figure 5.7: Histogram plots and KDEs for AGE real and synthetic data distributions in Scenario 6 for Metabric.

difficult to assess how well the shared synthetic data align with the real distribution. However, the second subfigure, which presents just KDEs, provides a much clearer picture. Here, the KDE of the *biased* synthetic data more closely follows the trend of the real data, aligning with its skewness. In contrast, the KDE of the *naive* synthetic data remains more similar to the overall synthetic dataset, reinforcing that randomly selecting synthetic samples does not account for local data biases. Although the histogram comparison does not make the effect immediately evident (due to representation issues), the KDE plots clearly demonstrate that the *biased* synthetic data-sharing approach better preserves local data characteristics. This translates into significant improvements in similarity metrics and FL performance, highlighting the advantage of using *biased* synthetic data sharing in non-IID settings.

→ *Special IID Scenario*

In Scenario 7, the evaluation focuses on a unique and challenging non-IID case where one of the two nodes lacks the *age* covariate. Due to this limitation, FedAvg is not applicable, leaving FedSDS-based techniques as the only viable approach. Three distinct configurations are tested to assess the performance and adaptability of FedSDS, each leveraging synthetic data to compensate for the missing information and enhance the performance of the model.

The first approach, labeled *Imputation*, involves generating synthetic data at the 'good' node (the one with complete information) and training a predictor to estimate the missing *'age'* column for the 'bad' node (the one lacking the *'age'* covariate). These imputed data are incorporated into the 'bad' node training process. The second configuration, *Imputation + Synthetic data naive*, extends the first by augmenting the dataset in the 'bad' node with synthetic data generated by the 'good' node and aggregated using the *naive* technique. The third configuration, *Imputation + Synthetic data biased*, further refines the second by employing the *biased* aggregation technique to ensure that the synthetic data align closely with the local data distribution of the 'bad' node.

The results in Table 5.9 present the adjusted p-values for the C-index improvement in the 'bad' node across the three configurations. All datasets—Metabric and GBSG—demonstrate significant improvements in the C-index for each configuration compared to the isolated case, with p-values consistently below 0.05. Notably, the datasets achieve robust improvements, with p-values of 0.000 across all configurations, showcasing the effectiveness of FedSDS even in the most challenging non-IID scenarios.

Dataset	Isolated	Imputation	Imputation + Synthetic Data *naive*	Imputation + Synthetic Data *biased*	Adjusted p-values
Metabric	(0.544 - 0.572 - 0.602)	(0.602 - 0.634 - 0.669)	(0.594 - 0.629 - 0.666)	(0.602 - 0.631 - 0.661)	**0.000** / **0.000** / **0.000**
GBSG	(0.598 - 0.625 - 0.652)	(0.629 - 0.663 - 0.697)	(0.629 - 0.660 - 0.692)	(0.639 - 0.670 - 0.697)	**0.000** / **0.000** / **0.000**

Table 5.9: C-index comparison in Scenario 7 of the three different FedSDS settings. Average C-index results are shown with CIs. Adjusted p-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in **bold**.

→ ***Discussion***
The results of this study highlight the effectiveness of FL approaches, particularly the proposed FedSDS framework, in addressing challenges such as imbalanced data distributions, heterogeneity, and missing critical covariates. Across both IID and non-IID scenarios, FedSDS consistently outperforms traditional FL methods like FedAvg, especially when using the *biased* aggregation strategy, which shows clear advantages over the *naive* approach. FedSDS *biased* consistently performs better than its *naive* counterpart due to its ability to align synthetic data more closely with the local distributions of each node. The *naive* strategy aggregates synthetic data randomly, which can introduce noise or misaligned distributions, particularly in nodes with highly skewed or heterogeneous data. This misalignment may lead to suboptimal model updates and hinder performance improvement. In contrast, the *biased* strategy filters synthetic samples based on their similarity to the local data in the node in the latent space, ensuring that the aggregated synthetic data complements the unique characteristics of the node. This targeted alignment enhances model convergence and improves the ability of FedSDS to handle nodes with extreme data imbalances or biases.

In IID scenarios, where data are assumed to be distributed identically across nodes, FedSDS shows clear advantages in cases of imbalance. While no significant improvements are observed in scenarios with equal sample distribution across nodes (Scenario 1), FedSDS, particularly the *biased* aggregation strategy, consistently improves performance in scenarios where nodes face data scarcity (Scenarios 2 and 3). These improvements are most pronounced for nodes with fewer samples, reflecting the ability of the framework to mitigate the impact of data imbalance by effectively leveraging synthetic data generated during training.

The performance gap becomes even more evident in non-IID scenarios, where nodes exhibit biased distributions in key covariates. FedSDS *biased* consistently provides the most significant improvements, effectively addressing data heterogeneity. Nodes with skewed covariate distributions, particularly those with fewer samples, benefit from the tailored alignment of synthetic data with local distributions. This aligns with the design of FedSDS, which is explicitly aimed at handling heterogeneity through its advanced aggregation strategies.

Scenario 7, where one node lacks a critical covariate, is particularly challenging. In this setting, only FedSDS-based techniques are applicable. The results clearly show the effectiveness of leveraging synthetic data to compensate for missing information, with all configurations of FedSDS yielding significant improvements in model performance.

These findings underscore the importance of advanced FL techniques such as FedSDS for real-world applications, where data are often imbalanced, heterogeneous, or incomplete. FedSDS provides a robust and scalable solution for FL by integrating SDG and aggregation tailored to local distributions. This is particularly relevant in privacy-sensitive domains such as healthcare, where data sharing is restricted, and addressing data heterogeneity is crucial for building effective predictive models.

5.4.3 Conclusions

This study demonstrates the effectiveness of the FedSDS framework in addressing the challenges of data scarcity, heterogeneity, and missing critical covariates in SA. By integrating SDG

with FL, FedSDS offers a robust, scalable, and privacy-preserving approach for collaborative model training in decentralized settings. Across both IID and non-IID scenarios, FedSDS, particularly the *biased* aggregation strategy, consistently outperforms traditional methods such as FedAvg, highlighting its capacity to mitigate data imbalances and align synthetic data with local distributions. The results consistently demonstrate that FedSDS, particularly when employing the *biased* aggregation strategy, outperforms traditional methods like FedAvg, effectively addressing issues related to imbalanced and heterogeneous data distributions. Furthermore, the ability of the framework to align synthetic data with local distributions enhances convergence and ensures significant improvements, even in the most challenging scenarios.

The potential of FedSDS to transform SA methodologies is evident, but it also paves the way for further exploration. Enhancing the similarity of synthetic data remains a critical area for improvement, with future efforts potentially leveraging advanced generative models to better capture complex distributions in low-data settings. Additionally, the flexible architecture of the framework allows for the experimentation of alternative SA models. These models must incorporate the latent space representation characteristic of encoder-decoder architectures, essential for the *biased* aggregation strategy. By combining different methodologies beyond SAVAE, the generalizability and applicability of FedSDS could be further validated. Moreover, expanding the framework to accommodate multi-modal data, such as imaging, represents an exciting direction for research, given the growing prominence of multi-modal analytics in healthcare. Another important avenue for future work is adapting FedSDS to dynamic SA environments, where real-world data evolves due to changes in demographics, treatment protocols, or institutional practices. This adaptability would ensure the framework remains relevant and effective in ever-changing healthcare landscapes. Lastly, while synthetic data sharing inherently enhances privacy, integrating formal privacy guarantees, such as DP, would strengthen the appeal of the framework in highly sensitive domains like healthcare. These guarantees could ensure that FedSDS complies with stringent privacy regulations while maintaining the quality of collaborative model training. In addition, in future work, alternative proximity calculation techniques in the *biased* aggregation strategy could also be explored, such as using Kullback-Leibler or Jensen-Shannon divergences. Since these metrics compare probability distributions, they could be well-suited to the latent space representations used in FedSDS, potentially improving the alignment of synthetic data with local distributions.

In conclusion, FedSDS represents a significant advance in FL for SA. It addresses the critical limitations of traditional methods while enabling effective collaboration in privacy-sensitive settings. By continuing to refine and expand its capabilities, FedSDS has the potential to become a cornerstone methodology for decentralized healthcare analytics, ultimately improving patient outcomes and advancing precision medicine.

5.5 Chapter Conclusions

This chapter has explored the integration of SDG and FL in healthcare, focusing on addressing challenges such as data scarcity, heterogeneity, and privacy concerns. The findings demonstrate the effectiveness of frameworks like FedSDS in generating high-quality synthetic data while enabling collaborative and privacy-preserving model training across decentralized nodes. By leveraging advanced techniques, including VAEs and aggregation strategies, the research highlights the transformative potential of combining SDG and FL for healthcare analytics.

The studies consistently show that FedSDS outperforms traditional methods such as FedAvg and isolated training across both IID and non-IID scenarios. Notably, FedSDS effectively mitigates issues related to imbalanced and heterogeneous data distributions, ensuring that synthetic data closely aligns with local real-world distributions. This alignment is evidenced by improved metrics, such as lower values of $D_{\mathbb{JS}}$ and enhanced clinical utility validation, confirming the practical applicability of synthetic data in downstream machine learning tasks. The robustness of FedSDS in non-IID scenarios, especially in data-scarce and underrepresented settings, underscores its scalability and adaptability in diverse healthcare environments.

Future research directions include enhancing synthetic data similarity by incorporating advanced generative models to capture complex distributions in low-data settings. Adapting the FedSDS framework to handle multi-modal data, such as imaging or genomic sequences, is a promising avenue, given the increasing emphasis on multi-modal healthcare analytics. The flexibility of FedSDS also invites experimentation with alternative survival analysis models, potentially improving its generalizability and expanding its use cases. Given the increasing emphasis on multi-modal healthcare analytics, adapting the FedSDS framework to handle multi-modal data, such as imaging or genomic sequences, is a promising avenue.

Moreover, integrating formal privacy guarantees, such as DP or HE, could further bolster the utility of the framework in privacy-sensitive domains. These enhancements would ensure compliance with stringent privacy regulations while maintaining robust collaborative training across institutions. Alternative proximity metrics, such as Kullback-Leibler or Jensen-Shannon divergences, could also be explored within the biased aggregation strategy to refine the alignment of synthetic data with local distributions.

In conclusion, integrating SDG and FL through frameworks like FedSDS marks a significant step forward in healthcare analytics. By addressing critical challenges in decentralized environments, such frameworks enable effective collaboration, improve data diversity, and enhance model generalization, all while safeguarding patient privacy. These advancements not only pave the way for broader applications in healthcare research but also lay the foundation for precision medicine, ultimately improving patient outcomes and fostering equitable access to advanced analytics in under-resourced regions.

Chapter 6

Discussion

This thesis advances the field of healthcare analytics by addressing critical challenges through the innovative use of generative AI. The research spans SA, SDG, and FL, each contributing to the overarching goal of reducing healthcare inequities. These contributions are grounded in a robust methodological framework and supported by empirical validation, underscoring their relevance to modern healthcare challenges. This chapter critically examines the presented contributions, explores their broader implications for research and clinical practice, acknowledges inherent limitations, and discusses the ethical considerations essential for responsible and equitable implementation in this domain.

6.1 Overview of Key Findings and Contributions to the Field

This section synthesizes the principal contributions of this thesis and contextualizes their significance within the existing literature presented in Chapter 2. Rather than reiterating conclusions, this section critically analyzes how our findings address key challenges in SA, SDG, and FL, shedding light on their impact on clinical and data-centric applications.

By situating these results against state-of-the-art methodologies, we highlight how our proposed models advance the field, overcome existing limitations, and provide innovative tools for real-world problems.

6.1.1 Survival Analysis Models: SAVAE and CR-SAVAE

SA plays a fundamental role in healthcare, providing critical insights into time-to-event outcomes such as patient survival, disease progression, or treatment efficacy. Traditional SA models, including the KM estimator and the Cox-PH model, have long served as gold standards. However, these methods rely on strong assumptions—such as proportional hazards and linear relationships—that limit their ability to handle complex, high-dimensional, and heterogeneous medical datasets. Contemporary ML methods like DeepSurv and DeepHit have addressed some limitations, yet challenges remain, particularly in handling censoring,

non-linear relationships, and CR.

In response to these challenges, this thesis introduces two novel frameworks for SA: SAVAE and its extension, CR-SAVAE. These models leverage VAEs to combine flexibility, robustness, and interpretability, addressing key barriers outlined in Section 2.2.6.

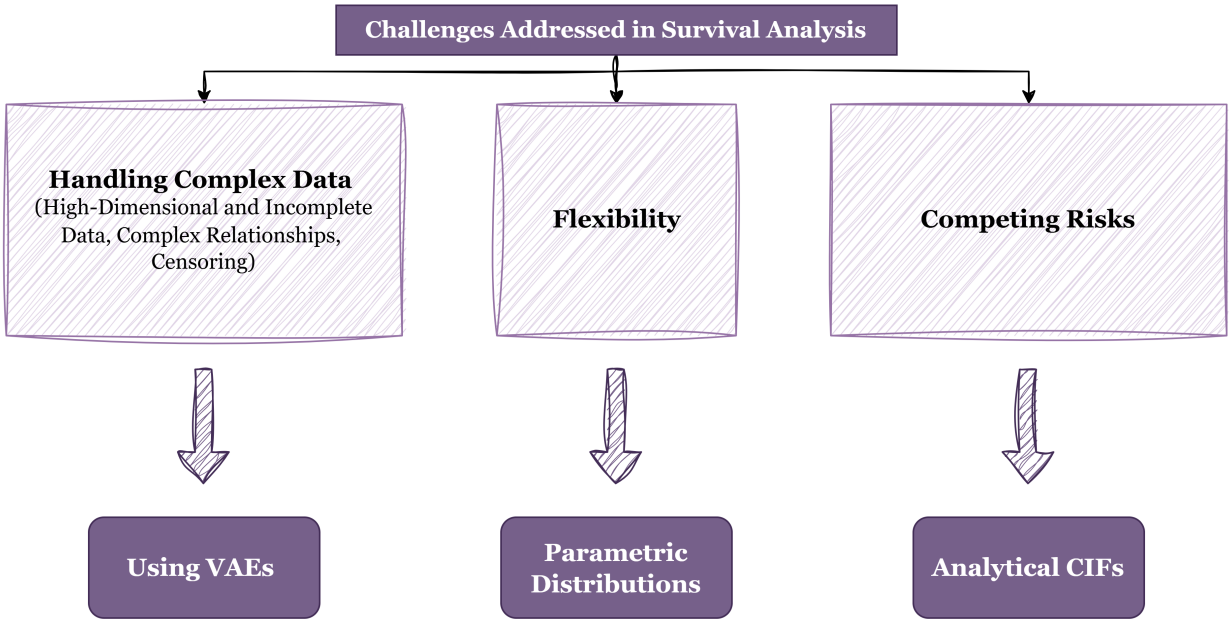


Figure 6.1: Overview of the key challenges addressed in SA. This figure connects the key challenges previously identified in Section 2.2.6 with the solutions proposed in SAVAE and CR-SAVAE, demonstrating how each methodological improvement directly addresses specific limitations in SA.

As illustrated in Figure 6.1, SAVAE and CR-SAVAE specifically tackle three fundamental challenges in SA:

1. **Flexibility and Robustness in SA:** SAVAE introduces significant improvements over traditional and contemporary methods. Unlike the Cox-PH model, which assumes proportional hazards and struggles with non-linear covariate relationships, SAVAE leverages VAEs for greater flexibility. By analytically modeling continuous or discrete time-to-event data with differentiable parametric distributions, SAVAE avoids rigid assumptions, adapting to diverse datasets.

   This design addresses the challenge of designing models capable of handling non-linear relationships and high-dimensional data without overfitting. Our results demonstrate competitive performance against state-of-the-art methods across multiple benchmarks (C-index and IBS), offering a robust alternative for clinical applications. Furthermore, unlike models trained with C-index optimization, the ability of SAVAE to optimize likelihood-based training ensures statistical consistency.

2. **Handling Censoring and Incomplete Data:** One of the persistent limitations in SA lies in the effective handling of censored data. SAVAE excels in incorporating

right-censored observations during training, ensuring unbiased predictions. Compared to DeepHit, which discretizes time and is limited by the requirement for large datasets, SAVAE efficiently models survival times under continuous and discrete settings. This capacity is particularly valuable in medical research, where sample sizes are often small, and censoring is prevalent.

By introducing this flexibility and robustness, SAVAE positions itself as a practical solution for complex healthcare datasets, addressing key challenges in model adaptability and censorship handling.

3. **CR with CR-SAVAE:** While significant progress has been made in SA, handling CR remains an unresolved challenge. DeepHit offers non-parametric estimates of CIFs, but its numerical nature lacks interpretability and statistical rigor. CR-SAVAE overcomes these limitations by providing parametric, analytical CIF estimates, enabling accurate risk prediction and robust statistical analyses such as CIs and hypothesis testing.

   The results confirm that CR-SAVAE achieves performance comparable to DeepHit on key metrics like the C-index and IBS despite the computational advantages of its parametric nature. By seamlessly handling both continuous and discrete time while remaining free from the assumption of proportional hazards, CR-SAVAE addresses the major research gaps that have been identified. Its application in real-world clinical settings, where competing events (e.g., multiple causes of mortality) are prevalent, demonstrates its utility for precision medicine.

In conclusion, the introduction of SAVAE and CR-SAVAE represents a substantial advancement in SA, offering flexible, robust, and interpretable solutions that effectively address the limitations of existing methods and meet the challenges outlined in Section 2.2.6, thereby paving the way for improved clinical decision-making and precision medicine.

6.1.2 Synthetic Data Generation

Given the multiple contributions of this work to SDG, this section first delineates the specific advancements made in each component. Subsequently, it connects them to the challenges identified in Section 2.3.5. By addressing model limitations, data validation gaps, and challenges of data scarcity, these contributions provide robust and practical solutions for SDG, particularly in healthcare contexts.

Variational Autoencoder with Bayesian Gaussian Mixture (VAE-BGM)

- **Improved Generative Capabilities:** The proposed VAE-BGM model overcomes the limitations of existing approaches, such as TVAE, which rely on Gaussian assumptions in the latent space. The model effectively captures complex, non-Gaussian data distributions by incorporating a BGM into the latent space of the VAE. Empirical results demonstrate superior performance over state-of-the-art models like CTGAN and TVAE across various datasets, including real-world medical applications.
- **Joint Distribution Preservation:** VAE-BGM explicitly models both marginal and joint feature distributions, ensuring that feature interdependencies—critical for clinical

tasks—are preserved. This is particularly relevant for healthcare datasets, where relationships such as age-disease interactions are essential for downstream predictive modeling.

- **Practical Validation Through Machine Learning Utility:** The generated synthetic data were validated using RF classifiers, demonstrating that VAE-BGM preserves the statistical and predictive properties of real data. Accuracy metrics comparable to real-world benchmarks highlight the practical utility of the model for tasks like classification and prediction.

Divergence-Based Validation for Synthetic Data

- **A Robust Framework for Data Quality Assessment:** A divergence-based validation framework is proposed, leveraging $D_{\mathbb{JS}}$ and $D_{\mathbb{KL}}$ to evaluate the similarity between real and synthetic data. Unlike traditional methods focusing on individual attributes, this approach captures joint probability distributions, offering a more holistic validation metric.
- **Probabilistic Discriminator for Density Ratio Estimation:** To estimate divergences efficiently, we proposed a probabilistic discriminator network that approximates density ratios between real and synthetic data distributions. This method improves computational efficiency and enhances the interpretability of the results.
- **Practical Impact:** Our framework sets a standard for validating SDG models, ensuring that generated data not only mimic marginal distributions but also retain complex interdependencies. Adopting $D_{\mathbb{JS}}$ as a bounded, symmetric metric facilitates easier comparison and interpretation of results.

Inductive Bias Techniques for Scarce-Data Scenarios

1. **General-Purpose Data:**
   - **Addressing Small Dataset Challenges:** We tackled the limitations of DGMs in scenarios with limited data by introducing inductive bias techniques. These include pre-training, model averaging, MAML, and DRS. Results show that incorporating inductive biases significantly improves the quality of synthetic data under low-data conditions.
   - **Performance Gains and Trade-offs:** Experimental results highlight that model averaging achieves the largest improvements in $D_{\mathbb{JS}}$ and $D_{\mathbb{KL}}$ across datasets, with up to 50% gains in $D_{\mathbb{JS}}$. Pre-training is also effective, particularly when applied to VAEs and CTGANs. In contrast, MAML exhibits limited benefits, likely due to its reliance on larger task variability and higher computational overhead.
   - **Broader Applicability:** The proposed methodology provides a robust solution for SDG in resource-constrained environments, broadening its applicability to healthcare and finance, where small datasets are common.

2. **Medical Data:**
    - **Tackling Data Scarcity in Medical Tabular Data:** Healthcare datasets, particularly those for rare diseases and SA, are often limited in size and highly heterogeneous. The proposed methodology addresses this challenge by introducing inductive biases, enabling the synthetic generation of small, complex datasets without compromising quality.
    - **Clinical Utility Validation and Real-World Applicability:** Experiments on medical classification datasets (e.g., Heart) and cancer-related survival datasets (e.g., Metabric, GBSG, NWTco) demonstrated that synthetic data generated using the proposed approach achieve performance comparable to real data. Clinical utility metrics like accuracy, C-index, and IBS confirm that synthetic data are reliable for downstream tasks, such as diagnosis or prognosis modeling.
    - **Kaplan-Meier Estimations for Survival Data:** KM survival curves further validated the methodology, showing that synthetic data closely replicate real-world survival probabilities, narrowing deviations under data scarcity. This ensures applicability to realistic clinical studies, where patient privacy and limited data availability are major challenges.
    - **Reusability Across Tasks:** By repurposing datasets for alternative tasks (e.g., transforming survival time into classification labels), the proposed methodology proved its versatility in modeling complex relationships and generating synthetic data adaptable to varying clinical applications.

The contributions outlined above address the key challenges identified in Section 2.3.5 for SDG, as illustrated in Figure 6.2:

1. **Complex Data Structures and Heterogeneity:** The VAE-BGM model improves generative performance by effectively handling complex, heterogeneous tabular data distributions.
2. **Lack of Robust Validation Techniques:** The introduction of the divergence-based framework with $D_{\mathbb{JS}}$ and $D_{\mathbb{KL}}$ ensures reliable assessment of synthetic data quality, extending validation beyond marginal statistics to capture joint feature relationships.
3. **Data Scarcity in Healthcare:** Using inductive bias techniques and tailored SDG methodologies significantly enhances model performance in low-data scenarios, enabling reliable SDG for small, resource-limited medical datasets.
4. **Clinical Utility in Healthcare Applications:** By validating synthetic data through downstream tasks (e.g., classification, SA) and clinical metrics (e.g., C-index, IBS, KM curves), this work ensures that generated data maintain their utility in critical healthcare applications.

The contributions to SDG presented in this thesis represent a significant step forward in overcoming key challenges in the state-of-the-art. By improving generative architectures, establishing robust validation frameworks, and addressing data scarcity through inductive biases, this work delivers reliable solutions for SDG. Focused on healthcare applications, it

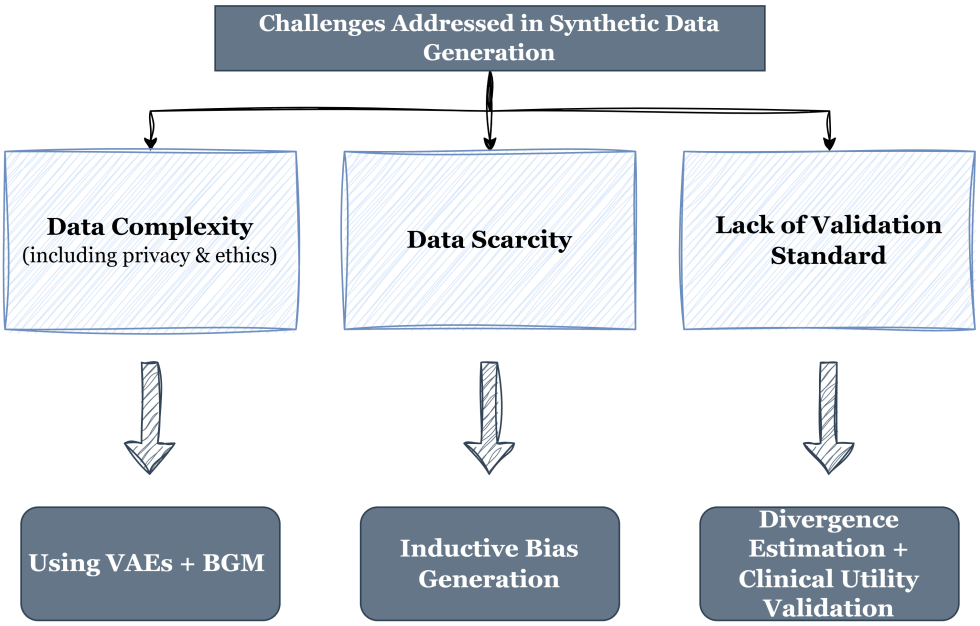


Figure 6.2: Overview of the key challenges addressed in SDG. This figure connects the key challenges previously identified in Section 2.3.5 with the solutions proposed, demonstrating how each methodological improvement directly addresses specific limitations in SDG.

ensures that synthetic data can serve as a practical alternative to real data for critical tasks, advancing research while preserving patient privacy and mitigating data limitations.

6.1.3 Federated Learning

The proposed research introduces significant advancements to address the key challenges of FL in healthcare, as identified in Section 2.4.4. These challenges include statistical heterogeneity, data scarcity, communication bottlenecks, and privacy preservation. The contributions of this work are summarized below, highlighting how each challenge is systematically tackled through the proposed FedSDS framework.

As illustrated in Figure 2.12, the FedSDS framework systematically addresses these challenges:

1. **Addressing Statistical Heterogeneity:** Statistical heterogeneity arises due to the non-IID nature of healthcare data, where different institutions capture datasets with varying distributions influenced by demographics, diseases, and regional conditions. Traditional FL techniques like FedAvg struggle to converge in diverse environments.

   The proposed FedSDS framework introduces a novel approach to mitigate statistical heterogeneity by replacing model parameter sharing with synthetic data exchange. Key innovations include:

   - **SDG using the VAE-BGM model**, which captures complex, multi-modal data distributions and generates high-quality synthetic tabular data.

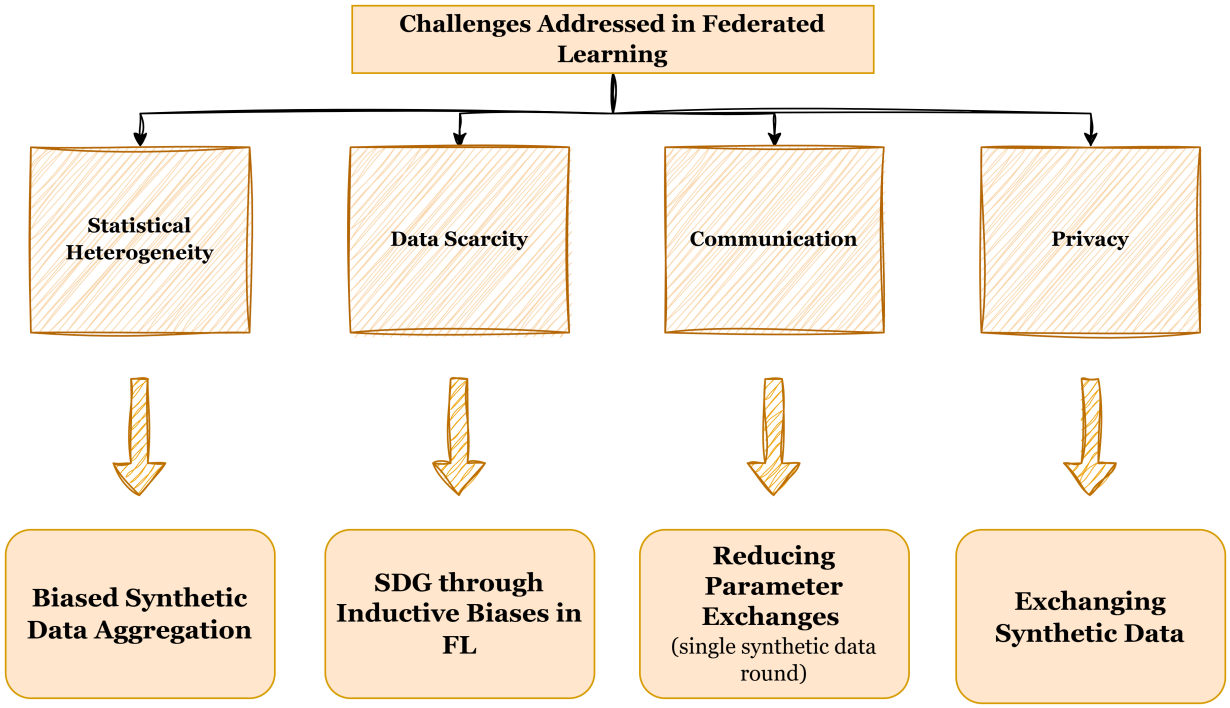


Figure 6.3: Overview of the key challenges addressed in FL. This figure connects the key challenges previously identified in Section 2.4.4 with the solutions proposed, demonstrating how each methodological improvement directly addresses specific limitations in FL.

- **Bias-Aware Aggregation (*biased* technique):** Synthetic data shared among nodes are filtered based on their proximity to local datasets in the latent space. This ensures that only the most relevant synthetic samples are integrated, aligning local models with the global representation and minimizing the impact of heterogeneity.

The results demonstrate that FedSDS significantly reduces divergence metrics like $D_{\mathbb{JS}}$ and improves model performance, even under extreme non-IID conditions. The framework balances data contributions across institutions by leveraging tailored synthetic data sharing, ensuring robust and fair global models.

2. **Mitigating Data Scarcity:** Healthcare datasets, particularly for rare diseases, are often small, incomplete, or imbalanced, hindering effective training of AI models.

   The integration of SDG into FL addresses data scarcity by augmenting local datasets with synthetic samples:

   - **VAE-BGM with Artificial Inductive Bias:** The framework generates realistic synthetic patient data even in low-data scenarios by incorporating model averaging and latent space refinements.
   - **Decentralized Data Augmentation:** Nodes with limited or biased datasets benefit from high-quality synthetic data shared from other nodes, improving their

ability to train local models effectively.

Through extensive experiments, FedSDS demonstrates significant gains in both statistical similarity (measured via $D_{\mathbb{JS}}$) and clinical utility (accuracy and C-index scores). Nodes with the scarcest data experience the most notable improvements, confirming the efficacy of the proposed method in resource-constrained settings.

3. **Efficient Communication:** Communication bottlenecks are a significant challenge in FL, especially in large-scale, resource-constrained healthcare environments with limited bandwidth.

   FedSDS minimizes communication overhead by replacing iterative parameter exchanges with a single round of synthetic data sharing. Key benefits include:

   - **Reduced Communication Rounds:** Unlike traditional methods such as FedAvg, which require multiple rounds of updates, FedSDS achieves convergence in fewer iterations.
   - **Scalability:** The framework is well-suited for deployments across geographically distributed institutions with varying network capacities.

   This design ensures that FedSDS remains efficient and scalable while maintaining performance improvements, even in non-IID and low-data environments.

4. **Privacy Preservation:** FL inherently enhances privacy by training models locally without sharing raw data. However, privacy remains a concern, especially when considering risks such as model inversion attacks or auxiliary information leakage.

   The FedSDS framework prioritizes privacy preservation by exchanging synthetic data instead of model parameters or raw patient records. Synthetic data generated via VAE-BGM are not exact duplicates of real patient data. Sampling from the latent space ensures that sensitive information is not directly replicated. Additionally, privacy risks are further mitigated through evaluations of minimum distances between real and synthetic samples, confirming that synthetic data remain statistically distinct while preserving utility.

   This approach reduces vulnerabilities associated with traditional FL techniques and aligns with privacy regulations like GDPR and HIPAA, fostering trust among institutions.

The proposed FedSDS framework combines SDG and FL to provide a robust solution to the critical challenges of FL in healthcare. It effectively addresses statistical heterogeneity, data scarcity, communication bottlenecks, and privacy concerns, making it a promising tool for collaborative, privacy-preserving AI in healthcare. By leveraging advanced aggregation techniques and high-quality synthetic data, FedSDS ensures improved model performance and generalizability, even in heterogeneous and low-data environments.

6.2 Limitations of the Work

While the proposed methodologies and findings presented in this thesis demonstrate significant advancements in SDG, SA, and FL for healthcare applications, it is essential to analyze their limitations critically. This section reflects on the constraints associated with the datasets, computational complexity, validation, and the proposed SDG approaches.

6.2.1 Limitations of the Datasets

The availability and diversity of datasets remain a critical limitation in the healthcare domain. Despite employing benchmark medical datasets, such as Metabric, GBSG, and NWTco, the scope of experiments is confined to a finite set of publicly available data. SA, in particular, relies on longitudinal datasets with time-to-event information, yet such datasets are scarce, especially when censoring and CR are included. This restricts the generalizability of the proposed approaches to broader clinical scenarios.

In FL experiments, statistical heterogeneity was introduced artificially to simulate non-IID conditions, primarily by skewing specific feature distributions, such as age covariates. While this controlled setting provides a valuable approximation, it may fail to fully reflect the complexity of real-world heterogeneity in clinical institutions, where variations in demographics, disease prevalence, and regional practices influence data distributions in far more intricate ways. Additionally, datasets for rare diseases, often small, incomplete, or highly imbalanced, pose significant challenges. Although the SDG techniques presented in this work address data scarcity to some extent, extremely small datasets limit the ability of models to learn representative latent distributions, leading to a decline in synthetic data quality and utility.

6.2.2 Computational Complexity

While the proposed models, such as VAE-BGM and SAVAE, are computationally efficient for state-of-the-art deep generative approaches, certain challenges remain when applied in resource-constrained environments. The training times for these models are not excessive under standard computational setups; however, their deployment in FL scenarios introduces additional considerations. In FL, where models are trained locally across multiple institutions, nodes with limited computational resources, such as small healthcare clinics or low-power devices, may face challenges in efficiently training generative models or performing iterative updates.

This limitation becomes particularly relevant in large-scale federated networks with significant heterogeneity in hardware capabilities among participating nodes. Institutions with older infrastructure or reduced access to high-performance computing resources may struggle to keep pace with nodes equipped with more advanced systems. This disparity can result in imbalanced contributions to the global model or delays in the FL process, affecting overall performance and scalability.

Furthermore, while the FedSDS framework reduces communication overhead by sharing synthetic data instead of model parameters, generating high-quality synthetic data—particularly when incorporating techniques like model averaging or inductive biases—still requires a

baseline level of computational capacity. Although this requirement is not prohibitive for most modern systems, it may limit the applicability of the proposed methodologies in environments where access to computational infrastructure is minimal.

Addressing these challenges in future work could involve exploring model compression, lightweight generative architectures, or adaptive FL strategies that account for hardware disparities across nodes, ensuring broader accessibility and scalability of the proposed frameworks.

6.2.3 Validation in Clinical Settings

The evaluation and validation of the proposed methodologies have primarily been conducted using publicly available datasets and simulated FL environments. While promising results, additional validation in real-world clinical settings remains necessary. The experiments rely on solid statistical metrics, such as the C-index, IBS, and $D_{\mathbb{JS}}$, providing robust performance evidence. However, these evaluations do not capture the complexities and nuances of real clinical workflows or patient populations.

The lack of clinical deployment prevents a thorough assessment of the models' impact on clinical decision-making processes, such as diagnosis, prognosis, or treatment planning. Real-world implementation would require collaboration with healthcare institutions to test the methodologies on real patient data under operational conditions. Additionally, while the proposed frameworks demonstrate robustness in controlled settings, their ability to handle real-world noise, missing values, and inconsistencies within medical datasets has not been fully explored. Validation incorporating these challenges would further strengthen the practical applicability and reliability of the methods.

6.2.4 Synthetic Data Generation Quality

Although significantly improved through VAE-BGM and inductive bias techniques, the quality of the generated synthetic data presents certain limitations. In extremely low-data scenarios, where the available real data are highly scarce or biased, the quality and diversity of the synthetic data may deteriorate. The ability of the model to capture sufficient information from very small datasets is inherently constrained despite introducing artificial inductive biases, such as model averaging, pre-training, and domain randomization techniques. This limitation may impact downstream tasks, such as classification, SA, or model generalization, particularly in rare disease studies.

Another important consideration relates to potential biases within the synthetic data. Since the generation of synthetic samples depends on the quality and representativeness of the original dataset, biases or gaps in the real data can propagate into the synthetic data, reinforcing pre-existing disparities. This issue highlights the need to carefully evaluate fairness and diversity in the generated datasets, particularly in clinical applications where biases may have ethical implications.

Finally, while synthetic data reduces privacy risks compared to raw data sharing, this work does not incorporate formal privacy guarantees such as DP. Although the proposed methodology

ensures that synthetic samples are not exact replicas of real patient data, future enhancements could integrate DP mechanisms to provide quantifiable privacy assurances. This would further align the SDG approach with strict privacy regulations, such as GDPR and HIPAA, ensuring broader trust and adoption in healthcare settings.

6.2.5 Summary of Limitations

In conclusion, the primary limitations of this work include constraints in the availability and diversity of datasets, the computational complexity of the proposed models, the need for real-world clinical validation, and challenges associated with synthetic data quality under extreme scarcity. While the presented methodologies successfully address key challenges in SA, FL, and SDG, further efforts are needed to enhance their scalability, applicability, and robustness in operational clinical environments. By addressing these limitations in future research, the proposed frameworks can be refined to unlock their full potential for advancing collaborative, privacy-preserving AI in healthcare.

6.3 Ethical Considerations

The ethical implications of this research are particularly pertinent due to its focus on healthcare and FL, where sensitive patient information, fairness, and trust play a central role. Although this work provides privacy-preserving frameworks and SDG methodologies, several ethical concerns must be critically examined to ensure their responsible adoption in clinical and collaborative AI environments.

6.3.1 Privacy of Patient Data

FL and SDG offer promising approaches to preserving patient privacy by avoiding centralized data aggregation and replacing raw data with generated alternatives. However, residual privacy risks remain, particularly in scenarios involving small and unique datasets. Even when models, such as VAE-BGM, generate synthetic data that are statistically distinct from real patient data, there remains a possibility of reidentification when auxiliary information or advanced inference attacks are leveraged. This risk is heightened in rare disease datasets, where limited sample sizes and unique patient features may unintentionally reveal sensitive information. Ensuring compliance with privacy regulations such as GDPR and HIPAA requires further integration of formal privacy guarantees, such as DP, to protect against these vulnerabilities rigorously.

6.3.2 Bias and Fairness in Models

Bias in AI models remains a significant ethical concern, particularly in the healthcare domain, where model outputs can directly influence patient diagnoses, prognoses, or treatment plans. The heterogeneity of real and synthetic data used in this work introduces the potential for bias, especially when datasets are skewed toward majority populations. For instance, models trained on data that disproportionately represent specific demographics, such as age groups,

ethnicities, or geographic regions, may perform poorly for underrepresented populations. This limitation could exacerbate health disparities, leading to less accurate predictions and potentially inequitable clinical outcomes. While the proposed FL methodologies, such as FedSDS, aim to mitigate data scarcity and heterogeneity, achieving fairness across diverse healthcare settings requires careful evaluation of model biases and deliberate strategies to ensure representativeness in the training data.

Future work should prioritize fairness-aware techniques, such as bias correction methods and fairness metrics, to systematically address inequities in model performance across subpopulations.

6.3.3 Clinical Implications of Synthetic Data

The use of synthetic data for training AI models raises important questions about the reliability and safety of these models in clinical practice. While this research demonstrates that synthetic data can closely replicate real data distributions, achieving comparable performance in downstream tasks such as classification, SA, and risk prediction, the potential risks associated with synthetic data remain. Synthetic datasets, particularly those generated from limited real data, may fail to capture rare events, subtle correlations, or the full complexity of clinical conditions. Deploying models trained on synthetic data without thorough validation in real-world clinical environments could lead to inaccurate predictions, impacting patient care and decision-making.

To address these concerns, extensive clinical validation must accompany the adoption of synthetic data-driven models. Validation processes should involve healthcare professionals, incorporate real patient outcomes, and assess the safety of the model safety, accuracy, and robustness across diverse and unseen populations. Ensuring that synthetic data do not distort clinical insights is essential to building trust among clinicians, patients, and institutions.

6.3.4 Transparency and Accountability

Given the high stakes of AI in healthcare, transparency in model design, data generation, and evaluation processes is crucial. Healthcare practitioners and stakeholders must clearly understand the methods used to generate synthetic data, the limitations of the resulting models, and the risks associated with their deployment. Transparency also extends to accountability frameworks that ensure models can be audited, explained, and validated to prevent misuse or misinterpretation of AI predictions.

This research advocates for the responsible deployment of FL and synthetic data methodologies by adhering to transparency, fairness, and privacy principles. Ensuring the ethical use of these models in clinical settings requires ongoing collaboration between AI researchers, ethicists, and medical practitioners to address emerging challenges and uphold patient-centric values.

6.4 Final Remarks

This thesis highlights the critical role of generative AI and FL in addressing some of the most pressing challenges in modern healthcare analytics. At a time when technological advancements are reshaping the landscape of medical research and clinical practice, the presented work contributes meaningfully to overcoming key barriers such as data scarcity, privacy preservation, and data heterogeneity. By advancing methodologies for SA, SDG, and FL, this research provides practical tools that empower institutions to collaborate securely, generate high-quality insights from limited datasets, and develop models capable of generalizing across diverse patient populations.

The proposed frameworks offer innovative solutions that address long-standing issues while aligning with evolving ethical and regulatory considerations. These contributions are particularly relevant in healthcare, where privacy concerns, limited data availability, and heterogeneity across institutions often hinder the development of robust predictive models. By leveraging synthetic data and decentralized learning, this work not only advances the state of the art but also ensures that innovations remain accessible and equitable, especially for under-resourced environments.

Looking ahead, integrating these solutions into real-world clinical workflows can transform healthcare systems by enabling more accurate, privacy-preserving, and scalable analytics. Ultimately, this thesis reaffirms that generative AI and FL are not just tools of technological progress but catalysts for addressing critical inequities in healthcare, paving the way for improved patient outcomes, better resource allocation, and a more inclusive approach to personalized precision medicine.

Chapter 7

Conclusions

This thesis addresses some of the most pressing challenges in modern healthcare analytics, focusing on reducing inequities through advanced AI-driven methodologies. Inspired by the need to bridge healthcare disparities, this work explored how generative AI, particularly VAEs, can mitigate data scarcity, privacy concerns, and statistical heterogeneity. By centering on SA, SDG, and FL, the research aligned with improving access to equitable and inclusive healthcare solutions.

In the domain of SA, the development of SAVAE and CR-SAVAE models represents a significant advancement over traditional and contemporary approaches. By relaxing assumptions such as proportional hazards and incorporating CR analysis, these models provide greater flexibility and robustness in handling censored and complex survival data. This ensures more accurate predictions of time-to-event outcomes, critical for personalized patient care and clinical decision-making.

SDG emerged as a revolutionary tool for overcoming data scarcity, particularly in resource-limited environments and rare disease research. By integrating a BGM into the latent space of VAEs, this work demonstrated the ability to generate high-quality synthetic tabular data while preserving the underlying statistical and clinical relationships. Such synthetic datasets enable broader participation in AI model development and ensure that healthcare innovation is not restricted to regions with abundant data resources.

The FedSDS framework highlights the potential of FL to foster decentralized collaboration among institutions. By leveraging synthetic data instead of raw records, this work introduced a novel method for mitigating statistical heterogeneity and communication bottlenecks while preserving patient privacy. FedSDS offers a scalable, privacy-aware solution for collaborative research, bridging the gap between data-rich and resource-constrained institutions.

Ultimately, this thesis demonstrates that integrating generative AI techniques can overcome critical barriers to equitable healthcare innovation. The synergy between SA, SDG, and FL underscores the versatility of VAEs in solving complex healthcare challenges. By focusing on real-world applications and aligning with global priorities, this work contributes to the ongoing efforts to reduce disparities, empower underrepresented communities, and ensure that no one is excluded from the benefits of AI-driven healthcare advancements.

7.1 Future Research Directions

While this thesis presents significant contributions, it also opens several avenues for future research to advance AI-driven healthcare solutions further:

1. **Frailty Models in SA:** Future work can integrate frailty models into SA frameworks to account for unobserved patient heterogeneity. Frailty models enable the inclusion of random effects, improving the precision of predictions by capturing latent risk factors that may vary across individuals. This extension would be particularly valuable for modeling complex healthcare scenarios involving diverse patient populations.
2. **Privacy Validation in SDG:** Although this thesis addresses privacy concerns through synthetic data, further research is needed to develop robust privacy validation frameworks. Techniques such as DP or HE could provide formal privacy guarantees, ensuring synthetic datasets remain trustworthy and secure for use in sensitive medical applications.
3. **Expansion to Multi-Modal Data:** The methodologies proposed primarily focus on tabular data. Future research should explore the integration of multi-modal datasets, including medical imaging, genomic sequences, and clinical text. Combining multi-modal data with generative models would enable a more comprehensive understanding of patient health, aligning with the increasing emphasis on precision medicine.
4. **Adaptive FL Strategies:** Exploring adaptive FL strategies for dynamic healthcare environments represents another promising direction. As data distributions evolve over time due to demographic changes, treatment innovations, or institutional shifts, adaptive learning techniques will ensure that FL models remain robust, efficient, and generalizable.
5. **Establishing Public Repositories for Synthetic Data:** A critical step toward broader adoption of SDG involves the creation of publicly accessible repositories of high-quality synthetic datasets. Such repositories would enable collaboration among institutions, facilitate benchmarking of AI models, and promote innovation while maintaining patient privacy and regulatory compliance.

Addressing these research directions, future studies can build on the contributions of this thesis to further enhance the scalability, fairness, and reliability of AI solutions in healthcare. The continued evolution of generative AI, coupled with collaborative frameworks like FL, holds the potential to transform healthcare systems worldwide, ensuring equitable access to advanced medical tools and fostering meaningful improvements in patient outcomes.

In conclusion, this thesis contributes to a growing body of research demonstrating how AI can be harnessed to address systemic healthcare inequities. Through innovative methods for survival modeling, SDG, and federated collaboration, this work paves the way for AI-driven solutions that are not only technologically advanced but also ethically sound, equitable, and globally inclusive. Moving forward, integrating these approaches into real-world clinical workflows and their continued refinement will help realize the vision of a future where healthcare is accessible, fair, and personalized for all.

References

[1] O. Commission, *Health at a Glance: Europe 2024: State of Health in the EU Cycle*. OECD Publishing, 2024, p. 233. DOI: 10.1787/b3704e14-en.

[2] S. M. McKinney, M. Sieniek, V. Godbole, *et al.*, "International evaluation of an ai system for breast cancer screening", *Nature*, vol. 577, no. 7788, pp. 89–94, 2020. DOI: 10.1038/s41586-019-1799-6.

[3] T. Burki, "A new paradigm for drug development", *The Lancet Digital Health*, vol. 2, no. 5, e226–e227, 2020. DOI: 10.1016/S2589-7500(20)30088-1.

[4] M. Ibrahim, Y. A. Khalil, S. Amirrajab, *et al.*, "Generative ai for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges", *ArXiv Preprint*, 2024. DOI: 10.48550/arXiv.2407.00116.

[5] V. C. Pezoulas, D. I. Zaridis, E. Mylona, *et al.*, "Synthetic data generation methods in healthcare: A review on open-source tools and methods", *Computational and Structural Biotechnology Journal*, vol. 23, pp. 2892–2920, 2024. DOI: 10.1016/j.csbj.2024.07.005.

[6] A. H. Z. Nik, M. A. Riegler, P. Halvorsen, and A. M. Storås, "Generation of synthetic tabular healthcare data using generative adversarial networks", in *Multimedia Modeling*, Springer, 2023, pp. 434–446. DOI: 10.1007/978-3-031-27077-2_34.

[7] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6679–6687, 2021. DOI: 10.1609/aaai.v35i8.16826.

[8] M. Hernadez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions", *Methods of Information in Medicine*, vol. 62, no. S 01, e19–e38, 2023. DOI: 10.1055/s-0042-1760247.

[9] C. Umesh, M. Mahendra, S. Bej, O. Wolkenhauer, and M. Wolfien, "Challenges and applications in generative ai for clinical tabular data in physiology", *Pflügers Archiv-European Journal of Physiology*, pp. 1–12, 2024. DOI: 10.1007/s00424-024-03024-w.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes", *Arxiv Preprint*, 2013. DOI: 10.48550/arXiv.1312.6114.

[11] A. Turing, "Computing machinery and intelligence", *Mind*, vol. 59, pp. 433–60, 1950. DOI: 10.1093/mind/LIX.236.433.

[12] S. L. Andresen, "John mccarthy: Father of ai", *IEEE Intelligent Systems*, vol. 17, no. 5, pp. 84–85, 2002. DOI: 10.1109/MIS.2002.1039837.

[13] W. van Melle, "Mycin: A knowledge-based consultation program for infectious disease diagnosis", in *International Journal of Man-Machine Studies*, vol. 10, 1978, pp. 313–322. DOI: 10.1016/S0020-7373(78)80049-2.

[14] N. J. Nilsson, C. A. Rosen, and B. Raphael, *Application of intelligent automata to reconnaissance*. Rome Air Development Center, 1969.

[15] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine", *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966. DOI: 10.1145/365153.365168.

[16] F. B. Rogers, "The development of medlars", *Bulletin of the Medical Library Association*, vol. 52, no. 1, p. 150, 1964.

[17] G. Freiherr, *The seeds of artificial intelligence: SUMEX-AIM*. US Department of Health, Education, and Welfare, Public Health Service, 1980.

[18] S. M. Weiss, C. A. Kulikowski, S. Amarel, and A. Safir, "A model-based method for computer-aided medical decision-making", *Artificial Intelligence*, vol. 11, no. 1, pp. 145–172, 1978. DOI: 10.1016/0004-3702(78)90015-2.

[19] R. A. Miller, H. E. Pople Jr, and J. D. Myers, *Internist-I, an experimental computer-based diagnostic consultant for general internal medicine*. Springer, 1985, pp. 139–158. DOI: 10.1007/978-1-4612-5108-8_8.

[20] G. O. Barnett, J. J. Cimino, J. A. Hupp, and E. P. Hoffer, "Dxplain: An evolving diagnostic decision-support system", *Jama*, vol. 258, no. 1, pp. 67–74, 1987. DOI: 10.1001/jama.258.1.67.

[21] L. T. Powell, G. A. Diamond, P. K. Shah, and J. G. Ferguson, "Corsage: A critiquing system for coronary care", in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, American Medical Informatics Association, 1989, pp. 152–156.

[22] G. Edwards, R. Malor, A. Srinivasan, L. Lazarus, and P. Compton, "Peirs: A pathologist-maintained expert system for the interpretation of chemical pathology reports", *Pathology*, vol. 25, no. 1, pp. 27–34, 1993. DOI: 10.3109/00313029309068898.

[23] C. Kulikowski, "An opening chapter of the first generation of artificial intelligence in medicine: The first rutgers aim workshop, june 1975", *Yearbook of medical informatics*, vol. 24, no. 01, pp. 227–233, 2015. DOI: 10.15265/IY-2015-016.

[24] A. Esteva, B. Kuprel, R. A. Novoa, *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks", *Nature*, vol. 542, no. 7639, pp. 115–118, 2017. DOI: 10.1038/nature21056.

[25] D. Ferrucci, E. Brown, J. Chu-Carroll, *et al.*, "Building watson: An overview of the deepqa project", *AI magazine*, vol. 31, no. 3, pp. 59–79, 2010. DOI: 10.1609/aimag.v31i3.2303.

[26] E. Callaway, "It will change everything: Deepmind's ai makes gigantic leap in solving protein structures", *Nature*, vol. 588, no. 7837, pp. 203–205, 2020. DOI: 10.1038/d41586-020-03348-4.

[27] K. H. Goh, L. Wang, A. Y. K. Yeow, *et al.*, "Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare", *Nature communications*, vol. 12, no. 1, p. 711, 2021. DOI: 10.1038/s41467-021-20910-4.

[28] S. D. Mohanty, D. Lekan, T. P. McCoy, M. Jenkins, and P. Manda, "Machine learning for predicting readmission risk among the frail: Explainable ai for healthcare", *Patterns*, vol. 3, no. 1, 2022. DOI: 10.1016/j.patter.2021.100395.

[29] A. Alamgir, O. Mousa, Z. Shah, *et al.*, "Artificial intelligence in predicting cardiac arrest: Scoping review", *JMIR Medical Informatics*, vol. 9, no. 12, e30798, 2021. DOI: 10.2196/30798.

[30] R. Daneshjou, K. Vodrahalli, R. A. Novoa, *et al.*, "Disparities in dermatology ai performance on a diverse, curated clinical image set", *Science advances*, vol. 8, no. 31, eabq6147, 2022. DOI: 10.1126/sciadv.abq6147.

[31] P. J. Bevan and A. Atapour-Abarghouei, "Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification", in *MICCAI Workshop on Domain Adaptation and Representation Transfer*, Springer, 2022, pp. 1–11. DOI: 10.1007/978-3-031-16852-9_1.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets", in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014. DOI: 10.1145/3422622.

[33] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners", in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[34] OpenAI, *Gpt-3.5: Generative pre-trained transformer*, https://openai.com, 2022.

[35] J. Haemmerli, L. Sveikata, A. Nouri, *et al.*, "Chatgpt in glioma adjuvant therapy decision making: Ready to assume the role of a doctor in the tumour board?", *BMJ Health & Care Informatics*, vol. 30, no. 1, 2023. DOI: 10.1136/bmjhci-2023-100775.

[36] Y. Skandarani, P.-M. Jodoin, and A. Lalande, "Gans for medical image synthesis: An empirical study", *Journal of Imaging*, vol. 9, no. 3, p. 69, 2023. DOI: 10.3390/jimaging9030069.

[37] S. Pati, U. Baid, B. Edwards, *et al.*, "Federated learning enables big data for rare cancer boundary detection", *Nature communications*, vol. 13, no. 1, p. 7346, 2022. DOI: 10.1038/s41467-022-33407-5.

[38] Z. Papalamprakopoulou, D. Stavropoulos, S. Moustakidis, D. Avgerinos, M. Efremidis, and P. N. Kampaktsis, "Artificial intelligence-enabled atrial fibrillation detection using smartwatches: Current status and future perspectives", *Frontiers in Cardiovascular Medicine*, vol. 11, p. 1 432 876, 2024. DOI: 10.3389/fcvm.2024.1432876.

[39] T. Mishra, M. Wang, A. A. Metwally, *et al.*, "Pre-symptomatic detection of covid-19 from smartwatch data", *Nature Biomedical Engineering*, vol. 4, pp. 1208–1220, 2020. DOI: 10.1038/s41551-020-00640-6.

[40] D. Glass, "Graunt's life table", *Journal of the Institute of Actuaries*, vol. 76, no. 1, pp. 60–64, 1950.

[41] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations", *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958. DOI: 10.2307/2281868.

[42] J. P. Klein and M. L. Moeschberger, *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006. DOI: 10.1007/978-1-4757-2728-9.

[43] E. Lee, *Statistical methods for survival data analysis*. John Wiley & Sons, 2003. DOI: 10.1002/0471458546.

[44] E. Marubini and M. G. Valsecchi, *Analysing survival data from clinical trials and observational studies*. John Wiley & Sons, 2004, vol. 15. DOI: 10.1093/oxfordjournals.aje.a008913.

[45] D. R. Cox, "Regression models and life-tables", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972. DOI: 10.1111/j.2517-6161.1972.tb00899.x.

[46] W. Nelson, "Theory and applications of hazard plotting for censored failure data", *Technometrics*, vol. 14, no. 4, pp. 945–966, 1972. DOI: 10.1080/00401706.1972.10488991.

[47] O. Aalen, "Nonparametric inference for a family of counting processes", *The Annals of Statistics*, vol. 6, no. 2, pp. 701–726, 1978. DOI: 10.1214/aos/1176344247.

[48] S. J. Cutler and F. Ederer, "Maximum utilization of the life table method in analyzing survival", *Journal of chronic diseases*, vol. 8, no. 6, pp. 699–712, 1958. DOI: 10.1016/0021-9681(58)90126-7.

[49] N. E. Breslow, "Discussion of professor cox's paper", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, pp. 202–220, 1972. DOI: 10.1111/j.2517-6161.1972.tb00900.x.

[50] R. Tibshirani, "The lasso method for variable selection in the cox model", *Statistics in medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[51] P. J. Verweij and H. C. Van Houwelingen, "Penalized likelihood in cox regression", *Statistics in medicine*, vol. 13, no. 23-24, pp. 2427–2436, 1994. DOI: 10.1002/sim.4780132307.

[52] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005. DOI: 10.1111/j.1467-9868.2005.00503.x.

[53] H. Binder and M. Schumacher, "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models", *BMC bioinformatics*, vol. 9, pp. 1–10, 2008. DOI: 10.1186/1471-2105-9-14.

[54] D. G. Kleinbaum and M. Klein, *Survival analysis, a self-learning text*. Springer, 1996. DOI: 10.1007/978-1-4419-6646-9.

[55] H. Ishwaran and U. B. Kogalur, "Random survival forests for r", *R news*, vol. 7, no. 2, pp. 25–31, 2007.

[56] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers", *Machine learning*, vol. 29, pp. 131–163, 1997. DOI: 10.1023/A:1007465528199.

[57] F. M. Khan and V. B. Zubek, "Support vector regression for censored data (svrc): A novel tool for survival analysis", in *2008 Eighth IEEE International Conference on Data Mining*, IEEE, 2008, pp. 863–868. DOI: 10.1109/ICDM.2008.50.

[58] D. Faraggi and R. Simon, "A neural network model for survival data", *Statistics in medicine*, vol. 14, no. 1, pp. 73–82, 1995. DOI: 10.1002/sim.4780140108.

[59] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network", *BMC medical research methodology*, vol. 18, pp. 1–12, 2018. DOI: 10.1186/s12874-018-0482-1.

[60] H. Kvamme, Ø. Borgan, and I. Scheel, "Time-to-event prediction with neural networks and cox regression", *Journal of machine learning research*, vol. 20, no. 129, pp. 1–30, 2019.

[61] T. Ching, X. Zhu, and L. X. Garmire, "Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data", *PLoS computational biology*, vol. 14, no. 4, e1006076, 2018. DOI: 10.1371/journal.pcbi.1006076.

[62] J. Hao, Y. Kim, T. Mallavarapu, J. H. Oh, and M. Kang, "Cox-pasnet: Pathway-based sparse deep neural network for survival analysis", in *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*, IEEE, 2018, pp. 381–386. DOI: 10.1109/BIBM.2018.8621345.

[63] M. Luck, T. Sylvain, H. Cardinal, A. Lodi, and Y. Bengio, "Deep learning for patient-specific kidney graft survival analysis", *ArXiv Preprint*, 2017. DOI: 10.48550/arXiv.1705.10245.

[64] C. Lee, W. Zame, J. Yoon, and M. Van Der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks", in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018. DOI: 10.1609/aaai.v32i1.11842.

[65] C. Lee, J. Yoon, and M. Van Der Schaar, "Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data", *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 122–133, 2019. DOI: 10.1109/TBME.2019.2909027.

[66] E. Giunchiglia, A. Nemchenko, and M. van der Schaar, "Rnn-surv: A deep recurrent model for survival analysis", in *rtificial Neural Networks and Machine Learning – ICANN 2018*, Springer, 2018, pp. 23–32. DOI: 10.1007/978-3-030-01424-7_3.

[67] E. Martinsson, "Wtte-rnn: Weibull time to event recurrent neural network", M.S. thesis, Chalmers University of Technology & University of Gothenburg, 2016.

[68] J. P. Fine and R. J. Gray, "A proportional hazards model for the subdistribution of a competing risk", *Journal of the American statistical association*, vol. 94, no. 446, pp. 496–509, 1999. DOI: 10.1080/01621459.1999.10474144.

[69] Y. Li, W. Jia, Y. Kang, *et al.*, "Deepcomp: Which competing event will hit the patient first?", in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2020, pp. 629–636. DOI: 10.1109/BIBM49941.2020.9313333.

[70] G. Gupta, V. Sunder, R. Prasad, and G. Shroff, "Cresa: A deep learning approach to competing risks, recurrent event survival analysis", in *Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, PAKDD 2019, Macau, China, April 14-17, 2019, Proceedings, Part II 23*, Springer, 2019, pp. 108–122. DOI: 10.1007/978-3-030-16145-3_9.

[71] S. Chi, Y. Tian, F. Wang, Y. Wang, M. Chen, and J. Li, "Deep semisupervised multitask learning model and its interpretability for survival analysis", *IEEE journal of biomedical and health informatics*, vol. 25, no. 8, pp. 3185–3196, 2021. DOI: 10.1109/JBHI.2021.3064696.

[72] P. Huang, Y. Liu, *et al.*, "Deepcompete: A deep learning approach to competing risks in continuous time domain", in *AMIA annual symposium proceedings*, vol. 2020, 2021, p. 177.

[73] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests", *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982. DOI: 10.1001/jama.1982.03320430047030.

[74] L. Antolini, P. Boracchi, and E. Biganzoli, "A time-dependent discrimination index for survival data", *Statistics in medicine*, vol. 24, no. 24, pp. 3927–3944, 2005. DOI: 10.1002/sim.2427.

[75] J. M. Robins *et al.*, "Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers", in *Proceedings of the biopharmaceutical section, American statistical association*, San Francisco CA, vol. 24, 1993, p. 3. DOI: 10.1007/978-1-4757-1229-2_14.

[76] H. Murtaza, M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar, and A. Bano, "Synthetic data generation: State of the art in health care domain", *Computer Science Review*, vol. 48, p. 100 546, 2023. DOI: 10.1016/j.cosrev.2023.100546.

[77] D. B. Rubin, "Statistical disclosure limitation", *Journal of official Statistics*, vol. 9, no. 2, pp. 461–468, 1993.

[78] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695. DOI: 10.1109/CVPR52688.2022.01042.

[79] A. Ramesh, M. Pavlov, G. Goh, *et al.*, "Zero-shot text-to-image generation", in *International conference on machine learning*, PMLR, 2021, pp. 8821–8831.

[80] OpenAI, *Chatgpt: Optimizing language models for dialogue*, 2022.

[81] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression", in *Proceedings of the IEEE Symposium on Research in Security and Privacy*, Technical Report, SRI International, 1998.

[82] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity", *Acm transactions on knowledge discovery from data (tkdd)*, vol. 1, no. 1, 3–es, 2007. DOI: 10.1109/ICDE.2006.1.

[83] V. Puri, S. Sachdeva, and P. Kaur, "Privacy preserving publication of relational and transaction data: Survey on the anonymization of patient data", *Computer Science Review*, vol. 32, pp. 45–61, 2019. DOI: 10.1016/j.cosrev.2019.02.001.

[84] M. Jayabalan and M. E. Rana, "Anonymizing healthcare records: A study of privacy preserving data publishing techniques", *Advanced Science Letters*, vol. 24, no. 3, pp. 1694–1697, 2018. DOI: 10.1166/asl.2018.11139.

[85] K. El Emam and R. Hoptroff, *The synthetic data paradigm for using and sharing data*, 2019.

[86] A. Oganian, "V-dispersed synthetic data based on a mixture model with constraints", in *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2014, Ibiza, Spain, September 17-19, 2014. Proceedings*, Springer, 2014, pp. 200–212. DOI: 10.1007/978-3-319-11257-2_16.

[87] A. Oganian and J. Domingo-Ferrer, "Local synthesis for disclosure limitation that satisfies probabilistic k-anonymity criterion", *Transactions on data privacy*, vol. 10, no. 1, p. 61, 2017.

[88] P. Kumar and M. M. Shoukri, "Copula functions for modelling dependence structure with applications in the analysis of clinical data", *Journal of Indian Soc. Agric. Statist*, vol. 61, no. 2, pp. 179–191, 2007.

[89] Z. Wang, P. Myles, and A. Tucker, "Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy", *Computational Intelligence*, vol. 37, no. 2, pp. 819–851, 2021. DOI: 10.1111/coin.12427.

[90] S. Demarta and A. J. McNeil, "The t copula and related copulas", *International statistical review*, vol. 73, no. 1, pp. 111–129, 2005. DOI: 10.1111/j.1751-5823.2005.tb00254.x.

[91] Y. Park, J. Ghosh, and M. Shankar, "Perturbed gibbs samplers for generating large-scale privacy-safe synthetic health data", in *2013 IEEE International Conference on Healthcare Informatics*, IEEE, 2013, pp. 493–498. DOI: 10.1109/ICHI.2013.76.

[92] Y. Park and J. Ghosh, "Pegs: Perturbed gibbs samplers that generate privacy-compliant synthetic data", *Trans. Data Priv.*, vol. 7, no. 3, pp. 253–282, 2014.

[93] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via bayesian networks", *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 4, pp. 1–41, 2017. DOI: 10.1145/3134428.

[94] D. Kaur, M. Sobiesk, S. Patil, *et al.*, "Application of bayesian networks to generate synthetic health data", *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 801–811, 2021. DOI: 10.1093/jamia/ocaa303.

[95] R. E. Foraker, S. C. Yu, A. Gupta, *et al.*, *Spot the difference: Comparing results of analyses from real patient data and synthetic derivatives*, 2020. DOI: 10.1093/jamiaopen/ooaa060.

[96] R. Foraker, A. Guo, J. Thomas, *et al.*, "The national covid cohort collaborative: Analyses of original and computationally derived electronic health record data", *Journal of medical Internet research*, vol. 23, no. 10, e30697, 2021. DOI: 10.2196/30697.

[97] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique", *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002. DOI: 10.1613/jair.953.

[98] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning", in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, IEEE, 2008, pp. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.

[99] J. Vaidya, B. Shafiq, M. Asani, N. Adam, X. Jiang, and L. Ohno-Machado, "A scalable privacy-preserving data generation methodology for exploratory analysis", in *AMIA Annual Symposium Proceedings*, vol. 2017, 2018, p. 1695.

[100] K. E. Emam, L. Mosquera, and C. Zheng, "Optimizing the synthesis of clinical trial data using sequential trees", *Journal of the American Medical Informatics Association*, vol. 28, no. 1, pp. 3–13, 2021. DOI: 10.1093/jamia/ocaa249.

[101] K. El Emam, L. Mosquera, E. Jonker, and H. Sood, "Evaluating the utility of synthetic covid-19 case data", *JAMIA open*, vol. 4, no. 1, ooab012, 2021. DOI: 10.1093/jamiaopen/ooab012.

[102] T. Das, Z. Wang, and J. Sun, "Twin: Personalized clinical trial digital twin generation", in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 402–413. DOI: 10.1145/3580305.3599534.

[103] A. Koloi, V. S. Loukas, A. Sakellarios, *et al.*, "A comparison study on creating simulated patient data for individuals suffering from chronic coronary disorders", in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2023, pp. 1–4. DOI: 10.1109/EMBC40787.2023.10340194.

[104] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review", *Medical image analysis*, vol. 58, p. 101 552, 2019. DOI: 10.1016/j.media.2019.101552.

[105] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks", *ArXiv Preprint*, 2015. DOI: 10.48550/arXiv.1511.06732.

[106] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks", in *Machine learning for healthcare conference*, PMLR, 2017, pp. 286–305.

[107] M. K. Baowaly, C.-L. Liu, and K.-T. Chen, "Realistic data synthesis using enhanced generative adversarial networks", in *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, IEEE, 2019, pp. 289–292. DOI: 10.1109/AIKE.2019.00057.

[108] M. K. Baowaly, C.-C. Lin, C.-L. Liu, and K.-T. Chen, "Synthesizing electronic health records using improved generative adversarial networks", *Journal of the American Medical Informatics Association*, vol. 26, no. 3, pp. 228–241, 2019. DOI: 10.1093/jamia/ocy142.

[109] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan", *Advances in neural information processing systems*, vol. 32, 2019.

[110] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen, "Ctab-gan: Effective table data synthesizing", in *Asian Conference on Machine Learning*, PMLR, vol. 157, 2021, pp. 97–112.

[111] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks", *Proc. VLDB Endow.*, vol. 11, no. 10, pp. 1071–1083, 2018. DOI: 10.14778/3231751.3231757.

[112] C. Yan, Z. Zhang, S. Nyemba, and B. A. Malin, "Generating electronic health records with multiple data types and constraints", in *AMIA annual symposium proceedings*, vol. 2020, 2021, p. 1335.

[113] J. Jordon, J. Yoon, and M. Van Der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees", in *International conference on learning representations*, 2019.

[114] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data", *ArXiv Preprint*, 2016. DOI: 10.48550/arXiv.1610.05755.

[115] M. L. Fang, D. S. Dhami, and K. Kersting, "Dp-ctgan: Differentially private medical data generation using ctgans", in *International Conference on Artificial Intelligence in Medicine*, Springer, 2022, pp. 178–188. DOI: 10.1007/978-3-031-09342-5_17.

[116] J. Yoon, L. N. Drumright, and M. Van Der Schaar, "Anonymization through data synthesis using generative adversarial networks (ads-gan)", *IEEE journal of biomedical and health informatics*, vol. 24, no. 8, pp. 2378–2388, 2020. DOI: 10.1109/JBHI.2020.2980262.

[117] A. Torfi and E. A. Fox, "Corgan: Correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records", in *The thirty-third international flairs conference*, 2020.

[118] L. Wang, W. Zhang, and X. He, "Continuous patient-centric sequence generation via sequentially coupled adversarial learning", in *Database Systems for Advanced Applications: 24th International Conference, DASFAA 2019, Chiang Mai, Thailand, April 22–25, 2019, Proceedings, Part II 24*, Springer, 2019, pp. 36–52. DOI: 10.48550/arXiv.1610.05755.

[119] Z. Lin, A. Khetan, G. Fanti, and S. Oh, "Pacgan: The power of two samples in generative adversarial networks", *Advances in neural information processing systems*, vol. 31, 2018. DOI: 10.1109/JSAIT.2020.2983071.

[120] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans", *Advances in neural information processing systems*, vol. 30, 2017.

[121] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning internal representations by error propagation*. MIT Press, 1986, vol. 71, pp. 318–362.

[122] C. Ma, S. Tschiatschek, R. Turner, J. M. Hernández-Lobato, and C. Zhang, "Vaem: A deep generative model for heterogeneous mixed type data", *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 237–11 247, 2020.

[123] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling incomplete heterogeneous data using vaes", *Pattern Recognition*, vol. 107, p. 107 501, 2020. DOI: 10.1016/j.patcog.2020.107501.

[124] H. Akrami, S. Aydore, R. M. Leahy, and A. A. Joshi, "Robust variational autoencoder for tabular data with beta divergence", *ArXiv Preprint*, 2020. DOI: 10.48550/arXiv.2006.08204.

[125] S. M. Tazwar, M. Knobbout, E. H. Quesada, and M. Popa, "Tab-vae: A novel vae for generating synthetic tabular data", in *ICPRAM*, 2024, pp. 17–26. DOI: 10.5220/0012302400003654.

[126] L. V. H. Vardhan and S. Kok, "Generating privacy-preserving synthetic tabular data using oblivious variational autoencoders", in *Proceedings of the Workshop on Economics of Privacy and Data Labor at the 37 th International Conference on Machine Learning*, 2020.

[127] H. Li, S. Yu, and J. Principe, "Causal recurrent variational autoencoder for medical time series generation", in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, 2023, pp. 8562–8570. DOI: 10.1609/aaai.v37i7.26031.

[128] B. Dai and D. Wipf, "Diagnosing and enhancing vae models", *Preprint ArXiv*, 2019. DOI: 10.48550/arXiv.1903.05789.

[129] J. Tomczak and M. Welling, "Vae with a vampprior", in *International conference on artificial intelligence and statistics*, PMLR, 2018, pp. 1214–1223.

[130] J. Wu, K. Plataniotis, L. Liu, E. Amjadian, and Y. Lawryshyn, "Interpretation for variational autoencoder used to generate financial synthetic tabular data", *Algorithms*, vol. 16, no. 2, p. 121, 2023. DOI: 10.3390/a16020121.

[131] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models", *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[132] J. Kim, C. Lee, and N. Park, "Stasy: Score-based tabular data synthesis", *ArXiv Preprint*, 2022. DOI: 10.48550/arXiv.2210.04018.

[133] C. Lee, J. Kim, and N. Park, "Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis", in *International Conference on Machine Learning*, PMLR, 2023, pp. 18 940–18 956.

[134] H. Zhang, J. Zhang, B. Srinivasan, *et al.*, "Mixed-type tabular data synthesis with score-based diffusion in latent space", *ArXiv Preprint*, 2023. DOI: 10.48550/arXiv.2310.09656.

[135] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko, "Tabddpm: Modelling tabular data with diffusion models", in *International Conference on Machine Learning*, PMLR, 2023, pp. 17 564–17 579.

[136] S. Zheng and N. Charoenphakdee, "Diffusion models for missing value imputation in tabular data", *ArXiv Preprint*, 2022. DOI: 10.48550/arXiv.2210.17128.

[137] M. Villaizán-Vallelado, M. Salvatori, C. Segura, and I. Arapakis, "Diffusion models for tabular data imputation and synthetic data generation", *ArXiv Preprint*, 2024. DOI: 10.48550/arXiv.2407.02549.

[138] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners", *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[139] V. Borisov, K. Seßler, T. Leemann, M. Pawelczyk, and G. Kasneci, "Language models are realistic tabular data generators", *ArXiv Preprint*, 2022. DOI: 10.48550/arXiv.2210.06280.

[140] Z. Zhao, R. Birke, and L. Chen, "Tabula: Harnessing language models for tabular data synthesis", *ArXiv Preprint*, 2023. DOI: 10.48550/arXiv.2310.12746.

[141] J. Dahmen and D. Cook, "Synsys: A synthetic data generation system for healthcare applications", *Sensors*, vol. 19, no. 5, p. 1181, 2019. DOI: 10.3390/s19051181.

[142] H. Ping, J. Stoyanovich, and B. Howe, "Datasynthesizer: Privacy-preserving synthetic datasets", in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 2017, pp. 1–5. DOI: 10.1145/3085504.3091117.

[143] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods", *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969. DOI: 10.2307/1912791.

[144] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans", *Advances in neural information processing systems*, vol. 29, 2016.

[145] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium", *Advances in neural information processing systems*, vol. 30, 2017.

[146] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, *Bertscore: Evaluating text generation with bert*, 2019. DOI: 10.48550/arXiv.1904.09675.

[147] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation", in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135.

[148] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, "Boosting deep learning risk prediction with generative adversarial networks for electronic health records", in *2017 IEEE*

International Conference on Data Mining (ICDM), IEEE, 2017, pp. 787–792. DOI: 10.1109/ICDM.2017.93.

[149] F. Yang, Z. Yu, Y. Liang, *et al.*, "Grouped correlational generative adversarial networks for discrete electronic health records", in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2019, pp. 906–913. DOI: 10.1109/BIBM47256.2019.8983215.

[150] S. Rashidian, F. Wang, R. Moffitt, *et al.*, "Smooth-gan: Towards sharp and smooth synthetic ehr data generation", in *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18*, Springer, 2020, pp. 37–48. DOI: 10.1007/978-3-030-59137-3_4.

[151] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data", *BMC medical research methodology*, vol. 20, pp. 1–40, 2020. DOI: 10.1186/s12874-020-00977-1.

[152] S. Dash, A. Yale, I. Guyon, and K. P. Bennett, "Medical time-series data generation using generative adversarial networks", in *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18*, Springer, 2020, pp. 382–391. DOI: 10.1007/978-3-030-59137-3_34.

[153] J. Snoke, G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic, *General and specific utility measures for synthetic data*, 2018. DOI: 10.1111/rssa.12358.

[154] Z. Wang, P. Myles, and A. Tucker, "Generating and evaluating synthetic uk primary care data: Preserving data utility & patient privacy", in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, 2019, pp. 126–131. DOI: 10.1109/CBMS.2019.00036.

[155] S. Imtiaz, M. Arsalan, V. Vlassov, and R. Sadre, "Synthetic and private smart health care data generation using gans", in *2021 International Conference on Computer Communications and Networks (ICCCN)*, IEEE, 2021, pp. 1–7. DOI: 10.1109/ICCCN52240.2021.9522203.

[156] S. Bourou, A. El Saer, T.-H. Velivassaki, A. Voulkidis, and T. Zahariadis, "A review of tabular data synthesis using gans on an ids dataset", *Information*, vol. 12, no. 09, p. 375, 2021. DOI: 10.3390/info12090375.

[157] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles, "Generating high-fidelity synthetic patient data for assessing machine learning healthcare software", *NPJ digital medicine*, vol. 3, no. 1, pp. 1–13, 2020. DOI: 10.1038/s41746-020-00353-9.

[158] K. Bhanot, S. Dash, J. Pedersen, I. Guyon, and K. P. Bennett, "Quantifying resemblance of synthetic medical time-series", in *ESANN*, 2021. DOI: 10.14428/esann/2021.ES2021-108.

[159] D. Hazra and Y.-C. Byun, "Synsiggan: Generative adversarial networks for synthetic biomedical signal generation", *Biology*, vol. 9, no. 12, p. 441, 2020. DOI: 10.3390/biology9120441.

[160] D. Rankin, M. Black, R. Bond, J. Wallace, M. Mulvenna, G. Epelde, *et al.*, "Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing", *JMIR medical informatics*, vol. 8, no. 7, e18910, 2020. DOI: 10.2196/18910.

[161] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test", *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[162] C. Zhang, S. R. Kuppannagari, R. Kannan, and V. K. Prasanna, "Generative adversarial network for synthetic time series data generation in smart grids", in *2018 IEEE international conference on communications, control, and computing technologies for smart grids (SmartGridComm)*, IEEE, 2018, pp. 1–6. DOI: 10.1109/SmartGridComm.2018.8587464.

[163] Z. Zhang, C. Yan, D. A. Mesa, J. Sun, and B. A. Malin, "Ensuring electronic medical record simulation through better training, modeling, and evaluation", *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 99–108, 2020. DOI: 10.1093/jamia/ocz161.

[164] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. P. Bennett, "Assessing privacy and quality of synthetic health data", in *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*, 2019, pp. 1–4. DOI: 10.1145/3359115.3359124.

[165] J. Walonoski, S. Klaus, E. Granger, *et al.*, "Synthea™ novel coronavirus (covid-19) model and synthetic data set", *Intelligence-based medicine*, vol. 1, p. 100 007, 2020. DOI: 10.1016/j.ibmed.2020.100007.

[166] M. Hittmeir, A. Ekelhart, and R. Mayer, "On the utility of synthetic data: An empirical evaluation on machine learning tasks", in *Proceedings of the 14th International Conference on Availability, Reliability and Security*, 2019, pp. 1–6. DOI: 10.1145/3339252.3339281.

[167] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional gans", *ArXiv Preprint*, 2017. DOI: 10.48550/arXiv.1706.02633.

[168] R. Venugopal, N. Shafqat, I. Venugopal, B. M. J. Tillbury, H. D. Stafford, and A. Bourazeri, "Privacy preserving generative adversarial networks to model electronic health records", *Neural Networks*, vol. 153, pp. 339–348, 2022. DOI: 10.1016/j.neunet.2022.06.022.

[169] D. Lee, H. Yu, X. Jiang, *et al.*, "Generating sequential electronic health records using dual adversarial autoencoder", *Journal of the American Medical Informatics Association*, vol. 27, no. 9, pp. 1411–1419, 2020. DOI: 10.1093/jamia/ocaa119.

[170] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, *et al.*, "Privacy-preserving generative deep neural networks support clinical data sharing", *Circulation: Cardiovascular Quality and Outcomes*, vol. 12, no. 7, e005122, 2019. DOI: 10.1161/CIRCOUTCOMES.118.005122.

[171] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney, "Privacy preserving synthetic data release using deep learning", in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, Springer, 2019, pp. 510–526. DOI: 10.1007/978-3-030-10925-7_31.

[172] C. A. Libbi, J. Trienes, D. Trieschnigg, and C. Seifert, "Generating synthetic training data for supervised de-identification of electronic health records", *Future Internet*, vol. 13, no. 5, p. 136, 2021. DOI: 10.3390/fi13050136.

[173] J. Guan, R. Li, S. Yu, and X. Zhang, "A method for generating synthetic electronic medical record text", *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 18, no. 1, pp. 173–182, 2019. DOI: 10.1109/TCBB.2019.2948985.

[174] Y. Du, S. Lin, and Z. Huang, "Generation of semantic patient data for depression", in *Health Information Science: 6th International Conference, HIS 2017, Moscow, Russia, October 7-9, 2017, Proceedings 6*, Springer, 2017, pp. 102–112. DOI: 10.1007/978-3-319-69182-4_11.

[175] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. P. Bennett, "Generation and evaluation of privacy preserving synthetic health data", *Neurocomputing*, vol. 416, pp. 244–255, 2020. DOI: 10.1016/j.neucom.2019.12.136.

[176] N. Rieke, J. Hancox, W. Li, *et al.*, *The future of digital health with federated learning*, 2020. DOI: 10.1038/s41746-020-00323-1.

[177] T. Haritha and A. Anitha, "Asymmetric consortium blockchain and homomorphically polynomial-based pir for secured smart parking systems", *Computers, Materials & Continua*, vol. 75, no. 2, 2023. DOI: 10.32604/cmc.2023.036278.

[178] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation", in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, Springer, 2019, pp. 92–104. DOI: 10.1007/978-3-030-11723-8_9.

[179] F. Cremonesi, V. Planat, V. Kalokyri, *et al.*, "The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform", *Journal of Biomedical Informatics*, vol. 141, p. 104 338, 2023. DOI: 10.1016/j.jbi.2023.104338.

[180] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge ai: On-demand accelerating deep neural network inference via edge computing", *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 447–457, 2019. DOI: 10.1109/TWC.2019.2946140.

[181] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge ai: Algorithms and systems", *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2167–2191, 2020. DOI: 10.1109/COMST.2020.3007787.

[182] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data", in *Artificial intelligence and statistics*, PMLR, 2017, pp. 1273–1282.

[183] P. Kairouz, H. B. McMahan, B. Avent, *et al.*, "Advances and open problems in federated learning", *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021. DOI: 10.1561/2200000083.

[184] B. Liu, N. Lv, Y. Guo, and Y. Li, "Recent advances on federated learning: A systematic survey", *Neurocomputing*, p. 128 019, 2024. DOI: 10.1016/j.neucom.2024.128019.

[185] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics", *Journal of healthcare informatics research*, vol. 5, pp. 1–19, 2021. DOI: 10.1007/s41666-020-00082-4.

[186] P. Foley, M. J. Sheller, B. Edwards, *et al.*, "Openfl: The open federated learning library", *Physics in Medicine & Biology*, vol. 67, no. 21, p. 214 001, 2022. DOI: 10.1088/1361-6560/ac97d9.

[187] Y. Liu, L. Zhang, N. Ge, and G. Li, "A systematic literature review on federated learning: From a model quality perspective", *ArXiv Preprint*, 2020. DOI: 10.48550/arXiv.2012.01973.

[188] Prayitno, C.-R. Shyu, K. T. Putra, *et al.*, "A systematic review of federated learning in the healthcare area: From the perspective of data properties and applications", *Applied Sciences*, vol. 11, no. 23, p. 11 191, 2021. DOI: 10.3390/app112311191.

[189] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, "Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results", *Medical image analysis*, vol. 65, p. 101 765, 2020. DOI: 10.1016/j.media.2020.101765.

[190] B. Pfitzner, N. Steckhan, and B. Arnrich, "Federated learning in a medical context: A systematic literature review", *ACM Transactions on Internet Technology (TOIT)*, vol. 21, no. 2, pp. 1–31, 2021. DOI: 10.1145/3412357.

[191] D. Cha, M. Sung, Y.-R. Park, *et al.*, "Implementing vertical federated learning using autoencoders: Practical application, generalizability, and utility study", *JMIR medical informatics*, vol. 9, no. 6, e26598, 2021. DOI: 10.2196/26598.

[192] Y. Otoum, Y. Wan, and A. Nayak, "Federated transfer learning-based ids for the internet of medical things (iomt)", in *2021 IEEE Globecom Workshops (GC Wkshps)*, IEEE, 2021, pp. 1–6. DOI: 10.1109/GCWkshps52748.2021.9682118.

[193] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, "Fedhealth: A federated transfer learning framework for wearable healthcare", *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 83–93, 2020. DOI: 10.1109/MIS.2020.2988604.

[194] Y. Liu, T. Fan, T. Chen, Q. Xu, and Q. Yang, "Fate: An industrial grade platform for collaborative learning with data protection", *Journal of Machine Learning Research*, vol. 22, no. 226, pp. 1–6, 2021.

[195] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?", *ArXiv Preprint*, 2019. DOI: 10.48550/arXiv.1911.07963.

[196] A. Ziller, A. Trask, A. Lopardo, *et al.*, "Pysyft: A library for easy federated learning", *Federated Learning Systems: Towards Next-Generation AI*, pp. 111–139, 2021. DOI: 10.1007/978-3-030-70604-3_5.

[197] H. Ludwig, N. Baracaldo, G. Thomas, *et al.*, "Ibm federated learning: An enterprise framework white paper", *ArXiv Preprint*, 2020. DOI: 10.48550/arXiv.2007.10987.

[198] D. J. Beutel, T. Topal, A. Mathur, *et al.*, "Flower: A friendly federated learning research framework", *ArXiv Preprint*, 2020. DOI: 10.48550/arXiv.2007.14390.

[199] D. Zeng, S. Liang, X. Hu, H. Wang, and Z. Xu, "Fedlab: A flexible federated learning framework", *Journal of Machine Learning Research*, vol. 24, no. 100, pp. 1–7, 2023.

[200] C. He, S. Li, J. So, *et al.*, "Fedml: A research library and benchmark for federated machine learning", *ArXiv Preprint*, 2020. DOI: 10.48550/arXiv.2007.13518.

[201] F. Lai, Y. Dai, S. Singapuram, *et al.*, "Fedscale: Benchmarking model and system performance of federated learning at scale", in *International conference on machine learning*, PMLR, 2022, pp. 11 814–11 827. DOI: 10.1145/3477114.3488760.

[202] S. Silva, A. Altmann, B. Gutman, and M. Lorenzi, "Fed-biomed: A general open-source frontend framework for federated learning in healthcare", in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction*

with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2, Springer, 2020, pp. 201–210. DOI: 10.1007/978-3-030-60548-3_20.

[203] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks", *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[204] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging", *ArXiv Preprint*, 2020. DOI: 10.48550/arXiv.2002.06440.

[205] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning", in *International conference on machine learning*, PMLR, 2020, pp. 5132–5143.

[206] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization", *Advances in neural information processing systems*, vol. 33, pp. 7611–7623, 2020.

[207] F. Zhang, D. Kreuter, Y. Chen, *et al.*, "Recent methodological advances in federated learning for healthcare", *Patterns*, vol. 5, no. 6, 2024. DOI: 10.1016/j.patter.2024.101006.

[208] S. Banerjee, R. Misra, M. Prasad, E. Elmroth, and M. H. Bhuyan, "Multi-diseases classification from chest-x-ray: A federated deep learning approach", in *AI 2020: Advances in Artificial Intelligence: 33rd Australasian Joint Conference, AI 2020, Canberra, ACT, Australia, November 29–30, 2020, Proceedings 33*, Springer, 2020, pp. 3–15. DOI: 10.1007/978-3-030-64984-5_1.

[209] S. H. Khan and M. G. R. Alam, "A federated learning approach to pneumonia detection", in *2021 international conference on engineering and emerging technologies (ICEET)*, IEEE, 2021, pp. 1–6. DOI: 10.1109/ICEET53442.2021.9659591.

[210] S. Baghersalimi, T. Teijeiro, D. Atienza, and A. Aminifar, "Personalized real-time federated learning for epileptic seizure detection", *IEEE journal of biomedical and health informatics*, vol. 26, no. 2, pp. 898–909, 2021. DOI: 10.1109/JBHI.2021.3096127.

[211] S. Liu, X. Guo, S. Qi, H. Wang, and X. Chang, "Learning personalized brain functional connectivity of mdd patients from multiple sites via federated bayesian networks", *ArXiv Preprint*, 2023. DOI: 10.48550/arXiv.2301.02423.

[212] A. E. Cetinkaya, M. Akin, and S. Sagiroglu, "A communication efficient federated learning approach to multi chest diseases classification", in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, IEEE, 2021, pp. 429–434. DOI: 10.1109/UBMK52708.2021.9558913.

[213] A. Jiménez-Sánchez, M. Tardy, M. A. G. Ballester, D. Mateus, and G. Piella, "Memory-aware curriculum federated learning for breast cancer classification", *Computer Methods and Programs in Biomedicine*, vol. 229, p. 107 318, 2023. DOI: 10.1016/j.cmpb.2022.107318.

[214] Y. Wang, K. Wang, X. Peng, *et al.*, "Deepsdm: Boundary-aware pneumothorax segmentation in chest x-ray images", *Neurocomputing*, vol. 454, pp. 201–211, 2021. DOI: 10.1016/j.neucom.2021.05.029.

[215] G. Elmas, S. U. Dar, Y. Korkmaz, *et al.*, "Federated learning of generative image priors for mri reconstruction", *IEEE Transactions on Medical Imaging*, vol. 42, no. 7, pp. 1996–2009, 2022. DOI: 10.1109/TMI.2022.3220757.

[216] D. K. Zhang, F. Toni, and M. Williams, "A federated cox model with non-proportional hazards", in *Multimodal AI in healthcare: A paradigm shift in health intelligence*, Springer, 2022, pp. 171–185. DOI: 10.1007/978-3-031-14771-5_12.

[217] C. R. Hansen, G. Price, M. Field, *et al.*, "Larynx cancer survival model developed through open-source federated learning", *Radiotherapy and Oncology*, vol. 176, pp. 179–186, 2022. DOI: 10.1016/j.radonc.2022.09.023.

[218] Y. Zhang, Z. Li, X. Han, *et al.*, "Pseudo-data based self-supervised federated learning for classification of histopathological images", *IEEE Transactions on Medical Imaging*, 2023. DOI: 10.1109/TMI.2023.3323540.

[219] C. Dwork, "Differential privacy: A survey of results", in *International conference on theory and applications of models of computation*, Springer, 2008, pp. 1–19. DOI: 10.1007/978-3-540-79228-4_1.

[220] C. Gentry, "Fully homomorphic encryption using ideal lattices", in *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 2009, pp. 169–178. DOI: 10.1145/1536414.1536440.

[221] W. Dai, A. Kumar, J. Wei, Q. Ho, G. Gibson, and E. Xing, "High-performance distributed ml at scale through parameter server consistency models", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 2015.

[222] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge", *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pp. 1–7, 2019. DOI: 10.1109/ICC.2019.8761315.

[223] Z. Charles and D. Papailiopoulos, "Gradient coding using the stochastic block model", in *IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 1998–2002. DOI: 10.1109/ISIT.2018.8437887.

[224] S. Caldas, J. Konecný, H. McMahan, and A. Talwalkar, "Expanding the reach of federated learning by reducing client resource requirements", *ArXiv Preprint*, 2018. DOI: /10.48550/arXiv.1812.07210.

[225] X. Lian, C. Zhang, H. Zhang, C. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent", in *NIPS*, 2017.

[226] R. Ranganath, A. Perotte, N. Elhadad, and D. Blei, "Deep survival analysis", in *Machine Learning for Healthcare Conference*, PMLR, 2016, pp. 101–114.

[227] S. Liverani, L. Leigh, I. L. Hudson, and J. E. Byles, "Clustering method for censored and collinear survival data", *Computational Statistics*, vol. 36, pp. 35–60, 2021. DOI: 10.1007/s00180-020-01000-3.

[228] R. Ranganath, D. Tran, J. Altosaar, and D. Blei, "Operator variational inference", *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[229] J. W. Tukey, "The philosophy of multiple comparisons", *Statistical science*, pp. 100–116, 1991. DOI: 10.1214/ss/1177011945.

[230] E. L. Lehmann and J. P. Romano, *Generalizations of the familywise error rate*. Springer, 2012. DOI: 10.1214/009053605000000084.

[231] M. J. Van der Laan, S. Dudoit, and K. S. Pollard, "Multiple testing. part ii. step-down procedures for control of the family-wise error rate", *Statistical applications in genetics and molecular biology*, vol. 3, no. 1, 2004. DOI: 10.2202/1544-6115.1041.

[232] S. Holm, "A simple sequentially rejective multiple test procedure", *Scandinavian journal of statistics*, pp. 65–70, 1979.

[233] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library", *Advances in neural information processing systems*, vol. 32, 2019.

[234] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "Scipy 1.0: Fundamental algorithms for scientific computing in python", *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[235] W. B. Nelson, *Applied life data analysis*. John Wiley & Sons, 2005.

[236] K.-L. Lim, X. Jiang, and C. Yi, "Deep clustering with variational autoencoder", *IEEE Signal Processing Letters*, vol. 27, pp. 231–235, 2020. DOI: 10.1109/LSP.2020.2965328.

[237] J. T. McCoy, S. Kroon, and L. Auret, "Variational autoencoders for missing data imputation with application to a simulated milling circuit", *IFAC-PapersOnLine*, vol. 51, no. 21, pp. 141–146, 2018. DOI: 10.1016/j.ifacol.2018.09.406.

[238] C. Chadebec and S. Allassonniere, "Data augmentation with variational autoencoders and manifold sampling (2021)", in *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections*, Cham: Springer International Publishing, 2021, pp. 184–192. DOI: 10.1007/978-3-030-88210-5_17.

[239] Z. Gu, L. He, P. Li, P. Sun, J. Shi, and Y. Yang, "Frepd: A robust federated learning framework on variational autoencoder.", *Comput. Syst. Sci. Eng.*, vol. 39, no. 3, pp. 307–320, 2021. DOI: 10.32604/csse.2021.017969.

[240] M. Polato, "Federated variational autoencoder for collaborative filtering", in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–8. DOI: 10.1109/IJCNN52387.2021.9533358.

[241] C. Nagpal, X. Li, and A. Dubrawski, "Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks", *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3163–3175, 2021. DOI: 10.1109/JBHI.2021.3052441.

[242] P. C. Austin, D. S. Lee, and J. P. Fine, "Introduction to the analysis of survival data in the presence of competing risks", *Circulation*, vol. 133, no. 6, pp. 601–609, 2016. DOI: 10.1161/CIRCULATIONAHA.115.017719.

[243] B. L. Welch, "The generalization of 'student's'problem when several different population variances are involved", *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 1947. DOI: 10.2307/2332510.

[244] I. Covert, S. M. Lundberg, and S.-I. Lee, *Understanding global feature contributions with additive importance measures*, 2020.

[245] N. Kostantinos, "Gaussian mixtures and their applications to signal processing", *Advanced signal processing handbook: theory and implementation for radar, sonar, and medical imaging real time systems*, pp. 3–1, 2000.

[246] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. Springer, 2006, vol. 4.

[247] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault", in *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 399–410. DOI: 10.1109/DSAA.2016.49.

[248] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 2000, pp. 17–33. DOI: 10.1007/978-1-4757-3264-1.

[249] L. C. Tiao, "Density Ratio Estimation for KL Divergence Minimization between Implicit Distributions", *tiao.io*, 2018.

[250] K. Choi, C. Meng, Y. Song, and S. Ermon, "Density ratio estimation via infinitesimal classification", in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 151, PMLR, 2022, pp. 2552–2573.

[251] B. Rhodes, K. Xu, and M. U. Gutmann, "Telescoping density-ratio estimation", in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 4905–4916.

[252] M. Elasri, O. Elharrouss, S. Al-Maadeed, and H. Tairi, "Image generation: A review", *Neural Processing Letters*, vol. 54, no. 5, pp. 4609–4646, 2022. DOI: 10.1007/s11063-022-10777-x.

[253] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models", *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–37, 2023. DOI: 10.1145/3617680.

[254] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, "Video transformers: A survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. DOI: 10.1109/TPAMI.2023.3243465.

[255] M. Goldblum, M. Finzi, K. Rowan, and A. G. Wilson, "The no free lunch theorem, kolmogorov complexity, and the role of inductive biases in machine learning", *ArXiv Preprint*, 2023. DOI: 10.48550/arXiv.2304.05366.

[256] A. Goyal and Y. Bengio, "Inductive biases for deep learning of higher-level cognition", *Proceedings of the Royal Society A*, vol. 478, no. 2266, 2022. DOI: 10.1098/rspa.2021.0068.

[257] S. Müller, N. Hollmann, S. P. Arango, J. Grabocka, and F. Hutter, "Transformers can do bayesian inference", *ArXiv Preprint*, 2021. DOI: 10.48550/arXiv.2112.10510.

[258] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "Tabpfn: A transformer that solves small tabular classification problems in a second", *ArXiv Preprint*, 2022. DOI: 10.48550/arXiv.2207.01848.

[259] S. J. Pan and Q. Yang, "A survey on transfer learning", *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010. DOI: 10.1109/TKDE.2009.191.

[260] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review", *BMC Medical Imaging 2022 22:1*, vol. 22, no. 1, pp. 1–13, 2022. DOI: 10.1186/S12880-022-00793-7.

[261] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning", *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016. DOI: 10.1186/s40537-016-0043-6.

[262] F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu, "Cross-domain co-extraction of sentiment and topic lexicons", in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2012, pp. 410–419.

[263] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptation regularization: A general framework for transfer learning", *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, pp. 1076–1089, 2014. DOI: 10.1109/TKDE.2013.111.

[264] L. Duan, D. Xu, and S.-F. Chang, "Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1338–1345, 2012. DOI: 10.1109/CVPR.2012.6247819.

[265] Z. Wang and J. Sun, "Transtab: Learning transferable tabular transformers across tables", *Advances in Neural Information Processing Systems*, vol. 35, pp. 2902–2915, 2022.

[266] B. Zhu, X. Shi, N. Erickson, M. Li, G. Karypis, and M. Shoaran, "Xtab: Cross-table pretraining for tabular transformers", in *International Conference on Machine Learning*, 2023.

[267] R. Winkler and S. Makridakis, "The combination of forecasts", *Journal of the Royal Statistical Society. Series A (General)*, vol. 146, pp. 150–157, Jan. 1983. DOI: 10.2307/2982011.

[268] R. T. Clemen and R. L. Winkler, "Combining economic forecasts", *Journal of Business & Economic Statistics*, vol. 4, no. 1, pp. 39–46, 1986. DOI: 10.2307/1391385.

[269] S. Thrun and L. Pratt, *Learning to Learn: Introduction and Overview*. Springer US, 1998, pp. 3–17. DOI: 10.1007/978-1-4615-5529-2_1.

[270] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks", in *International conference on machine learning*, PMLR, 2017, pp. 1126–1135.

[271] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, "Meta-learning in neural networks: A survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 5149–5169, 2020. DOI: 10.1109/TPAMI.2021.3079209.

[272] J. Gordon, J. F. Bronskill, M. Bauer, S. Nowozin, and R. E. Turner, "Meta-learning probabilistic inference for prediction", in *International Conference on Learning Representations*, 2018.

[273] K. Smith-Miles, "Cross-disciplinary perspectives on meta-learning for algorithm selection", *ACM Comput. Surv.*, vol. 41, 6:1–6:25, 2009. DOI: 10.1145/1456650.1456656.

[274] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, "Bilevel programming for hyperparameter optimization and meta-learning", in *International Conference on Machine Learning*, 2018.

[275] K. Gao and O. Sener, "Modeling and optimization trade-off in meta-learning", in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 11 154–11 165.

[276] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world", *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017. DOI: 10.1109/IROS.2017.8202133.

[277] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Standardised metrics and methods for synthetic tabular data evaluation", *Authorea Preprints*, 2023. DOI: 10.1055/s-0042-1760247.

[278] M. Ye, X. Fang, B. Du, P. C. Yuen, and D. Tao, "Heterogeneous federated learning: State-of-the-art and research challenges", *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–44, 2023. DOI: 10.1145/3625558.

[279] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data", *ArXiv Preprint*, 2018. DOI: 10.48550/arXiv.1806.00582.

[280] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data", *ArXiv Preprint*, 2019. DOI: 10.48550/arXiv.1907.02189.

[281] A. Boles, R. Kandimalla, and P. H. Reddy, "Dynamics of diabetes and obesity: Epidemiological perspective", *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1863, no. 5, pp. 1026–1036, 2017. DOI: 10.1016/j.bbadis.2017.01.016.

[282] J. K. Alexander, "Obesity and coronary heart disease", *The American journal of the medical sciences*, vol. 321, no. 4, pp. 215–224, 2001. DOI: 10.1097/00000441-200104000-00002.

[283] J. A. Seiglie, M.-E. Marcus, C. Ebert, *et al.*, "Diabetes prevalence and its relationship with education, wealth, and bmi in 29 low-and middle-income countries", *Diabetes care*, vol. 43, no. 4, pp. 767–775, 2020. DOI: 10.2337/dc19-1782.

[284] C. Dwork, "Differential privacy", in *International colloquium on automata, languages, and programming*, Springer, 2006, pp. 1–12. DOI: 10.1007/11787006_1.

[285] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes", in *International conference on the theory and applications of cryptographic techniques*, Springer, 1999, pp. 223–238. DOI: 10.1007/3-540-48910-X_16.

[286] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", *science*, vol. 313, no. 5786, pp. 504–507, 2006. DOI: 10.1126/science.1127647.

[287] D. W. Hosmer Jr, S. Lemeshow, and S. May, *Applied survival analysis: regression modeling of time-to-event data*. John Wiley & Sons, 2008, vol. 618. DOI: 10.1002/9780470258019.

[288] W. A. Knaus, F. E. Harrell, J. Lynn, *et al.*, "The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults", *Annals of internal medicine*, vol. 122, no. 3, pp. 191–203, 1995. DOI: 10.7326/0003-4819-122-3-199502010-00007.

[289] J. A. Foekens, H. A. Peters, M. P. Look, *et al.*, "The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients", *Cancer research*, vol. 60, no. 3, pp. 636–43, 2000.

[290] M. Schumacher, G. Bastert, H. Bojar, *et al.*, "Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group", *Journal of Clinical Oncology*, vol. 12, no. 10, pp. 2086–2093, 1994. DOI: 10.1200/JCO.1994.12.10.2086.

[291] A. Dispenzieri, J. A. Katzmann, R. A. Kyle, *et al.*, "Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population", in *Mayo Clinic Proceedings*, Elsevier, vol. 87, 2012, pp. 517–523. DOI: 10.1016/j.mayocp.2012.03.009.

[292] N. E. Breslow and N. Chatterjee, "Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis", *Journal of the Royal Statistical Society:*

Series C (Applied Statistics), vol. 48, no. 4, pp. 457–468, 1999. DOI: 10.1111/1467-9876.00165.

[293] B. Pereira, S.-F. Chin, O. M. Rueda, *et al.*, "The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes", *Nature communications*, vol. 7, no. 1, pp. 1–16, 2016. DOI: 10.1038/ncomms11479.

[294] T. M. Therneau, "Extending the cox model", in *Proceedings of the first Seattle symposium in biostatistics: survival analysis*, Springer, 1997, pp. 51–84. DOI: 10.1007/978-1-4684-6316-3_5.

[295] Angelo Canty and B. D. Ripley, *Boot: Bootstrap r (s-plus) functions*, R package version 1.3-30, 2024.

[296] T. M. Therneau, *A package for survival analysis in r*, R package version 3.5-8, 2024.

[297] L. C. de Wreede, M. Fiocco, and H. Putter, "mstate: An R package for the analysis of competing risks and multi-state models", *Journal of Statistical Software*, vol. 38, no. 7, pp. 1–30, 2011. DOI: 10.18637/jss.v038.i07.

[298] B. Becker and R. Kohavi, *Adult*, UCI Machine Learning Repository, 1996. DOI: 10.24432/C5XW20.

[299] R. Kohavi, *Census income*, UCI Machine Learning Repository, 1996. DOI: 10.24432/C5GP7S.

[300] K. Fernandes, P. Vinagre, P. Cortez, and P. Sernadela, *Online news popularity*, UCI Machine Learning Repository, 2015. DOI: 10.24432/C5NS3V.

[301] S. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. Chan, *Kdd cup 1999 data*, UCI Machine Learning Repository, 1999. DOI: 10.24432/C51C7N.

[302] K. Woźnica, P. Wilczyński, and P. Biecek, "Sefnet: Bridging tabular datasets with semantic feature nets", *ArXiv Preprint*, 2023. DOI: 10.48550/arXiv.2306.11636.

[303] L. S. Shapley, "A value for n-person games", *Contribution to the Theory of Games*, vol. 2, 1953.

[304] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", in *Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2017.

[305] S. K. Bechis, P. R. Carroll, and M. R. Cooperberg, "Impact of age at diagnosis on prostate cancer treatment and survival", *Journal of Clinical Oncology*, vol. 29, no. 2, pp. 235–241, 2011. DOI: 10.1200/JCO.2010.30.2075.

[306] Y. B. Cheung, F. Gao, and K. S. Khoo, "Age at diagnosis and the choice of survival analysis methods in cancer epidemiology", *Journal of clinical epidemiology*, vol. 56, no. 1, pp. 38–43, 2003. DOI: 10.1016/S0895-4356(02)00536-X.

[307] E. E. van Eeghen, S. D. Bakker, A. van Bochove, and R. J. Loffeld, "Impact of age and comorbidity on survival in colorectal cancer", *Journal of gastrointestinal oncology*, vol. 6, no. 6, p. 605, 2015. DOI: 10.3978/j.issn.2078-6891.2015.070.

[308] C. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilita", *Pubblicazioni del R istituto superiore di scienze economiche e commericiali di firenze*, vol. 8, pp. 3–62, 1936.

[309] G. W. Brier, "Verification of forecasts expressed in terms of probability", *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[310] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, "Assessment and comparison of prognostic classification schemes for survival data", *Statistics in medicine*, vol. 18, no. 17-18, pp. 2529–2545, 1999. DOI: 10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5.

Appendices

Appendix A

Variational Autoencoder

A.1 Vanilla Variational Autoencoder

The original VAE, introduced in 2013 by [10], provides a robust approach for performing Bayesian inference using DNNs. This method addresses the problem of modeling a dataset consisting of N IID samples x_i of a continuous or discrete variable, where $i \in 1, 2, ..., N$. The data are assumed to be generated through the following random process (depicted in Figure A.1):

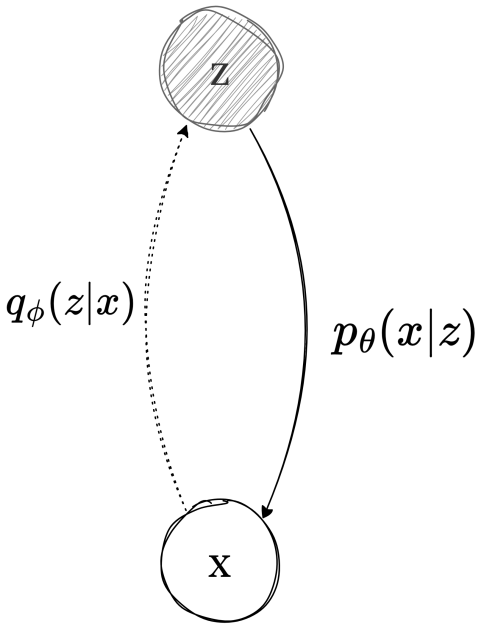


Figure A.1: Bayesian VAE vanilla model. The shaded circle represents the latent variable z, while the white circle denotes the observable variable x. The probabilities $p_\theta(x|z)$ and $q_\phi(z|x)$ refer to the generative model and the variational approximation to the posterior, respectively, as the true posterior $p_\theta(z|x)$ is unknown.

1. A latent variable z_i is sampled from a prior probability distribution $p(z)$.

   The original research [10] assumes a form $p_\theta(z)$, i.e., the prior depends on some parameters θ, but its main result drops this dependence. Therefore, a simple prior $p(z)$ is assumed in this research.

2. A conditional distribution, $p_\theta(x|z)$, with parameters θ generates the observed values, x_i. A generative model governs this process. Certain assumptions are made, including the differentiability of pdfs, $p(z)$, and $p_\theta(x|z)$, regarding θ and z.

The latent variable z and the parameters θ are unknown. Without simplifying assumptions, evaluating the marginal likelihood $p_\theta(x) = \int p(z)p_\theta(x|z)dz$ is infeasible. The true posterior density $p_\theta(z|x)$, which serves as the target for approximation, can be defined as Equation (A.1) using the theorem of Bayes:

$$p_\theta(z|x) = \frac{p_\theta(x|z)p(z)}{p_\theta(x)}. \tag{A.1}$$

However, since the marginal likelihood $p_\theta(x)$ is often intractable, directly computing the true posterior $p_\theta(z|x)$ becomes impractical.

Variational methods offer a solution by introducing a variational approximation, $q_\phi(z|x)$, to the true posterior. This approximation involves optimizing the best parameters for a chosen family of distributions. The quality of the approximation depends on the expressiveness of this parametric family.

A.1.1 Evidence Lower BOund Derivation

To effectively address the optimization problem, developing a comprehensive target function is essential. Assuming the data points x_i are IID, the marginal likelihood of a set of points $\{x_i\}_{i=1}^N$ can be written as

$$\log p_\theta(x_1, x_2, ..., x_N) = \sum_{i=1}^{N} \log p_\theta(x_i), \tag{A.2}$$

where each term can be expressed as

$$p_\theta(x) = \int p_\theta(x, z)dz = \int p_\theta(x, z)\frac{q_\phi(z|x)}{q_\phi(z|x)}dz = \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x, z)}{q_\phi(z|x)}\right]. \tag{A.3}$$

Using Jensen's inequality, the following lower bound can be derived:

$$\log p_\theta(x) = \log\left[\mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x, z)}{q_\phi(z|x)}\right]\right] \geq \mathbb{E}_{q_\phi(z|x)}\left[\log\frac{p_\theta(x, z)}{q_\phi(z|x)}\right]. \tag{A.4}$$

Rearranging Equation (A.4), the bound becomes

$$\begin{aligned}
&\mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)}{q_\phi(z|x)}\right)\right] \\
&= \int q_\phi(z|x)\log\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}dz \\
&= \int q_\phi(z|x)\log\frac{p(z)}{q_\phi(z|x)}dz + \int q_\phi(z|x)\log p_\theta(x|z)dz \\
&= -\int q_\phi(z|x)\log\frac{q_\phi(z|x)}{p(z)}dz + \int q_\phi(z|x)\log p_\theta(x|z)dz \\
&= -D_{\mathbb{KL}}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \\
&= \mathcal{L}(x,\theta,\phi),
\end{aligned} \tag{A.5}$$

where $D_{\mathbb{KL}}(p||q)$ is the $D_{\mathbb{KL}}$ between distributions p and q, and $\mathcal{L}(x,\theta,\phi)$ represents the ELBO. The ELBO is derived from Equation (A.4) and is given as:

$$\log p_\theta(x) \geq -D_{\mathbb{KL}}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] = \mathcal{L}(x,\theta,\phi), \tag{A.6}$$

where the ELBO is a lower bound for the marginal log-likelihood of the data; thus, maximizing the ELBO maximizes the log-likelihood, framing the optimization problem that must be solved.

Implementation

The ELBO derived from Equation (A.6) can be effectively implemented using a DNN-based architecture. However, challenges arise when computing its gradient with respect to ϕ, primarily due to the presence of ϕ within the expectation term (the second part of the ELBO in Equation (A.6)). To address this issue, the original research [10] introduced the reparameterization trick, a method that transforms the latent space sampling process to make it differentiable. This transformation enables the use of gradient-based optimization techniques. Instead of directly sampling from the latent space distribution, VAEs sample ϵ from a simpler distribution, typically a standard normal distribution. Subsequently, a deterministic transformation g_ϕ is applied to ϵ, resulting in $z = g_\phi(x,\epsilon)$, where $z \sim q_\phi(z|x)$ and $\epsilon \sim p(\epsilon)$. This approach allows the ELBO to be estimated as follows:

$$\hat{\mathcal{L}}(x,\theta,\phi) = \frac{1}{N}\sum_{i=1}^{N}\Big(-D_{\mathbb{KL}}(q_\phi(z|x_i)||p(z)) + \log p_\theta(x_i|g_\phi(x_i,\epsilon_i))\Big). \tag{A.7}$$

This reformulation simplifies the calculation of the ELBO gradient with respect to both θ and ϕ, enabling the application of standard optimization methods.

Equation (A.7) is implemented using DNNs, where functions are parameterized by ϕ and θ. Gradients are conveniently computed using the Backpropagation algorithm, which various programming libraries automate. The term VAE derives from the fact that Equation (A.7) resembles the architecture of an AE [286], as illustrated in Figure A.2. In this implementation,

the variational distribution q_ϕ is modeled using a DNN with parameters ϕ. This DNN takes an input sample x_i and outputs parameters for the deterministic transformation g_ϕ. The latent space of the VAE comprises the latent variable z distribution, which is a deterministic transformation g_ϕ of the output of the encoder combined with random ancillary noise ϵ. A sampled value z_i from the latent distribution is passed to a second DNN, parameterized by θ, which acts as the decoder. This decoder generates the parameters for the distribution $p_\theta(x|z)$, ultimately reconstructing the input data.

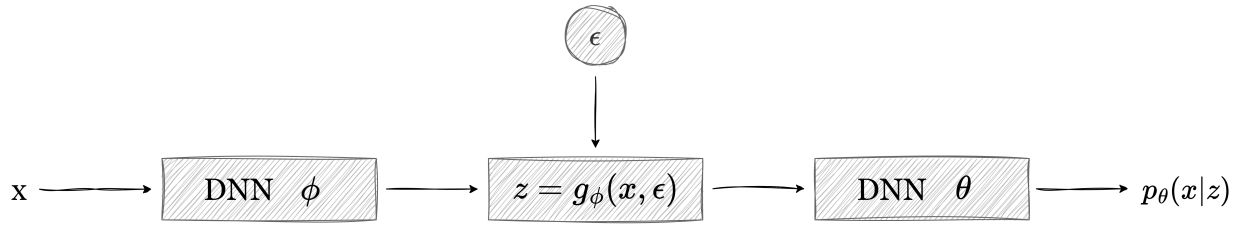


Figure A.2: VAE vanilla model implementation using DNNs.

Two key observations highlight the uniqueness of VAEs:

1. The ELBO loss in Equation (A.7) consists of a regularization term, which penalizes deviations from the prior in the latent space, and a reconstruction error term, which enforces similarity between generated samples and original inputs.
2. Unlike standard AEs, VAEs incorporate intermediate sampling, making them inherently non-deterministic. This property is particularly beneficial for applications requiring the estimation of output variable distributions, as it allows the derivation of input distribution parameters.

Appendix B

Data

This appendix provides a comprehensive overview of all datasets used in this thesis to evaluate and benchmark the proposed models. By centralizing this information, the goal is to enhance clarity and minimize redundancy in the main chapters. Each dataset is described alongside its corresponding experimental sections to facilitate cross-referencing.

B.1 Survival Analysis Datasets

SA datasets capture whether an event of interest occurred during a study period, categorizing patients as censored or uncensored while recording their follow-up times. Although these datasets are primarily designed for SA tasks, they have been repurposed for additional analysis in this thesis, highlighting their versatility.

The datasets vary widely in size, number of covariates, censoring proportions, and CR settings, ensuring thorough evaluation across diverse real-world scenarios. Preprocessing was conducted according to established methodologies to ensure fairness and comparability with state-of-the-art benchmarks.

The selection includes datasets from various domains, such as cardiovascular disease, oncology, hematological disorders, and infectious diseases, providing a comprehensive foundation for systematic model evaluation across different contexts.

B.1.1 Dataset Descriptions

Each dataset is described below, highlighting its characteristics, original purpose, and integration into this research. A detailed summary of their properties is provided in Table B.1.

- **WHAS [287]:** The Worcester Heart Attack Study (WHAS) focuses on patients with acute myocardial infarction, providing demographic and clinical data.
- **Support [288]:** The Study to Understand Prognoses Outcomes and Risks of Treatment (Support) captures data from seriously ill hospitalized adults, including demographics, comorbidities, and physiological measurements.

Dataset	# Samples	# Covariates	CR (#/no)	Event (or CR) Proportions (%)	Data Types	Used In
WHAS	1,638	7	No	57.78	Continuous, categorical	SAVAE
Support	9,104	16	No	31.89	Continuous, categorical	SAVAE
GBSG	2,232	9	No	43.23	Binary, continuous	SAVAE, SDG Methodology (medical data)
FLChain	6,524	10	No	69.92	Binary, continuous	SAVAE
NWTco	4,028	8	No	85.82	Binary, discrete	SAVAE, SDG Methodology (medical data)
Metabric	1,980	23	No	56.18	Binary, continuous	SAVAE, VAE-BGM
PBC	418	19	No	61.48	Continuous, categorical	SAVAE
STD	877	23	No	60.43	Categorical, integer, binary	SAVAE VAE-BGM
Pneumon	3,470	15	No	97.9	Binary, continuous	SAVAE
Melanoma	205	9	2	0.65/0.28/0.07	Continuous, categorical	CR-SAVAE
MGUS2	1,348	7	2	0.29/0.08/0.63	Continuous, categorical	CR-SAVAE
EBMT	8,966	7	6	0.63/0.13/0.09/ 0.02/0.02/0.01 0.1	Continuous, categorical	CR-SAVAE

Table B.1: Summary of SA datasets used in this thesis. For each dataset, the table provides key statistics, data types, and the corresponding experimental sections where the dataset was used.

- **GBSG [289], [290]:** The Rotterdam & German Breast Cancer Study Group (GBSG) combines data from node-positive breast cancer patients and chemotherapy trials with genomic and clinical information.
- **FLChain [291]:** Examines the relationship between serum immunoglobulin-free light chains (FLChains) and mortality in haematological disorders.
- **NWTco [292]:** The National Wilms Tumor Study (NWTco) focuses on Wilms tumors in children with binary and discrete covariates.
- **Metabric [293]:** The Molecular Taxonomy of Breast Cancer International Consortium (Metabric) dataset includes genomic and clinical data from breast cancer patients.
- **PBC [294]:** Focuses on Primary Biliary Cholangitis (PBC), a chronic liver disease

with survival outcomes and various clinical covariates.

- **STD** [42]**:** Provides information on Sexually Transmitted Diseases (STD) with categorical, integer, and binary data types.
- **Pneumon** [42]**:** Examines infant pneumonia with highly censored observations.
- **Melanoma** [295]**:** Contains data on patients with malignant melanoma, with two CR: death from melanoma and death from other causes.
- **MGUS2** [296]**:** Tracks patients with the Monoclonal Gammopathy of Unknown Significance (MGUS2), with two competing risks: progression to plasma cell malignancy or death from other causes.
- **EBMT** [297]**:** The European Society for Blood and Marrow Transplantation (EBMT) records six competing post-transplant complications or death risks.

B.2 Other-Task Datasets

This section describes datasets that are not specifically designed for SA. Instead, these datasets come from various tasks, including classification and regression, spanning a wide range of domains, feature types, and complexities. Their inclusion enables a comprehensive evaluation of models across diverse experimental scenarios.

The datasets address tasks such as income prediction, network intrusion detection, and heart disease classification, showcasing the broader applicability of the thesis contributions beyond SA. The selection incorporates both medical and non-medical datasets, ensuring a robust assessment of the versatility of the model.

B.2.1 Dataset Descriptions

Each dataset is described below, highlighting its characteristics, original purpose, and integration into this research. A detailed summary of their properties is provided in Table B.2.

- **Adult** [298]**:** The Adult Census Income dataset is extracted from the 1994 U.S. Census [299]. The dataset is used to predict whether the annual income of an individual exceeds $50,000.
- **News** [300]**:** The News Popularity Prediction dataset contains information about articles published on the Mashable news blog over two years. The objective is to predict the popularity of an article, measured by the number of social media shares.
- **King**[1]**:** The King County House Sales dataset is a regression dataset containing house sale prices for King County, Washington, including Seattle.
- **Intrusion** [301]**:** The KDD Cup 1999 Data was used for The Third Knowledge Discovery and Data Mining Competition to classify connections in a military network environment.

[1]Source: King Dataset (Accessed on December 8th, 2024)

- **Heart**[2]**:** The Heart dataset sourced from [302] originates from the cleaned Behavioral Risk Factor Surveillance System (BRFSS) 2015 survey and focuses on binary classification of heart disease presence. BRFSS is an annual health-related telephone survey designed to gather information on health conditions and risk factors.
- **Diabetes_H**[3]**:** This dataset comprises survey responses collected by the Centers for Disease Control and Prevention from the BBRFSS 2015. The target variable includes three classes: 0 for no diabetes or diabetes during pregnancy, 1 for prediabetes, and 2 for diabetes.

Dataset	# Samples	# Features	Data Types	Original Task	Used In
Adult	32,561	14	Categorical, binary, integer	Classification	VAE-BGM, Divergence Estimation, SDG Methodology (general-purpose data)
News	39,644	58	Categorical, continuous	Classification	SDG Methodology (general-purpose data)
King	21,613	20	Integer, continuous	Regression	SDG Methodology (general-purpose data)
Intrusion	494,021	39	Categorical, binary, integer	Classification	SDG Methodology (general-purpose data)
Heart	253,680	39	Categorical, binary	Classification	SDG Methodology (medical data)
Diabetes_H	253,680	22	Categorical, binary, integer	Classification	FedVAE

Table B.2: Summary of other-purpose datasets used in this thesis. For each dataset, the table provides key statistics, data types, the original task of the dataset, and the corresponding experimental sections where the dataset was used.

[2]Source: Heart Dataset (Accessed on December 8th, 2024)
[3]Source: Diabetes_H Dataset (Accessed on December 8th, 2024)

Appendix C

Survival Analysis

C.1 SAVAE

C.1.1 Sensitivity Analysis

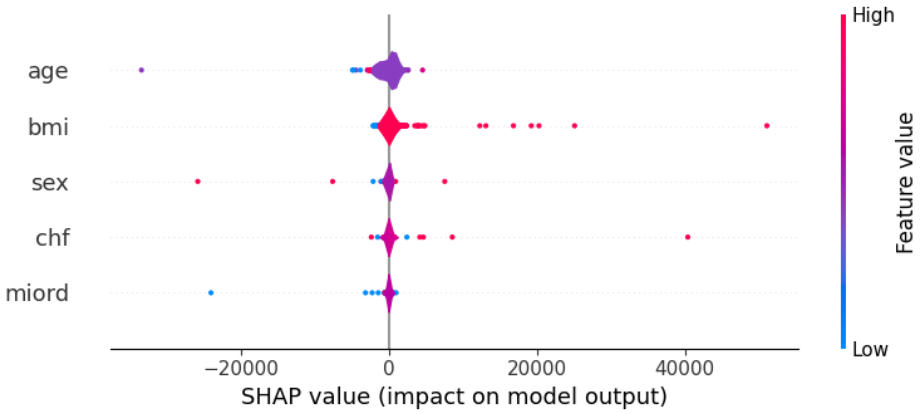
In AI models, particularly in complex DL architectures, sensitivity and robustness analyses are essential to understand how variations in input data influence the model predictions. These analyses play a crucial role in assessing the stability and reliability of the output of the model, providing insight into whether small changes in input features could significantly impact the results. This is especially important in SA tasks, where accurate and stable predictions are critical for applications in healthcare and other fields.

We have implemented two strategies to ensure the robustness of our proposed model, SAVAE. First, we trained the model on each dataset using ten different random seeds, followed by a 5-fold cross-validation. This approach allows us to evaluate the variability in model performance across different training-test splits and random initializations, minimizing the dependence on a single set of conditions. By averaging results across multiple runs, we provide more reliable and stable performance metrics that account for the inherent variability in the data and model training process.

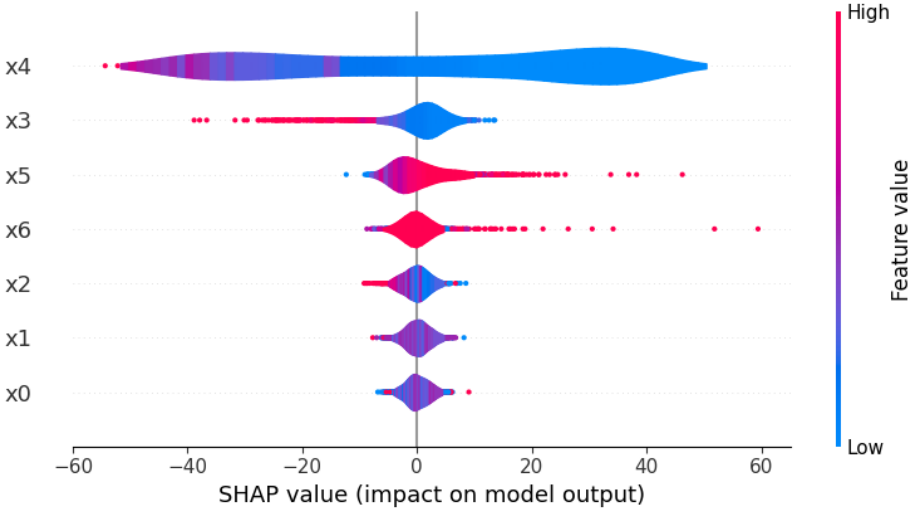
In addition to these measures, we conducted a Shapley value-based sensitivity analysis to explore how input features impact predicted survival times. Shapley values [303], originating from cooperative game theory, offer a mathematically grounded method to determine the contribution of each input feature to the final prediction. By computing Shapley values, we can assess the relative importance of each feature in influencing the output of the model. This analysis enhances the transparency and interpretability of the model, providing a clear understanding of how specific features drive survival time predictions.

These combined methods ensure a thorough sensitivity and robustness assessment, confirming that the model performs consistently across different conditions while offering insight into the role of input features in shaping predictions. These analyses strengthen the overall validity and interpretability of our model. The detailed results of the sensitivity analysis are presented below.

Specifically, our analysis focuses on quantifying the importance of each input feature. For this purpose, we used SHAP (SHapley Additive exPlanations) [304], an approach that enables global interpretability analyses for each dataset and the model. Using the Shap module from Python, we computed the importance of features following the official documentation, guaranteeing rigorous and consistent evaluation across datasets.



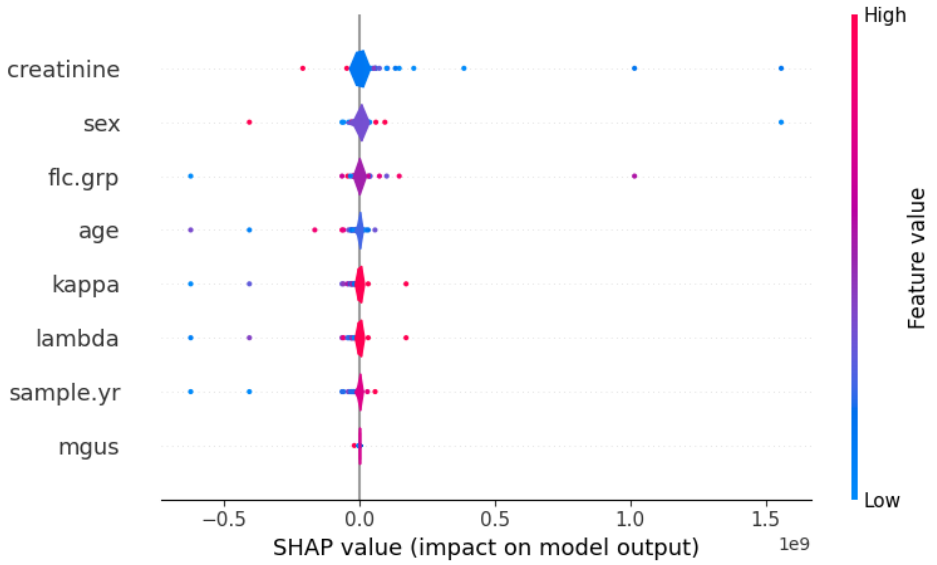
(a) WHAS dataset



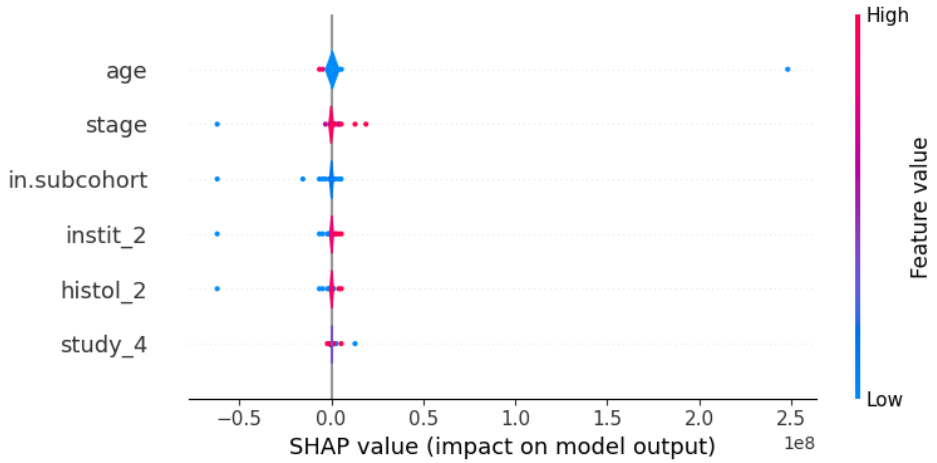
(b) GBSG dataset

Figure C.1: Feature Importance using SHAP in SAVAE datasets I (WHAS and GBSG.

Figure C.1 and Figure C.2 depict the feature importance for four selected datasets, chosen for their relatively small number of features to enhance visualization and analysis. To evaluate the robustness and generalizability of our model, we computed SHAP values using the training set of each dataset. The study presented corresponds to a single fold from one training seed. Recall that the model was trained for each dataset using ten different random seeds and employed a K-Fold cross-validation technique, where the data was divided into five subsets. Thus, the SHAP analysis shown here represents only a fraction of the complete robustness assessment.



(a) FLChain dataset



(b) NWTco dataset

Figure C.2: Feature Importance using SHAP in SAVAE datasets II (FLChain and NWTco.

Each violin plot in Figure C.1 and Figure C.2 ranks features from highest to lowest based on their contribution to the predicted survival times. The horizontal axis represents SHAP values, indicating the magnitude of the influence of each feature on the prediction. The individual dots within the plots correspond to specific observations, while the color gradient reflects whether the value of a feature is high or low relative to the dataset. This visual representation facilitates the interpretation of how each feature affects survival predictions in the training set. For example, in the GBSG dataset, lower values of the *'X4'* feature positively impact the survival time.

Furthermore, beyond providing an analytical perspective, this approach enables the evaluation of consistency between the interpretation of the model and established clinical knowledge. As

observed, the age feature emerges as one of the most influential factors in survival predictions across all datasets, which is an expected outcome given its well-documented significance in SA [305]–[307]. This alignment between model interpretation and clinical understanding reinforces the trustworthiness and applicability of the predictions of the model in real-world healthcare scenarios.

Interpretability is particularly critical in the medical domain, where understanding the correlation between clinical features and predicted survival times provides valuable insights for practitioners. For example, certain features, such as age, comorbidities, and treatment variables, may strongly influence survival predictions, as reflected in their SHAP values. These insights enable healthcare professionals to assess how specific features impact individual patient outcomes, leading to better-informed decisions.

Thus, this analysis serves as an additional validation step, demonstrating that the model is both accurate and interpretable while offering valuable insights consistent with clinical expectations. The ability to explain the predictions of the model further reinforces its robustness and enhances its applicability in high-impact domains such as healthcare.

C.1.2 Ablation Study

To evaluate the contribution of each component in the proposed SAVAE model and justify their inclusion, we conducted a comprehensive ablation study. This analysis aims to assess the impact of various architectural choices on model performance by systematically modifying key elements and measuring their effects on performance metrics, specifically the C-index and the IBS.

The SAVAE architecture consists of three DNNs: one encoder and two decoders, each responsible for inferring covariates and time parameters. The ablation study focused on the following key components:

- **Latent Space Dimensionality:** The default latent space dimensionality was set to 5. To examine its impact, we varied the dimensionality (e.g., reducing to 3 or increasing to 10) and analyzed its effect on the ability of the model to capture underlying data patterns.
- **Neurons:** Each module includes two hidden layers with 50 neurons. We examined whether reducing or increasing the number of neurons negatively influenced the performance of the model, particularly in capturing complex non-linear relationships between covariates and survival times.
- **Dropout Rate:** A dropout rate of 20% is applied in the first hidden layer of each decoder. To assess its role in preventing overfitting and enhancing generalization, we examined the impact of removing dropouts and experimented with an increased dropout rate (50%).

The Metabric dataset was selected for this study due to its high-dimensional feature space, particularly relevant for latent space generation in the VAE framework. Performance was measured using C-index and IBS values, obtained through 5-fold cross-validation for each

Latent Space Dimension	Number of Neurons	Dropout Rate	C-index	IBS
3	10	0	0.594	0.186
		0.2	0.590	0.185
		0.5	0.589	0.187
	50	0	0.598	0.184
		0.2	0.596	0.188
		0.5	0.590	0.187
	500	0	0.595	0.186
		0.2	0.593	0.185
		0.5	0.592	0.185
5	10	0	0.596	0.185
		0.2	0.593	0.186
		0.5	0.588	0.188
	50	0	0.604	0.185
		0.2	0.598	0.185
		0.5	0.594	0.185
	500	0	0.598	0.185
		0.2	0.595	0.185
		0.5	0.599	0.187
50	10	0	0.587	0.187
		0.2	0.582	0.189
		0.5	0.575	0.189
	50	0	0.604	0.187
		0.2	0.602	0.186
		0.5	0.589	0.190
	500	0	0.602	0.185
		0.2	0.606	0.184
		0.5	0.598	0.186

Table C.1: Ablation study results for Metabric with SAVAE. The impact of latent space dimensionality, number of neurons, and dropout rate on C-index and IBS.

Latent Space Dimension	Number of Neurons	Dropout Rate	C-index	IBS
3	10	0	0.575	0.213
		0.2	0.570	0.212
		0.5	0.571	0.213
	50	0	0.576	0.210
		0.2	0.573	0.210
		0.5	0.573	0.212
	500	0	0.573	0.213
		0.2	0.573	0.211
		0.5	0.568	0.211
5	10	0	0.574	0.210
		0.2	0.566	0.212
		0.5	0.584	0.211
	50	0	0.579	0.208
		0.2	0.585	0.209
		0.5	0.571	0.212
	500	0	0.585	0.209
		0.2	0.581	0.212
		0.5	0.576	0.209
50	10	0	0.566	0.213
		0.2	0.570	0.212
		0.5	0.584	0.212
	50	0	0.581	0.209
		0.2	0.577	0.211
		0.5	0.562	0.212
	500	0	0.584	0.208
		0.2	0.588	0.208
		0.5	0.574	0.208

Table C.2: Ablation study results for STD with SAVAE. The impact of latent space dimensionality, number of neurons, and dropout rate on C-index and IBS.

of the ten training seeds, ensuring robust and reliable results. The findings, summarized in Table C.1, indicate minimal variation in performance across different configurations. Given this stability, a latent space dimensionality of 5 was chosen as it exhibited the least variability and provided a balanced trade-off between model complexity and performance. Similarly, a moderate number of neurons (50) in the hidden layers was selected, resulting in low variability and competitive performance across both metrics. Notably, these final settings maintain a relatively low model complexity, leading to faster execution times without compromising accuracy, thereby enhancing the practical usability of the model model.

As observed in the Metabric results, no significant differences concerning the dropout rate were found. To investigate its impact further, we extended the analysis to the dataset with the fewest samples, STD, to determine whether dropout helps control overfitting in smaller datasets. Table C.2 presents the results using the same format as the Metabric table.

Overall, the findings for the STD dataset follow a similar pattern to those observed in Metabric, with performance remaining relatively consistent across different hyperparameter configurations. An intermediate latent space dimension (5) consistently yields higher C-index values and lower IBS values, indicating better model performance. A similar trend is observed for the number of neurons in the hidden layers, where intermediate values provide optimal results. While dropout does not significantly impact overall performance, slightly improved results are obtained when included.

Based on these findings, the component values specified in Section 3.2.3 were selected. However, it is essential to highlight the robustness of the model, as it maintains stable performance across various hyperparameter settings, demonstrating its resilience and adaptability. These decisions align with our goal of balancing model complexity and performance, ensuring that the final architecture remains computationally efficient while delivering robust and reliable predictions.

C.1.3 Computational Runtime Comparison

To comprehensively assess the computational complexity of our proposed model, SAVAE, we compared the execution times for training and validation across different models. This comparison encompasses the entire process, from training to validation on the test set, ensuring consistency in evaluating computational demands.

Two key factors motivated the decision to treat training and validation as a single process: (1) performance metrics for the test set are computed after each epoch to monitor model behavior over time, and (2) the computation of the C-index and IBS, the evaluation metrics used in this study, is independent of the model architecture. By considering the entire workflow together, we ensure that all models undergo an equivalent evaluation process.

To maintain a fair comparison, we trained all models without Early Stopping mechanisms, allowing each model to run for the full number of epochs. This approach ensured that all models were evaluated under identical training conditions. While the final results presented in Section 3.2.3 were obtained using Early Stopping with identical patience values, different models converged at varying points based on various optimization criteria (e.g., maximizing

a performance metric or minimizing a loss function). Running all models without Early Stopping ensures a standardized training environment for comparison purposes.

All models were trained for consistency using the same number of epochs (3,000) and batch size (64). To provide a clear comparison, we report the average execution time for each fold, using a 5-fold cross-validation setup across all models. Although execution times are expected to remain relatively stable across folds, averaging provides a more robust measure of computational demands.

For SAVAE, given the stochastic nature of VAEs, we ran ten seeds for each fold to ensure reliable results. Therefore, the reported average runtime for SAVAE corresponds to a single fold and a single seed. It is important to note that the total runtime for SAVAE can vary depending on the parallelization strategy and computational resources available. Nonetheless, this comparison method ensures that execution time differences are attributed solely to model complexity rather than external environmental factors.

Dataset	CoxPH	DeepSurv	DeepHit	SAVAE
WHAS	165.21	250.07	204.85	294.80
Support	18.26	24.52	48.58	123.44
GBSG	93.83	141.25	265.49	558.46
FLChain	305.18	442.95	933.32	1880.55
NWTco	236.73	325.04	638.76	922.92
Metabric	76.52	110.50	225.07	1058.42
PBC	28.14	38.58	118.53	198.13
STD	40.02	58.89	115.63	397.00
Pneumon	188.09	273.14	382.50	1860.52
Average	128.00	184.99	325.86	810.47

Table C.3: Average execution times (in seconds) for training and validating different models, including CoxPH, DeepSurv, DeepHit, and SAVAE, across various datasets. The total time reflects the duration of a single fold in a 5-fold cross-validation setting.

Table C.3 presents the average execution times in seconds for each model, providing insights into their computational efficiency. These times reflect the total duration required to train and validate a model on a single fold. As expected, the execution times for SAVAE are generally higher than those of other models due to its inherent architectural complexity. Including a VAE latent space and additional decoders increases computational demand, particularly when modeling complex non-linear relationships in survival data. However, while SAVAE requires longer runtimes, its enhanced flexibility and capacity to model intricate data patterns more effectively justify the increased computational cost. Importantly, execution times remain within reasonable limits, with all datasets requiring only a few minutes to complete a full

training and validation cycle. This balance ensures that the improved performance of SAVAE does not come at the expense of excessive computational costs, making it practical for real-world applications.

C.1.4 Statistical Significance and Multiple Testing Adjustment

To enhance the statistical rigor of our performance evaluation, we conducted multiple hypothesis tests to compare the mean C-index and IBS values of SAVAE against state-of-the-art models across various folds in a 5-fold cross-validation setup. Specifically, we tested the null hypothesis that the benchmark models outperform SAVAE regarding these performance metrics. Statistical significance was assessed using p-values, with a conventional significance threshold of 0.05. However, conducting multiple hypothesis tests increases the FWER, the probability of making one or more Type I errors (false positives) when the null hypotheses are true. As noted in the statistical literature [229]–[231], the FWER increases with the number of tests, inflating the likelihood of incorrectly rejecting true null hypotheses.

To address this issue and control the FWER, we applied the Holm-Bonferroni method [232], a step-down procedure that adjusts p-values to maintain the overall significance level. Compared to the traditional Bonferroni correction [308], the Holm method offers greater statistical power and does not require the assumption of test independence, making it well-suited for our analysis. This adjustment was implemented using the *statsmodels* package in Python, which provides robust tools for multiple testing corrections.

The adjusted p-values for both the C-index and IBS metrics are presented in Table C.4, respectively. This table compares the original and Holm-adjusted p-values for each dataset. By applying these adjustments, we account for multiple comparisons and prevent inflation of the FWER. Each subtable reports results as pairs of *original p-value - adjusted p-value*, where the Holm method has been applied across all model comparisons within each dataset. The comparison allows us to assess whether the adjusted p-values differ significantly from the original values and whether the adjustments substantially impact statistical significance. As expected, the adjusted p-values are higher due to the correction applied for multiple testing. However, the overall conclusions remain unchanged, confirming the robustness of our findings. As indicated by the bold entries in the tables, several instances still show p-values below the threshold of 0.05, reinforcing that SAVAE significantly outperforms state-of-the-art models in these cases.

These findings demonstrate that applying the Holm adjustment preserves the conclusions drawn from the original p-values. In particular, the results confirm that SAVAE outperforms existing models across multiple datasets, as evidenced by consistently low p-values. This further highlights the effectiveness of our model and its ability to generalize well across diverse datasets.

Model	WHAS	Support	GBSG	FLChain	NWTco	Metabric	PBC	STD	Pneumon
CoxPH	0.579 - 1.000	0.058 - 0.058	**0.000 - 0.000**	**0.000 - 0.000**	0.268 - 0.536	**0.003 - 0.006**	0.450 - 0.683	0.887 - 1.000	**0.003 - 0.009**
DeepSurv	1.000 - 1.000	**0.020 - 0.040**	0.149 - 0.149	**0.000 - 0.000**	0.135 - 0.406	0.549 - 0.549	0.280 - 0.683	0.927 - 1.000	0.382 - 0.765
DeepHit	1.000 - 1.000	**0.000 - 0.001**	**0.000 - 0.000**	**0.001 - 0.001**	0.644 - 0.645	**0.000 - 0.000**	0.228 - 0.683	0.727 - 1.000	0.935 - 0.935

(a) C-index

Model	WHAS	Support	GBSG	FLChain	NWTco	Metabric	PBC	STD	Pneumon
CoxPH	1.000 - 1.000	0.470 - 1.000	0.998 - 1.000	1.000 - 1.000	**0.000 - 0.000**	0.995 - 1.000	0.888 - 1.000	0.575 - 1.000	**0.000 - 0.000**
DeepSurv	**0.000 - 0.000**	0.341 - 1.000	0.561 - 1.000	1.000 - 1.000	**0.000 - 0.000**	1.000 - 1.000	0.868 - 1.000	0.746 - 1.000	**0.000 - 0.000**
DeepHit	**0.000 - 0.000**	0.950 - 1.000	1.000 - 1.000	1.000 - 1.000	**0.000 - 0.000**	1.000 - 1.000	1.000 - 1.000	0.995 - 1.000	**0.000 - 0.000**

(b) IBS

Table C.4: Comparison of original and Holm-adjusted p-values to assess whether the mean validation metrics values for SAVAE are better than those of state-of-the-art models across cross-validation folds. The results are presented as *original p-value - adjusted p-value*. **Bold** indicates a p-value below the threshold of 0.05, confirming that SAVAE performs significantly better than the other models.

Appendix D

Synthetic Data Generation

D.1 Divergence Estimation in Synthetic Data Validation

D.1.1 Experiment 2 Additional Analysis

Experiment 2 follows the same methodology and validation process as Experiment 1. This section presents additional analyses that further support the findings of Experiment 2.

Table D.1 extends the results from the previous experiment by evaluating the performance of the proposed divergence estimation approach with a limited number of training and validation samples. The results indicate a convergence trend towards the known ground truth divergence values as the number of training and validation samples increases.

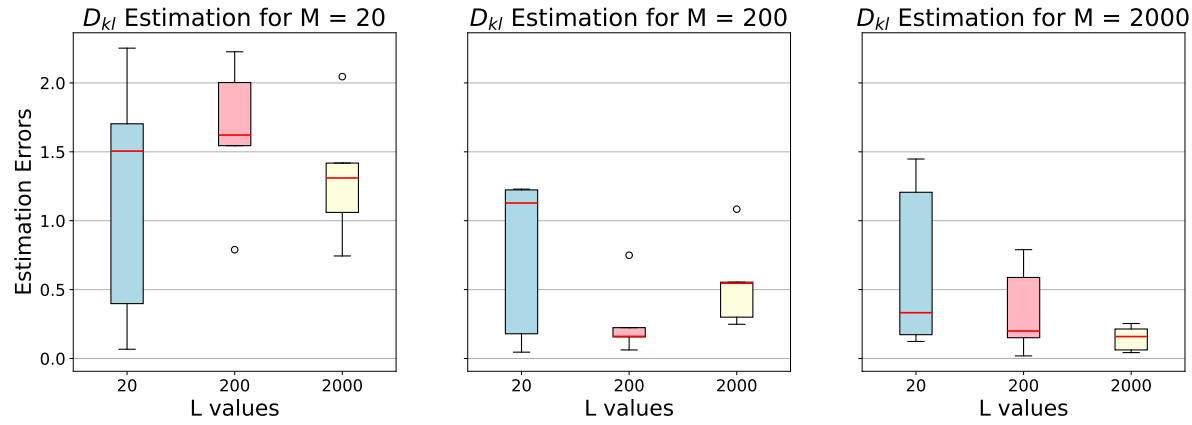
The estimation errors for $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ are presented in Figure D.1. These subfigures illustrate how errors decrease as training and validation samples increase, confirming the expected convergence trend.

Additionally, Figure D.2 examines the behavior of the loss function of the discriminator across different sample size configurations. This analysis provides insights into how the training process is affected by sample size constraints and the potential impact on the accuracy of divergence estimation.

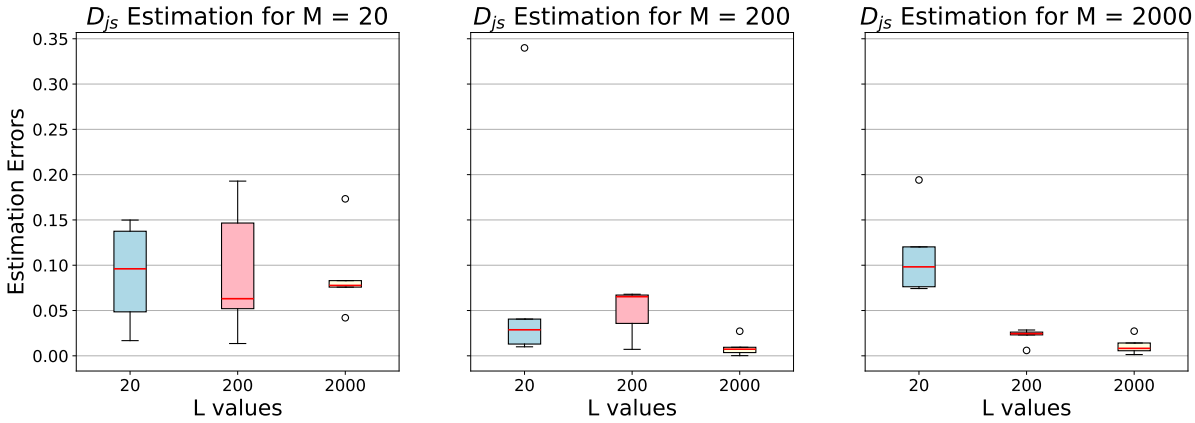
These additional analyses confirm the robustness of the proposed divergence estimation method. The results demonstrate a consistent trend of convergence toward ground truth values as the number of training and validation samples increases. Furthermore, the behavior of the discriminator underscores the importance of sample size in preventing overfitting and ensuring accurate divergence estimation.

M	L	MC Estimated $\boldsymbol{D}_{\mathbb{KL}}$	Discriminator Estimated $\boldsymbol{D}_{\mathbb{KL}}$	MC Estimated $\boldsymbol{D}_{\mathbb{JS}}$	Discriminator Estimated $\boldsymbol{D}_{\mathbb{JS}}$
20	20	3.245 ± 0.820	1.832 ± 0.912	0.490 ± 0.083	0.371 ± 0.078
	200	2.699 ± 0.095	1.193 ± 0.430	0.439 ± 0.017	0.328 ± 0.058
	2000	2.833 ± 0.032	1.512 ± 0.381	0.423 ± 0.006	0.331 ± 0.039
200	20	2.639 ± 0.404	3.101 ± 0.779	0.405 ± 0.066	0.340 ± 0.114
	200	2.669 ± 0.118	2.686 ± 0.294	0.392 ± 0.013	0.374 ± 0.020
	2000	2.887 ± 0.045	3.037 ± 0.517	2.887 ± 0.045	3.037 ± 0.517
2000	20	2.974 ± 0.626	2.771 ± 0.751	0.481 ± 0.090	0.474 ± 0.096
	200	2.710 ± 0.100	2.628 ± 0.357	0.415 ± 0.020	0.412 ± 0.019
	2000	2.850 ± 0.071	2.913 ± 0.129	0.423 ± 0.014	0.420 ± 0.013

Table D.1: Impact of training and validation samples on $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ estimation for Experiment 2. Analytical $D_{\mathbb{KL}}$ along with MC $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ estimations, as well as proposed discriminator estimations for both divergences. Results are displayed for various combinations of training samples M and validation samples L, showing a clear correlation between the number of samples and the estimation error.



(a) $D_{\mathbb{KL}}$



(b) $D_{\mathbb{JS}}$

Figure D.1: Estimation error representation for $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ in Experiment 2. Results are shown for different combinations of training sample sizes M and validation sample sizes L. As expected, a decrease and precision in the error is observed with increasing values of M and L.

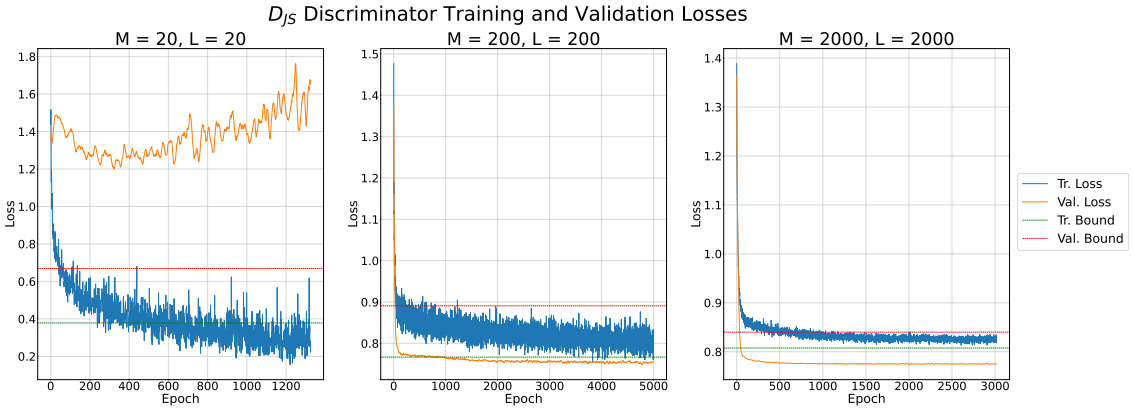


Figure D.2: Discriminator loss curves for Experiment 2. The loss curves show a clear overfitting due to low sample sizes. Green and red dashed lines represent theoretical convergence values.

D.2 Synthetic Data Generation in Scarce-Data Settings

D.2.1 Impact of Sample Size on Divergence Metrics

To further analyze the impact of sample size on divergence metric performance, we conducted an additional validation experiment using a reduced number of samples for both M and L. Specifically, we set $M = 100$ and $L = 100$, representing a scenario with severely limited data availability. As highlighted several times in this thesis, sufficient samples are crucial for accurate distribution comparisons. Therefore, this low-sample setting was expected to lead to less reliable divergence estimations.

The results of this validation experiment are presented in Table D.2 and Table D.3. Under these constrained conditions, the improvements observed in previous experiments do not hold consistently across all datasets. Furthermore, the obtained divergences and MMD values are significantly lower than expected, suggesting that the underlying distributions are not adequately captured when the number of samples is insufficient. This, in turn, leads to an underestimation of the actual divergence between distributions.

These findings highlight the critical role of sample size in evaluating SDG methods. When data is scarce, metrics may not accurately reflect distributional differences, potentially leading to misleading conclusions about the quality of synthetic data. Therefore, it is essential to account for sample size when interpreting divergence and MMD results to ensure meaningful and reliable evaluations.

Scenario	N	M	L	VAE $D_{\mathbb{JS}}$	CTGAN $D_{\mathbb{JS}}$	VAE $D_{\mathbb{KL}}$	CTGAN $D_{\mathbb{KL}}$	VAE MMD	CTGAN MMD
Big data	10000	7500	1000	0.079 (0.001)	0.150 (0.002)	0.153 (0.019)	0.420 (0.025)	0.0007 (0.0002)	0.0047 (0.0003)
Low data	300	7500	1000	0.331 (0.004)	0.563 (0.002)	0.697 (0.018)	1.653 (0.015)	0.0032 (0.0004)	0.0148 (0.0007)
Low data	300	100	100	0.030 (0.012)	0.333 (0.027)	0.084 (0.101)	2.197 (0.257)	0.0051 (0.0020)	0.0157 (0.0046)
Pre-train	300	7500	1000	**0.171 (0.004)**	0.563 (0.002)	**0.427 (0.021)**	1.753 (0.040)	**0.0013 (0.0003)**	0.0174 (0.0007)
Pre-train	300	100	100	**-0.002 (0.004)**	**0.237 (0.014)**	0.056 (0.103)	**0.978 (0.307)**	**0.0020 (0.0010)**	0.0137 (0.0024)
AVG	300	7500	1000	**0.157 (0.004)**	N/A	**0.380 (0.043)**	N/A	**0.0019 (0.0002)**	N/A
AVG	300	100	100	**0.002 (0.002)**	N/A	0.049 (0.107)	N/A	0.0041 (0.0028)	N/A
MAML	300	7500	1000	**0.300 (0.002)**	N/A	0.686 (0.037)	N/A	**0.0007 (0.0001)**	N/A
MAML	300	100	100	**0.002 (0.008)**	N/A	0.054 (0.116)	N/A	0.0029 (0.0012)	N/A
DRS	300	7500	1000	**0.189 (0.006)**	N/A	**0.427 (0.043)**	N/A	0.0036 (0.0003)	N/A
DRS	300	100	100	**0.015 (0.008)**	N/A	0.089 (0.051)	N/A	0.0040 (0.0013)	N/A

(a) Adult dataset

Scenario	N	M	L	VAE $D_{\mathbb{JS}}$	CTGAN $D_{\mathbb{JS}}$	VAE $D_{\mathbb{KL}}$	CTGAN $D_{\mathbb{KL}}$	VAE MMD	CTGAN MMD
Big data	10000	7500	1000	0.253 (0.009)	0.463 (0.003)	0.647 (0.045)	1.506 (0.031)	0.0006 (0.0001)	0.0014 (0.0002)
Low data	300	7500	1000	0.840 (0.003)	0.962 (0.002)	4.582 (0.136)	8.994 (0.909)	0.0024 (0.0002)	0.0122 (0.0001)
Low data	300	100	100	0.003 (0.004)	0.482 (0.093)	-0.062 (0.113)	2.290 (1.100)	0.0054 (0.0017)	0.0132 (0.0017)
Pre-train	300	7500	1000	**0.746 (0.003)**	**0.937 (0.003)**	**3.516 (0.082)**	8.603 (0.463)	0.0024 (0.0002)	0.0137 (0.0003)
Pre-train	300	100	100	0.010 (0.006)	0.843 (0.013)	0.061 (0.051)	4.599 (0.629)	0.0048 (0.0011)	0.0149 (0.0017)
AVG	300	7500	1000	**0.609 (0.003)**	N/A	**2.596 (0.060)**	N/A	0.0026 (0.0001)	N/A
AVG	300	100	100	-0.001 (0.004)	N/A	0.073 (0.088)	N/A	0.0047 (0.0014)	N/A
MAML	300	7500	1000	0.851 (0.001)	N/A	5.176 (0.242)	N/A	0.0028 (0.0002)	N/A
MAML	300	100	100	0.179 (0.071)	N/A	2.256 (0.927)	N/A	0.0047 (0.0013)	N/A
DRS	300	7500	1000	**0.645 (0.006)**	N/A	**2.449 (0.057)**	N/A	0.0026 (0.0002)	N/A
DRS	300	100	100	0.002 (0.004)	N/A	-0.010 (0.114)	N/A	0.0046 (0.0015)	N/A

(b) News dataset

Table D.2: Resemblance metrics results across scenarios III. 'Big data' represents the ideal case where many samples ($N = 10,000$) are available to generate reliable synthetic data. 'Low data' represents a more realistic scenario in which a limited number of samples ($N = 300$) are available, posing a challenge for SDG. Two low-data configurations are tested: a more reliable case ($M = 7,500$, $L = 1,000$) and a highly constrained case ($M = 100$, $L = 100$). The next rows compare the divergences obtained by each methodology ('Pre-train', 'AVG', 'MAML', and 'DRS') applied to the 'Low data' scenarios. **Bold** indicates improvements. Results are presented as *mean (standard deviation)*, with lower values being preferable.

Scenario	N	M	L	VAE $D_{\mathbb{JS}}$	CTGAN $D_{\mathbb{JS}}$	VAE $D_{\mathbb{KL}}$	CTGAN $D_{\mathbb{KL}}$	VAE MMD	CTGAN MMD
Big data	10000	7500	1000	0.862 (0.002)	0.777 (0.003)	4.768 (0.072)	3.124 (0.115)	0.0225 (0.0016)	0.0006 (0.0001)
Low data	300	7500	1000	0.927 (0.002)	0.940 (0.003)	13.763 (0.696)	7.470 (0.392)	0.0029 (0.0003)	0.0109 (0.0009)
Low data	300	100	100	0.533 (0.062)	0.682 (0.029)	3.264 (0.555)	3.731 (0.279)	0.0021 (0.0011)	0.0110 (0.0046)
Pre-train	300	7500	1000	0.862 (0.002)	0.945 (0.002)	**5.286 (0.327)**	9.533 (0.453)	**0.0020 (0.0003)**	0.0137 (0.0010)
Pre-train	300	100	100	**0.228 (0.018)**	0.622 (0.070)	**0.698 (0.197)**	3.455 (0.426)	0.0028 (0.0025)	0.0127 (0.0050)
AVG	300	7500	1000	**0.740 (0.002)**	N/A	**3.489 (0.209)**	N/A	**0.0010 (0.0003)**	N/A
AVG	300	100	100	**-0.001 (0.002)**	N/A	**0.020 (0.108)**	N/A	0.0015 (0.0010)	N/A
MAML	300	7500	1000	**0.910 (0.002)**	N/A	**6.436 (0.496)**	N/A	0.0057 (0.0006)	N/A
MAML	300	100	100	**0.322 (0.046)**	N/A	**1.030 (0.127)**	N/A	0.0089 (0.0048)	N/A
DRS	300	7500	1000	**0.809 (0.003)**	N/A	**4.321 (0.215)**	N/A	0.0028 (0.0003)	N/A
DRS	300	100	100	**0.068 (0.006)**	N/A	**0.447 (0.093)**	N/A	0.0071 (0.0026)	N/A

(a) King dataset

Scenario	N	M	L	VAE $D_{\mathbb{JS}}$	CTGAN $D_{\mathbb{JS}}$	VAE $D_{\mathbb{KL}}$	CTGAN $D_{\mathbb{KL}}$	VAE MMD	CTGAN MMD
Big data	10000	7500	1000	0.760 (0.013)	0.531 (0.033)	2.744 (0.084)	2.623 (0.537)	0.0001 (0.0000)	0.0057 (0.0006)
Low data	300	7500	1000	0.920 (0.003)	0.961 (0.002)	6.216 (0.154)	8.841 (0.710)	0.0644 (0.0011)	0.0793 (0.0020)
Low data	300	100	100	0.050 (0.009)	0.681 (0.052)	0.182 (0.150)	4.365 (0.175)	0.0121 (0.0052)	0.0088 (0.0016)
Pre-train	300	7500	1000	**0.793 (0.004)**	0.959 (0.001)	**3.831 (0.151)**	8.443 (0.630)	**0.0600 (0.0017)**	**0.0682 (0.0022)**
Pre-train	300	100	100	0.067 (0.004)	0.604 (0.039)	0.286 (0.091)	**3.818 (0.189)**	0.0185 (0.0038)	**0.0023 (0.0010)**
AVG	300	7500	1000	**0.867 (0.007)**	N/A	5.798 (0.295)	N/A	**0.0617 (0.0013)**	N/A
AVG	300	100	100	0.055 (0.003)	N/A	0.167 (0.132)	N/A	0.0117 (0.0024)	N/A
MAML	300	7500	1000	**0.913 (0.003)**	N/A	6.359 (0.054)	N/A	0.0639 (0.0017)	N/A
MAML	300	100	100	0.049 (0.002)	N/A	0.261 (0.093)	N/A	0.0118 (0.0037)	N/A
DRS	300	7500	1000	**0.835 (0.009)**	N/A	**4.587 (0.166)**	N/A	0.0612 (0.0026)	N/A
DRS	300	100	100	0.066 (0.005)	N/A	0.310 (0.173)	N/A	0.0128 (0.0022)	N/A

(b) Intrusion dataset

Table D.3: Resemblance metrics results across scenarios IV. 'Big data' represents the ideal case where many samples ($N = 10,000$) are available to generate reliable synthetic data. 'Low data' represents a more realistic scenario in which a limited number of samples ($N = 300$) are available, posing a challenge for SDG. Two low-data configurations are tested: a more reliable case ($M = 7,500$, $L = 1,000$) and a highly constrained case ($M = 100$, $L = 100$). The next rows compare the divergences obtained by each methodology ('Pre-train', 'AVG', 'MAML', and 'DRS') applied to the 'Low data' scenarios. **Bold** indicates improvements. Results are presented as *mean (standard deviation)*, with lower values being preferable.

D.3 Synthetic Data Generation in Medical Scarce-Data Settings

D.3.1 Additional Results with Other Medical Datasets

This section presents supplementary experiments conducted to further validate our findings using additional medical datasets. The results follow the same format as the main experiments in Section 4.4.3 to maintain consistency and facilitate comparison.

To test the methodology under varying data availability conditions, we first used two large classification datasets related to diabetes: Diabetes_H and Diabetes_130. These datasets provided sufficient samples to assess model performance under both optimal and data-scarce conditions.

Subsequently, we extended our research to SA datasets with significantly fewer samples, reflecting real-world data constraints better. Unlike previous cancer-related datasets, these new publicly available datasets focus on other medical conditions: WHAS (heart attack cases), PBC (autoimmune liver disease), and STD (sexually transmitted diseases).

Table D.4 summarizes the additional datasets used in these experiments, highlighting their key characteristics. By incorporating these datasets, we comprehensively evaluate the SDG methodology, ensuring its robustness and generalizability across different medical domains and data availability conditions.

Dataset	Number of samples	Number of features	Data types	Task
Diabetes_H	253,680	22	Binary and discrete	Classification
Diabetes_130	101,766	45	Binary and discrete	Classification
WHAS	1,638	7	Binary, continuous and discrete	Survival Analysis
PBC	418	19	Binary, continuous and discrete	Survival Analysis
STD	877	23	Binary and discrete	Survival Analysis

Table D.4: Additional medical datasets for scarce-data SDG. Overview of supplementary datasets used to assess the methodology under different data availability conditions. Classification datasets (Diabetes_H and Diabetes_130) contain a large number of samples, while SA datasets (WHAS, PBC, and STD) have fewer samples, reflecting real-world constraints. The table provides details on the number of samples, features, data types, and associated tasks.

Classification datasets

The results for the Diabetes_H dataset are presented in the subtable (a) in Table D.5. In the 'Big data' scenario ($N = 10,000$), $D_{\mathbb{JS}}$ reaches an upper bound of 0.278 ± 0.094, while in the 'Low Data' scenario ($N = 100$), the divergence increases significantly to 0.860 ± 0.001, indicating substantial room for improvement. Consistent with the findings from the Heart dataset, 'AVG' and 'DRS' techniques achieve the lowest $D_{\mathbb{JS}}$, demonstrating their effectiveness. Similarly, $D_{\mathbb{KL}}$ improves across all techniques, with 'DRS' yielding the most significant decrease.

Regarding clinical utility validation, the classification accuracy obtained with limited real data (0.500 ± 0.114) closely aligns with that of the 'Big data' scenario (0.606 ± 0.016). This trend, which is observed consistently across all cases, suggests that SDG does not degrade classification performance. Instead, maintaining accuracy might depend on correctly generating a few critical variables, even if some variables are not as accurately represented.

Scenario	SIMILARITY VALIDATION			CLINICAL UTILITY VALIDATION		
	$D_{\mathbb{JS}}$	$D_{\mathbb{KL}}$	MMD	Real Acc	Synth Acc	Synth Fine-Tuned Acc
Big data	0.258 (0.122)	0.724 (0.267)	0.0002 (0.0000)	0.605 (0.018)	0.598 (0.023)	0.621 (0.016)
Low data	0.841 (0.005)	7.094 (0.936)	0.0068 (0.0002)	0.529 (0.143)	0.634 (0.113)	0.607 (0.067)
Pre-train	**0.723 (0.015)**	**3.765 (0.075)**	**0.0054 (0.0002)**	N/A	0.540 (0.274)	0.602 (0.049)
AVG	**0.697 (0.013)**	**3.833 (0.192)**	**0.0030 (0.0001)**	N/A	0.548 (0.046)	0.571 (0.050)
DRS	**0.709 (0.010)**	**3.611 (0.148)**	**0.0051 (0.0002)**	N/A	0.528 (0.060)	0.532 (0.055)

(a) Diabetes_H dataset

Scenario	SIMILARITY VALIDATION			CLINICAL UTILITY VALIDATION		
	$D_{\mathbb{JS}}$	$D_{\mathbb{KL}}$	MMD	Real Acc	Synth Acc	Synth Fine-Tuned Acc
Big data	0.469 (0.054)	0.968 (0.181)	0.0003 (0.0000)	0.452 (0.019)	0.452 (0.014)	0.455 (0.019)
Low data	0.979 (0.002)	16.217 (2.972)	0.0086 (0.0004)	0.298 (0.160)	0.252 (0.160)	0.256 (0.156)
Pre-train	**0.953 (0.002)**	21.535 (1.580)	**0.0064 (0.0002)**	N/A	0.386 (0.051)	0.381 (0.051)
AVG	**0.943 (0.002)**	15.335 (2.071)	**0.0061 (0.0002)**	N/A	0.279 (0.153)	0.292 (0.155)
DRS	**0.944 (0.003)**	19.544 (1.908)	**0.0048 (0.0002)**	N/A	0.444 (0.103)	0.438 (0.098)

(b) Diabetes_130 dataset

Table D.5: Validation results for the Diabetes datasets across different scenarios. The 'Big data' scenario ($N = 10,000$) represents an ideal condition with abundant samples, enabling reliable synthetic data. The 'Low data' scenario ($N = 100$) reflects a more realistic constraint setting. The similarity validation section reports $D_{\mathbb{JS}}$, $D_{\mathbb{KL}}$, and MMD values for different techniques ('Pre-train', 'AVG', and 'DRS') applied to the 'Low data' scenario, where lower values indicate better similarity. **Bold** indicates improvements. The clinical utility validation section compares classification accuracy between models trained on real, synthetic, and fine-tuned synthetic data, where higher values indicate better performance. **Bold** denotes that the adjusted p-value is below the significance threshold of 0.01. All results are expressed as *mean (standard deviation)*.

The MMD results for the Diabetes_H dataset align with the divergence metrics, where higher MMD values correspond to higher divergences. In the 'Big data' scenario, MMD is 0.0002 ± 0.0000, reflecting a strong alignment between real and synthetic data distributions. However, in the 'Low data' scenario, MMD increases significantly to 0.0068 ± 0.0002, consistent with the sharp rise observed in $D_{\mathbb{JS}}$ and $D_{\mathbb{KL}}$. The applied techniques reduce MMD, with 'AVG' achieving the lowest value (0.0030 ± 0.0001), reinforcing its capacity to enhance synthetic data quality under low-data constraints.

The Diabetes_130 dataset results, shown in subtable (b) in Table D.5, reveal higher divergences across all scenarios. This is likely due to the increased complexity and higher dimensionality (45 variables) of the dataset. Despite this, the $D_{\mathbb{JS}}$ divergence improves with the applied techniques, confirming their effectiveness even in challenging cases. Conversely, no improvement is observed in $D_{\mathbb{KL}}$, highlighting the inherent difficulty of generating high-quality synthetic data when dataset dimensionality increases. Capturing complex inter-variable dependencies becomes increasingly challenging. The MMD values follow a pattern similar to the divergence metrics. The 'Big data' scenario achieves the lowest MMD (0.0003 ± 0.0000), while the 'Low data' scenario presents a much higher MMD (0.0086 ± 0.0004), mirroring the increased $D_{\mathbb{JS}}$ and $D_{\mathbb{KL}}$ values. Techniques such as DRS achieve the most substantial MMD reduction (0.0048 ± 0.0002), reinforcing its role in improving synthetic data resemblance. Clinical utility validation results align with those from the Diabetes_H, indicating no significant difference between classifiers trained on real or synthetic data. Accuracy metrics remain consistent across all scenarios, reinforcing the idea that $D_{\mathbb{JS}}$ is a reliable metric for similarity validation, while clinical utility validation alone does not suffice to assess synthetic data quality.

Survival Analysis data

The additional SA datasets used in this study preserve the heterogeneity observed in the previously analyzed datasets in Section 4.4.3. Specifically, the WHAS, PBC, and STD datasets were sourced from the SAVAE experiments. This dataset selection aligns with the common challenges in SA research, which typically involve limited sample sizes and complex variable interactions.

Table D.6 presents the C-index and IBS results across different settings (Real, Synthetic, and Synthetic Fine-Tuned), following the same structure as previous experiments. The findings indicate no significant difference in performance metrics when applying the SDG methodologies compared to baseline models. Moreover, performance remains consistent regardless of whether a larger or smaller sample size is used, reinforcing the notion that clinical utility validation alone is insufficient to assess the effectiveness of SDG in scarce-data settings.

To further evaluate the fidelity of the generated synthetic data, we provide KM estimations with CIs in Figure D.3. These plots compare survival functions for real and synthetic data under varying sample sizes. Notably, while the survival curves appear more similar in the PBC and STD datasets, this is likely due to their inherently small sample sizes. However, in all cases, the synthetic survival functions generated using our methodology closely approximate the upper-bound survival functions obtained from larger sample sizes, whereas the lower-bound curves (green) deviate more significantly. This supports the conclusion that our SDG methodology enhances the accuracy of time-to-event estimations in SA, even under data-scarce conditions.

Scenario	Real CI	Synth CI	Synth Fine-Tuned CI	Real IBS	Synth IBS	Synth Fine-Tuned IBS
Big data	0.731 (0.035)	0.715 (0.037)	0.714 (0.036)	0.178 (0.030)	0.179 (0.030)	0.185 (0.030)
Low data	0.703 (0.036)	0.698 (0.036)	0.696 (0.038)	0.177 (0.030)	0.180 (0.030)	0.182 (0.030)
Pre-train	N/A	0.689 (0.037)	0.675 (0.037)	N/A	0.183 (0.030)	0.185 (0.030)
AVG	N/A	0.725 (0.036)	0.720 (0.037)	N/A	0.175 (0.030)	0.174 (0.030)
DRS	N/A	0.710 (0.037)	0.708 (0.036)	N/A	0.178 (0.030)	0.179 (0.030)

(a) WHAS dataset

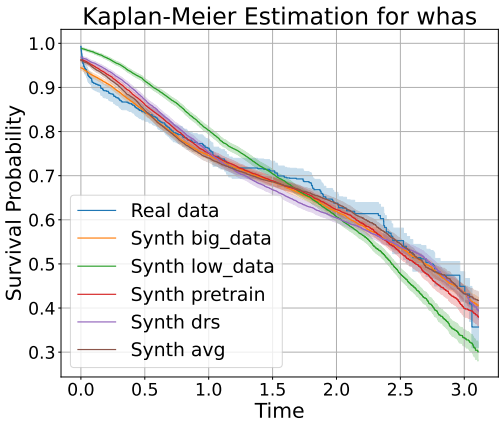
Scenario	Real CI	Synth CI	Synth Fine-Tuned CI	Real IBS	Synth IBS	Synth Fine-Tuned IBS
Big data	0.815 (0.069)	0.815 (0.065)	0.826 (0.063)	0.174 (0.062)	0.156 (0.058)	0.159 (0.058)
Low data	0.541 (0.063)	0.521 (0.055)	0.530 (0.054)	0.242 (0.051)	0.246 (0.046)	0.249 (0.046)
Pre-train	N/A	0.829 (0.063)	0.834 (0.059)	N/A	0.153 (0.060)	0.136 (0.054)
AVG	N/A	0.817 (0.062)	0.836 (0.059)	N/A	0.153 (0.058)	0.147 (0.056)
DRS	N/A	0.826 (0.063)	0.840 (0.061)	N/A	0.163 (0.058)	0.157 (0.060)

(b) PBC dataset

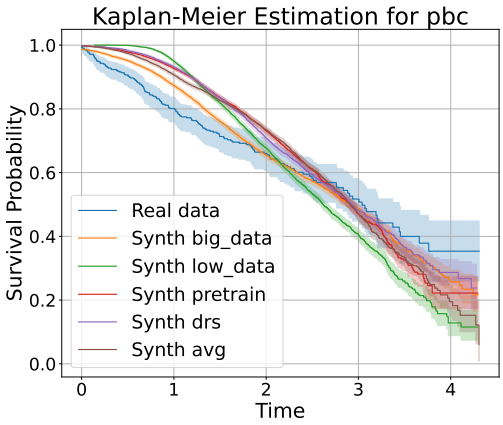
Scenario	Real CI	Synth CI	Synth Fine-Tuned CI	Real IBS	Synth IBS	Synth Fine-Tuned IBS
Big data	0.544 (0.055)	0.577 (0.057)	0.608 (0.052)	0.224 (0.046)	0.215 (0.044)	0.214 (0.044)
Low data	0.733 (0.079)	0.789 (0.072)	0.804 (0.066)	0.176 (0.061)	0.159 (0.060)	0.191 (0.069)
Pre-train	N/A	0.540 (0.057)	0.536 (0.058)	N/A	0.229 (0.045)	0.234 (0.046)
AVG	N/A	0.553 (0.060)	0.553 (0.057)	N/A	0.225 (0.045)	0.230 (0.045)
DRS	N/A	0.530 (0.057)	0.523 (0.054)	N/A	0.227 (0.045)	0.232 (0.045)

(c) STD dataset

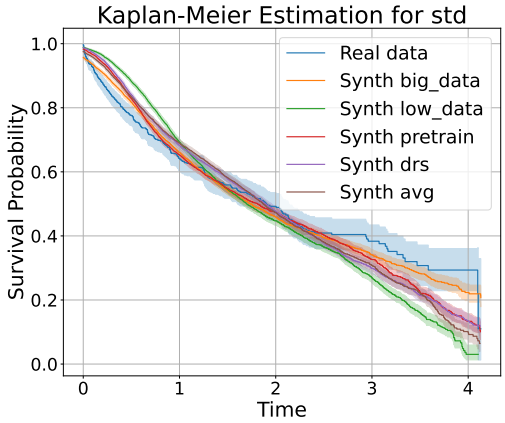
Table D.6: Validation results for the additional SA datasets across different scenarios. The 'Big data' scenario represents an ideal condition with a larger sample size ($N = 1,311$, $N = 335$, and $N = 702$, 80% of the data for WHAS, PBC and STD, respectively), enabling reliable SDG. The 'Low data' scenario reflects a more realistic constraint with a smaller sample size ($N = 100$), posing challenges for SDG. The table presents SA metrics (C-index and IBS) comparing models trained on real data, synthetic data, and synthetic data fine-tuned on real data. Higher C-index values indicate better predictive performance, while lower IBS values are preferable. **Bold** highlights significant improvements using the methodology, while * indicates a significant disadvantage. Results are reported as *mean (standard deviation)*.



(a) WHAS dataset



(b) PBC dataset



(c) STD dataset

Figure D.3: KM estimations with CIs using real and synthetic data in additional datasets. The survival functions for the upper bounds (blue and orange) illustrate the survival probabilities of real data and synthetic data generated from a large number of samples. The survival functions of synthetic data generated using the proposed methodology (red, purple, and brown) show convergence toward these upper bounds. In contrast, the lower-bound survival function (green) deviates significantly, highlighting the challenges of generating reliable survival estimates with limited data.

Different data utility results

Finally, Table D.7 and Table D.8 additional clinical utility validation results for the supplementary classification and SA. In this validation, we modified the final predictive task by selecting different target variables for classification datasets and using alternative variables (not time) for prediction in SA datasets.

Diabetes_H			Diabetes_130		
Feature	**Low data**	**DRS**	**Feature**	**Low data**	**DRS**
HighBP	0.710 (0.011)	0.708 (0.006)	**race**	0.179 (0.206)	0.343 (0.279)
HighChol	0.646 (0.005)	0.598 (0.011)*	**gender**	0.444 (0.132)	0.476 (0.017)
CholCheck	0.697 (0.209)	0.445 (0.070)	**num_procedures**	0.258 (0.049)	0.153 (0.066)
Smoker	0.530 (0.035)	0.540 (0.015)	**metformin**	0.056 (0.082)	0.064 (0.069)
Stroke	0.514 (0.063)	**0.761 (0.032)**	**repaglinide**	0.632 (0.379)	0.016 (0.019)
HeartDiseaseorAttack	0.635 (0.026)	0.638 (0.030)	**nateglinide**	0.552 (0.060)	0.417 (0.180)
PhysActivity	0.679 (0.010)	0.625 (0.049)	**chlorpropamide**	0.392 (0.234)	0.335 (0.107)
Fruits	0.384 (0.008)	0.388 (0.014)	**glimepiride**	0.166 (0.194)	**0.692 (0.197)**
Veggies	0.706 (0.090)	0.635 (0.030)	**glipizide**	0.216 (0.221)	0.157 (0.242)
HvyAlcoholConsump	0.757 (0.166)	0.534 (0.194)	**glyburide**	0.342 (0.237)	0.107 (0.063)
AnyHealthcare	0.471 (0.093)	**0.668 (0.022)**	**tolbutamide**	0.787 (0.112)	0.867 (0.025)
NoDocbcCost	0.741 (0.210)	0.630 (0.062)	**pioglitazone**	0.446 (0.393)	0.340 (0.304)
DiffWalk	0.783 (0.015)	0.785 (0.004)	**rosiglitazone**	0.218 (0.144)	0.505 (0.391)
Sex	0.531 (0.045)	0.529 (0.043)	**acarbose**	0.444 (0.263)	0.199 (0.335)
Diabetes	0.545 (0.029)	0.494 (0.087)	**miglitol**	0.123 (0.212)	0.240 (0.365)
			tolazamide	0.938 (0.027)	0.941 (0.014)
			insulin	0.284 (0.023)	**0.345 (0.020)**
			glyburide-metformin	0.235 (0.250)	0.480 (0.321)
			change	0.575 (0.007)	**0.615 (0.007)**
			diabetesMed	0.560 (0.045)	0.555 (0.047)
			readmitted	0.379 (0.149)	0.308 (0.165)

Table D.7: Different clinical utility validation results for the additional classification datasets. Accuracy comparison between the 'Low data' ($N = 100$) scenario and the 'DRS' technique applied to the lower bound case for each feature used as a classification label. Higher values indicate better performance. **Bold** indicates a significant improvement with the methodology, while * indicates a significant decline. Results are reported as *mean (standard deviation)*.

Whas			Pbc			Std		
Feature	**Low data**	**DRS**	**Feature**	**Low data**	**DRS**	**Feature**	**Low data**	**DRS**
sex	0.532 (0.060)	0.642 (0.006)	**treatment**	0.626 (0.027)	0.579 (0.072)	**race**	0.439 (0.101)	0.600 (0.051)
chf	0.645 (0.007)	**0.680 (0.009)**	**sex**	0.448 (0.173)	0.617 (0.106)	**marital**	0.641 (0.027)	0.622 (0.054)
miord	0.552 (0.031)	0.574 (0.067)	**ascites**	0.414 (0.363)	0.533 (0.363)	**iinfct**	0.370 (0.112)	0.282 (0.088)
event	0.736 (0.015)	0.754 (0.001)	**hepatom**	0.588 (0.012)	0.621 (0.028)	**os12m**	0.501 (0.059)	0.581 (0.031)
			spiders	0.681 (0.056)	0.714 (0.023)	**os30d**	0.537 (0.209)	0.613 (0.015)
			edema	0.457 (0.072)	0.517 (0.034)	**rs12m**	0.674 (0.075)	0.651 (0.000)
			stage	0.362 (0.012)	0.367 (0.030)	**rs30d**	0.754 (0.162)	0.712 (0.347)
			event	0.836 (0.009)	0.833 (0.011)	**abdpain**	0.790 (0.086)	0.622 (0.247)
						discharge	0.512 (0.011)	0.507 (0.015)
						dysuria	0.539 (0.033)	0.651 (0.195)
						condom	0.339 (0.056)	0.259 (0.149)
						itch	0.690 (0.251)	0.567 (0.299)
						lesion	0.600 (0.045)	0.634 (0.033)
						rash	0.552 (0.207)	0.639 (0.234)
						lymph	0.659 (0.203)	0.486 (0.271)
						vagina	0.658 (0.146)	0.647 (0.136)
						dchexam	0.493 (0.233)	0.518 (0.292)
						event	0.561 (0.035)	0.565 (0.025)

Table D.8: Different clinical utility validation results for the additional SA datasets. Accuracy comparison between the 'Low data' ($N = 100$) scenario and the 'DRS' technique applied to the lower bound case for each feature used as a classification label. Higher values indicate better performance. **Bold** indicates a significant improvement with the methodology, while * indicates a significant decline. Results are reported as *mean (standard deviation)*.

These results align with previous findings, indicating that while the 'DRS' technique does not consistently yield substantial improvements, it enhances feature generation and inter-variable relationships in specific cases. In particular, several features exhibit improved performance when using the 'DRS' methodology, suggesting its potential for refining synthetic data quality in data-scarce scenarios. Across all datasets, improvements outnumber declines, reinforcing the viability of applying this approach to small datasets for enhancing data utility.

D.3.2 Discrepancy Between Similarity and Clinical Utility Validation

This section explores a potential explanation for the observed divergence between similarity validation and clinical utility validation outcomes.

Proposed Analysis

To investigate this discrepancy, we experimented using classification datasets to examine the relationship between feature importance and classification performance. Specifically, we assess how removing less important features impacts classification accuracy. This analysis employs the same MLP classifier used for clinical utility validation in this study (Section 4.4.3).

Unlike decision tree-based models, such as RFs, which inherently provide feature importance measures, MLP classifiers do not offer this information directly. To address this limitation, we utilize the SHAP framework, an interpretability tool based on cooperative game theory principles (already explained in Appendix C). SHAP assigns Shapley values to features, quantifying their contribution to model predictions. This enables us to estimate feature importance even in black-box models like MLP.

The analysis follows these steps:

1. **Baseline classification:** Train the classifier using the complete dataset, including all available features.
2. **Feature importance estimation:** Compute feature importance using the SHAP framework.
3. **Iterative feature removal:** Identify and remove the feature with the lowest absolute SHAP value.
4. **Reclassification:** Retrain the classifier using the reduced feature set.
5. **Repetition:** Repeat steps 2–4 iteratively, eliminating one feature at a time until only the most critical feature remains.

This experiment aims to demonstrate that classification accuracy remains stable or does not decline significantly even as many features are removed until only the most critical features are left. This finding suggests that only a subset of features is crucial for achieving high classification performance, while many features contribute minimally to the task.

This behavior provides a plausible explanation for the discrepancy between similarity validation and clinical utility validation results:

- **Similarity validation:** This method assesses the statistical fidelity of the entire dataset, evaluating how well all features and their dependencies are preserved in synthetic data. It considers the dataset holistically, treating all features equally important for generative quality.
- **Clinical utility validation:** This method evaluates synthetic data based on its performance in specific tasks (e.g., classification or SA). In these scenarios, a subset of

features (those with high SHAP values) must be well-represented for the synthetic data to perform adequately. Features with low SHAP values contribute minimally to the task, meaning their poor generation has little impact on classification outcomes.

Thus, synthetic data may not perfectly replicate the statistical properties of all features (as indicated by higher $D_{\mathbb{JS}}$ or other similarity validation metrics). However, if the critical task-relevant features are accurately generated, synthetic data can still achieve strong performance in clinical utility validation.

This distinction underscores the importance of aligning validation methods with task-specific requirements. It also highlights why clinical utility validation may yield favorable results even when similarity validation reveals deficiencies in feature generation. Understanding this relationship is crucial for accurately assessing synthetic data quality across different evaluation frameworks.

Results

This proposed analysis provides empirical evidence explaining the observed discrepancy between similarity and clinical utility validation results. We examine the Heart dataset, a binary classification problem to achieve this. Unlike the other two datasets, which involve three-class classification, the Heart dataset offers a simpler structure, making interpreting SHAP values more intuitive. This focused approach ensures clarity while maintaining the methodological robustness of the study.

Table D.9 presents the classification accuracy of MLP models as features are removed iteratively based on their SHAP importance scores. The initial model, trained on all 22 features, achieves an accuracy of 0.643. As less important features are gradually removed, accuracy remains relatively stable, with only minor variations until a significant reduction in the number of features occurs. Notably, even after eliminating 12 features, accuracy remains around 0.650, comparable to the initial performance of the model. A noticeable drop in accuracy emerges only when fewer than five features remain, with the final two-feature model reaching 0.560. These findings suggest that while all features contribute to the dataset, most have minimal impact on classification accuracy.

These results confirm that only a subset of features is essential for classification performance. Despite the removal of many features, the stable accuracy highlights that the accurate generation of key features with high SHAP importance primarily influences clinical utility validation. This aligns with the hypothesis that while similarity validation metrics may indicate poor overall feature generation (e.g., high divergence scores), clinical utility validation can still yield favorable results if the critical features for the task are accurately reproduced.

This behavior reinforces the distinct objectives of the two validation approaches:

- **Similarity validation:** Assesses the generation of all features and their dependencies, making it sensitive to overall dataset discrepancies.
- **Clinical utility validation:** Prioritizes task-specific performance, where the accurate generation of critical features compensates for deficiencies in less impactful ones.

Feature removed	Number of features	Acc
None	22	0.643
AnyHealthcare	21	0.637
MentHlth	20	0.630
Education	19	0.616
CholCheck	18	0.632
Veggies	17	0.638
HvyAlcoholConsump	16	0.634
NoDocbcCost	15	0.645
BMI	14	0.650
Fruits	13	0.645
PhysActivity	12	0.655
Stroke	11	0.640
PhysHlth	10	0.608
Diabetes	9	0.619
DiffWalk	8	0.612
Smoker	7	0.593
Income	6	0.640
Sex	5	0.630
HighBP	4	0.635
HighChol	3	0.650
GentHlth	2	0.560

Table D.9: MLP classification accuracy on the Heart dataset with feature reduction. The initial model is trained on the complete dataset, and subsequent models are retrained after iteratively removing the least important feature based on SHAP feature importance rankings.

These findings emphasize the importance of integrating both validation approaches for comprehensively evaluating synthetic data. While similarity validation ensures overall dataset fidelity, clinical utility validation confirms the effectiveness of synthetic data in specific downstream tasks, particularly in data-scarce scenarios. Together, they provide a balanced assessment of SDG methodologies.

D.3.3 Statistical Significance and Multiple Testing Adjustment

To enhance the statistical rigor of our validation results, we conducted hypothesis testing across all similarity and clinical utility validation tables in this study. Specifically, we formulated the null hypothesis (H_0): The performance metrics in the 'Low data' scenario are better than those of the scenarios where the methodology is applied. To test this hypothesis, we calculated p-values as a measure of statistical significance and set a threshold of 0.01, a standard practice in hypothesis testing. A p-value below this threshold indicates sufficient evidence to reject H_0, suggesting a significant performance improvement when applying the methodology. Conversely, if the p-value exceeds 0.01, no statistically significant difference is

observed between the methodology-applied and 'Low data' scenarios.

Scenario	SIMILARITY VALIDATION		CLINICAL UTILITY VALIDATION		
	JS	$D_{\mathbb{KL}}$	Real Acc	Synth Acc	Synth Fine-Tuned Acc
Big data	N/A	N/A	N/A	N/A	N/A
Low data	N/A	N/A	N/A	N/A	N/A
Pre-train	0.000 / **0.000**	0.000 / **0.000**	N/A	0.012 / 0.083	0.010 / 0.080
AVG	0.000 / **0.000**	0.000 / **0.001**	N/A	0.006 / 0.057	0.032 / 0.192
DRS	0.000 / **0.001**	0.000 / **0.000**	N/A	0.130 / 0.520	0.050 / 0.249

(a) Heart dataset

Scenario	SIMILARITY VALIDATION		CLINICAL UTILITY VALIDATION		
	JS	$D_{\mathbb{KL}}$	Real Acc	Synth Acc	Synth Fine-Tuned Acc
Big data	N/A	N/A	N/A	N/A	N/A
Low data	N/A	N/A	N/A	N/A	N/A
Pre-train	0.000 / **0.000**	0.001 / **0.005**	N/A	0.938 / 1.000	0.328 / 1.000
AVG	0.000 / **0.000**	0.001 / **0.005**	N/A	0.785 / 1.000	0.558 / 1.000
DRS	0.000 / **0.000**	0.001 / **0.005**	N/A	0.991 / 1.000	0.960 / 1.000

(b) Diabetes_H dataset

Scenario	SIMILARITY VALIDATION		CLINICAL UTILITY VALIDATION		
	JS	$D_{\mathbb{KL}}$	Real Acc	Synth Acc	Synth Fine-Tuned Acc
Big data	N/A	N/A	N/A	N/A	N/A
Low data	N/A	N/A	N/A	N/A	N/A
Pre-train	0.000 / **0.000**	0.012 / 0.048	N/A	0.210 / 1.000	0.235 / 1.000
AVG	0.000 / **0.000**	0.603 / 0.603	N/A	0.839 / 1.000	0.952 / 1.000
DRS	0.000 / **0.000**	0.074 / 0.170	N/A	0.084 / 0.757	0.092 / 0.757

(c) Diabetes_130 dataset

Table D.10: Statistical validation for the classification datasets across different scenarios. The tables present p-values for similarity and clinical utility validation metrics, reported as *original p-value / adjusted p-value* after applying the Holm correction for multiple testing. **Bold** indicates cases where H_0 was rejected, signifying statistically significant differences from the 'Low data' scenario. N/A denotes metrics that were not applicable or calculated for a given scenario.

Since multiple hypothesis tests were conducted across different validation scenarios, we recognized the need to control the FWER—the probability of making one or more Type I errors (false positives). As the number of tests increases, the likelihood of incorrectly rejecting a true null hypothesis also rises. This issue, extensively documented in statistical literature [229]–[231], needs multiple testing corrections to maintain statistical validity.

To mitigate this problem, we applied the Holm-Bonferroni correction [232], a step-down adjustment method designed to control the overall significance level while preserving statistical power. Unlike the traditional Bonferroni correction [308], which is overly conservative, the

Holm method offers greater sensitivity and does not assume independence among tests. This makes it particularly suitable for datasets with interdependent features and metrics, as in our study.

The Holm adjustment was implemented using the *statsmodels* package in Python, a well-established library for multiple testing corrections. The original and adjusted p-values for all hypothesis tests are presented in Table D.10. By applying this correction, we ensure that statistical significance thresholds remain consistent across multiple comparisons, reducing the risk of false-positive results and improving the reliability of our conclusions.

Our analysis confirms that after applying the Holm adjustment, the adjusted p-values for $D_{\mathbb{JS}}$ in similarity validation consistently fall below the 0.01 threshold in scenarios where the methodology was applied. This result provides strong statistical evidence that the method significantly improves over the 'Low data' scenario regarding data similarity. However, as previously observed, no significant differences emerge in clinical utility validation, indicating that the methodology does not yield substantial advantages in downstream predictive performance. This reinforces the earlier findings that similarity validation and clinical utility validation capture distinct aspects of synthetic data quality, with the former evaluating global data fidelity and the latter focusing on task-specific utility.

Appendix E

Federated Learning

E.1 Extended Evaluation of FedVAE

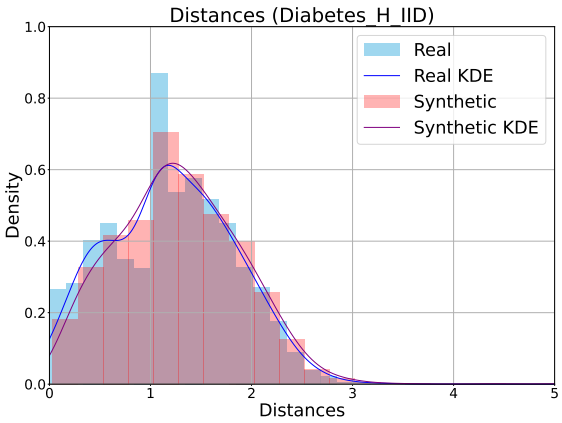
E.1.1 Privacy Concerns

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.

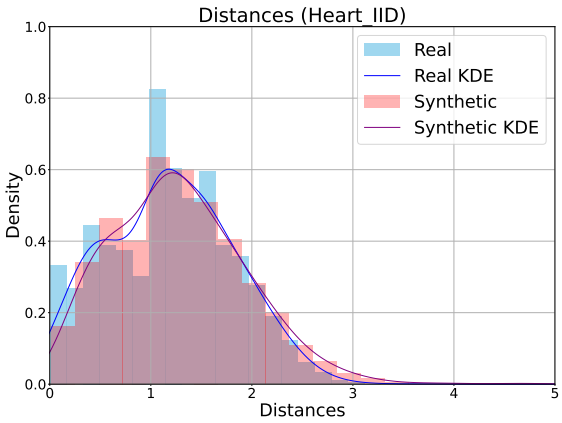
We designed the study as follows:

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:
    - 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 KS tests to compare the two distance distributions.

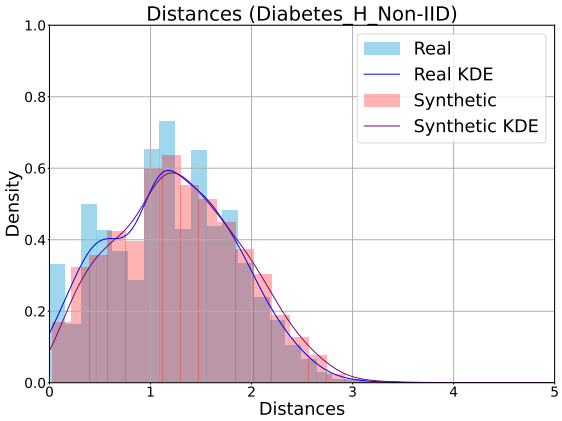
The results of the privacy evaluation are presented in Figure E.1. This evaluation was specifically performed on Node 3 in both IID and non-IID scenarios for the two datasets used in the study (Heart and Diabetes_H), resulting in four distinct plots. These analyses comprehensively evaluate privacy preservation under different data distributions and experi-



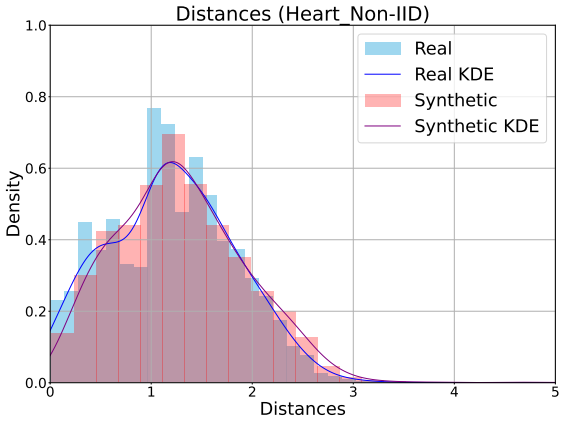
(a) Diabetes_H IID Scenario



(b) Heart IID Scenario



(c) Diabetes_H non-IID Scenario



(d) Heart non-IID Scenario

Figure E.1: 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.

mental 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.

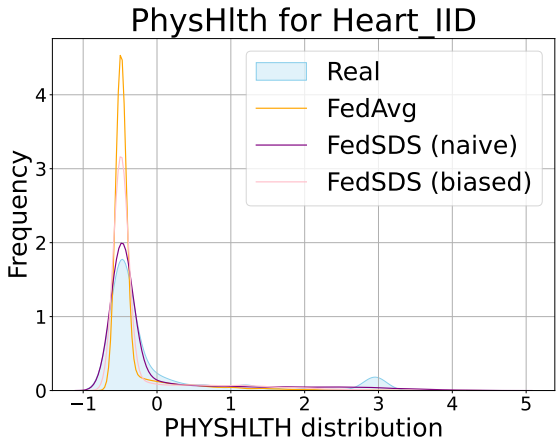
These results demonstrate that the synthetic data provide sufficient privacy protection. Specifically, even in the event of an attack such as man-in-the-middle during the federated data sharing process, the exposure of synthetic data would present a far lower risk than raw data transmission, as the synthetic 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.

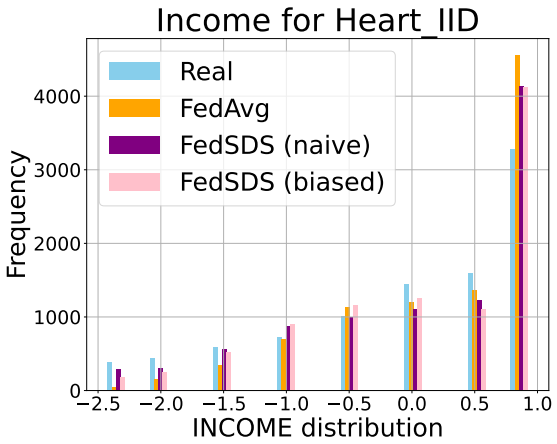
E.1.2 Comparison of Feature Distributions

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 Heart 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 FedSDS (*naive* and *biased* aggregation strategies), aiming to assess the quality of the synthetic data generated by each technique.

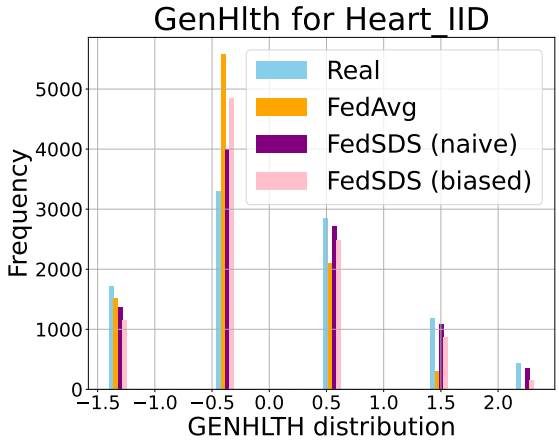
Figures E.2 and E.3 show the KDEs and histograms of the selected features for both IID and non-IID scenarios. These comparisons provide insights into how effectively each SDG technique captures the distributions of the real data. We can confirm that FedSDS captures critical continuous features more accurately. FedSDS better approximates the real data distribution for features such as *Age* and *BMI* compared to FedAvg. This suggests that FedSDS can model complex continuous variables more effectively. In addition, FedSDS 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 FedSDS to generate accurate feature distributions remains evident in both of them.



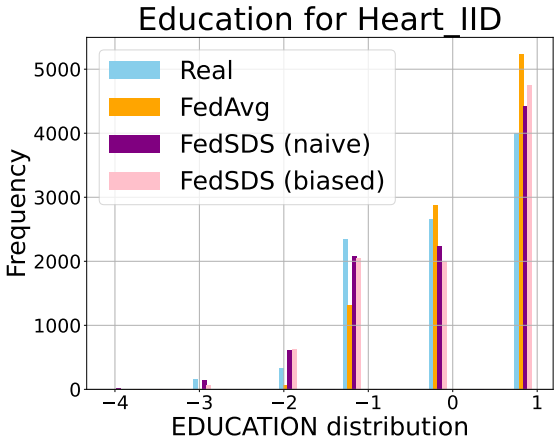
(a) PhysHlth (continuous)



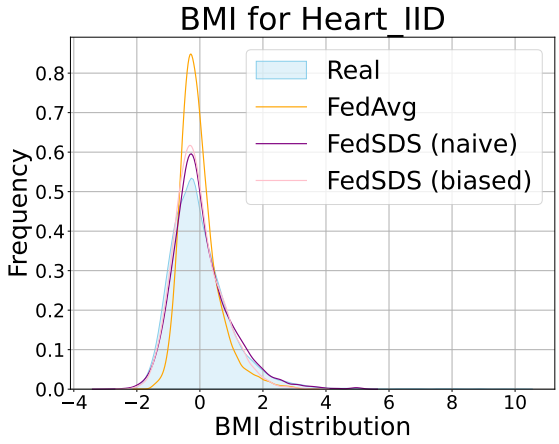
(b) Income (categorical)



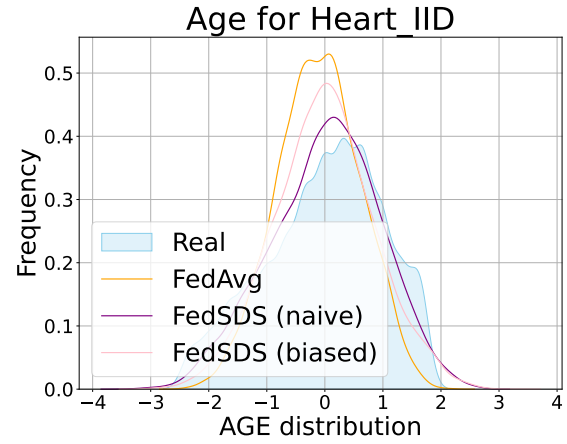
(c) GenHlth (categorical)



(d) Education (categorical)

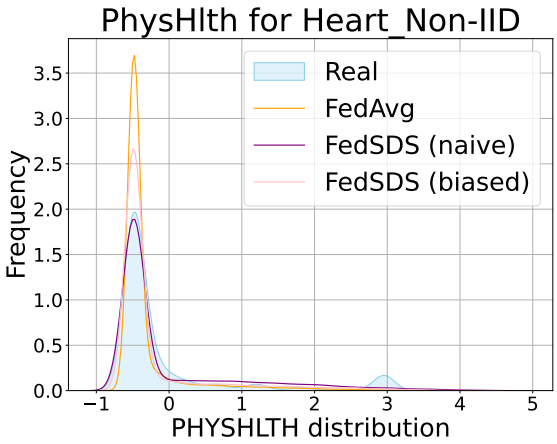


(e) BMI (continuous)

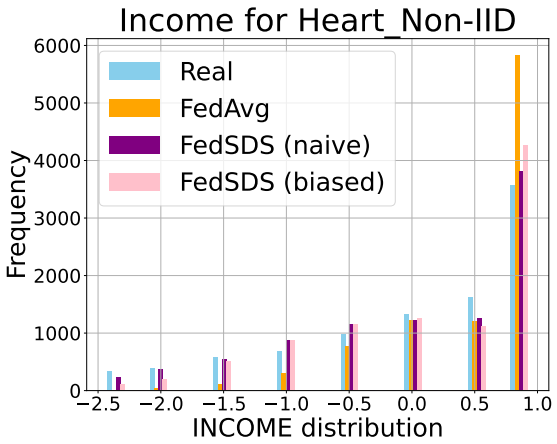


(f) Age (continuous)

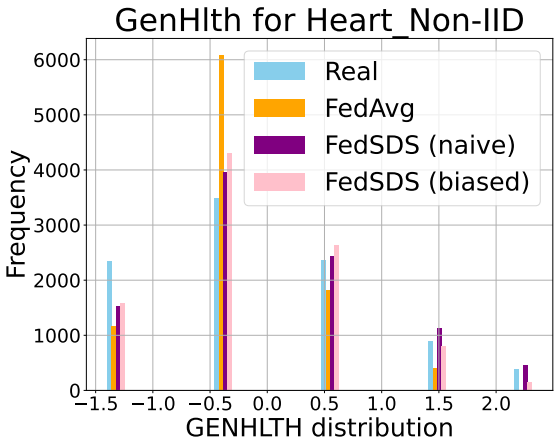
Figure E.2: Normalized distribution plots of selected features. Comparison based on features from the real data, synthetic data generated by FedAvg, and synthetic data generated by FedSDS in the IID scenario for the Heart dataset.



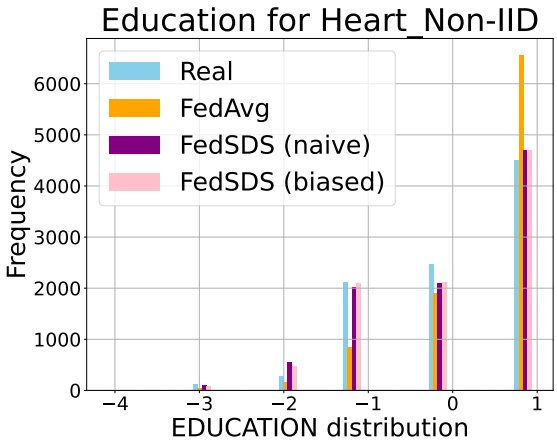
(a) PhysHlth (continuous)



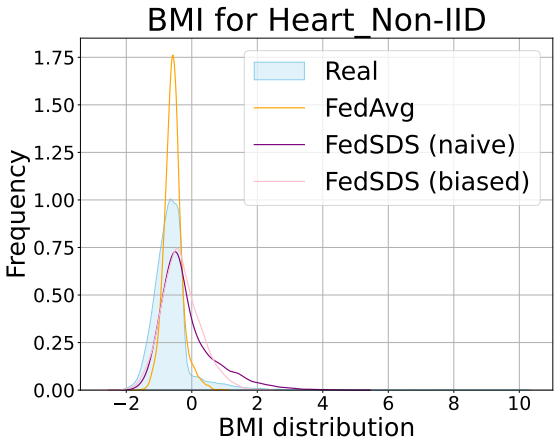
(b) Income (categorical)



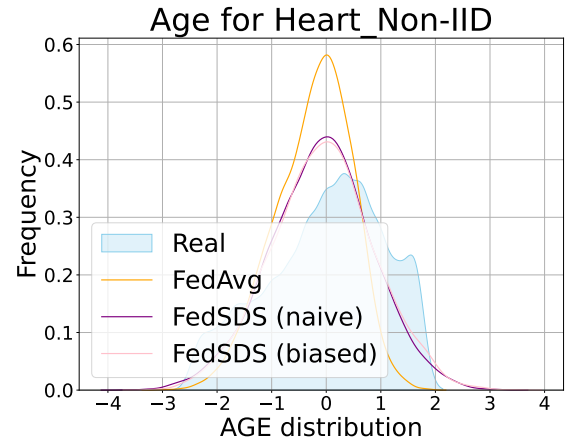
(c) GenHlth (categorical)



(d) Education (categorical)



(e) BMI (continuous)



(f) Age (continuous)

Figure E.3: Normalized distribution plots of selected features. Comparison based on features from the real data, synthetic data generated by FedAvg, and synthetic data generated by FedSDS in the non-IID scenario for the Heart dataset.

E.2 Extended Evaluation of FedSAVAE: Integrating IBS

We introduce in this approach an additional validation measure, the IBS, to complement the C-index metric used in our primary analysis. The following tables present IBS results in the same structured format as in Section 5.4.2 to ensure consistency and facilitate direct comparison. This supplementary metric offers deeper insights into model performance, particularly in terms of predictive accuracy.

Recall that the BS, a widely recognized measure of predictive accuracy [309], [310], is the foundation for this evaluation. BS quantifies the squared prediction error, incorporating IPCW [75] to account for censored data. IPCW assigns greater weights to uncensored samples, ensuring that censored observations do not bias the estimation. To provide a time-independent evaluation, we use the IBS, defined as:

$$iBS(t_{max}) = \frac{1}{t_{max}} \int_0^{t_{max}} BS(t)dt. \tag{E.1}$$

In addition, to ensure robust statistical comparisons, the Holm method [232] was applied to adjust p-values when evaluating the statistical significance of differences across FL methods. This correction mitigates the risk of false positives when performing multiple hypothesis tests.

By integrating IBS alongside the C-index, this extended analysis provides a more comprehensive assessment of model performance, capturing both predictive accuracy and survival estimation quality across different scenarios.

E.2.1 IID Scenarios

The following tables present IBS results under IID scenarios, where nodes vary in both sample size and the percentage of missing data across different settings. The IBS is a complementary metric to the C-index, providing a more stringent evaluation of model performance by assessing both discrimination and calibration. Unlike the C-index, which focuses primarily on ranking survival probabilities, IBS is highly sensitive to miscalibrations and prediction errors, making it a more challenging metric for improvement.

Key findings are:

- In Metabric (first subtable in Table E.1), significant IBS improvements are observed in Node 3 for Scenarios 2 and 3, which correspond to the nodes with the smallest sample size and the highest proportion of missing data. These results highlight the ability of federated approaches to mitigate data scarcity effects. Notably, in Scenario 2, FedSDS methods—particularly the *naive* approach—demonstrate a greater relative decrease in IBS compared to isolated training.
- The IBS improvements are more limited under IID scenarios for GBSG in the second subtable in Table E.1. In Scenario 1, significant IBS reductions are observed for Nodes 2 and 3 using FedAvg, similar to the pattern seen with the C-index. In Scenarios 2

and 3, notable reductions occur for Node 3 (Scenario 2) and Node 2 (Scenario 3), both driven by FedAvg.

Scenario	Nodes	Isolated	FedAvg	FedSDS *naive*	FedSDS *biased*	Adjusted p-values
Centralized	**Node 1**	(0.127 - 0.146 - 0.166)	-	-	-	-
	Node 1	(0.134 - 0.153 - 0.175)	(0.132 - 0.152 - 0.173)	(0.133 - 0.153 - 0.174)	(0.133 - 0.153 - 0.173)	0.504 / 1.000 / 0.834
Scenario 1	**Node 2**	(0.144 - 0.164 - 0.185)	(0.144 - 0.163 - 0.184)	(0.141 - 0.165 - 0.190)	(0.143 - 0.164 - 0.188)	0.945 / 1.000 / 1.000
	Node 3	(0.135 - 0.154 - 0.175)	(0.133 - 0.153 - 0.173)	(0.133 - 0.157 - 0.182)	(0.134 - 0.154 - 0.177)	0.078 / 1.000 / 1.000
	Node 1	(0.133 - 0.153 - 0.173)	(0.134 - 0.153 - 0.174)	(0.135 - 0.154 - 0.176)	(0.134 - 0.153 - 0.175)	1.000 / 1.000 / 1.000
Scenario 2	**Node 2**	(0.151 - 0.174 - 0.199)	(0.150 - 0.173 - 0.196)	(0.144 - 0.170 - 0.200)	(0.152 - 0.172 - 0.195)	1.000 / 0.970 / 1.000
	Node 3	(0.161 - 0.190 - 0.223)	(0.143 - 0.167 - 0.191)	(0.131 - 0.154 - 0.180)	(0.139 - 0.160 - 0.184)	**0.012** / **0.001** / **0.005**
	Node 1	(0.134 - 0.153 - 0.174)	(0.132 - 0.152 - 0.173)	(0.134 - 0.153 - 0.174)	(0.133 - 0.153 - 0.175)	0.530 / 1.000 / 1.000
Scenario 3	**Node 2**	(0.151 - 0.173 - 0.195)	(0.151 - 0.173 - 0.198)	(0.144 - 0.168 - 0.193)	(0.151 - 0.172 - 0.195)	1.000 / 0.527 / 1.000
	Node 3	(0.180 - 0.212 - 0.250)	(0.147 - 0.172 - 0.196)	(0.135 - 0.161 - 0.188)	(0.149 - 0.173 - 0.198)	**0.005** / **0.001** / **0.006**

(a) Metabric dataset

Scenario	Nodes	Isolated	FedAvg	FedSDS *naive*	FedSDS *biased*	Adjusted p-values
Centralized	**Node 1**	(0.174 - 0.196 - 0.219)	-	-	-	-
	Node 1	(0.163 - 0.186 - 0.211)	(0.163 - 0.185 - 0.207)	(0.166 - 0.187 - 0.210)	(0.165 - 0.188 - 0.212)	1.000 / 1.000 / 1.000
Scenario 1	**Node 2**	(0.159 - 0.181 - 0.204)	(0.151 - 0.173 - 0.196)	(0.160 - 0.183 - 0.211)	(0.156 - 0.181 - 0.212)	**0.001** / 1.000 / 1.000
	Node 3	(0.150 - 0.173 - 0.198)	(0.145 - 0.166 - 0.188)	(0.150 - 0.172 - 0.199)	(0.152 - 0.172 - 0.194)	**0.010** / 1.000 / 1.000
	Node 1	(0.167 - 0.188 - 0.211)	(0.165 - 0.187 - 0.210)	(0.165 - 0.186 - 0.209)	(0.166 - 0.188 - 0.211)	0.436 / 0.055 / 1.000
Scenario 2	**Node 2**	(0.165 - 0.190 - 0.217)	(0.162 - 0.185 - 0.213)	(0.171 - 0.197 - 0.229)	(0.164 - 0.191 - 0.220)	0.919 / 1.000 / 1.000
	Node 3	(0.172 - 0.197 - 0.230)	(0.158 - 0.180 - 0.204)	(0.161 - 0.189 - 0.217)	(0.163 - 0.188 - 0.216)	**0.021** / 0.518 / 0.366
	Node 1	(0.166 - 0.187 - 0.210)	(0.163 - 0.186 - 0.209)	(0.164 - 0.186 - 0.211)	(0.163 - 0.186 - 0.210)	1.000 / 1.000 / 1.000
Scenario 3	**Node 2**	(0.160 - 0.186 - 0.215)	(0.162 - 0.187 - 0.210)	(0.165 - 0.191 - 0.220)	(0.164 - 0.192 - 0.223)	1.000 / 1.000 / 1.000
	Node 3	(0.184 - 0.208 - 0.236)	(0.178 - 0.203 - 0.236)	(0.187 - 0.217 - 0.246)	(0.190 - 0.215 - 0.242)	0.988 / 1.000 / 1.000

(b) GBSG dataset

Table E.1: IBS comparison of isolated, FedAvg, and FedSDS (*naive* and *biased*) methods for Metabric and GBSG in IID scenarios. The tables present mean IBS values along with CIs. Adjusted p-values below 0.05, indicating statistically significant differences compared to the isolated training scenario, are highlighted in **bold**.

These findings suggest that while FL techniques can enhance IBS performance, the degree of improvement depends on dataset characteristics, including its structure and survival distribution. The Metabric dataset, with its greater heterogeneity and missing data challenges,

exhibits more substantial benefits from federated approaches than GBSG.

E.2.2 Non-IID Scenarios

In non-IID settings, where heterogeneity is introduced in the distribution of the *age* covariate across Nodes 2 and 3, the IBS results reveal distinct improvement patterns that differ notably from those observed with the C-index. This contrast underscores the complementary nature of these two evaluation metrics.

Key findings are:

- In Metabric (first subtable in Table E.2), significant IBS improvements are observed in Scenarios 5 and 6, primarily in Node 2, whereas C-index improvements were predominantly concentrated in Node 3. In Scenario 5, both FedAvg and FedSDS *biased* methods exhibit performance gains, with FedSDS *biased* demonstrating the most substantial enhancements. In Scenario 6, FedSDS *biased* is the only method to yield significant IBS improvements, highlighting its robustness in handling data heterogeneity.
- In GBSG (second subtable in Table E.2), the FedAvg technique is the only approach associated with IBS improvements, particularly in Node 2 of Scenario 6. No additional significant gains are observed, emphasizing the challenges of reducing IBS in non-IID settings, where discrepancies in data distribution introduce additional complexity.

These results suggest that FL methods (particularly FedSDS *biased*) can effectively mitigate the impact of non-IID data, although the degree of improvement varies across datasets and scenarios. While C-index captures ranking consistency, IBS provides a more nuanced measure of model calibration, making it a crucial complementary evaluation metric.

E.2.3 Special IID Scenario

Table C.5 presents IBS results for Scenario 7, where only FedSDS-based techniques can be applied due to the removal of the *age* column in Node 2. This scenario provides insight into how federated approaches handle missing critical covariates while maintaining model performance.

Key findings are:

- Significant IBS improvements are observed in both the Metabric and GBSG datasets, demonstrating the effectiveness of FedSDS methods in mitigating missing data effects.
- Methods that combine column prediction with synthetic data augmentation yield the most substantial performance gains.
- The *biased* synthetic data aggregation method produces the biggest improvements, aligning the synthetic data distribution more closely with local data in Node 2 and reducing the impact of missing covariates.

The *biased* FedSDS approach achieves the most significant improvements for both datasets, with adjusted p-values below 0.01. These findings underscore the effectiveness of integrating imputation with synthetic data aggregation in addressing data heterogeneity and compensating

Scenario	Nodes	Isolated	FedAvg	FedSDS *naive*	FedSDS *biased*	Adjusted p-values
Centralized	**Node 1**	(0.127 - 0.146 - 0.166)	-	-	-	-
	Node 1	(0.133 - 0.152 - 0.173)	(0.133 - 0.152 - 0.174)	(0.134 - 0.153 - 0.174)	(0.132 - 0.153 - 0.173)	1.000 / 1.000 / 1.000
Scenario 4	**Node 2**	(0.154 - 0.175 - 0.197)	(0.153 - 0.174 - 0.196)	(0.154 - 0.176 - 0.203)	(0.154 - 0.175 - 0.197)	0.493 / 1.000 / 1.000
	Node 3	(0.135 - 0.156 - 0.180)	(0.134 - 0.153 - 0.174)	(0.132 - 0.155 - 0.179)	(0.134 - 0.153 - 0.174)	0.452 / 1.000 / 0.573
	Node 1	(0.134 - 0.154 - 0.178)	(0.133 - 0.153 - 0.175)	(0.134 - 0.153 - 0.174)	(0.134 - 0.154 - 0.178)	1.000 / 1.000 / 1.000
Scenario 5	**Node 2**	(0.153 - 0.174 - 0.198)	(0.148 - 0.168 - 0.190)	(0.141 - 0.167 - 0.194)	(0.141 - 0.161 - 0.185)	**0.006** / 0.187 / **0.000**
	Node 3	(0.137 - 0.161 - 0.187)	(0.135 - 0.157 - 0.183)	(0.136 - 0.172 - 0.206)	(0.145 - 0.170 - 0.198)	1.000 / 1.000 / 1.000
	Node 1	(0.134 - 0.153 - 0.173)	(0.133 - 0.153 - 0.174)	(0.133 - 0.153 - 0.174)	(0.134 - 0.154 - 0.174)	1.000 / 1.000 / 1.000
Scenario 6	**Node 2**	(0.154 - 0.174 - 0.197)	(0.150 - 0.172 - 0.193)	(0.140 - 0.165 - 0.190)	(0.142 - 0.163 - 0.186)	0.068 / 0.054 / **0.000**
	Node 3	(0.142 - 0.166 - 0.192)	(0.169 - 0.198 - 0.230)	(0.166 - 0.205 - 0.256)	(0.175 - 0.212 - 0.243)	1.000 / 1.000 / 1.000

(a) Metabric dataset

Scenario	Nodes	Isolated	FedAvg	FedSDS *naive*	FedSDS *biased*	Adjusted p-values
Centralized	**Node 1**	(0.174 - 0.196 - 0.219)	-	-	-	-
	Node 1	(0.164 - 0.186 - 0.209)	(0.164 - 0.186 - 0.208)	(0.166 - 0.187 - 0.211)	(0.164 - 0.186 - 0.209)	1.000 / 1.000 / 1.000
Scenario 4	**Node 2**	(0.150 - 0.171 - 0.194)	(0.147 - 0.167 - 0.188)	(0.148 - 0.172 - 0.201)	(0.147 - 0.170 - 0.192)	0.061 / 1.000 / 1.000
	Node 3	(0.149 - 0.171 - 0.195)	(0.147 - 0.167 - 0.189)	(0.148 - 0.171 - 0.198)	(0.149 - 0.171 - 0.194)	0.169 / 1.000 / 1.000
	Node 1	(0.165 - 0.187 - 0.210)	(0.165 - 0.185 - 0.209)	(0.164 - 0.186 - 0.210)	(0.167 - 0.188 - 0.211)	0.561 / 1.000 / 1.000
Scenario 5	**Node 2**	(0.170 - 0.193 - 0.218)	(0.170 - 0.192 - 0.214)	(0.172 - 0.194 - 0.217)	(0.171 - 0.194 - 0.220)	1.000 / 1.000 / 1.000
	Node 3	(0.177 - 0.207 - 0.240)	(0.169 - 0.197 - 0.224)	(0.170 - 0.199 - 0.233)	(0.167 - 0.192 - 0.224)	0.505 / 1.000 / 0.136
	Node 1	(0.165 - 0.188 - 0.212)	(0.165 - 0.186 - 0.208)	(0.166 - 0.187 - 0.209)	(0.165 - 0.188 - 0.212)	1.000 / 1.000 / 1.000
Scenario 6	**Node 2**	(0.171 - 0.194 - 0.217)	(0.170 - 0.191 - 0.214)	(0.174 - 0.198 - 0.225)	(0.167 - 0.193 - 0.218)	**0.043** / 1.000 / 1.000
	Node 3	(0.177 - 0.204 - 0.236)	(0.170 - 0.193 - 0.217)	(0.178 - 0.205 - 0.234)	(0.177 - 0.201 - 0.229)	0.169 / 1.000 / 1.000

(b) GBSG dataset

Table E.2: IBS comparison of isolated, FedAvg, and FedSDS (*naive* and *biased*) methods for Metabric and GBSG in non-IID scenarios. The table presents mean IBS values along with CIs. Adjusted p-values below 0.05, indicating statistically significant differences compared to the isolated training scenario, are highlighted in **bold**.

for missing variables, further supporting the robustness of FedSDS techniques in federated survival analysis.

Dataset	Imputation	Imputation	Imputation + Synthetic Data *naive*	Imputation + Synthetic Data *biased*	Adjusted p-values
Metabric	(0.154 - 0.176 - 0.199)	(0.156 - 0.178 - 0.201)	(0.139 - 0.168 - 0.196)	(0.136 - 0.160 - 0.183)	0.992 / **0.044** / **0.000**
GBSG	(0.175 - 0.197 - 0.220)	(0.180 - 0.216 - 0.255)	(0.157 - 0.190 - 0.223)	(0.169 - 0.192 - 0.218)	0.992 / 0.122 / **0.009**

Table C.5: IBS comparison in Scenario 7 across different FedSDS settings. Average IBS results with CIs are presented. Adjusted p-values below 0.05 indicate statistically significant differences from the isolated case and are highlighted in **bold**.

E.2.4 Discussion

Improving IBS is inherently more challenging than enhancing the C-index due to its dual emphasis on discrimination and calibration. While the C-index solely measures the ability of a model to rank survival probabilities correctly, IBS also penalizes miscalibration, assessing the alignment of predicted probabilities with actual event outcomes. This enhanced sensitivity makes IBS particularly vulnerable to model misalignments in scenarios involving heterogeneous data distributions, censoring effects, or limited sample sizes. As a result, achieving significant IBS improvements requires methodologies that preserve ranking accuracy while ensuring well-calibrated survival probabilities. FedSDS *biased* consistently delivers superior IBS performance among the tested approaches. This technique aligns synthetic data distributions more closely with local node characteristics, reducing the impact of heterogeneity and missing critical covariates and ultimately leading to better-calibrated survival probability estimations.

Appendix F

Code Availability

This appendix provides direct links to the repositories containing all the necessary code and resources to ensure full reproducibility of the experiments conducted in this thesis. These repositories include implementations of the proposed methodologies, experimental configurations, and validation frameworks used throughout this research. By making these materials available, we aim to facilitate transparency, foster collaboration, and enable further advancements in the field. Each repository contains a comprehensive *README.m* file detailing the execution process, available scripts, trained models, and the results obtained. Additionally, a requirements file specifying the necessary packages is included to facilitate the installation of dependencies. All experiments were conducted using Python 3.8 or 3.9, ensuring compatibility with the provided code.

- **Survival Analysis:**
  - SAVAE (Section 3.2): https://github.com/Patricia-A-Apellaniz/savae
  - CR-SAVAE (Section 3.3): https://github.com/Patricia-A-Apellaniz/cr-savae
- **Synthetic Data Generation:**
  - VAE-BGM (Section 4.2): https://github.com/Patricia-A-Apellaniz/vae-bgm_data_generator
  - Divergence Estimation (Section 4.3): https://github.com/Patricia-A-Apellaniz/divergence_estimator
  - Generation Methodology (Section 4.4): For general-purpose data, the code is available in https://github.com/Patricia-A-Apellaniz/low_sample_data_generator. For medical data, the code is available in https://github.com/Patricia-A-Apellaniz/medical_low_sample_generator.
- **Federated Learning:**
  - FedVAE (Section 5.3): https://github.com/Patricia-A-Apellaniz/fed_vae
  - FedSAVAE (Section 5.4): https://github.com/Patricia-A-Apellaniz/fed_savae