Synthetic tabular data generation in Federated Learning environments: A practical use case for Acute Myeloid Leukemia

Imanol Isasa, Mikel Catalina, Gorka Epelde, Naiara Aginako, Andoni Beristain

Submitted to: JMIR Medical Informatics
on: March 20, 2025

Table of Contents

Synthetic tabular data generation in Federated Learning environments: A practical use case for Acute Myeloid Leukemia

Imanol Isasa[1, 2] MSc; Mikel Catalina[1] MSc; Gorka Epelde[1, 3] PhD; Naiara Aginako[2] PhD; Andoni Beristain[1, 2, 3] PhD

[1]Digital Health & Biomedical Technologies Department Foundation (BRTA) Vicomtech Donostia-San Sebastián ES
[2]Computer Science and Artificial Intelligence Department University of the Basque Country Donostia-San Sebastián ES
[3]eHealth Group Biogipuzkoa Health Research Institute Donostia-San Sebastián ES

Corresponding Author:
Imanol Isasa MSc
Digital Health & Biomedical Technologies Department
Foundation (BRTA)
Vicomtech
Mikeletegi Pasealekua 57
Donostia-San Sebastián
ES

Abstract

Background: Data scarcity and dispersion pose significant obstacles in biomedical research, particularly when addressing rare diseases. In such scenarios, Synthetic Data Generation (SDG) has emerged as a promising path to mitigate the first issue. Concurrently, Federated Learning (FL) is a machine learning paradigm where multiple nodes collaborate to create a centralized model with knowledge that is distilled from the data in different nodes, but without the need for sharing it. This research explores the combination of SDG and FL technologies in the context of Acute Myeloid Leukemia, a rare hematological disorder, evaluating their combined impact and the quality of the generated artificial datasets.

Objective: To evaluate the privacy- and fidelity-related impact of federating a SDG model in different data distribution scenarios and with different numbers of nodes, comparing them with a centralized baseline SDG model.

Methods: A state-of-the-art Generative Adversarial Network architecture was trained considering four different scenarios: a (1) non-federated baseline with all the data available, a (2) federated scenario where the data was evenly distributed among different nodes, a (3) federated scenario where the data was unevenly and randomly distributed (imbalanced data), and a (4) federated scenario with non-IID data distributions. For each of the federated scenarios, a fixed set of node quantities (3, 5, 7, 10) was considered to assess its impact, and the generated data was evaluated attending to a fidelity-privacy trade-off.

Results: The computed fidelity metrics exhibited statistically significant deteriorations ($P < 0.001$) ranging from 0.21% to 21.23% due to the federation process. When comparing federated experiments trained with diverse numbers of nodes, no strong tendencies were observed, even if specific comparisons resulted in significative differences. Privacy metrics were mainly maintained while obtaining maximum improvements of 55.17% and maximum deteriorations of 26.23, although they were not statistically significant.

Conclusions: Within the scope of the use case scenario in this paper, the act of federating an SDG algorithm results in a loss of data fidelity compared to the non-federated baseline while maintaining privacy levels. However, this deterioration does not significantly increase as the number of nodes used to train the models grows, even though significative differences were found in specific comparisons. The fact that the amount of data was differently distributed was neither significant for most experiments nor metrics, as similar tendencies were found for all scenarios.

(JMIR Preprints 20/03/2025:74116)
DOI: https://doi.org/10.2196/preprints.74116

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Synthetic tabular data generation in Federated Learning environments: A practical use case for Acute Myeloid Leukemia

Abstract

Background: Data scarcity and dispersion pose significant obstacles in biomedical research, particularly when addressing rare diseases. In such scenarios, Synthetic Data Generation (SDG) has emerged as a promising path to mitigate the first issue. Concurrently, Federated Learning (FL) is a machine learning paradigm where multiple nodes collaborate to create a centralized model with knowledge that is distilled from the data in different nodes, but without the need for sharing it. This research explores the combination of SDG and FL technologies in the context of Acute Myeloid Leukemia, a rare hematological disorder, evaluating their combined impact and the quality of the generated artificial datasets.
Objective: To evaluate the privacy- and fidelity-related impact of federating a SDG model in different data distribution scenarios and with different numbers of nodes, comparing them with a centralized baseline SDG model.
Methods: A state-of-the-art Generative Adversarial Network architecture was trained considering four different scenarios: a (1) non-federated baseline with all the data available, a (2) federated scenario where the data was evenly distributed among different nodes, a (3) federated scenario where the data was unevenly and randomly distributed (imbalanced data), and a (4) federated scenario with non-IID data distributions. For each of the federated scenarios, a fixed set of node quantities (3, 5, 7, 10) was considered to assess its impact, and the generated data was evaluated attending to a fidelity-privacy trade-off.
Results: The computed fidelity metrics exhibited statistically significant deteriorations ($P < 0.001$) ranging from 0.21% to 21.23% due to the federation process. When comparing federated experiments trained with diverse numbers of nodes, no strong tendencies were observed, even if specific comparisons resulted in significative differences. Privacy metrics were mainly maintained while obtaining maximum improvements of 55.17% and maximum deteriorations of 26.23, although they were not statistically significant.
Conclusions: Within the scope of the use case scenario in this paper, the act of federating an SDG algorithm results in a loss of data fidelity compared to the non-federated baseline while maintaining privacy levels. However, this deterioration does not significantly increase as the number of nodes used to train the models grows, even though significative differences were found in specific comparisons. The fact that the amount of data was differently distributed was neither significant for most experiments nor metrics, as similar tendencies were found for all scenarios.

Keywords: Rare diseases; Privacy; Machine Learning; Federated Learning; Synthetic Data Generation; Leukemia; Data Fidelity; Trade-off.

Introduction

Acute Myeloid Leukemia (AML) is a group of bone marrow (BM) stem cell cancers that causes an extreme proliferation of clonal hematopoietic cells. This abnormal growth is caused by multiple cytogenetic and genetic malformations, resulting in a poorly differentiated myeloid cell accumulation in the BM and the consequent spread to the blood [1].

Despite the latest scientific advances and the at least 10 recently approved therapies, it is still causing 250,000 deaths yearly worldwide [2]. Moreover, even if AML accounts for about 80% of all

diagnosed leukemias in adults, there are just 4.2 new cases per 100,000 subjects in the United States yearly, which makes it classifiable as a rare hematological disease [3]. On top of that, the proportion of AML cases among all the diagnosed leukemias worldwide increased from 18% in 1990 to 23.1% in 2017, augmenting their incidence and suggesting a potential upcoming major global public health concern [4].

According to the World Health Organization (WHO), AML can be classified into several categories: those that are derived from (1) genetic abnormalities, (2) myelodysplasia-related changes, those that are related to (3) previous chemotherapy or radiation therapies, (4) myeloid sarcomas, (5) myeloid proliferations related to Down syndrome, (6) undifferentiated and biphenotypic leukemias, and those that (7) are not otherwise specified [5, 6]. In them, symptoms include bleeding, bruising, infections, fatigue, and bone pain.

The rarity of the AML as a prevalent form of leukemia brings with it inherent limitations with regard to the data exploitation and consequent improvement in terms of Artificial Intelligence (AI) models and their application in real-world environments. First, the necessity of data is leading to the emergence of various repositories that encompass information of the disease [7], but it is important to highlight that these are often limited in size [8, 9], revealing an underlying problem of data scarcity. Besides, the data protection legislation, such as the General Data Protection Regulation in Europe or the Health Insurance Portability and Accountability Act in the United States, adds a layer of complexity to the process of data sharing due to the sensitive nature of health data, as it typically consists of Electronic Health Records (EHR) that may contain extensive clinical or even genomic data. As a consequence, even if AML data records exist, they are unevenly distributed across different institutions, balking any intention to make use of big amounts of data. This uneven distribution does not only refer to the amount of data points available in each data silo but also to biases in them, such as racial and ethnic disparities on AML prevalence statistics [10], especially in pediatric patients [11]. This makes it even more difficult to access quality data that can be used to infer knowledge using AI.

Synthetic Data (SD) is defined as artificial information that is generated from original data and a model that is trained to reproduce its characteristics and structure [12]. Thus, SD Generation (SDG) is a widely employed tool for creating data that mimics real-world datasets, which has been found to be helpful for augmentation tasks, as a class balancing tool, and as a Privacy Enhancing Technology (PET) [13]. Therefore, SD is often evaluated in terms of its fidelity with respect to the real data, its utility for downstream applications, and the privacy that it offers, the last one being one of the main topics of research in literature. In light of the current situation regarding the AML use case, SD can be considered a suitable approach for improving the current paradigm by increasing the quantity of data institution-wise. However, while SD aims to replicate real-world distributions by capturing the same range and structure as the input data, its primary focus is on addressing data scarcity rather than mitigating the problem of scattering. In this regard, SDG would be able to learn based on the local distributions and attending to the data variability found within an institution, possibly limiting the learning process and not being able to sufficiently represent a global population [14].

In order to address the scattering issue, Federated Learning (FL) is a Machine Learning (ML) computation framework that seeks to address data governance by training the algorithms without exchanging the data itself [15]. In a canonical FL environment, a model is trained in a central server using the weights each client shares after training local models on local real data [16]. Even if that local real data does not leave the node, the learnt information is shared and a global model is created, covering all the local distributions among the federated nodes and better adapting to a theoretical global distribution.

However, the adoption of technologies that combine both elements and the posterior validation of those should go hand in hand with thorough prior analyses. To do that, the contributions of this paper are:

- An evaluation of the impact of federating an SDG algorithm with respect to having a model trained on all data available on the same site (centralized).
- An evaluation of the impact of the number of federated nodes on the performance of an SDG model.
- An evaluation of the impact of having a randomly sampled imbalanced quantity of data in each federated node.
- An evaluation of the impact of having an imbalanced quantity of data that constitutes non-IID distributions in each federated node.

The remainder of this article is organized as follows: The *Methods* section provides information about the utilized materials and the methodology, describing the dataset that was used, the generative model, the FL setup and the evaluation metrics that were implemented. The next section presents the *Results*, while the *Discussion* section shows principal results, the limitations of this work, a comparison with prior existing work and final conclusions.

Background

Over the last few years, the usage of SD has gained momentum in several contexts. In healthcare, simulations and prediction research, educational and training content creation, and investigation including release of data have benefited from SD usage [17]. In this sense, SDG must be understood as a spectrum of possibilities regarding model selection, parameter tuning and use case-specific contextualization. However, generating data inevitably involves sophisticated techniques, which may include Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), or diffusion models, among other AI approaches.

One of the most promising usages of SD, as it was mentioned previously, is the generation of artificial information that does not compromise patient privacy. Therefore, the assessment and analyses of SD are expanding to such an extent that the expectations are surpassing those of traditional anonymization techniques, even attempting to combine them. In this sense, it is important to mention the Differential Privacy (DP), which is a technique that ensures privacy of individuals by adding random noise to the data, making it nearly impossible to determine whether any individual's data is included in a dataset or not. DP is the most widely used PET that is being implemented in generative models to enhance data privacy, and its implementation in them depends on the intended usage for the resulting SD.

Regarding AI models, a GAN is composed of two networks, a generator and a discriminator, working in an adversarial manner. While the former one is supposed to minimize the loss function by generating samples that are as similar as possible to the training set, the latter is tasked with differentiating original samples from synthetic ones, trying to maximize the loss. GANs were first presented by Goodfellow et al. in 2014 [18], and since then several modifications have been proposed in order to cover a wide range of use cases. The work developed by Zhao et al. in [19], for example, represents the latest generation of GAN architectures, named CTAB-GAN+, which includes state-of-the-art features such as conditional generation of samples, improved loss functions, possibility to handle both categorical and continuous data simultaneously, and DP. Narrowing down to tabular data generation for healthcare-related use cases, GANs were identified to be the most widely used architectures [20]. Additionally, several recent publications have addressed various

unresolved questions within the field. For instance, Ramachandranpillai et al. [21] introduced the Bias-transforming GAN (Bt-GAN), which addresses the challenge of biased data generation in the healthcare domain by incorporating several information-constraints inside the generation process. Moreover, various GANs are currently being experimentally tested for several use cases, as demonstrated by Akiya et al. [22] in oncological clinical trials, Khan et al. [23] in cardiovascular disease mortality predictions, or Dhawan and Nijhawan [24] in brain Magnetic Resonance Imaging and chest X-ray data.

As for VAEs in the context of SDG, they are also composed of two fundamental components: the encoder and the decoder. In this context, the encoder is responsible for mapping the input training data into a latent space with a lower dimensionality, while the decoder samples new values from this latent space to reconstruct data that imitates the original inputs. Starting from the original architecture, VAEs have also undergone several modifications that help cover diverse use cases [25]. As for the latest research, Biswas and Talukdar [26] researched the enhancement of clinical documentation using both GAN and VAE-generated synthetic data with the aim of improving patient care. Li et al. [27] implemented the causal recurrent VAE (CR-VAE) aiming for generating medical time series data. Other applications that are being investigated and include VAEs are drug dosing determinants, such as in Titar et al. [28].

Lastly, diffusion models create SD by gradually transforming simple, noise-like data into complex data structures that were used during the training process. Even if this type of generative models has mostly been focused on image generation, currently they are able to support different data types, too [28, 29]. For example, Naseer et al. [31] presented ScoEHR, a continuous-time diffusion model able to generate artificial EHRs. Digital pathology data was also generated with diffusion models by Pozzi et al. in [32].

As for generative modelling applied to rare hematological diseases, recently in [33] D’Amico et al. trained a CTGAN with the aim of generating Myelodysplastic Syndromes (MDS) and AML data. Additionally, Eckardt et al. in [34] made their synthetic AML dataset publicly available after having considered both utility and privacy thresholds. The published synthetic dataset comprises 1,606 patients generated using a CTAB-GAN+ [19]. In [35], Licandro et al. utilized a Wasserstein GAN for two distinct scenarios where differently sized datasets were used. In their research, the primary objective was defined to discern the embeddings of the data, enabling subsequent differentiation between blast and non-blast cells. The results show that using the generator model to learn embeddings outperforms the results obtained with baseline models, improving the AUC for both dataset sizes. The study carried out in [36] by Rupapara et al. made use of the ADASYN [37] SD generator to balance the dataset and enhance prediction outcomes. The dataset encompassed data from various blood-related cancers, including AML. By employing the ADASYN resampler, the classification models demonstrated improved accuracy.

Regarding FL-related studies, several experiments have been conducted to overcome the data scattering issue. For example, the study carried out in [38] by Linardos et al. simulates a federated environment consisting of four nodes. The study aimed to help diagnose hypertrophic cardiomyopathy diseases the results supporting the effectiveness of FL by achieving better AUC results than with a Collaborative Data Sharing framework. The work presented by Liu et al. in [39] focused on employing FL to achieve a deep learning model that makes use of EHRs to predict patient mortality, which they called FADL. The work presented by Azizi et al. [40] also utilized EHR information scattered among 50 nodes, each of them containing 560 patients, to predict mortality. However, in this case, they employed a clustering method and utilized Community-Based FL, surpassing the performance of the canonical FL environment across various scenarios.

Both techniques, SDG and FL, have demonstrated their effectiveness in various healthcare-related topics and use cases. To serve as an example, in [40] a framework for cardiovascular data based on a FL architecture of two nodes and a generative model using sequential trees is shown. The study presented in [41] by Behera et al. demonstrates the implementation of a GAN within a federated environment, called FedSyn. In addition to applying DP, thereby enhancing data protection, the researchers utilized the CIFAR10 and MNIST datasets for their analyses. The research outlined in [42] by Xin et al. employs a federated GAN augmented with DP, trained on both MNIST and CelebA datasets. The authors analyzed the privacy the generated data offered against the original one, concluding an improved privacy against Membership Inference Attacks (MIA). However, despite the combination of both SDG and FL explored in different studies, many aspects of this mixture still require evaluation.

Methods

AML dataset

The AML dataset used to perform the research of this paper was accessed from the work developed by Tazi et al. in [43] and its associated GitHub repository [44]. That study was conducted following the completion of informed consent forms by all the included subjects. Also, all the relevant ethical guidelines were followed, and necessary Institutional Review Board and ethics committee approvals were obtained. The trial was conducted in accordance with the tenets of the Helsinki Declaration, and it was sponsored by Cardiff University and approved by the Wales research ethics committee under the protocol number 08/MRE09/29. The analysis of the data in the original study was approved by the Memorial Sloan Kettering Cancer Center Institutional Review Board protocol number x20-064. All the raw data was deposited in the European Genome Phenome Archive under the reference number EGAS00001000570.

Among the available datasets in the repository, the *paper_full_data_validation* dataset was chosen for this research. All the genetic mutation-related variables were discarded, preserving clinical, demographic, and disease-related information (Table 1). The variable selection was carried out to maintain acceptable sample-to-feature ratios across various federated configurations regarding node quantities, as the original dataset comprised 130 features. Having a low number of samples and too many variables would have limited the experiments in this regard as highly overfitted models would appear. The resulting dataset consisted of 1,540 samples and 12 features, from which the categorical ones were label encoded in the preprocessing stage.

Table 1. Description of the variables that were included in this study.

Variable		Type	Description
Clinical			
	Age (yr)	Continuous	Age of the patient
	BM_blasts (%)	Continuous	Number of bone marrow blasts
	HB (g/dl)	Continuous	Hemoglobin
	PLT (10^{-9}/l)	Continuous	Platelet count
	WBC (10^{-9}/l)	Continuous	Number of white blood cells
	OS	Continuous	Overall survival
	Perf_status (ECOG)	Categorical	Performance status in Eastern Cooperative Oncology Group (ECOG) scale
	AHD	Binary	Antecedent hematologic disease

	OS_status	Binary	Overall survival status
Demographic			
	Gender	Binary	Gender of the patient
Disease-related			
	Secondary	Categorical	Secondary AML
	Eln_2017	Categorical	European LeukemiaNet 2017 risk classification

Generative model

The generative model that was selected for this experiment is the Conditional Tabular GAN (CTGAN) [45] as it was recently reported to have one of the most appropriate generators among different GAN and VAE architectures [46]. Additionally, it is implemented in a way that it models the relationships between imbalanced variable distributions [47]. The Synthetic Data Vault [48] implementation was used in this work, even if some modifications had to be implemented for the correct usage of the model in a federated environment. Diffusion models were not considered for this selection as they were not yet considered sufficiently mature for practical usage and are still in the earlier stages of development and research [49].

In order for each participant node to transform the data in the same manner and to avoid averaging mismatches, One Hot Encoding for discrete columns and Gaussian Mixtures transformations for continuous variables were fit using the whole dataset, also being able to avoid unseen classes in federated nodes. The objects were then included in each client with the aim of transforming each data partition *in situ* and using the same mapping.

As for the model parameters, the default configuration presented in the Synthetic Data Vault implementation was utilized. The same architecture was set for both the discriminator and the generator with a two hidden layer structure, both containing 256 units each. The learning rates of both objects were set to $2x10^{-4}$, with a decay fixed to $1x10^{-6}$. A batch size of 500 samples was defined along with an embedding dimension of 128 samples. The discriminator was updated along with the generator at every training step, and a 10-sample group (*pac* parameter) was introduced into the discriminator each time it was applied.

Regarding the number of epochs to be performed during the training process of the models, different experiments were empirically conducted on the baseline model with 500, 1,000, 1,500, 2,000, and 3,000 epochs. The optimal configuration was proven to be 500 epochs as increasing the iterations did not show any significant improvement in the generated synthetic sample quality. All the federated models were trained for 500 epochs for the experiments to be comparable. Also, the number of federation rounds was set to 500.

Experimental setup

In this experiment a comparison between non-federated and federated generative models was performed for three different federated scenarios. In the first one, the data quantity was assumed to be evenly distributed across the participant nodes (from now on, B scenario, for balanced), while for the second the data points were randomly split creating partitions with uneven sample quantities (from now on, IB scenario). In the third federated scenario, non-IID (non-independent and non-identically distributed) distributions (Figure 1) were built depending on the age variable (from now on, $IB_{non\text{-}iid}$ scenario).

For the three scenarios the data was partitioned prior to the model training phase, allowing for traceability and higher results reproducibility. The partitions for the B scenario were created by randomly selecting n/N samples for each of the nodes, n being the total number of samples in the original dataset and N being the number of nodes that participate in a specific federated experiment. On the other hand, the IB scenario partitions were created so that for a specific N, $N-1$ nodes were trained on 5% of the data that was chosen randomly, and the N^{th} node was trained on the remaining samples. The $IB_{non\text{-}iid}$ scenario was created by sampling age-dependent data points from Dirichlet ($\alpha=10$¿ distributions. These are typically used as prior distributions in Bayesian statistics, constituting an appropriate choice to simulate real-world data [50]. In this analysis, the federation was evaluated for a set of $N \in \{3,5,7,10\}$ (Figure 2).

The federated models were trained by averaging the model weights coming from each node and attending to the number of samples each one contained (Federated Average).

The Flower 1.7.0 Framework [51] was used in order to federate the models using the simulation module. The experiments were performed in a High-Performance Computing cluster, and they were allocated for 10 CPU and 2 GPU jobs, having initiated a Ray actor cluster prior executing federation rounds. Figure 3 represents the workflow that was carried out during the experiment execution, where data partition, model training, and evaluation phases are shown.

Evaluation metrics

The generated SD was analyzed to gauge its fidelity and privacy with respect to the real data. In the scope of this work, fidelity is defined as the degree to which the generated SD replicates the characteristics, patterns, correlations, and distributions of the real data. While a high fidelity means the SD resembles the real data well, a low fidelity would indicate poor learning by the model generators. On the other hand, privacy is defined as the extent to which the generated data protects sensitive information from being disclosed in the original dataset. In this section the methods and metrics to evaluate the SD are described.

Considering a simulated FL scenario, the comparison was performed against the whole real data, thus being able to compare the performance of each setup against the non-federated scenario as a baseline. Inspired by usual ML cross-validation, 10 different synthetic datasets were generated with each model, allowing a separate evaluation for each of them. The results were then averaged to provide a more robust perspective on their generalizability. The set of metrics calculated in each fold also enabled the execution of statistical tests for significance.

In order to assess intervariable correlations, the ϕ_k coefficient [52] and the Vendi Score (VS) [53] metrics were implemented. On the one hand, the ϕ_k coefficient is based on refinement to the Pearson's hypothesis tests. However, unlike Pearson's hypothesis, ϕ_k can calculate correlations with both numerical and categorical variables, the higher the values suggesting better intervariable relations. Moreover, ϕ_k can capture non-linear relationships. Correlation matrices were generated for both real and synthetic versions of the datasets using the ϕ_k coefficient (Multimedia Appendix 1), and the Cosine Similarity (CS) metric was employed to obtain a quantitative measure that compares them, which is defined as $1-d_{\cos}$:

$$d_{\cos}(x, x') = \frac{\sum_{i=1}^{n} x_i \cdot x_i'}{\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i'}$$

where x_i is a real sample and x_i' is a synthetic counterpart. A low cosine similarity metric suggests the two matrices do not look alike, while higher values imply higher similarities between them.

On the other hand, the VS is a novel metric that computes the diversity of a given dataset without the need to compare it against another set of data. This score requires to define a positive semidefinite similarity function, which was set to be the cosine similarity in this case [54]. Accordingly, the VS of just the numerical attributes was computed due to the cosine similarity only being applicable to numerical features. In the following equation the mathematical expression for the VS can be observed.

$$VS_k(x_1, \ldots, x_n) = \exp\left(-\sum_{i=1}^{n} \lambda_i \cdot \log \lambda_i\right)$$

where k is a given similarity function and λ_n are the eigenvalues of K/n and K is the kernel matrix.

Furthermore, a Data Labelling Analysis (DLA) was performed. In this procedure, an ML classifier is trained to ascertain its ability to differentiate between synthetic and real samples, mimicking the functionality of a GAN discriminator. Due to the characteristics of the analysis, the classification process was evaluated using the F1-score metric since it is sensitive to the class distributions making it a reliable metric when labels are imbalanced, and the recall score as it returns the number of correctly identified synthetic samples. On top of that, the AUC curve was calculated. Regarding the trained ML models for the DLA, the LazyPredict classifier [55] object was used to train various models per iteration. The best classifier was chosen for each fold to account for the most restrictive case, while the mean and standard deviation were calculated in the process.

With respect to privacy metrics, two types of attacks were conducted: Membership Inference Attacks (MIA), and Attribute Inference Attacks (AIA). In MIAs, an adversary is simulated to assess whether a specific data point was part of the training dataset used to train a generative model, thereby posing potential privacy risks. The attack methodology involves computing distance measures between pairs of records and applying a threshold to distinguish between high-risk matches and those considered safe. In the context of this experiment, a Gower distance of 0.05 was defined, which is a similarity measure that may be used to handle multi-type data within a same dataset.

In contrast, AIAs occur when an adversary attempts to infer sensitive information that was not originally disclosed with the dataset. AIAs seek to extract additional private information about individuals, even if their membership is already known or assumed. In this case, risks are calculated variable-wise as each one may pose differently ranked sensitivities.

Results

The results of the study are presented in this section, comparing the three presented federated scenarios with the baseline model. All statistical tests were performed for a significance level of 0.05 using the averaged results, while some variable-specific metrics can be checked in the supplementary material (Multimedia Appendix 1).

Starting with the baseline non-federated CTGAN model fidelity evaluation, the ϕ_k coefficient results showed a mean cosine similarity of 0.930. Regarding the DLA execution, the obtained AUC was 0.796, the F1-score was 0.872 and the recall metric was 0.958. The VS resulted in a mean value of 1.405 in the SD, comparing it to the VS obtained in the real dataset of 1.406. To finish, the average Hellinger distance was 0.223. Regarding the privacy evaluation, the MIA demonstrated no significant membership inference risk, while the averaged AIA resulted in a 4.5% risk for attribute information to be inferred.

Regarding the federated models that were trained with balanced datasets (the B scenario), most of the performed experiments showed statistically significant differences in fidelity metrics with respect to the baseline scenario $(P<.001)$, even with variations in the number of nodes (Table 2). No statistical significance was found among the calculated privacy metrics, suggesting no improvement despite the fidelity loss.

Table 2. Fidelity and privacy metric results for the B scenario.

Metric			3N	5N	7N	10N
Fidelity						
	CS ϕ_k					
		$\mu(\sigma)$	0.842 (0.005)	0.846 (0.003)	0.845 (0.003)	0.849 (0.003)
		t	45.525	67.770	74.556	78.915
		P	<.001	<.001	<.001	<.001
	DLA AUC					
		$\mu(\sigma)$	0.965 (0.008)	0.962 (0.008)	0.946 (0.008)	0.965 (0.009)
		t	19.818	19.333	17.391	19.402
		P	<.001	<.001	<.001	<.001
	DLA F1					
		$\mu(\sigma)$	0.965 (0.007)	0.961 (0.008)	0.945 (0.008)	0.964 (0.009)
		t	17.940	16.931	13.853	16.935
		P	<.001	<.001	<.001	<.001
	DLA Recall					
		$\mu(\sigma)$	0.969 (0.012)	0.948 (0.009)	0.940 (0.012)	0.968 (0.010)
		t	2.402	2.644	3.791	2.316
		P	.02	.02	.001	.03
	VS					
		$\mu(\sigma)$	1.437 (0.002)	1.442 (0.001)	1.337 (0.004)	1.251 (0.006)
		t	24.969	31.451	39.839	66.590
		P	<.001	<.001	<.001	<.001
	$d_{hellinger}$					
		$\mu(\sigma)$	0.229 (0.010)	0.223 (0.002)	0.220 (0.002)	0.221 (0.002)
		t	1.767	0.165	2.930	2.517
		P	.09	.87	.009	.02
Privacy						
	MIA					
		$\mu(\sigma)$	0 (0)	0 (0)	0 (0)	0 (0)
		t	-	-	-	-
		P	-	-	-	-
	AIA					

		$\mu(\sigma)$	0.038 (0.018)	0.055 (0.031)	0.038 (0.019	0.039 (0.009)
		t	1.496	1.000	1.414	1.667
		P	.15	.33	.17	.11

Specifically, intervariable correlations were shown to be more distorted than the ones presented by the baseline model, and the DLA suggested that the synthetic samples that were generated by federated models are prone to be detected more easily than the ones generated by the baseline model, although variable-wise metrics such as the Hellinger distance did not demonstrate too different results.

In the IB scenario, most of the performed experiments showed high statistical significances $(P<.001)$ with respect to the baseline, too (Table 3). Intervariable correlations, DLA metrics and VS values were shown to be quite different to the baseline model attending the statistical tests, while the Hellinger distances did not show too big of a difference. In this case, the 10N experiment showed statistically significant differences with the AIA metric obtained in the baseline, suggesting an improvement on the privacy while deteriorating the fidelity values. However, no specific tendency can be observed while increasing the number of nodes in this sense.

Table 3. Fidelity and privacy metric results for the IB scenario.

Metric			3N	5N	7N	10N
Fidelity						
	CS ϕ_k					
		$\mu(\sigma)$	0.848 (0.002)	0.841 (0.003)	0.839 (0.005)	0.847 (0.001)
		t	87.050	84.898	56.581	110.235
		P	<.001	<.001	<.001	<.001
	DLA AUC					
		$\mu(\sigma)$	0.983 (0.004)	0.993 (0.004)	0.968 (0.005)	0.938 (0.009)
		t	23.187	23.700	20.550	16.342
		P	<.001	<.001	<.001	<.001
	DLA F1					
		$\mu(\sigma)$	0.988 (0.004)	0.993 (0.004)	0.968 (0.005)	0.938 (0.008)
		t	24.086	25.109	19.405	12.136
		P	<.001	<.001	<.001	<.001
	DLA Recall					
		$\mu(\sigma)$	0.979 (0.007)	0.986 (0.008)	0.964 (0.008)	0.928 (0.014)
		t	5.749	7.472	1.581	5.666
		P	<.001	<.001	.13	<.001
	VS					
		$\mu(\sigma)$	1.360 (0.011)	1.424 (0.003)	1.436 (0.003)	1.360 (0.011)
		t	27.714	11.833	21.772	12.155
		P	<.001	<.001	<.001	<.001
	$d_{hellinger}$					
		$\mu(\sigma)$	0.222 (0.001)	0.216 (0.002)	0.223 (0.002)	0.217 (0.002)
		t	0.976	7.892	0.119	5.931
		P	.34	<.001	.90	<.001
Privacy						
	MIA					

		$\mu(\sigma)$	0 (0)	0 (0)	0 (0)	0 (0)
		t	-	-	-	-
		P	-	-	-	-
	AIA					
		$\mu(\sigma)$	0.052 (0.019)	0.061 (0.029)	0.041 (0.016)	0.029 (0.017)
		t	0.076	0.768	1.400	2.437
		P	.94	.45	.27	.03

The $IB_{non\text{-}iid}$ scenario followed the same overall patterns found in the previous two scenarios, showing statistically significant differences in fidelity metrics but with no difference in the performed privacy metrics (Table 4).

Table 4. Fidelity and privacy metric results for the $IB_{non\text{-}iid}$ scenario.

Metric			3N	5N	7N	10N
Fidelity						
	CS ϕ_k					
		$\mu(\sigma)$	0.848 (0.003)	0.842 (0.001)	0.841 (0.004)	0.842 (0.003)
		t	82.533	111.330	69.045	83.862
		P	<.001	<.001	<.001	<.001
	DLA AUC					
		$\mu(\sigma)$	0.977 (0.006)	0.948 (0.010)	0.983 (0.005)	0.950 (0.006)
		t	21.572	17.233	22.408	18.339
		P	<.001	<.001	<.001	<.001
	DLA F1					
		$\mu(\sigma)$	0.976 (0.005)	0.947 (0.010)	0.982 (0.005)	0.950 (0.006)
		t	21.033	13.353	22.724	15.602
		P	<.001	<.001	<.001	<.001
	DLA Recall					
		$\mu(\sigma)$	0.960 (0.010)	0.938 (0.019)	0.985 (0.009)	0.952 (0.013)
		t	0.383	3.039	6.891	1.195
		P	.71	.007	<.001	.25
	VS					
		$\mu(\sigma)$	1.440 (0.001)	1.445 (0.001)	1.198 (0.010)	1.367 (0.004)
		t	28.834	33.888	62.753	21.978
		P	<.001	<.001	<.001	<.001
	$d_{hellinger}$					
		$\mu(\sigma)$	0.214 (0.002)	0.214 (0.003)	0.215 (0.002)	0.213 (0.002)
		t	8.801	8.147	8.333	11.928
		P	<.001	<.001	<.001	<.001
Privacy						
	MIA					
		$\mu(\sigma)$	0 (0)	0 (0)	0 (0)	0 (0)
		t	-	-	-	-
		P	-	-	-	-
	AIA					
		$\mu(\sigma)$	0.048 (0.022)	0.044 (0.023)	0.049 (0.024)	0.057 (0.032)
		t	0.366	0.799	0.325	0.362

		P	.72	.44	.75	.72

Gathering all the results in a single figure, similar tendencies can be observed in the three scenarios, where lower correlation values and higher DLA-related metrics can be found among the federated models with respect to the baseline model. The VS metric fluctuated most among the scenarios, offering insight regarding the variability of each generated synthetic data. Hellinger distances and privacy metrics were found not to fluctuate even when comparing centralized and federated models (Figure 4).

Now considering federated experiment pairs (that is, comparing 3N experiments with 5N, 5N with 7N and 7N with 10N) to evaluate if additional federated nodes impact the SD quality in terms of fidelity and privacy, no clear tendencies can be observed in neither scenario (Table 5). Privacy metrics did not show statistical significance, therefore assuming no improvement was achieved in terms of privacy, even though the baseline centralized model offered good results already. Among the fidelity metrics, the CS of the ϕ_k was found to differ between pairs of experiments in some cases, but no specific trend was detected. The same occurred for the DLA-related metrics, where some pairs pointed out significant differences. All the VS metrics were found to be different, even if no improvement or deterioration trend was found, and the Hellinger distance metric varied depending on the scenario.

Table 5. *t*-test results for experiment pairs.

Metric	3N-5N		5N-7N		7N-10N	
	t	*P*	*t*	*P*	*t*	*P*
B scenario						
Fidelity						
CS ϕ_k	1.858	.08	0.522	.61	2.939	.008
DLA AUC	0.895	.38	4.522	<.001	4.944	<.001
DLA F1	0.875	.39	4.495	<.001	4.951	<.001
DLA Recall	4.612	<.001	1.621	.12	5.553	<.001
VS	7.949	<.001	82.267	<.001	36.200	<.001
$d_{hellinger}$	1.704	.11	2.815	.01	0.714	.48
Privacy						
MIA	-	-	-	-	-	-
AIA	1.000	.33	1.000	.33	0.045	.96
IB scenario						
Fidelity						
CS ϕ_k	6.296	<.001	1.447	.16	5.676	<.001
DLA AUC	2.165	.04	11.306	<.001	9.083	<.001
DLA F1	2.192	.04	11.485	<.001	9.026	<.001
DLA Recall	2.088	.05	6.129	<.001	6.967	<.001
VS	14.629	<.001	9.680	<.001	21.062	<.001
$d_{hellinger}$	8.239	<.001	7.517	<.001	5.708	<.001
Privacy						
MIA	-	-	-	-	-	-
AIA	0.743	.47	1.796	.09	1.592	.13
$IB_{non\text{-}iid}$ scenario						
Fidelity						

		CS ϕ_k	7.211	<.001	0.985	.34	0.476	.64
		DLA AUC	7.895	<.001	9.805	<.001	14.043	<.001
		DLA F1	7.705	<.001	9.718	<.001	14.039	<.001
		DLA Recall	3.192	.005	7.086	<.001	6.784	<.001
		VS	15.145	<.001	79.749	<.001	50.058	<.001
		$d_{hellinger}$	0.120	.91	0.585	.57	2.085	.05
	Privacy							
		MIA	-	-	-	-	-	-
		AIA	0.435	.67	0.441	.66	0.615	.54

Discussion

Principal Results

In these experiments a comparison between a GAN based centralized SDG model and federated implementations of the same model was performed using AML data with the aim of evaluating the SD fidelity and privacy trade-off in each of the scenarios, assessing the SD generation techniques over a FL approach to address the data scattering issue while addressing data scarcity. Three different scenarios were considered for the federated models: the one in which the number of samples in each node was evenly distributed (B), the one where the node-wise data quantity was randomly and unevenly distributed (IB), and the one where non-IID data distributions were created ($IB_{non\text{-}iid}$).

For the B scenario the ϕ_k metric deteriorated to a maximum of 9.46% with respect to the baseline, while the DLA showed an average difference of 17.04% in the AUC, a difference of 9.07% for the F1-score, and a difference of 0.21% for the recall metric. The VS showed a difference between 2.22% and 12.31% showing that the diversity of the generated samples varied among experiments. For this scenario, the Hellinger distance varied to a maximum of 2.62%. The most privacy-preserving experiments in terms of the AIA were 3N and 7N with a risk reduction of 18.42% with respect to the baseline, while the worst one (5N) performed 18.18% below, although not statistically significant.

In the IB scenario the maximum deterioration of the ϕ_k metric was 10.84%, and the DLA showed average values of 16.73% in the case of the AUC, 10.26% for the F1-score and 0.62% for the recall metric. The VS showed numerical variables are diversely generated achieving differences between 1.33% and 3.31%, while the Hellinger distances varied to a maximum of 3.24%. AIA analyses showed a maximum improvement of 55.17% on data privacy, while the maximum deterioration was 26.23%, although not statistically significant.

Lastly, the $IB_{non\text{-}iid}$ scenario showed a maximum deterioration of 10.58% in the ϕ_k metric, while DLA AUC scores showed average differences of 21.23%, the F1-score varied 9.54%, and the recall varied 0.10%. The VS showed maximum variations of 17.27%, and the Hellinger distance varied to 4.69%. Regarding privacy metrics, the AIA showed a maximum improvement of 2.27% and a maximum deterioration of 21.05%, although not being statistically significant in this case either.

Limitations

As it was pointed out in the beginning, this work is aimed to provide insight into a specific use case of federated SDG for an AML dataset. Therefore, extending the analysis by using more state-of-the-art SDG models and more extensive aggregation functions for the FL framework may result in more generalizable conclusions, which will be prioritized in future work. Linked with that, models

incorporating more novel tools like DP and the implementation of advanced FL security frameworks should be covered in future extensions of this research.

Furthermore, the three scenarios resulted in similar overall tendencies for all the calculated metrics, suggesting the scenario proposals in this work may not have that much of an impact on the results. Further research may uncover differences for various data dispersion schemas and non-IID distribution types, which may have much more impact on the calculated metrics and the methodology to be followed.

Lastly, the generative aspect of this research should be taken into account for future research, as expanding the calculations to a higher number of datasets may derive in more robust and scalable optimizations on federated SDG. Related to the issue, a more extensive set of metrics should also be considered in the future to match state-of-the-art literature, as this is one of the most evolving topics on the matter.

Comparison with Prior Work

While there are a few articles which analyze the combination of both FL and SDG, this is, to the best of the authors' knowledge, the first research work that evaluates their usage by analyzing them against baseline centralized models, using different numbers of nodes, and considering different real-world data distribution scenarios in AML.

Expanding the literature search, incorporating DP to federated SDG models has been widely investigated as well as successfully integrated into several use cases, improving privacy metrics under those conditions [42], [56]. Our research, however, has shown reasonable privacy guarantees both for centralized and federated scopes without the need of using DP, suggesting the incorporation of it may depend on the final use case as it deteriorates SD fidelity, even though more extensive evaluations should be performed.

Furthermore, in the scope of federated learning, previous works on classification and regression models have hardly shown any deterioration due to the federation process with regards to a centralized model [57]. Instead, when SDG models were compared to centralized results in this AML use case, statistically significant differences appeared, suggesting that SDG models may be more sensible to the federation step than usual ML cases, such as classification or regression.

To finish, the data distribution scenarios considered for this research demonstrated robustness against non-IID distributions, which is in line with other experiments performed in the literature [58].

Conclusions

The results for the three scenarios showed a considerable data fidelity loss after the model federation process, while no significant deterioration or improvement was found in the privacy metrics. The number of federated nodes did not show any significant trend, even though specific comparisons resulted in statistically relevant differences in some cases.

Acknowledgments

Funded by the European Union's Horizon Europe Research and Innovation Program under project SYNTHEMA with Grant Agreement no 101095530. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for

them.

Funded by the European Union

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

AI: Artificial intelligence
AML: Acute Myeloid Leukemia
AUC: Area under the curve
CS: Cosine similarity
CTGAN: Conditional tabular generative adversarial network
DLA: Data labelling analysis
DP: Differential privacy
ECOG: Eastern cooperative oncology group
EHR: Electronic health record
FL: Federated learning
GAN: Generative adversarial network
MDS: Myelodysplastic syndrome
MIA: Membership inference attack
ML: Machine learning
MNIST: Modified National Institute of Standards and Technology
PET: Privacy enhancing technology
RCT: Randomized controlled trial
SD: Synthetic data
VAE: Variational autoencoder
VS: Vendi score
WHO: World Health Organization

Data availability

The AML data that supports the conclusion of this study was taken from the original public repository in https://github.com/papaemmelab/Tazi_NatureC_AML with doi:10.5281/zenodo.6878209.

References

[1] A. Dozzo, A. Galvin, J.-W. Shin, S. Scalia, C. M. O'Driscoll, and K. B. Ryan, "Modelling acute myeloid leukemia (AML): What's new? A transition from the classical to the modern," *Drug Deliv. and Transl. Res.*, vol. 13, no. 8, pp. 2110–2141, Aug. 2023, doi: 10.1007/s13346-022-01189-4.

[2] A. Dhall, B. M. Zee, F. Yan, and M. A. Blanco, "Intersection of Epigenetic and Metabolic Regulation of Histone Modifications in Acute Myeloid Leukemia," *Front. Oncol.*, vol. 9, May 2019, doi: 10.3389/fonc.2019.00432.

[3] A. Vakiti and P. Mewawalla, "Acute Myeloid Leukemia," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2024. Accessed: Jan. 31, 2024. [Online]. Available:

http://www.ncbi.nlm.nih.gov/books/NBK507875/

[4] Y. Dong *et al.*, "Leukemia incidence trends at the global, regional, and national level between 1990 and 2017," *Exp Hematol Oncol*, vol. 9, p. 14, 2020, doi: 10.1186/s40164-020-00170-6.

[5] D. A. Arber *et al.*, "The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia," *Blood*, vol. 127, no. 20, pp. 2391–2405, May 2016, doi: 10.1182/blood-2016-03-643544.

[6] American Cancer Society, "Acute Myeloid Leukemia (AML) Subtypes and Prognostic Factors." Accessed: Jan. 31, 2024. [Online]. Available: https://www.cancer.org/cancer/types/acute-myeloid-leukemia/detection-diagnosis-staging/how-classified.html

[7] B. Benard, A. J. Gentles, T. Köhnke, R. Majeti, and D. Thomas, "Data mining for mutation-specific targets in acute myeloid leukemia," *Leukemia*, vol. 33, no. 4, pp. 826–843, Apr. 2019, doi: 10.1038/s41375-019-0387-y.

[8] A. Abhishek, R. K. Jha, R. Sinha, and K. Jha, "Automated classification of acute leukemia on a heterogeneous dataset using machine learning and deep learning techniques," *Biomedical Signal Processing and Control*, vol. 72, p. 103341, Feb. 2022, doi: 10.1016/j.bspc.2021.103341.

[9] K. Karami, M. Akbari, M.-T. Moradi, B. Soleymani, and H. Fallahi, "Survival prognostic factors in patients with acute myeloid leukemia using machine learning techniques," *PLoS One*, vol. 16, no. 7, p. e0254976, 2021, doi: 10.1371/journal.pone.0254976.

[10] S. Dong *et al.*, "Racial and Ethnic Disparities in Acute Myeloid Leukemia: 15-Year Experience at a Safety-Net Health System," *Blood*, vol. 140, no. Supplement 1, pp. 3194–3195, Nov. 2022, doi: 10.1182/blood-2022-165285.

[11] L. E. Winestone *et al.*, "Racial and ethnic disparities in acuity of presentation among children with newly diagnosed acute leukemia," *Pediatric Blood & Cancer*, vol. 71, no. 1, p. e30726, 2024, doi: 10.1002/pbc.30726.

[12] European Data Protection Supervisor, "Synthetic Data." [Online]. Available: https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data

[13] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic data generation for tabular health records: A systematic review," *Neurocomputing*, vol. 493, pp. 28–45, Jul. 2022, doi: 10.1016/j.neucom.2022.04.053.

[14] B. Van Breugel, Z. Qian, and M. Van Der Schaar, "Synthetic data, real errors: how (not) to publish and use synthetic data," in *Proceedings of the 40th International Conference on Machine Learning*, in ICML'23, vol. 202. Honolulu, Hawaii, USA: JMLR.org, Jul. 2023, pp. 34793–34808.

[15] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," Jan. 26, 2023, *arXiv*: arXiv:1602.05629. doi: 10.48550/arXiv.1602.05629.

[16] R. Shouval *et al.*, "Prediction of Hematopoietic Stem Cell Transplantation Related Mortality-Lessons Learned from the In-Silico Approach: A European Society for Blood and Marrow Transplantation Acute Leukemia Working Party Data Mining Study," *PLoS One*, vol. 11, no. 3, p. e0150637, 2016, doi: 10.1371/journal.pone.0150637.

[17] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review," *PLOS Digit Health*, vol. 2, no. 1, p. e0000082, Jan. 2023, doi: 10.1371/journal.pdig.0000082.

[18] I. J. Goodfellow *et al.*, "Generative Adversarial Networks," Jun. 10, 2014, *arXiv*: arXiv:1406.2661. doi: https://doi.org/10.48550/arXiv.1406.2661.

[19] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen, "CTAB-GAN+: Enhancing Tabular Data Synthesis," Apr. 01, 2022, *arXiv*: arXiv:2204.00401. doi: https://doi.org/10.48550/arXiv.2204.00401.

[20] P. A. Osorio-Marulanda, G. Epelde, M. Hernandez, I. Isasa, N. M. Reyes, and A. B. Iraola, "Privacy Mechanisms and Evaluation Metrics for Synthetic Data Generation: A Systematic Review," *IEEE Access*, vol. 12, pp. 88048–88074, 2024, doi: 10.1109/ACCESS.2024.3417608.
[21] R. Ramachandranpillai, M. F. Sikder, D. Bergström, and F. Heintz, "Bt-GAN: Generating Fair Synthetic Healthdata via Bias-transforming Generative Adversarial Networks," *Journal of Artificial Intelligence Research*, vol. 79, pp. 1313–1341, Apr. 2024, doi: 10.1613/jair.1.15317.
[22] I. Akiya, T. Ishihara, and K. Yamamoto, "Comparison of Synthetic Data Generation Techniques for Control Group Survival Data in Oncology Clinical Trials: Simulation Study," *JMIR Medical Informatics*, vol. 12, no. 1, p. e55118, Jun. 2024, doi: 10.2196/55118.
[23] S. A. Khan, H. Murtaza, and M. Ahmed, "Utility of GAN generated synthetic data for cardiovascular diseases mortality prediction: an experimental study," *Health Technol.*, vol. 14, no. 3, pp. 557–580, May 2024, doi: 10.1007/s12553-024-00847-6.
[24] K. Dhawan and S. S. Nijhawan, "Cross-Modality Synthetic Data Augmentation using GANs: Enhancing Brain MRI and Chest X-ray Classification," Jun. 10, 2024, *medRxiv*. doi: 10.1101/2024.06.09.24308649.
[25] V. C. Pezoulas *et al.*, "Synthetic data generation methods in healthcare: A review on open-source tools and methods," *Computational and Structural Biotechnology Journal*, vol. 23, pp. 2892–2910, Dec. 2024, doi: 10.1016/j.csbj.2024.07.005.
[26] A. Biswas and W. Talukdar, "Enhancing Clinical Documentation with Synthetic Data: Leveraging Generative Models for Improved Accuracy," *International Journal of Innovative Science and Research Technology (IJISRT)*, pp. 1553–1566, Jun. 2024, doi: 10.38124/ijisrt/IJISRT24MAY2085.
[27] H. Li, S. Yu, and J. Principe, "Causal Recurrent Variational Autoencoder for Medical Time Series Generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, Art. no. 7, Jun. 2023, doi: 10.1609/aaai.v37i7.26031.
[28] Raginee R. Titar and Murali Ramanathan, "Variational autoencoders for generative modeling of drug dosing determinants in renal, hepatic, metabolic, and cardiac disease states," *Clin Transl Sci.*, vol. 17, no. 7, Jul. 2024, doi: https://doi.org/10.1111/cts.13872.
[29] D. G. Saragih, A. Hibi, and P. N. Tyrrell, "Using diffusion models to generate synthetic labeled data for medical image segmentation," *Int J CARS*, Jun. 2024, doi: 10.1007/s11548-024-03213-z.
[30] Y. Yang *et al.*, "A Survey on Diffusion Models for Time Series and Spatio-Temporal Data," Jun. 11, 2024, *arXiv*: arXiv:2404.18886. doi: 10.48550/arXiv.2404.18886.
[31] A. A. Naseer *et al.*, "ScoEHR: Generating Synthetic Electronic Health Records using Continuous-time Diffusion Models," in *Proceedings of the 8th Machine Learning for Healthcare Conference*, PMLR, Dec. 2023, pp. 489–508. Accessed: Jul. 30, 2024. [Online]. Available: https://proceedings.mlr.press/v219/naseer23a.html
[32] M. Pozzi *et al.*, "Generating synthetic data in digital pathology through diffusion models: a multifaceted approach to evaluation," Nov. 22, 2023, *medRxiv*. doi: 10.1101/2023.11.21.23298808.
[33] S. D'Amico *et al.*, "Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology," *JCO Clin Cancer Inform*, vol. 7, p. e2300021, Jun. 2023, doi: 10.1200/CCI.23.00021.
[34] J.-N. Eckardt *et al.*, "Mimicking clinical trials with synthetic acute myeloid leukemia patients using generative artificial intelligence," *npj Digit. Med.*, vol. 7, no. 1, pp. 1–11, Mar. 2024, doi: 10.1038/s41746-024-01076-x.
[35] R. Licandro *et al.*, "WGAN Latent Space Embeddings for Blast Identification in Childhood Acute Myeloid Leukaemia," in *2018 24th International Conference on Pattern Recognition (ICPR)*, Aug. 2018, pp. 3868–3873. doi: 10.1109/ICPR.2018.8546177.
[36] V. Rupapara, F. Rustam, W. Aljedaani, H. F. Shahzad, E. Lee, and I. Ashraf, "Blood cancer

prediction using leukemia microarray gene data and hybrid logistic vector trees model," *Sci Rep*, vol. 12, no. 1, p. 1000, Jan. 2022, doi: 10.1038/s41598-022-04835-6.

[37] 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)*, Jun. 2008, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.

[38] A. Linardos, K. Kushibar, S. Walsh, P. Gkontra, and K. Lekadir, "Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease," *Sci Rep*, vol. 12, no. 1, p. 3551, Mar. 2022, doi: 10.1038/s41598-022-07186-4.

[39] D. Liu, T. Miller, R. Sayeed, and K. D. Mandl, "FADL:Federated-Autonomous Deep Learning for Distributed Electronic Health Record," Dec. 02, 2018, *arXiv*: arXiv:1811.11400. doi: 10.48550/arXiv.1811.11400.

[40] Z. Azizi *et al.*, "A comparison of synthetic data generation and federated analysis for enabling international evaluations of cardiovascular health," *Sci Rep*, vol. 13, no. 1, p. 11540, Jul. 2023, doi: 10.1038/s41598-023-38457-3.

[41] M. R. Behera, S. Upadhyay, S. Shetty, S. Priyadarshini, P. Patel, and K. F. Lee, "FedSyn: Synthetic Data Generation using Federated Learning," Apr. 05, 2022, *arXiv*: arXiv:2203.05931. doi: 10.48550/arXiv.2203.05931.

[42] B. Xin *et al.*, "Federated synthetic data generation with differential privacy," *Neurocomputing*, vol. 468, pp. 1–10, Jan. 2022, doi: 10.1016/j.neucom.2021.10.027.

[43] Y. Tazi *et al.*, "Unified classification and risk-stratification in Acute Myeloid Leukemia," *Nat Commun*, vol. 13, no. 1, p. 4622, Aug. 2022, doi: 10.1038/s41467-022-32103-8.

[44] Y. Tazi, *yanistazi/Tazi_NatureC_AML: Nature Paper release*. (Jul. 21, 2022). Zenodo. doi: 10.5281/zenodo.6878209.

[45] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular data using Conditional GAN," in *Advances in Neural Information Processing Systems*, Vancouver, Canada: Curran Associates, Inc., 2019. Accessed: Jan. 31, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

[46] A. Kiran and S. S. Kumar, "A Methodology and an Empirical Analysis to Determine the Most Suitable Synthetic Data Generator," *IEEE Access*, vol. 12, pp. 12209–12228, 2024, doi: 10.1109/ACCESS.2024.3354277.

[47] E. Fössing and J. Drechsler, "An Evaluation of Synthetic Data Generators Implemented in the Python Library Synthcity," *Privacy in Statistical Databases*, vol. 14915, pp. 178–193, 2024, doi: https://doi.org/10.1007/978-3-031-69651-0_12.

[48] N. Patki, R. Wedge, and K. Veeramachaneni, "The Synthetic Data Vault," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Montreal, QC, Canada: IEEE, Oct. 2016, pp. 399–410. doi: 10.1109/DSAA.2016.49.

[49] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion Models in Vision: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023, doi: 10.1109/TPAMI.2023.3261988.

[50] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated Learning on Non-IID Data Silos: An Experimental Study," Oct. 28, 2021, *arXiv*: arXiv:2102.02079. doi: 10.48550/arXiv.2102.02079.

[51] D. J. Beutel *et al.*, "Flower: A Friendly Federated Learning Research Framework," Mar. 05, 2022, *arXiv*: arXiv:2007.14390. Accessed: Mar. 07, 2024. [Online]. Available: http://arxiv.org/abs/2007.14390

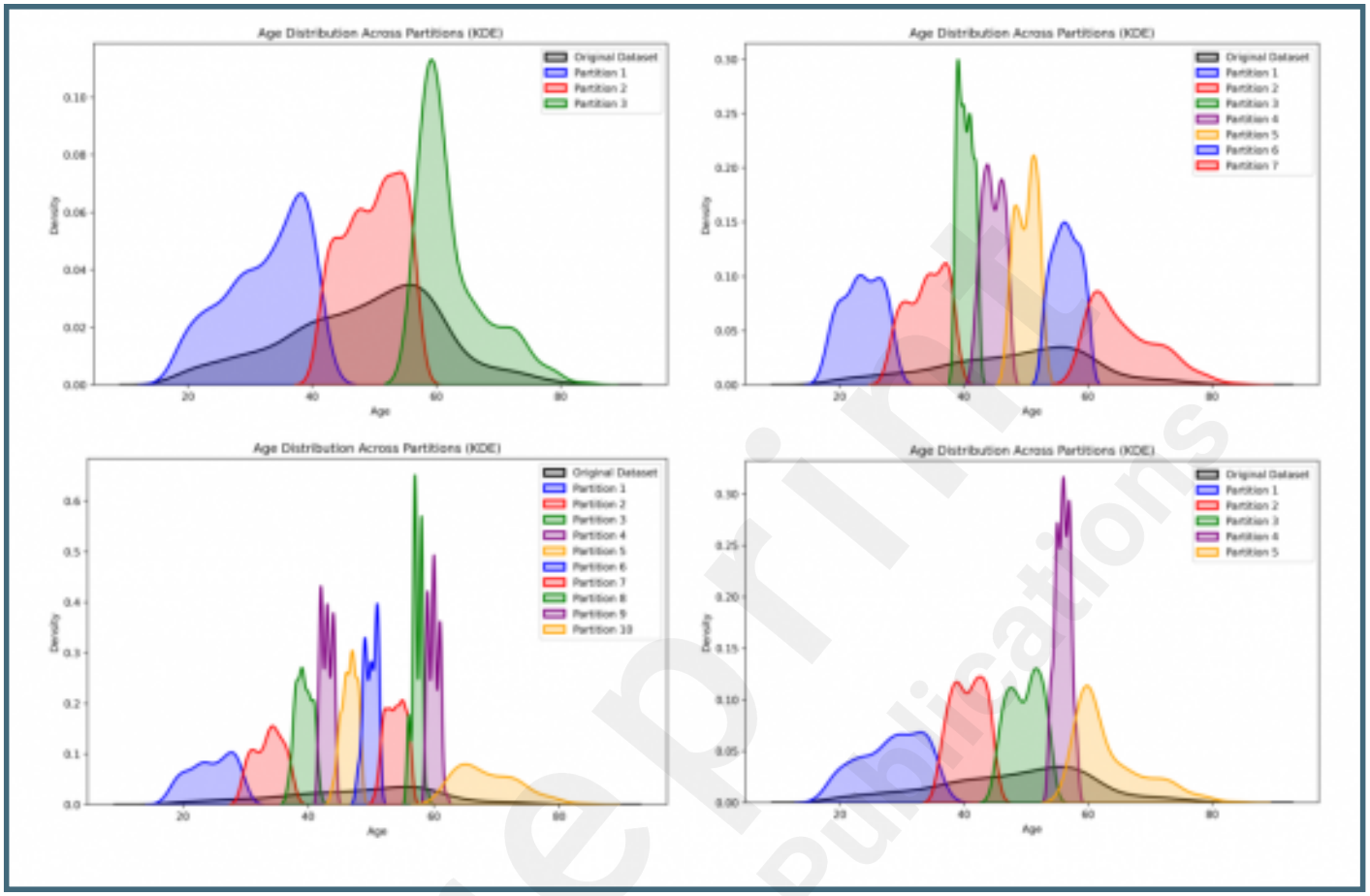
[52] M. Baak, R. Koopman, H. Snoek, and S. Klous, "A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics," *Computational Statistics & Data Analysis*, vol. 152, p. 107043, Dec. 2020, doi: 10.1016/j.csda.2020.107043.

[53] D. Friedman and A. B. Dieng, “The Vendi Score: A Diversity Evaluation Metric for Machine Learning,” Jul. 02, 2023, *arXiv*: arXiv:2210.02410. doi: 10.48550/arXiv.2210.02410.
[54] 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 Inf Med*, Jan. 2023, doi: 10.1055/s-0042-1760247.
[55] Shankar R. Pandala, “LazyPredict,” GitHub. Accessed: Sep. 04, 2024. [Online]. Available: https://github.com/shankarpandala/lazypredict
[56] B. Xin, W. Yang, Y. Geng, S. Chen, S. Wang, and L. Huang, “Private FL-GAN: Differential Privacy Synthetic Data Generation Based on Federated Learning,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 2927–2931. doi: 10.1109/ICASSP40776.2020.9054559.
[57] N. Rodríguez-Barroso *et al.*, “Federated Learning and Differential Privacy: Software tools analysis, the Sherpa.ai FL framework and methodological guidelines for preserving data privacy,” *Information Fusion*, vol. 64, pp. 270–292, Dec. 2020, doi: 10.1016/j.inffus.2020.07.009.
[58] P. A. Apellániz, J. Parras, and S. Zazo, “Improving Synthetic Data Generation Through Federated Learning in Scarce and Heterogeneous Data Scenarios,” *Big Data and Cognitive Computing*, vol. 9, no. 2, Art. no. 2, Feb. 2025, doi: 10.3390/bdcc9020018.

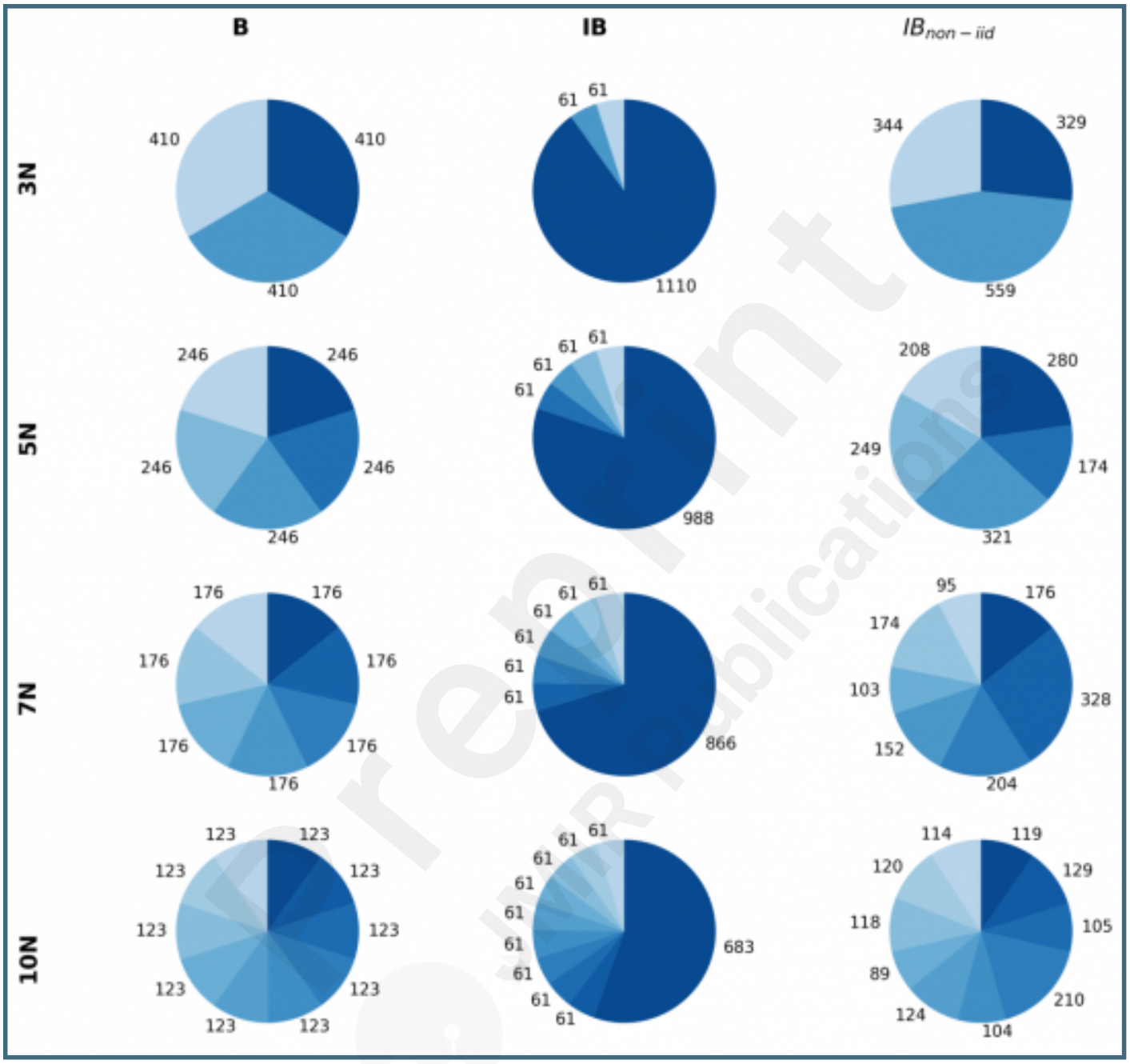
Supplementary Files

Figures

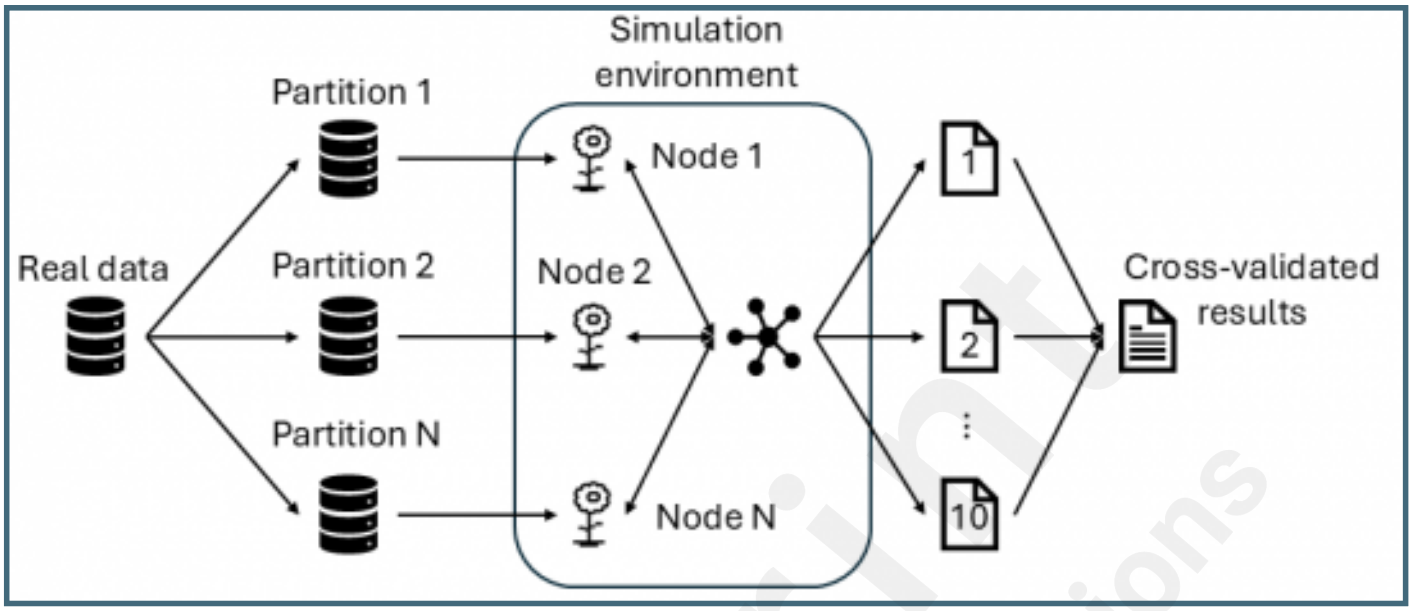
Generated non-IID age-dependant distribution plots. Top left shows the scenario with 3 nodes, top right shows the 7-node scenario, bottom left shows the 10-node scenario, and bottom right shows the 5-node one.



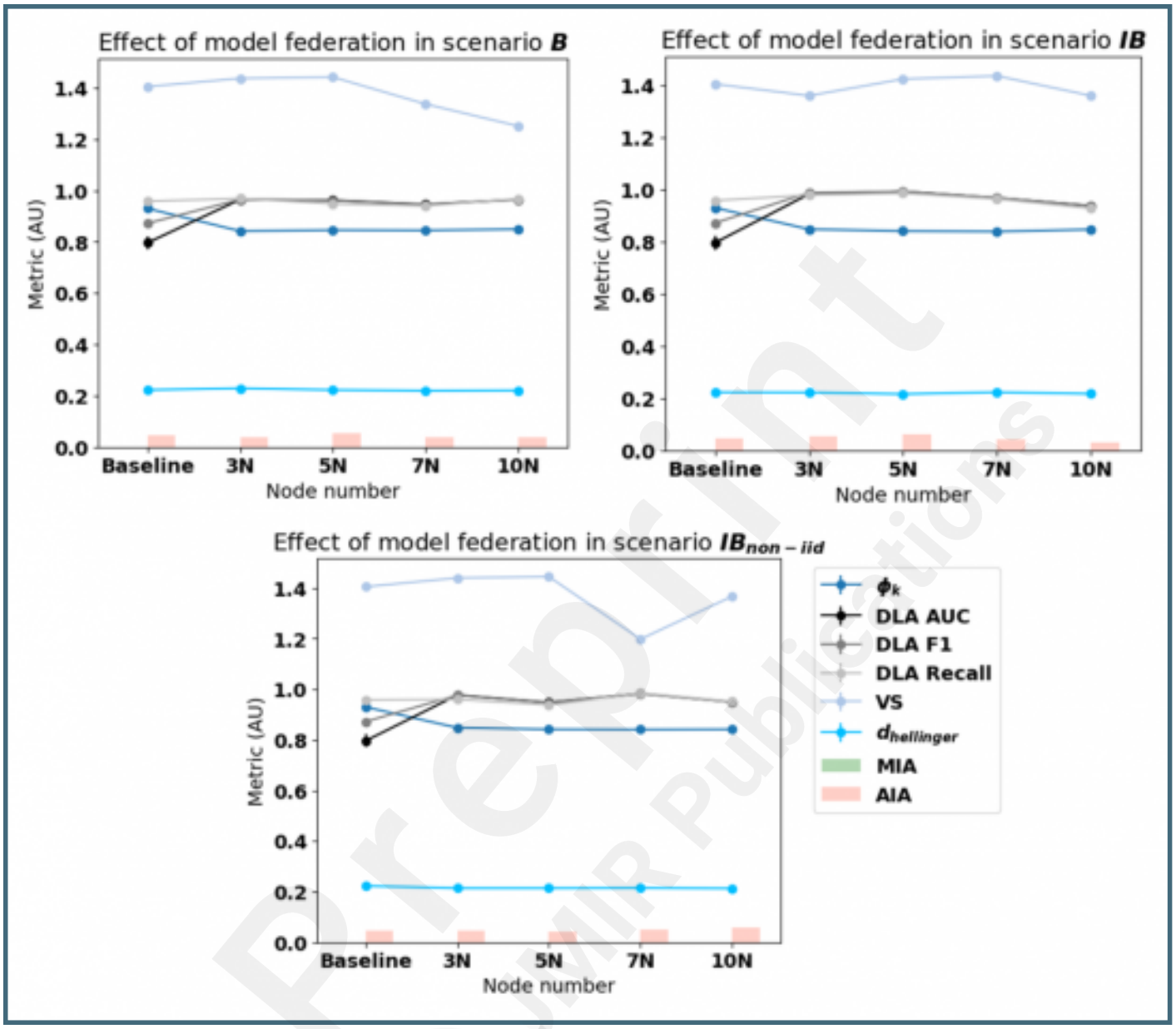
Number of samples used for each experiment. The plots are divided scenario-wise in the vertical axis and node quantity-wise in the horizontal axis.



Experiment execution framework, including data partitioning, model training and SD evaluation processes.



Graphical comparison of federated scenarios regarding fidelity and privacy metrics.



Multimedia Appendixes

Variable-wise metric results (Hellinger distance and AIA).
URL: http://asset.jmir.pub/assets/3888860eef5cb4793e1f09bebbcc5178.pdf