Poster: Federated Multimodal Medical Data Generation

Hajar Homayouni
Dept. of Computer Science
San Diego State University
hhomayouni@sdsu.edu

Maryam Mary Pourebadi
Dept. of Computer Science and Engineering
University of California, San Diego
pourebadi@ucsd.edu

Hossein Shirazi
Dept. of Management Information Systems
San Diego State University
hshirazi@sdsu.edu

***Abstract*—Deep learning in healthcare faces challenges such as limited annotated Electronic Health Records (EHR), data privacy concerns, and imbalanced datasets. Most synthetic data methods focus on single data types without privacy safeguards. To overcome these limitations, the proposed Federated Privacy-preserving Multimodal Generative (FPMG) framework integrates federated learning principles for decentralized training without direct data access. It proposes a multimodal generative adversarial model to create comprehensive synthetic data while preserving privacy. Silo Federated Learning is applied for decentralized training, maintaining patient privacy. Multimodal generative systems capture complex associations between health features, enhancing synthetic data quality and disease understanding. Preliminary research successfully generated synthetic images, including rare co-occurrences like COVID-19 and lung cancer.**

I. Introduction

Addressing the pivotal challenge of training generalizable deep learning models in the medical domain for accurate results requires a large volume of EHR with symmetric distributions. However, this data is scarce in medical applications due to legal and ethical constraints and the high costs associated with data annotation. Moreover, certain individuals being more heavily represented can lead to skewed and inaccurate outcomes that can disproportionately impact minority groups, emphasizing the need for balanced data in healthcare analysis.

Various analysis tasks require accessing a mixture of data types (i.e., modalities) to make precise decisions. For example, an automated diagnosis system requires a combination of information on patient health status, such as lab results, Oxygen level, and medical images, to make the right decisions. A grand challenge in multimodal data generation is capturing associations among features and generating high-quality data that adheres to these associations. For example, in generating mock patient profiles, the new data must accurately represent associations between visual features from images (e.g., skin color, clothes, height, and weight) and personal features from tabular data (e.g., age, gender, and occupation).

The existing synthetic data generation methods can replicate bias in the training data they learn from. Additionally, the gradients of deep networks can disclose private information in training, potentially allowing the reconstruction of the original data [1]. The centralization of patient data from diverse sources can elevate the risk of data breaches or unauthorized access, often clashing with data protection regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) or the General Data Protection Regulation (GDPR) [2]. Moreover, the process necessitates substantial data transfers between healthcare institutions and central repositories, introducing bandwidth, latency, and cost challenges. The inherent data heterogeneity across institutions can also lead to poorly generalized models across various demographics or medical conditions. Lastly, centralizing databases can amplify biases in the original datasets, resulting in skewed models with potential ramifications on clinical decision-making [3].

To address these gaps, adaptive Federated Learning (FL) emerges as a solution by preserving patient privacy and avoiding centralized data risks [4], [5]. Silo Federated Learning, a specialized branch of FL, emphasizes training within distinct healthcare organizational silos, such as hospitals, clinics, or research centers. This ensures that patient data remains within its originating environment, eliminating the need for direct data transfer [5]. Silo Federated Learning addresses the pivotal concerns about data privacy and regularity compliance [6]. Compliance with regulations like HIPAA and GDPR is imperative, necessitating the protection and localization of patient data. Furthermore, training models within their respective healthcare silos allows for harnessing the heterogeneity of patient data across institutions, potentially resulting in more robust and accurate diagnostic or predictive tools. This decentralized approach can also significantly reduce data transfer costs and latency, offering a more efficient and sustainable model-training paradigm.

To tackle these challenges, our research aims to generate balanced, correlated types of EHR data for precise and fair training of data-driven models. Our proposed FPMG framework, based on a novel differentially private deep generative adversarial model, enables multiple parties to contribute privately. This allows the discovery of complex associations among multi-modal features, and generates comprehensive balanced datasets while addressing privacy concerns. The objectives include (1) unbiased multimodal EHR data generation for precise decision-making and (2) decentralized privacy-preserving data augmentation by adapting cross-silo federate learning. To evaluate reliability and effectiveness, we aim to use real-world data accumulated from our collaborators at Anschutz Medical Campus at CU Denver and SDSU HealthLINK Center. The goal is to showcase the generated data's quality, diversity, and correlation, as well as assess the model's privacy preservation and effectiveness in training reliable models.

II. Unbiased Multimodal Data Generation

We take three stages to develop the FPMG in a way that its operations can apply to a mixture of different data types.

Stage 1. Encode real data into 1D representation vectors. The project employs representation learning to convert real inputs into continuous vectors $\{v_r1, \ldots, v_rm\}$, making the network differentiable across data types. Principal Component Analysis and autoencoders enable the system to encode numerical data, while transformer autoencoders handle multivariate time series and textual data, and Convolutional Neural Network transformers image and video data.

Stage 2. Synthesis of Multimodal Representation Vectors. The FPMG framework integrates a generative adversarial architecture composed of generator networks (G) and a discriminator (D) to produce encoded continuous synthetic vectors for diverse data types. The FPMG framework leverages joint deep

representations across various data types and adopts parallel conditional generators to address imbalanced datasets. These generators condition their outputs on semantic inputs, encapsulating expected class labels, which are then incorporated into D and G networks for enriched data generation and classification.

Stage 3. Decode the synthetic vectors into realistic representations. The FPMG framework then decodes the generated 1D representation vectors into their corresponding realistic representations. The framework employs the decoder components of the m autoencoder models trained in stage 1, each tailored to handle one of the m distinct data types. For instance, real medical images paired with patients' lab results may be used as input data, and the generated data could manifest as simulated images of patients with corresponding lab outcomes.

Evaluation Plan. Our evaluation plan includes the Reality Guarantee (RG), a metric assessing the closeness of generated data to genuine datasets and the consistent representation of associations among multimodal features [7]. Utilizing Euclidean distance, lower distances between shared representation vectors and the first subset of actual data indicate superior data realism. Additionally, the Diversity Guarantee (DG) measures the diversity of produced data by employing a similar Euclidean distance methodology with a second subset of real data [8]. Comparing the performance of a model trained on original data with one trained on synthetic data, smaller performance gaps signify the high quality of the synthetic dataset.

III. Decentralized Data Generation

To enable collaboration among multiple silos within a decentralized network, where each silo can only communicate with its neighboring silos, we aim to develop a decentralized federated learning algorithm for the FPMG framework using a ring-topology network in two stages. This approach enhances model effectiveness through weight aggregation, ultimately improving the accuracy of the global model.

Stage 1. Protocol Setup. We aim to establish a ring topology, interconnecting each silo in a circular manner, with each silo assigned two neighbors—one preceding and one succeeding—to inherently promote decentralized communication. In the Weight Exchange Protocol, we initiate a weight exchange mechanism based on the ring topology. Each silo sends its model weights only to its immediate neighbor, and this sequential transfer continues around the ring until each silo has both received and relayed weights once.

Stage 2. Training Strategy. Each silo initiates the training of its local model using its dataset by utilizing specific hyperparameters tailored to the silo's data characteristics. At every weight reception, each silo uses our proposed robust weight aggregation method to merge its local model weights with the received weights. This process iteratively refines the model, leveraging insights from neighboring silos. Once the model converges or reaches a specified number of iterations, we will assess its performance on various machine learning benchmarks. We aim to retrain or adjust the algorithm based on evaluation outcomes as necessary.

Evaluation Plan. Our evaluation plan includes introducing the Privacy Guarantee (PG) metric to assess the model's privacy measures. Concurrently, the Federated Consistency (FC) metric evaluates the model weights' consistency across nodes after each federated learning round with reduced variance, highlighting a unified federated process [9]. Additionally, the Communication Overhead (CO) metric, fundamental in federated setups, gauges communication expenses per learning round, where decreased values signify efficiency and adaptability [10]. Post-federated rounds, the Aggregated Model Performance (AMP) metric compares the aggregated model's proficiency to a centralized benchmark, with smaller disparities suggesting system robustness [11].

IV. Preliminary Results

HRCX framework [12] was introduced, combining adversarial networks and predictive learning for balanced, diverse, and high-resolution COVID-19 images. The framework also establishes a COVID-19 severity index aiding illness prediction. Over 3000 high-quality, diverse COVID-19 X-ray images were generated, and their severity scores were effectively demonstrated for both normal and infected cases, enhancing diagnostic capabilities. The research group is delving into the generation of rare occurrences, particularly images of patients with both COVID-19 and lung cancer. High-quality CT-Scan images were successfully generated and validated by certified radiologists, revealing predominant cancer features and reduced COVID-19 signs due to sequential training. Transformer-based models are being explored to generate distinctive CXR image and text representations, leveraging the expert reasoning embedded in radiologist reports.

V. Conclusions

The outcomes of this proposal will yield an innovative technical roadmap for unbiased EHR data generation, establishing a robust foundation for responsible data practices in the equitable training and validation of learning models. Moreover, this research contributes to advancing data science by developing models capable of leveraging a mixture of correlated data types, enhancing their performance in prediction and decision-making tasks.

References

[1] F. Mo *et al.*, "Quantifying and localizing private information leakage from neural network gradients," *arXiv e-prints*, pp. arXiv–2105, 2021.

[2] F. K. Dankar *et al.*, "The dilemma of patient's privacy versus public interest in health data," *Health Informatics Journal*, pp. 999–1011, 2019.

[3] Z. Obermeyer *et al.*, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, pp. 447–453, 2019.

[4] H. B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of AISTATS*, 2017.

[5] T. S. Brisimi *et al.*, "Federated learning of predictive models from federated electronic health records," *International journal of medical informatics*, vol. 112, pp. 59–67, 2018.

[6] N. Rieke *et al.*, "The future of digital health with federated learning," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[7] I. Goodfellow *et al.*, *Deep Learning*. MIT Press, 2016.

[8] H. Zhang *et al.*, "The limitations of deep learning in adversarial settings," *IEEE Access*, vol. 7, pp. 100384–100394, 2019.

[9] P. Kairouz and H. B. e. a. McMahan, "Advances and open problems in federated learning," 2019.

[10] K. Bonawitz *et al.*, "Towards federated learning at scale: System design," in *Proceedings of the 2nd SysML Conference*, 2019.

[11] V. Smith *et al.*, "Federated multi-task learning," *Advances in Neural Information Processing Systems*, pp. 4424–4434, 2017.

[12] S. Kaur *et al.*, "Synthetic high-resolution covid-19 chest x-ray generation," in *Proceedings of the ACSW*, p. 151–159, 2023.

Federated Multimodal Medical Data Generation

[1]Hajar Homayouni, [2]Maryam Mary Pourebadi, [1]Hossein Shirazi
[1]San Diego State University, [2]University of California San Diego

Introduction

- Deep learning integration in healthcare offers significant improvements in diagnostics, treatment, and patient care, reshaping healthcare delivery and research.
- Challenges include the need for diverse and annotated Electronic Health Records (EHRs), data availability, privacy, and ethical concerns, particularly regarding data biases and patient representation.
- This research aims to develop the **F**ederated **P**rivacy-preserving **M**ultimodal **G**enerative (**FPMG**) framework for generating synthetic, balanced, diverse, and privacy-conscious EHR data, addressing data generation challenges.
- FPMG utilizes federated learning and differential privacy, ensuring data privacy and security in a decentralized approach. It represents a significant advancement in ethical AI for equitable healthcare outcomes.

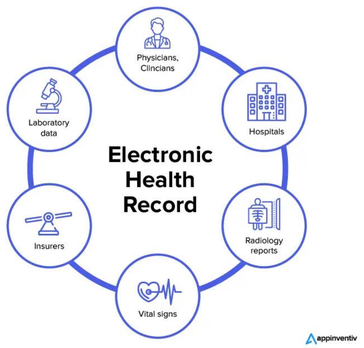
Limitations of Existing Approaches

- Existing synthetic data generation methods focus on single data types and can replicate bias.
- Deep networks' gradients may reveal private information in training data.
- Centralized learning in healthcare AI raises privacy, security, and compliance concerns.

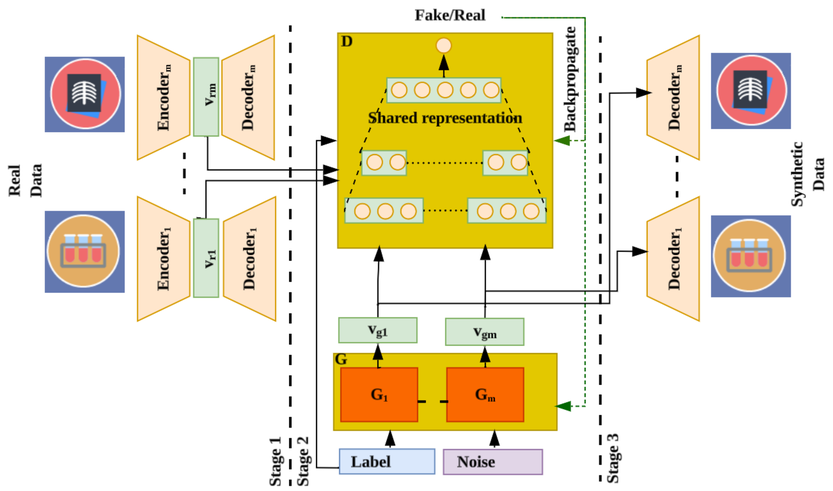
Research Goals

1. **Enhance Data Generalizability and Reduce Bias in Medical Data Analysis:** Address the challenge of insufficient and biased data in healthcare.

2. **Facilitate Collaborative and Privacy-preserving Research in Healthcare AI:** Enable collaborative development of AI models across healthcare institutions without compromising patient data privacy.



FPMG Framework



Approach

1. **Generative Multimodal Model:** Develop a novel generative model using advanced generative adversarial networks (GANs) that integrate various modalities of EHR data to produce realistic, balanced, and diverse synthetic data. Ensure differential privacy to protect patient confidentiality and retain statistical properties.
2. **Federated Learning:** Establish a federated learning framework allowing multiple healthcare entities to contribute to a shared generative AI model while keeping data decentralized and preserving privacy. Implement cutting-edge algorithms to optimize model performance while complying with legal and ethical standards (e.g., HIPAA, GDPR).

Evaluation

Goals

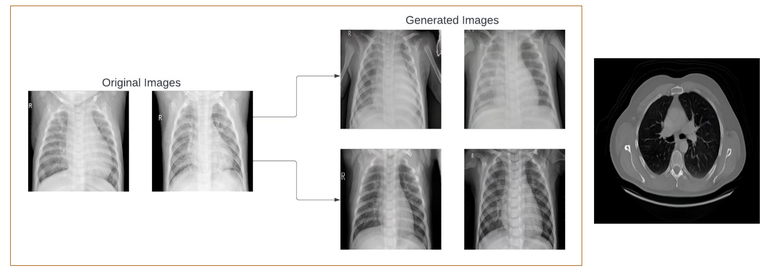
- Generated data should strike a balance between visual quality and diversity.
- Proposed model must not reveal any sensitive information about the original data.

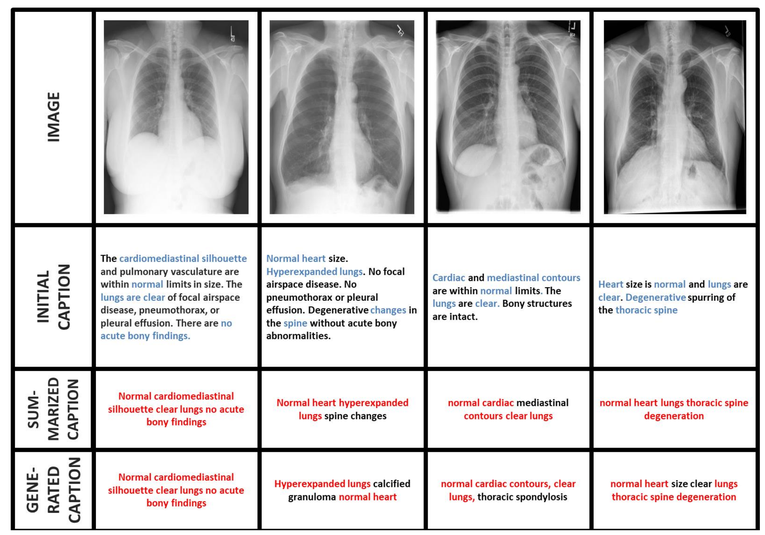
Metrics

- **Reality Guarantee (RG):** Measures data similarity to genuine datasets.
- **Diversity Guarantee (DG):** Measures diversity in generated data compared to genuine data.
- **Privacy Guarantee (PG):** Quantifies privacy preservation in the trained model.
- **Federated Consistency (FC):** Assesses model weight consistency across different nodes in federated learning.

Preliminary Results

- Generation of high-resolution COVID-19 images, combining adversarial networks and predictive learning.
- Generation of rare occurrences, generating images of patients with both COVID-19 and lung cancer using CT-Scans.
- Creation of text representations resembling radiologist reports for CXR images, leveraging Transformer-based models to enhance diagnostic capabilities.





Conclusions

- The research aims to create an innovative technical roadmap for generating unbiased EHR data, ensuring responsible data practices in training and validating learning models in the medical field.
- This research will contribute to data science by developing models capable of utilizing a combination of correlated data types, leading to improved performance in critical prediction and decision-making tasks in healthcare.

Acknowledgement

This project is supported by Microsoft Research AFMR Minority Serving Institutions grant.