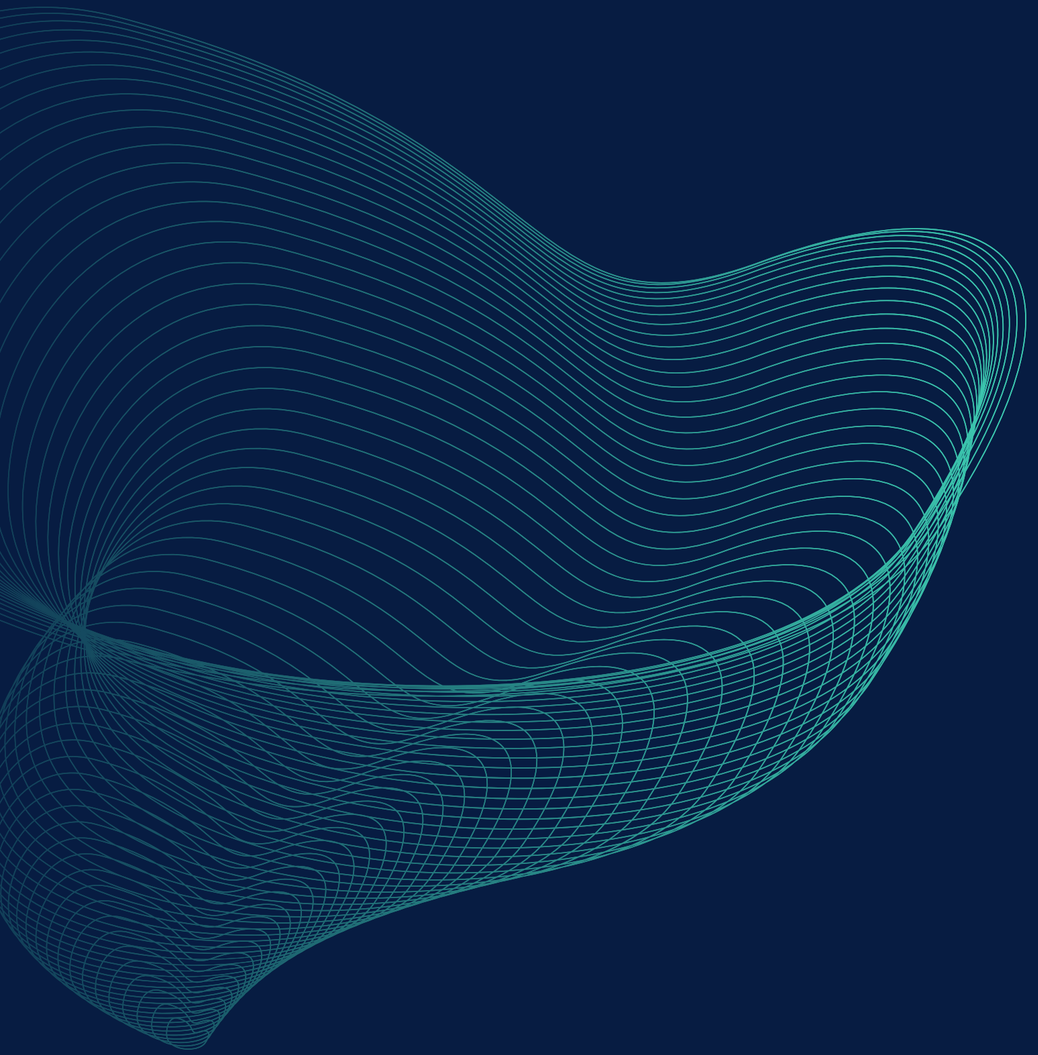
SJSU SAN JOSÉ STATE UNIVERSITY

Presented by **Kelly Nguyen**

Revolutionizing Patient Data Privacy:

Federated Learning & Clustering in Healthcare

Deep dive into “Privacy-preserving patient clustering for personalized federated learning” written by Ahmed Elhussein and Gamze Gursoy and presented at the Machine Learning for Healthcare 2023 conference



Overview

1. Research
2. Introduction to Healthcare Data Privacy
3. Deep Learning in Healthcare
4. Challenges in Patient Data Clustering
5. Federated Learning as a Privacy Beacon
6. Innovations of PCBFL
7. Methodology
8. Results and Significance
9. Conclusion
10. Personal Insights

Research

- Paper: **Privacy-preserving patient clustering for personalized federated learning**
- Written by Ahmed Elhussein and Gamze Gursoy
- Presented at Machine Learning for Healthcare 2023
- Focus on non-IID data and privacy in federated learning

Proceedings of Machine Learning Research 219:1–23, 2023 Machine Learning for Healthcare 2023

Privacy-preserving patient clustering for personalized federated learning

Ahmed Elhussein AE2722@CUMC.COLUMBIA.EDU
Department of Biomedical Informatics, Columbia University
New York City, NY, U.S.A

Gamze Gürsoy GAMZE.GURSOY@COLUMBIA.EDU
Department of Biomedical Informatics, Columbia University
New York Genome Center
New York City, NY, U.S.A

Abstract

Federated Learning (FL) is a machine learning framework that enables multiple organizations to train a model without sharing their data with a central server. However, it experiences significant performance degradation if the data is non-identically independently distributed (non-IID). This is a problem in medical settings, where variations in the patient population contribute significantly to distribution differences across hospitals. Personalized FL addresses this issue by accounting for site-specific distribution differences. Clustered FL, a Personalized FL variant, was used to address this problem by clustering patients into groups across hospitals and training separate models on each group. However, privacy concerns remained as a challenge as the clustering process requires exchange of patient-level information. This was previously solved by forming clusters using aggregated data, which led to inaccurate groups and performance degradation. In this study, we propose Privacy-preserving Community-Based Federated machine Learning (PCBFL), a novel Clustered FL framework that can cluster patients using patient-level data while protecting privacy. PCBFL uses Secure Multiparty Computation, a cryptographic technique, to securely calculate patient-level similarity scores across hospitals. We then evaluate PCBFL by training a federated mortality prediction model using 20 sites from the eICU dataset. We compare the performance gain from PCBFL against traditional and existing Clustered FL frameworks. Our results show that PCBFL successfully forms clinically meaningful cohorts of low, medium, and high-risk patients. PCBFL outperforms traditional and existing Clustered FL frameworks with an average AUC improvement of 4.3% and AUPRC improvement of 7.8%.

Introduction to Healthcare Data Privacy

In the realm of healthcare, the sanctity of patient data is paramount

- With the advent of digital record-keeping, the potential for sensitive information to be compromised has escalated, necessitating robust solutions to ensure privacy.
- The critical need for safeguarding healthcare data cannot be overstated
  - not just a matter of legal compliance, but of maintaining the trust at the very core of the patient-provider relationship.

Deep Learning in Healthcare

Enhances disease risk prediction, diagnostic support

Relies on large EHR datasets

Limitations

overfitting and poor generalization in small datasets

Challenges in Patient Data Clustering

1

Necessity for personalized FL models

2

Risks of traditional clustering methods

3

Privacy conflicts with data sharing

Federated Learning as a Privacy Beacon

A groundbreaking technique that promises to revolutionize the way medical data is utilized for research and analysis

- Approach allows for the development of powerful, predictive models by learning from decentralized datasets, without ever compromising the privacy of the individual's data.
- By processing data locally and sharing only model updates instead of raw information, federated learning provides a blueprint for harnessing the collective power of healthcare data while upholding the inviolable principle of patient privacy.

Innovations of PCBFL

Privacy-preserving Community-Based Federated Learning (PCBFL)

Utilizes Secure Multiparty Computation (SMPC) to enable the clustering of patient data without exposing individual patient information

Methodology

Cohort and Feature Extraction

Privacy Preserving CBFL Protocol

Model Training

Evaluation

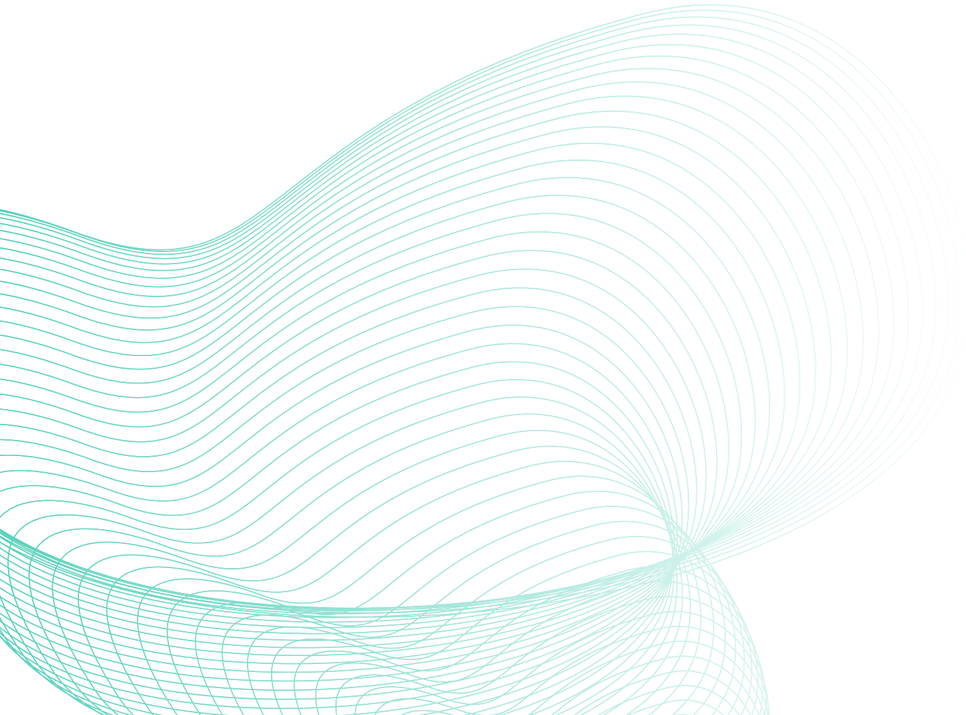
Cohort and Feature Extraction

In order to create a more realistic FL scenario, a subsampling approach was used, where the size of the dataset in each site was limited.

The study utilized the eICU collaborative research database, comprising data from over 200,000 patients across the U.S., to predict ICU mortality using variables like diagnosis, medications, and physical exams within the first 48 hours of admission.

Focused primarily on patients with complete data to avoid imputation bias, leading to a cohort of 5,000 patients from 20 sites, aligning with a realistic federated learning scenario with limited site data.

PCBFL Protocol



Creating Patient Embeddings

A federated autoencoder is trained to obtain latent variables for each feature domain.

Estimating Patient Similarity Securely

SMPC uses a secret sharing scheme to jointly calculate the dot product between pairs of vectors

Clustering Patients

Spectral clustering to cluster the patients using similarity matrix generated from pairwise cosine similarities of embeddings

Predicting Mortality

Cluster-based FL training. Each model is separately trained per cluster.

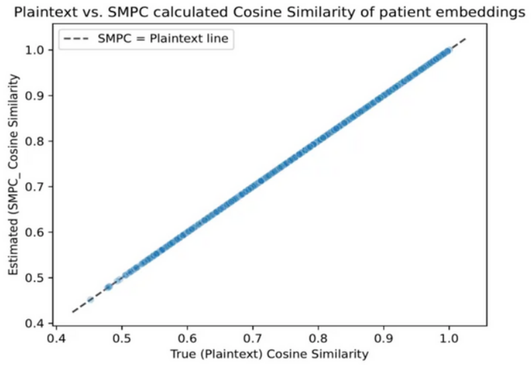
Model Training

Feed Forward models were implemented with ReLU activation in the hidden layers and sigmoid activation in the final output layers.

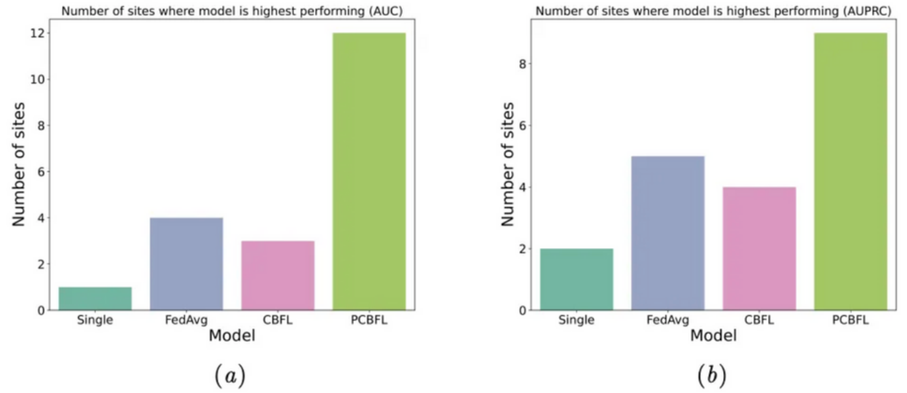
Evaluation

Analyzed patient clusters formed by Privacy-preserving Community-Based Federated Learning (PCBFL) and Community-Based Federated Learning (CBFL), assessing mortality rates and feature distributions

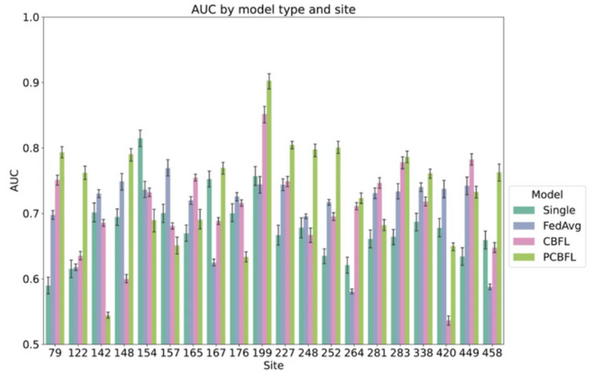
- The data was divided into training and testing sets at a 70:30 ratio.
- Given the imbalance in the dataset with only 20% positive labels, performance was evaluated using both AUC and AUPRC scores.
- The models were run 100 times to calculate mean scores and determine 95% confidence intervals with bootstrap estimation.



Comparison of cosine similarity scores



Number of sites where model has highest AUC (a) and AUPRC (b) for Single, Centralized, FedAvg, CBFL and PCBFL.



Model performance by site, AUC. Results for Single, FedAvg, CBFL and PCBFL.

Results

Things that PCBFL provided includes privacy-preserving and accurate patient similarity scores using secure multiparty computation, clinically meaningful clusters, and an increase in predictive performance.

Conclusion

- All in all, the paper presents a personalized federated learning (FL) framework that integrates data mining techniques like clustering to address data heterogeneity and privacy in healthcare.
- The study overall underscores PCBFL’s promise for collaborative healthcare data analysis which enhances patient privacy and future research aims to assess its generalizability across different healthcare datasets and domains

Personal Insights

- I found this research very interesting because not only does it explore traditional deep mining tactics such as clustering, but it goes into advanced modernized solutions such as federated learning.
- With personal experience working in biotech, the rise of using data mining and deep learning in healthcare is not only revolutionary, but also extraordinary in improving the quality of healthcare itself.
- I look forward to reading more research and articles on just that.

SJSU SAN JOSÉ STATE UNIVERSITY

Thank You

For further questions, feel free to email me at
kelly.nguyen01@sjsu.edu