Special collection: Privacy-preserving artificial intelligence in digital-healthcare

DIGITAL HEALTH

DIGITAL HEALTH
Volume 11: 1–14

DOI: 10.1177/20552076251360861
journals.sagepub.com/home/dhj

Sage

A personalized federated learning approach to enhance joint modeling for heterogeneous medical institutions

Hong Ye[1,2,*], Xiangzhou Zhang[1,3,*], Kang Liu[1,2], Ziyuan Liu[1,4], Weiqi Chen[5], Bo Liu[1,2], Eric WT Ngai[6] and Yong Hu[1,3]

Abstract

Background: Federated Learning (FL) offers a privacy-preserving solution for multi-party data collaboration in smart healthcare. However, the data heterogeneity among hospitals and among patients often results in suboptimal performance for some hospitals when applying a global FL model. Current clustering-based FL methods struggle to adapt to complex and diverse data distributions, negatively impacting model performance.

Methods: We propose a novel framework, Federated Gaussian Mixture Clustering (FedGMC), which leverages Gaussian Mixture Clustering to train personalized FL models. FedGMC determines the optimal number of clusters prior to the FL process, reducing the time and computational cost associated with traversing multiple clustering configurations in existing approaches.

Results: The FedGMC framework was evaluated using real-world eICU datasets with various classifiers and performance metrics. Experimental results show that FedGMC outperforms other baseline methods in terms of the overall performance of combining two classifiers and two performance metrics. Moreover, it mitigates the risk of performance degraded for participating hospitals following FL.

Conclusions: The FedGMC framework effectively addresses clinical heterogeneity, enhancing predictive performance and ensuring fairness among participating medical institutions. These improvements increase the willingness of data owners to engage in the collaboration FL initiatives.

Keywords

Federated learning, non-IID data, federated clustering, electronic health records, acute kidney injury, personalized federated learning

Received: 12 February 2025; accepted: 7 July 2025

Introduction

Applying machine learning to electronic health records (EHRs) has achieved remarkable outcomes in various medical and healthcare applications, furnishing substantial support for clinical decision-making.[1,2] However, accurate machine learning modeling often depends on large-scale data samples, which are typically distributed among diverse hospitals and institutions. EHR data contains extensive patient-related information, making its use subject to stringent privacy protection requirements.[3,4] Federated Learning (FL) has emerged as a promising solution to address these challenges. FL allows data owners (i.e. clients) to convey locally trained model parameters to a central server without

[1]Big Data Decision Institute, Jinan University, Guangzhou, China
[2]School of Management, Jinan University, Guangzhou, China
[3]School of Medicine, Jinan University, Guangzhou, China
[4]College of Information Science and Technology, Jinan University, Guangzhou, China
[5]School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China
[6]Faculty of Business, The Hong Kong Polytechnic University, Hong Kong, China

*These authors contributed equally to this study.

Corresponding authors:
Eric WT Ngai, Faculty of Business, The Hong Kong Polytechnic University, Hong Kong, China.
Email: eric.ngai@polyu.edu.hk
Yong Hu, Big Data Decision Institute, Jinan University, Guangzhou, China.
Email: henryhu200211@163.com

transferring the original data. The central server aggregates these parameters and generates a global model through iterative training across multiple rounds. By leveraging FL, the diversity and scale of data available for training can be significantly enhanced, leading to models with improved performance.[5–7] Given its ability to jointly develop machine learning models while preserving data privacy, FL is regarded as an ideal framework for privacy-sensitive data such as EHRs. It has been successfully applied to tasks such as disease risk prediction, clinical diagnosis, and medical image recognition, thereby demonstrating its potential to transform healthcare analytics,[8,9] directly contributing to Sustainable Development Goal (SDG) 3 (Good Health and Well-being) while addressing SDG 9 (Industry, Innovation, and Infrastructure) through AI frameworks.

Significant obstacles remain in application of FL in healthcare field. A primary challenge lies in the statistical data heterogeneity among clients, termed non-independent and identically distributed (non-IID) data, which is particularly pronounced in healthcare. This heterogeneity stems from factors such as geographic location, clinical practices, patient demographics, genetic diversity, and phenotypic differences. As a result, FL algorithms often face impaired performance, with the global model showing considerable variability in effectiveness across different hospitals. In some cases, the performance of the global model may even worse than local model in specific hospitals.[5,6] Furthermore, patient heterogeneity adds another layer of complexity, as a single global model may struggle to achieve desired performance across diverse patient cohorts.[10,11] If these issues can be overcome, it will facilitate the collaborative training of models among medical institutions while protecting data privacy, and further support SDG 17 (Partnerships for the Goals), thus addressing the problem of data silos that hinder global health programs.

Researchers have explored clustering-based FL approaches[12] to address these challenges. The research of Stallmann[13] divided these methods into two categories. The first and more prevalent type, called as clustered federation, divides clients into multiple clusters based on their similarity. Each client participates in federated training only within its respective cluster. The second category is called as Federated Clustering or Sample Clustered FL,[14] which is concerned with identifying global clusters to which samples belong in distributed data without sharing data.

Clustered federation approach assumes that data within each cluster is identically distributed or that each client contains a single type of data.[15–18] However, this assumption conflicts with the cross-silo scenarios typical in the medical domain, where heterogeneity among patients in each hospital may be significant. Consequently, these methods are better suited to cross-device scenarios, such as the Internet of Things or smart wearable devices and fail to adequately address the unique challenges posed by heterogeneity among hospitals in healthcare.[17,19–22]

Federated Clustering is particularly well-suited for cross-silo scenarios and effectively addresses the problem of data heterogeneity within a single data center.[13] However, research in this area remains relatively scarce. Most existing studies refined the federated clustering process by utilizing K-means and its derivative algorithms, such as C-means.[22–24] In the context of smart healthcare, research has shown that the misclassification rate using K-means clustering is approximately four times higher than that of clustering methods based on probabilistic models.[25] The K-means algorithm, despite its widespread use, faces inherent limitations that reduce its effectiveness in federated clustering scenarios, particularly in handling the complex and heterogeneous data common in healthcare. Key limitations include the following:

1. Poor recognition of non-convex and multi-peaked clusters. K-means assumes that the clusters are circular or nearly circular, and when the clusters in a dataset were irregularly shaped, it may struggle to accurately delineate the clusters.
2. A hard clustering method, not flexible enough. If a sample does not exactly match the characteristics of any of the clustering centers, it is still forced to be classified to a clustering center in the K-means algorithm.
3. Sensitive to noise. Since the K-means algorithm is based on a distance metric, outliers may significantly affect the location of the clustering centers and the quality of the clustering results.

To address these challenges, we propose the Federated Gaussian Mixture Clustering (FedGMC) framework. It comprises three sequential stages: Patient Encoding, Federated Clustering, and Personalized FL. It overcomes the limitations of traditional K-means by leveraging a probabilistic clustering approach and addresses the heterogeneity issue using a personalized FL method. We evaluated the effectiveness and stability of the framework using a real-world EHR dataset of acute kidney injury (AKI) patients. The experimental results demonstrated that FedGMC not only consistently outperformed baseline methods in terms of predictive performance metrics but also maintained fairness across participating institutions by significantly reducing risk of performance degradation.

Literature review

This study follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to conduct a systematic literature search on data heterogeneity solutions for federated learning based on clustering method, aiming to systematically evaluate the technical bottlenecks

of existing methods. We performed a structured search in the Google Scholar databases (January 2019 to May 2025) using the keyword "federated learning clustering." First, non-Original Research Articles such as reviews and conference abstracts were excluded; then, full-text quality evaluation was conducted on the remaining literature to exclude those lacking control groups or insufficient relevance to the theme. Finally, core research achievements were selected and are shown in Table 1.

The first category is clustered federation. This category of method performs clustering based on clients data distribution,[16,40] model parameters,[18] or update gradients.[17] The clustering methods used include K-means, its derivative algorithm fuzzy C-means, Hierarchical clustering, and GMM. Sahinet[26] and some researchers used K-means and its derivative algorithm fuzzy C-means for clustering. Briggs et al.[15] introduced a hierarchical clustering step and used the similarity between the client's local update and the global federated model to divide the client clusters. In 2024, Malekmohammad et al.[31] uses GMM clustering to achieve client clustering by combining model update and training loss. All these methods categorized as clustered federation usually assume that the data is evenly distributed among clients or that each client contains only a single type of data. It ignores the reality that there may be significantly different sample clusters between different clients/hospitals and within the same dataset.

The second category is Federated Clustering, which clusters samples. The clustering methods used include K-means, its derivative algorithm (fuzzy C-means, fuzzy C-means), Hierarchical clustering, and GMM. In algorithm k-FED,[23] each client performs K-means clustering on the local data and sends the clustering results to the server. The server performs K-means clustering on the centroids sent by all clients. Stallmann et al.[13] proposed the FedFCM algorithm in 2022, which is an algorithm similar to k-FED that utilizes fuzzy c-means.[41] In 2024 and 2025, latest researches still focus on improving algorithms with k-means and its derivative algorithm. Bárcena[36] and Stallmann[37] used c-Means and Fuzzy c-Means clustering, respectively. In the section Introduction, we have elaborated on the limitations of this type of algorithms. Ahmed et al.[38] proposed a framework PCBFL based on spectral clustering[42] in medical and healthcare area. It has a time complexity of $O(n^3)$. In contrast, the time complexity of the k-means algorithm and the Gaussian mixture clustering algorithm is linearly related to the number of samples, only $O(n)$. Some other studies that apply GMM to FL also differ significantly from ours. Valdeira et al.[39] applied GMM to the vertical FL scenario, where the participants have the same samples but different feature spaces. Our study is horizontal FL, where the datasets of the participants have similar or identical feature spaces but different samples. Although the FedGMM[43] algorithm also uses GMM to fit the joint data distribution of clients in FL training, it is significantly different from ours in terms of algorithm flow and function. FedGMM does not cluster samples, and its goal is to generate a personalized model for each client. Our FedGMC clusters samples but also builds a personalized FL model shared by all clients for each cluster generated by clustering.

Based on the above analysis, the clustered federation method cannot cope well with the data heterogeneity

Table 1. Clustering-based FL optimization method.

Category	Clustering methods	References
Clustered Federation	K-means	Sahin (2023)[26]; Huang (2019)[16]
	fuzzy C-means	Sun (2025)[27]; Nair (2022)[18];
	Hierarchical clustering	HaghighiFard (2025)[28]; Sun (2025)[29]; Vahidian (2023)[30]; Briggs (2020)[15]
	GMM	Malekmohammadi (2024)[31]; Long (2023)[32]; Sattler (2021)[17]
Federated Clustering	K-means	Alfawaz (2025)[33]; Wang (2025)[34]; DENNIS (2021)[23]; Huang (2019)[16]
	K-means derivative algorithm C-means, fuzzy C-means	Li (2025)[35]; Barcena (2024)[36]; Stallmann(2024)[37]; CHUNG (2022)[22]; PEDRYCZ (2022)[24]; Stallmann (2022)[13]
	Spectral clustering	Ahmed (2023)[38]
	GMM	Valdeira (2022)[39], Vertical FL scenario
		Our proposed FedGMC, Horizontal FL scenario

problem in the cross-silo scenario targeted by this study. Existing methods based on Federated Clustering have defects in accuracy and algorithm complexity or are not suitable for horizontal FL scenarios.

Methods

Framework overview

The FedGMC framework comprises three phases, as shown in Figure 1.

Patient encoding. This stage preprocesses patient data, creating robust feature representations while preserving data privacy. Each hospital uses the vector encoder provided by the central server to transform local patient cohorts into vector representations. Continuous data is directly encoded using its measurement values, while categorical data is processed using one-hot encoding, to maintain consistency and interpretability across the federated system.

Federated clustering. This phase involves clustering the entire patient cohort across hospitals without sharing raw data, preserving data privacy. The clustering process leverages the probabilistic Gaussian Mixture Model (GMM) to effectively capture complex data distributions and heterogeneity among patients. Detailed procedures are elaborated in section "Federated clustering."

Personalized FL for each cluster. The framework trains personalized FL models for each cluster, and all hospitals participate only in the FL of their respective clusters, ensuring that each cluster benefits from a tailored model optimized for its unique characteristics. Comprehensive details of this phase are provided in section "Personalized federated learning."

The FedGMC framework integrates these phases seamlessly to address the challenges of heterogeneous and privacy-sensitive data in healthcare, enabling effective and equitable predictive modeling across institutions.

Federated clustering

The GMM[44] is a probabilistic model capable of representing datasets that can be partitioned into multiple Gaussian distributions. Unlike hard clustering algorithms such as K-means, GMM assigns a likelihood to each sample belonging to a given cluster, offering a more nuanced approach to clustering. The model characterizes each cluster's position and shape using two key parameters: mean and covariance, enabling flexible, and precise clustering. In comparison to K-means and its derivative clustering algorithms,[25] GMM offers the following advantages. First, unlike K-means, which assumes spherical clusters, GMM can model clusters with diverse and irregular shapes. Second, as a soft clustering method based on probability density, it assigns each sample the probability of belonging to different clusters, which make it more adaptable to complex datasets where it is difficult to assign each sample to a hard cluster. Third, GMM's probabilistic approach makes it less sensitive to outliers and noisy data. Fourth, GMM accounts for not only the mean and centroids of clusters but also the variance, sample distribution shapes, and underlying statistical properties. Our study integrates the GMM method into federated clustering, enhancing the process by accounting not only for mean of samples and centroids but also for variance, sample distribution shapes, sample sizes, and feature differences across clusters and hospitals. This approach enables the central server to receive richer and more comprehensive distribution information about the patient cohort, resulting in significantly improved clustering performance. The improved GMM-based federated clustering phase (Phase 2) is detailed in Algorithm 1, which illustrates the step-by-step process. Compared to baseline algorithms, such as CBFL[16] and

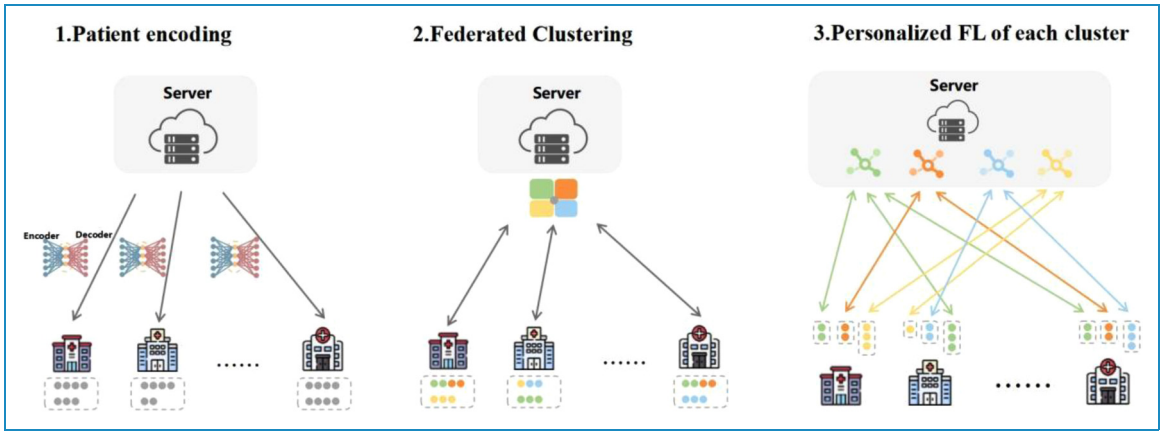

Figure 1. Schematic of FedGMC.

Algorithm 1. Federated Clustering

Input:

The number of clients: N;

The collection of samples from each client: $D = \{D_1, \ldots\ldots, D_N\}$;

The maximum number of iterations for the clustering model to learn: T_1;

The maximum number of clusters on the server side: K = 8;

The maximum number of clusters on the client side: k = 4.

Output: the best clusters of each client

1: procedure Train Federated Cluster Model:

2: Clients execute:

3: for each client D_i $(i = 1, 2, \ldots, N)$ in parallel do

4: $k_b^i, f_{GMC}^{k_b,i} \leftarrow$ *Select Best_k*(D_i, k)

5: obtain the mean μ_i and covariance matrix Σ_i through function $f_{GMC}^{k_b,i}$

6: Upload (μ_i, Σ_i) to server

7: end

8: Server executes:

9: for each (μ_i, Σ_i) $(i = 1, 2, \ldots, N)$ in parallel do

10: $X_{si} \leftarrow$ *Generate ten simulated samples using* (μ_i, Σ_i)

11: end

12: $X_s \leftarrow X_{s1} + X_{s2} + \ldots\ldots + X_{sN}$

13: $k_b, f_{GMC}^{k_b} \leftarrow$ *Select Best_k*(X_s, K)

14: obtain the final clustering model f_{GMC} and the number of clusters k

15:

16:

17: procedure Cluster Patients Cohort:

18: for each client D_i $(i = 1, 2, \ldots, N)$ in parallel do

19: Download f_{GMC}

20: partition the clusters using f_{GMC}: $\{C_1^i, C_2^i, \ldots, C_k^i\} \leftarrow f_{GMC}(D_i)$

21: end

Table 2. Comparison of phase 2 with baseline algorithms.

Phase 2	CBFL[16]	k-FED[23]	FedGMC
Step 1: Server	Confirm clusters number K.		
Step 2: Hospitals	Take the mean of all samples.	Use K-means to cluster local cohort.	Use GMM to cluster local cohort.
Step 3: Hospitals	Send mean of the feature vector to server.	Upload centroids to server.	Send probability distribution parameters to server.
Step 4: Server	Divide the feature vectors uploaded by all hospitals with k-means.	Divide the centroids uploaded by all hospitals with k-means.	Generate virtual samples according to parameters from hospitals, and cluster with GMM.
Step 5: Server	—	—	Take the centroids as initial setup of Step 2, iterate until preset rounds.
Step 6: Server	Output a federated clustering model.	Output a federated clustering model.	Jump to Step 1 to traverse possible clusters number until confirm the optimal. Output a federated clustering model
Step 7: Hospitals	Each hospital uses the federated clustering model f to classify all local samples.		

K-Fed,[23] the enhancements introduced in FedGMC focus on improving the performance of Phase 2. Table 2 summarizes the differences between these methods, highlighting how FedGMC's GMM-based approach addresses the limitations of existing federated clustering frameworks.

Algorithm 1. Federated Clustering

Determining the optimal number of clusters K is a critical challenge in clustering algorithms. Existing studies often require exhaustive traversal through all possible values of K, leading to significant time, communication, and computational overhead in FL settings. To address this, we introduce the silhouette coefficient in Phase 2 as an evaluation metric for clustering effectiveness.[45] The silhouette coefficient measures the difference in similarity between samples within clusters and between clusters, providing a comprehensive assessment of clustering quality. Its value range is $[-1, +1]$.[46] The larger the value, the better the clustering effect. The K value that provides the maximum silhouette coefficient is selected as the optimal number of clusters (see Appendix A. Algorithm 2).

Personalized federated learning

In phase 3, all hospitals engage exclusively in the training of personalized FL models for the clusters they host. The detailed training process is as follows:

1. The server initializes the prediction model for each cluster and distributes it to the hospitals containing patients belonging to that cluster.
2. Each hospital conducts FL model training locally for the clusters it contains, updating the model parameters using its data.
3. Hospitals encrypt the updated model parameters and transmit them securely to the server. The server aggregates the parameters received from all hospitals to update the cluster-specific models.
4. If the preset convergence criterion is not reached, the server sends a new round of updated cluster model to the client containing the cluster, and the process loops back to (2) for a new round of FL training. Always criteria for convergence of FL training are generally set to one of the following, such as the change in loss value or model parameters between iterations is less than the preset value, or the preset maximum number of training rounds is reached.[47]

Data and experiments design

Datasets and preprocessing

This study used a real-world eICU dataset to validate proposed method.[48] The dataset encompasses EHR data from over 200 hospitals and more than 100,000 ICU patients across the United States in 2014 and 2015. It includes diverse patient information such as demographics, medications, diagnoses, procedures, time-stamped vital signs, and laboratory test results.

The experimental prediction task was to forecast whether a patient would develop acute kidney injury (AKI) 48 h in advance. AKI is a potentially life-threatening condition that complicates treatment, impacts clinical trajectories,

and can significantly worsen outcomes for a substantial number of hospitalized or ICU patients.[49] Early prediction of AKI risk can enable timely interventions, improve patient outcomes, and significantly reduce hospitalization costs and mortality rates.[1,50] The definition of AKI followed the method proposed by the Kidney Disease Improving Global Outcomes (see Appendix A).[51] The study cohort excluded the following patient samples to ensure data reliability and relevance:

1. Patients with <2 measurements of Serum Creatinine (SCR);
2. Patients with an estimated Glomerular Filtration Rate (eGFR) < 15 mL/min/1.73 m^2 at admission;
3. Patients aged <18 years;
4. Patients with an ICU stay duration of <48 h from admission to discharge.

Selecting feature variables guided by expert knowledge has proven to be more effective for constructing data-driven machine learning models.[52] In this study, clinical features critical for predicting AKI were selected based on the expert recommendations from the Kidney Disease Improving Global Outcomes (KDIGO) guidelines.[51] Features with a high rate of missing data (occurrence <10%) were excluded to ensure data quality. This process resulted in the selection of 22 discrete traits and 71 continuous traits. We deleted discrete features from the clustering process for three reasons. First, as we employed the GMM algorithm, which assumes Gaussian-distributed data, only continuous features meet this requirement. Second, GMM characterizes correlations between features using covariance matrices. However, the covariance between binary and continuous variables lacks practical interpretability. Third, the discrete "0/1" transitions of binary features can distort probability density estimation, leading to ambiguous clustering boundaries. All 93 selected features were incorporated in the personalized FL module (see Table 3). Missing values were addressed using tailored imputation methods: for discrete features, a zero placeholder was used, whereas continuous features were imputed using the random forest-based filling method. We also standardized the continuous data; these preprocessing steps ensured a robust and reliable foundation for subsequent analysis.

In this study, we selected the 20 hospitals with the largest patient cohorts as participants, encompassing a total of 64,974 ICU stays. These hospitals exhibited variations in sample size, AKI proportion, and patient characteristics, ensuring a diverse dataset for analysis. Each hospital's cohort was randomly split into training and testing datasets using an 80:20 ratio, maintaining a consistent approach across all participants.

Table 3. Feature used in modeling.

Feature category	No. of features	Details
Demographics	3	Age, Race, Gender
Vital signs	5	BMI, Systemicsystolic, Temperature, Heart rate, Respiratory rate
Laboratory tests	64	SaO2, pH, Potassium, Calcium, Glucose, Sodium, HCO3, Methemoglobin etc.
Comorbidity	11	Cancer, Diabetes, Hypertension, Heart disease, Pulmonary disease etc.
Medication	2	Insulin, Lactulose
Procedures	9	Transfusion, CT Scan, X Ray, Insertion, Injection etc.

Compared methods

We compare our framework with the following methods.

Local: each hospital only uses its local data to train model.

Centralized: it aggregates data from all participants together to train a global model.

FedAvg[53]: Federated Averaging algorithm is the most commonly used model aggregation algorithm in federated learning. The core idea is that after multiple participants locally train models, the parameters of these models are weighted and averaged to obtain a global model.

FedProx[54]: A well-known algorithm used to solve the heterogeneity of FL data. The core idea is to introduce the proximal term in the client optimization process, constrain the difference between the local update and the global model, alleviate the divergence of the client optimization direction, and improve the generalization ability of FL under heterogeneous data.

CBFL[16]: classic clustering-based FL algorithm. It uses the k-means algorithm to cluster patients based on the average feature vector of samples in each hospital and then trains FL models for each cluster separately. This method ignores heterogeneity among patients in each hospital.

K-FED[23]: improved algorithm of CBFL. Each hospital uses the k-means algorithm to cluster local patients. The server uses the k-means to cluster patients based on all centroid's information uploaded by each hospital and then trains FL models for each cluster separately. K-Fed overlooks the weight differences between centroids.

FedGMC: algorithm proposed in this study.

Parameter settings

In this study, we employed two widely adopted classifiers: the logistic regression (LR) model, known for its interpretability, and the Multi-Layer Perceptron (MLP) model, a

common neural network architecture. Consistent classifier parameters were applied across different algorithms to ensure a fair comparison.

Model performance was evaluated using Recall and Area Under the Receiver Operating Characteristic Curve (AUC). Recall measures the model's ability to identify positive cases, making it particularly relevant for risk-sensitive scenarios such as EICU. While AUC, on the other hand, provides a comprehensive assessment of overall classification performance.

To compare various FL algorithms, key parameters such as the number of communication rounds, local training iterations, and batch size were standardized. Based on expert recommendations, the range for the number of clusters was set between 3 and 8.

Results

Cluster analysis

As shown in Figure 2, the clusters present in each hospital differ significantly, with notable variations in both the proportions and sample sizes of each cluster, even among hospitals with the same cluster types. This highlights the inherent heterogeneity among hospitals and among patient populations. Additionally, hospitals with larger patient cohorts tend to contain more cluster types, indicating that cluster centroids are more heavily influenced by these hospitals.

Taking the LR model as an example, we calculated the top-ranked features in each clustering model based on the absolute value of the regression coefficient[55] (see Table A.1). Tables in Figure 3 show the overlap rates of the Top5, Top10, and Top20 features of the clusters generated by different algorithms. It presents the feature overlap rates for all pairwise combinations (a total of 6 pairs) of the four clusters from cluster_0 to cluster_3 and also marks the overall average overlap rate. FedGMC exhibits values of 0.2, 0.35, and 0.35 for the top 5, 10, and 20 features, respectively. This is in contrast to CBFL, which shows average values of 0.3, 0.32, and 0.44 for the corresponding feature sets, and K-fed, with average values of 0.33, 0.40, and 0.42. A smaller overlap rate indicates a more significant difference between clusters. The differences between clusters in the FedGMC algorithm are more significant, demonstrating that it effectively captures heterogeneity with a better clustering performance. Table A.1 highlights the top 10 features of FedGMC that are unique to each cluster.

Predictive performance

We conducted 5-fold cross-validation to measure model performance. Table 4 presents the performance results of FedGMC and baseline algorithms. Considering the overall performance across the two classifiers and two performance metrics, FedGMC outperforms other baseline methods. Compared to FedAvg, CBFL, and K-Fed, FedGMC_LR exhibits improvements in mean Recall by 5.80%, 5.65%, and 2.06%, respectively, and in AUC by 2.20%, 1.03%, and 1.91%. Similarly, FedGMC_MLP demonstrates improvements in mean Recall of 4.30%, 3.77%, and 1.40%, and in AUC of 1.01%, 0.60%, and 1.63%. On the

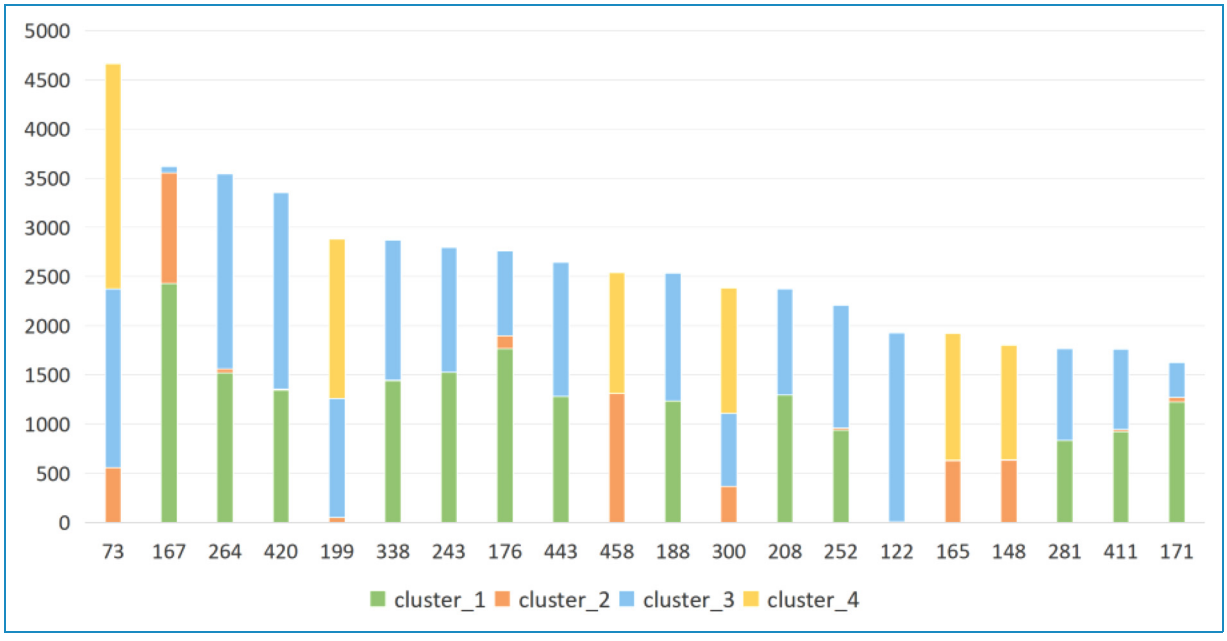

Figure 2. Distribution of patients among hospitals.

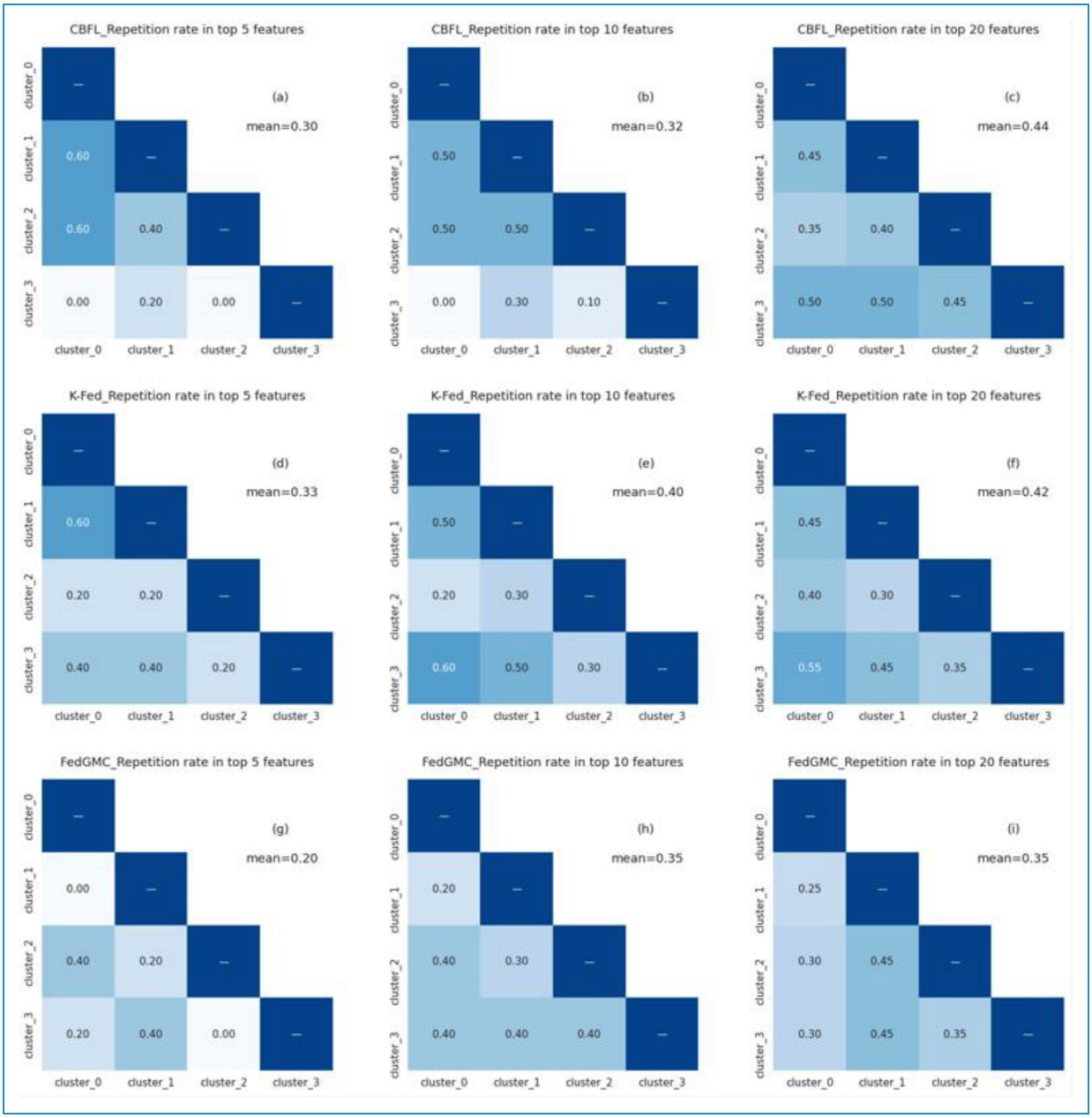

Figure 3. Comparison of inter-cluster repetition rates under different top features for CBFL, K-Fed, and FedGMC.

MLP classifier, although the AUC of FedProx leads FedGMC by 0.67%, it lags behind by 2.99% in terms of the Recall. On the LR classifier, the Recall and AUC metrics of FedProx lag behind those of the FedGMC by 7.17% and 2.05% even more significantly.

Figure 4 shows the number of hospitals that achieved optimal performance for each model. FedGMC_LR achieved the best performance in terms of Recall at 10 hospitals and AUC at 12 hospitals. FedGMC_MLP showed optimal performance in Recall in 12 hospitals and in AUC at 3 hospitals.

Discussion

The proposed personalized FL framework, FedGMC, which leverages a probabilistic modeling approach to overcome the limitations of existing methods. By utilizing sample distribution information, the server generates virtual samples to enhance the accuracy of federated patient clustering. Subsequently, a personalized FL model is trained for each cluster, leading to improved overall prediction performance.

Experimental results show that the FedGMC outperforms all baseline algorithms across both classifiers and

Table 4. Predictive performance on 20 hospitals.

Algorithm	Recall	Improved No. (Ratio)	AUC	Improved No. (Ratio)
Classifier:LR				
Local	0.7490 [0.7406,0.7573]	—	0.7324 [0.7283,0.7365]	—
FedAvg	0.7523 [0.7295,0.7750]	9 (45%)	0.7309 [0.7258,0.7360]	13 (65%)
FedProx	0.7386 [0.7136,0.7636]	8 (40%)	0.7325 [0.7262,0.7387]	12 (60%)
CBFL	0.7538 [0.7323,0.7753]	9 (45%)	0.7426 [0.7372,0.7479]	15 (75%)
K-FED	0.7897 [0.7618,0.8176]	15 (75%)	0.7338 [0.7279,0.7396]	12 (60%)
FedGMC	**0.8103 [0.7865,0.8342]**	**19 (95%)**	**0.7530 [0.7436,0.7624]**	**17 (85%)**
Classifier:MLP				
Local	0.7314 [0.7156,0.7472]	—	0.7125 [0.7083,0.7168]	—
FedAvg	0.7603 [0.7373,0.7832]	13 (65%)	0.7364 [0.7347,0.7380]	17 (85%)
FedProx	0.7734 [0.7500,0.7967]	16 (80%)	**0.7533 [0.7455,0.7612]**	**20 (100%)**
CBFL	0.7656 [0.7524,0.7788]	16 (80%)	0.7405 [0.7344,0.7465]	19 (95%)
K-FED	0.7892 [0.7702,0.8082]	18 (90%)	0.7302 [0.7231,0.7374]	14 (70%)
FedGMC	**0.8033 [0.7869,0.8198]**	**18 (90%)**	**0.7466 [0.7390,0.7541]**	**20 (100%)**

performance indicators, achieving the best predictive method in most hospitals. We also observe from Table 4 that while FL improves the overall average prediction performance of participants, it does not necessarily improve that of the vast majority of participants. For example, with FedAvg_LR and CBFL_LR, only 45% of hospitals show in recall. FedGMC significantly reduces the likelihood of performance loss for participants in joint modeling. This has important implications for enhancing data owners' willingness to participate and promoting the fairness of the algorithm.

Our study has several limitations that can be addressed in future studies. First, we used all continuous traits for clustering in our framework. Future research could explore whether selecting a subset of features would result in better clustering outcomes or be more aligned with clinical applications. Second, knowledge is not shared between clusters in our current framework. Future research could investigate whether transfer learning could be used to leverage useful information from other clusters. Third, the eICU dataset used in our study consists of hospitals exclusively from the United States, all participating in the Philips eICU program. This uniformity in data sources likely facilitated data standardization reduced the data statistical heterogeneity.[7] We suspect this is why, despite our optimization efforts, the performance of the MLP classifier still lags behind that of the LR model. Previous studies have also indicated that simpler FL algorithms, like FedAvg, may be more suitable for machine learning tasks on structured EHR data compared to more complex FL methods. We aim to validate our findings using additional datasets in the future. Fourth, the experiments show that the performance improvements of different FL algorithms for the participants vary considerably across participants, with some participants experiencing performance declines. Therefore, we recommend that further research focus more on ensuring the fairness of FL algorithms.

Conclusion

This study addresses the challenges posed by the heterogeneity of hospitals and patients in collaborative modeling. We propose a personalized FL framework FedGMC for the collaborative training of disease risk prediction models across medical institutions. This framework is designed to improve the performance of existing methods in complex datasets and scenarios. Experiments using EICU data show that FedGMC effectively captures heterogeneity. The personalized FL models generated by FedGMC not only outperform multiple baseline methods but also significantly reduce the likelihood of performance degradation among participants. The improvement is crucial for attracting

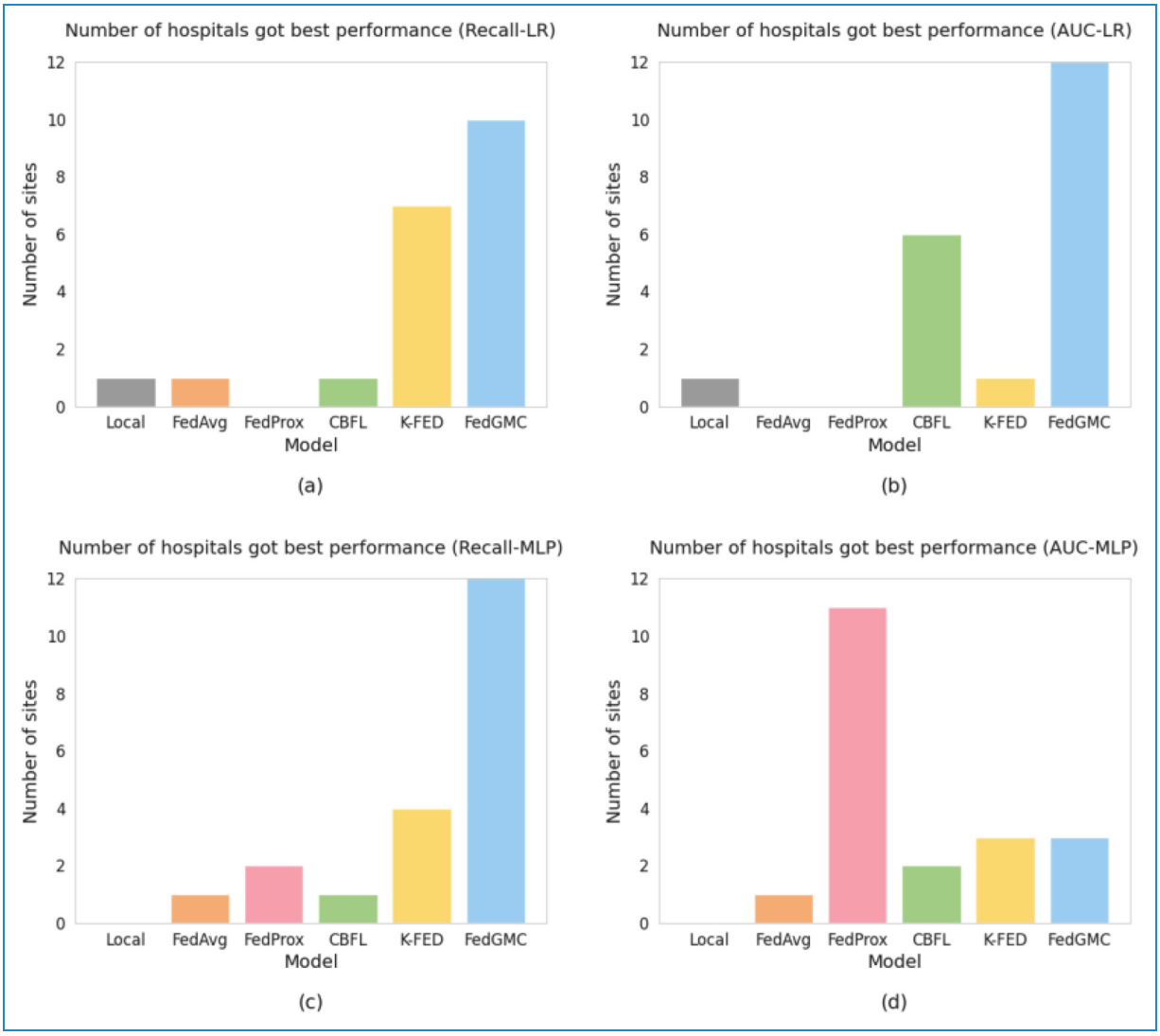

Figure 4. Number of hospitals achieving optimal performance for each model.

more data owners to join the collaborative efforts. Beyond healthcare, the proposed is also applicable to federated clustering and joint modeling for data privacy protection in other domains, such as finance and recommendation systems. This study plays a critical role in overcoming data silos and unlocking the value of data.

Acknowledgements

We thank the members of the MIT Laboratory for Computational Physiology for allowing access to and use of the eICU database.

ORCID iD

Hong Ye https://orcid.org/0009-0007-6466-5625

Ethical approval and consent to participate

The eICU database was accessed via the PhysioNet platform. Access to the database was approved after completing the Collaborative Institutional Training Initiative program "Data or Specimens Only Research" (certificate ID: 66813987), as well as signing the data usage agreement of the PhysioNet Review Board. The study was exempt from approval from the institutional review board of the Massachusetts Institute of Technology because of the retrospective design, lack of direct patient intervention, and the security schema, for which the re-identification risk was certified as meeting safe harbor standards by an independent privacy expert (Privacert) (Health Insurance Portability and Accountability Act Certification no. 1031219-2). The institutional review board of the Massachusetts Institute of Technology waived the need for informed consent for the same reason. The study was conducted following the Declaration of Helsinki. All methods used in this study were performed in accordance with the relevant guidelines and regulations.

Author contributions

Hong Ye: methodology, writing—original draft preparation, formal analysis, validation. Xiangzhou Zhang: project administration,

writing—original draft preparation, data curation. Kang Liu: writing—reviewing and editing, methodology, formal analysis. Ziyuan Liu: validation, visualization. Weiqi Chen: software, data curation. Bo Liu: investigation. Eric W. T. Ngai: writing—reviewing and editing, conceptualization, supervision. Yong Hu: conceptualization, supervision, methodology, funding acquisition.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article. This work was supported by the Major Research Plan of the National Natural Science Foundation of China (Key Program, Grant No. 91746204), the National Natural Science Foundation of China (Grant No. 72371116), the Science and Technology Development in Guangdong Province (Major Projects of Advanced and Key Techniques Innovation, Grant No. 2017B030308008), and Guangdong Engineering Technology Research Center for Big Data Precision Healthcare (Grant No. 603141789047).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Song X, Yu ASL, Kellum JA, et al. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nat Commun* 2020; 11: 5668.
2. Miotto R, Wang F, Wang S, et al. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* 2018; 19: 1236–1246.
3. Aldosari B. Patients' safety in the era of EMR/EHR automation. *Inform Med Unlocked* 2017; 9: 230–233.
4. Chen M, Qian Y, Chen J, et al. Privacy protection and intrusion avoidance for cloudlet-based medical data sharing. *IEEE Trans Cloud Comput* 2020; 8: 1274–1283.
5. Rajendran S, Xu Z, Pan W, et al. Data heterogeneity in federated learning with electronic health records: case studies of risk prediction for acute kidney injury and sepsis diseases in critical care. *PLOS Digit Health* 2023; 2: e0000117.
6. Nguyen DC, Pham Q-V, Pathirana PN, et al. Federated learning for smart healthcare: a survey. *ACM Comput Surv* 2023; 55: 1–37.
7. Dang TK, Lan X, Weng J, et al. Federated learning for electronic health records. *ACM Trans Intell Syst Technol* 2022; 13: 1–17.
8. Gunesli GN, Bilal M, Raza SEA, et al. A federated learning approach to tumor detection in colon histology images. *J Med Syst* 2023; 47: 99.
9. Xu J, Glicksberg BS, Su C, et al. Federated learning for healthcare informatics. *J Healthc Inform Res* 2021; 5: 1–19.
10. Liu K, Zhang X, Chen W, et al. Development and validation of a personalized model with transfer learning for acute kidney injury risk estimation using electronic health records. *JAMA Netw Open* 2022; 5: e2219776.
11. Dennis JM, Shields BM, Henley WE, et al. Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data. *Lancet Diabetes Endocrinol* 2019; 7: 442–451.
12. Ghosh A, Chung J, Yin D, et al. An efficient framework for clustered federated learning. *Adv Neural Inf Process Syst* 2020; 33: 19586–19597.
13. Stallmann M and Wilbik A. On a framework for federated cluster analysis. *Appl Sci* 2022; 12: 10455.
14. Manthe M, Lartizien C and Duffner S. Deep domain isolation and sample clustered federated learning for semantic segmentation. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2024, pp. 369–385. Springer.
15. Briggs C, Fan Z and Andras P. Federated learning with hierarchical clustering of local updates to improve training on non-IID data. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–9. Glasgow, United Kingdom: IEEE.
16. Huang L, Shea AL, Qian H, et al. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *J Biomed Inform* 2019; 99: 103291.
17. Sattler F, Muller K-R and Samek W. Clustered federated learning: model-agnostic distributed multitask optimization under privacy constraints. *IEEE Trans Neural Netw Learning Syst* 2021; 32: 3710–3722.
18. Nair DG, Nair JJ, Jaideep Reddy K, et al. A privacy preserving diagnostic collaboration framework for facial paralysis using federated learning. *Eng Appl Artif Intell* 2022; 116: 105476.
19. Ghosh A, Chung J, Yin D, et al. An efficient framework for clustered federated learning. *IEEE Trans Inf Theory* 2020; 68: 8076–8091.
20. Long G, Xie M, Shen T, et al. Multi-center federated learning: clients clustering for better personalization. *World Wide Web* 2023; 26: 481–500.
21. Li Z, Lu J, Luo S, et al. Towards effective clustered federated learning: a peer-to-peer framework with adaptive neighbor matching. *IEEE Trans Big Data* 2022; 10: 1–16.
22. Chung J, Lee K and Ramchandran K. Federated unsupervised clustering with generative models. In: AAAI 2022 International Workshop on Trustable, Verifiable and Auditable Federated Learning, 2022.
23. Dennis DK, Li T and Smith V. Heterogeneity for the Win: One-Shot Federated Clustering. In: The 38th International Conference on Machine Learning, pp.2611–2620.
24. Pedrycz W. Federated FCM: clustering under privacy requirements. *IEEE Trans Fuzzy Syst* 2022; 30: 3384–3388.
25. Magidson J and Vermunt J. Latent class models for clustering: a comparison with K-means. *Can J Market Res* 2002; 20: 36–43.

26. Şahin A. Wireless federated k-means clustering with non-coherent over-the-air computation. In: MILCOM 2023 - 2023 IEEE Military Communications Conference (MILCOM), pp. 339–344.
27. Sun L, Liu S and Muhammad G. FedWFC: federated learning with weighted fuzzy clustering for handling heterogeneous data in MIoT networks. *Alexandria Eng J* 2025; 111: 194–202.
28. HaghighiFard MS and Coleri S. Hierarchical federated learning in multi-hop cluster-based vanets. *IEEE Trans Veh Technol* 2025: 1–15.
29. Sun P, Liu E, Ni W, et al. Dual-segment clustering strategy for hierarchical federated learning in heterogeneous environments. *IEEE Wirel Commun Lett* 2025; 14: 1777–1781.
30. Vahidian S, Morafah M, Wang W, et al. Efficient distribution similarity identification in clustered federated learning via principal angles between client data subspaces. 2023, pp. 10043–10052.
31. Malekmohammadi S, Taik A and Farnadi G. Mitigating disparate impact of differential privacy in federated learning through robust clustering. *arXiv preprint arXiv:240519272*.
32. Long G, Xie M, Shen T, et al. Multi-center federated learning: clients clustering for better personalization. *World Wide Web* 2023; 26: 481–500.
33. Alfawaz O, El-Moursy AA, Saad M, et al. VFCkm: a federated clustering framework based on k-means algorithm for vertically partitioned data with shared attributes. *J Supercomput* 2025; 81: 855.
34. Wang Y, Pang W, Wang D, et al. One-Shot secure federated K-means clustering based on density cores. *IEEE Trans Neural Netw Learn Syst* 2025: 1–13.
35. Li L, Lu W and Pedrycz W. An adaptive federated fuzzy C-means clustering with nonindependently and identically distributed data. *IEEE Trans Syst Man Cybernet: Syst* 2025; 55: 4015–4028.
36. Bárcena JLC, Marcelloni F, Renda A, et al. Federated $c $-means and fuzzy c-means clustering algorithms for horizontally and vertically partitioned data. *IEEE Trans Artif Intell* 2024; 5: 6426–6441.
37. Stallmann M, Wilbik A and Weiss G. Towards unsupervised sudden data drift detection in federated learning with fuzzy clustering. In: 2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8. Yokohama, Japan: IEEE.
38. Elhussein A and Gürsoy G. Privacy-preserving patient clustering for personalized federated learnings. In: Machine Learning for Healthcare Conference, 2023, pp.150–166. PMLR.
39. Valdeira P, Soares C and Xavier J. Decentralized EM to learn Gaussian mixtures from datasets distributed by features. Epub ahead of print 2022. DOI: 10.48550/ARXIV.2201.09965.
40. Nguyen TM, Poh KL, Chong S-L, et al. FedDSS: a data-similarity approach for client selection in horizontal federated learning. *Int J Med Inf* 2024; 192: 105650.
41. Bezdek JC, Ehrlich R and Full W. FCM: the fuzzy c-means clustering algorithm. *Comput Geosci* 1984; 10: 191–203.
42. Von Luxburg U. A tutorial on spectral clustering. *Stat Comput* 2007; 17: 395–416.
43. Wu Y, Zhang S, Yu W, et al. Personalized federated learning under mixture of distributions. PMLR, 2023, pp. 37860–37879.
44. Reynolds DA. Gaussian Mixture models. *Encycl Biom* 2009; 741: 3.
45. Arbelaitz O, Gurrutxaga I, Muguerza J, et al. An extensive comparative study of cluster validity indices. *Pattern Recognit* 2013; 46: 243–256.
46. Dinh D-T, Fujinami T and Huynh V-N. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In: Knowledge and Systems Sciences: 20th International Symposium, KSS 2019, Da Nang, Vietnam, November 29–December 1, 2019, Proceedings 20, 2019, pp.1–17. Springer.
47. Yang Q, Liu Y, Chen T, et al. Federated machine learning: concept and applications. *ACM Trans Intell Syst Technol (TIST)* 2019; 10: 1–19.
48. Pollard TJ, Johnson AEW, Raffa JD, et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data* 2018; 5: 180178.
49. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA* 2016; 315: 762–774.
50. Vincent J-L, Pereira AJ, Gleeson J, et al. Early management of sepsis. *Clin Exp Emerg Med* 2014; 1: 3–7.
51. Khwaja A. KDIGO Clinical practice guidelines for acute kidney injury. *Nephron Clin Pract* 2012; 120: c179–c184.
52. Sun J, Hu J, Luo D, et al. Combining knowledge and data driven insights for identifying risk factors using electronic health records. *AMIA Annu Symp Proc* 2012; 2012: 901–910.
53. McMahan HB, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. Epub ahead of print 2016. DOI: 10.48550/ARXIV.1602.05629.
54. Li T, Sahu AK, Zaheer M, et al. Federated optimization in heterogeneous networks. *Proc Mach Learn Syst* 2020; 2: 429–450.
55. Hosmer DW Jr, Lemeshow S and Sturdivant RX. *Applied logistic regression*. USA: John Wiley & Sons, 2013.

Appendix

Appendix A. Definition of AKI

The occurrence of acute kidney injury (AKI) at the current time points of measurement for Serum Creatinine (SCr) or Urine Output (UO) is deemed present if either of the following criteria is met:

1. The SCr concentration exceeds 1.5 times the baseline value or demonstrates an increase of more than 0.3 mg/dL (26.5 μmol/L) within 48 h. The baseline value is established as the first SCr measurement post-admission and is dynamically updated with each subsequent SCr measurement; thus, the current SCr measurement serves as the baseline for future SCr assessments.
2. UO remains below 0.5 mL/kg/h for a continuous period of 6 h. Data collection spans from ICU admission until the defined prediction time point. For patient samples with documented AKI, the prediction time point is set 48 h prior to the AKI onset. For patients without AKI, the prediction time point is established 48 h before the last SCr measurement.

Appendix A. Algorithm 2

Algorithm 2. Determining optimal number of clusters

Input:
The dataset to be clustered: X;
The maximum number of clusters: K_{max};
The minimum number of clusters: K_{min};
Output:
The best number of clusters: k_b;
The best cluster model: $f_{GMM}^{k_b}$;

```
procedure Select Best_k(D, X)
    for k ← K_min to K_max do
        train GaussianMixture model f_GMM^k with X and k
        labels ← []
        score ← []
        labels ← predict labels of each sample in X using f_GMM^k
        score ← calculate the silhouette score using X and
        labels
    index ← the index of the maximum value in the score
k_b ← index + K_min
return k_b, f_GMM^{k_b}
```

Table A.1. Unique top 10 important features of each cluster model.

	cluster_0	cluster_1	cluster_2	cluster_3
1	race0: 0.3890	bun: 0.3192	**echocardiography: 0.2989**	age: 0.5310
2	age: 0.2096	bmi: 0.2355	chronic_kidney_disease: 0.2815	race0: 0.2331
3	bmi: 0.1539	**paco2: 0.2219**	bun: 0.2692	race4: 0.2225
4	**sex1: 0.1201**	**heart_disease: 0.2186**	race1: 0.1978	**paSystolic: 0.2215**
5	**mcv: 0.1035**	chronic_kidney_disease: 0.2116	**x_ray: 0.1862**	race1: 0.2003
6	bun: 0.1024	**fio2: 0.2091**	bmi: 0.1568	**hgb: 0.1817**
7	**bedside_glucose: 0.0986**	**lactulose: 0.2050**	age: 0.1462	bmi: 0.1585
8	race4: 0.0976	race0: 0.1987	race0: 0.1406	**st3: 0.1361**
9	**cpk: 0.0909**	**systemicsystolic: 0.1598**	**eos: 0.1269**	**respiration: 0.1323**
10	**mpv: 0.0760**	**diabetes: 0.1553**	**insulin: 0.1207**	**chloride: 0.1086**