Federated Fuzzy Clustering for Longitudinal Health Data

Salvador V Balkus
University of Massachusetts Dartmouth
Dartmouth, Massachusetts, U.S.A.
sbalkus@umassd.edu

Hua Fang
University of Massachusetts Dartmouth
Dartmouth, Massachusetts, U.S.A.
hfang2@umassd.edu

Honggang Wang
University of Massachusetts Dartmouth
Dartmouth, Massachusetts, U.S.A.
hwang1@umassd.edu

ABSTRACT

Traditional implementations of federated learning for preserving data privacy are unsuitable for longitudinal health data. To remedy this, we develop a federated enhanced fuzzy c-means clustering (FeFCM) algorithm that can identify groups of patients based on complex behavioral intervention responses. FeFCM calculates a global cluster model by incorporating data from multiple healthcare institutions without requiring patient observations to be shared. We evaluate FeFCM on simulated clusters as well as empirical data from four different dietary health studies in Massachusetts. Results find that FeFCM converges rapidly and achieves desirable clustering performance. As a result, FeFCM can promote pattern recognition in longitudinal health studies across hundreds of collaborating healthcare institutions while ensuring patient privacy.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Cluster analysis**; *Distributed algorithms*; • **Applied computing** → *Life and medical sciences.*

KEYWORDS

longitudinal, fuzzy, clustering, federated learning

ACM Reference Format:

Salvador V Balkus, Hua Fang, and Honggang Wang. 2022. Federated Fuzzy Clustering for Longitudinal Health Data. In *ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE' 22), November 17–19, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3551455.3559608

1 INTRODUCTION AND BACKGROUND

While modern healthcare institutions collect a vast array of data on their patients, its use is often restricted due to data privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and other laws [9, 12, 17]. Medical data is sensitive and highly regulated - anonymization is often insufficient to protect a patient's identity [15], and other technical challenges create difficulties in preserving privacy as well [3, 20]. As a result, healthcare institutions are often unable to share data with researchers who develop machine learning algorithms in the health domain [14, 22].

CHASE' 22, November 17–19, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9476-5/22/11…$15.00
https://doi.org/10.1145/3551455.3559608

Federated learning provides a solution. In lieu of data aggregation, the model's parameters are distributed to individual data sources (termed *clients*) and averaged [10]. In this way, a decentralized network of clients such as medical research centers can train models while circumventing the need to share data (see Figure 1), preserving patient privacy [14, 22]. This paper uses federated learning with fuzzy clustering to identify groups while allowing observations to be related to more than one cluster at a time [16]. This is necessary to group patients based on behavioral interventions, as it captures the complex relationship between patients, behaviors, and outcomes in longitudinal health studies [5, 8].

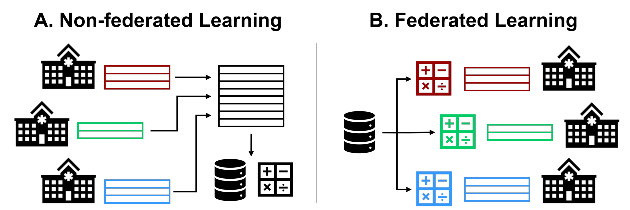

Figure 1: Comparison of federated learning. In A, healthcare institution data are aggregated to train the model. In B, parameters are distributed to each institution for training, keeping patient data private.

While federated clustering has been proposed [11, 13], the existing methods are not usable in longitudinal health studies. Firstly, they can require hundreds of rounds of communication. More importantly, they treat data from different clients as one aggregated set, rather than as individual trials, which may be undesirable when researchers seek to ensure that all clusters contain some observations from every client. Furthermore, they do not typically operate on clients whose datasets contain different unique features, which can occur if clients did not collect the same number of time points in their studies. Hence, current federated clustering methods are unsuitable for longitudinal trials.

In this paper, we develop and evaluate a federated fuzzy clustering method that maintains the privacy-preserving quality of federated learning while overcoming the aforementioned limitations on longitudinal health data. Specifically, we propose federated eFCM (FeFCM), a federated version of the enhanced fuzzy c-means clustering (eFCM) algorithm for longitudinal health studies [8].

Our work's novelty is that it extends federated learning to longitudinal health data clustering, a necessity that previous research has not yet addressed. Since the algorithm is new, we focus on evaluating its efficacy, rather than discussing specific system implementation details such as types of databases or communication techniques. Such details will be developed further in future work.

In FeFCM, after distributing starting centroids, each individual client runs eFCM simultaneously. The optimal number of clusters is selected through client voting, and final centroids are shared and aggregated using principal component analysis and weighted averaging to develop a global cluster model. Hence, FeFCM minimizes necessary communication, identifies clusters on longitudinal data with varying numbers of time points, and preserves patient privacy.

We evaluate FeFCM on both simulated clusters and empirical harmonized data from four longitudinal dietary studies in Massachusetts [6]. In the empirical data, each study represents a different client. We find that the FeFCM provides adequate global solutions based on multiple clustering metrics.

Our paper is structured as follows. First, Section 2 explains the FeFCM model. Then, Section 3 discusses our evaluation methods, including data and model parameters, while Section 4 evaluates the performance of FeFCM. Finally, Section 5 discusses the findings and future work, and Section 6 summarizes our conclusions.

2 FEDERATED ENHANCED FUZZY C-MEANS CLUSTERING (FEFCM)

This section outlines our proposed version of the eFCM algorithm which clusters data from multiple decentralized clients, such as healthcare institutions. Its input takes the form of a matrix or data frame; rows represent observations (patients or other individuals) and columns represent features (attributes) at a given time point in the longitudinal study. Figure 4 lists the FeFCM procedure.

In FeFCM, each client first performs eFCM individually. After sharing their solutions, each client then aggregates results from all other clients to create a global cluster model. We assume that each client contains an adequate number of sample observations - a number larger than the potential number of clusters in the data - and can communicate model parameters with every other client.

2.1 FeFCM Communication Architecture

The process of calculating the global model using data from across multiple clients is modeled as a decentralized parallel system. In this architecture, each client runs the algorithm simultaneously in parallel, until a round of communication is required. This allows the number of clients in practice to be easily scaled, since each client will contribute additional computing power.

Figure 2 demonstrates how clients transmit data in two rounds of communication. In the first, each client communicates its number of features (time points), and the client with the most generates the starting centroids, transmitting them to all other clients. In the second, each client transmits their updated centroids and their vote for the optimal number of clusters to every other client. These are then aggregated into the global model, as discussed below.

2.2 eFCM Clustering

In eFCM, each observation x_p possesses a degree of belonging to each cluster c_i [8], stored in a matrix U (example in Equation 1).

$$U = \begin{array}{c} \\ x_1 \\ \dots \\ x_p \end{array} \begin{array}{c} \begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\ \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ \dots & \dots & \dots \\ 0.3 & 0.4 & 0.3 \end{bmatrix} \end{array} \tag{1}$$

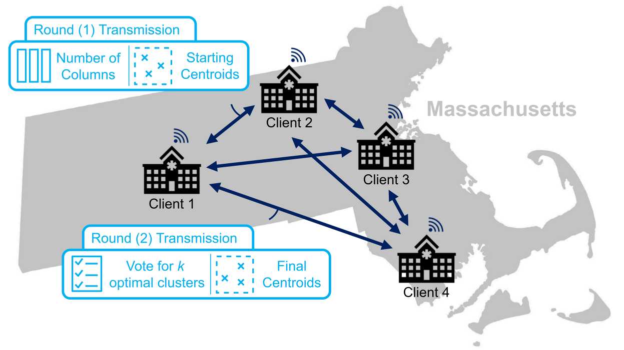

Figure 2: The decentralized communication process in FeFCM, shown on 4 clients across Massachusetts.

Two equations are used to calculate U. At iteration t, the degree of belonging u_{ij} between observation i and centroid j is calculated using Equation 2, where m represents the fuzzifier [2] (typically between 1.5 and 3). u_{ij} represents the ijth entry of the matrix U. Equation 3 calculates each updated centroid V_j at the next iteration. These calculations repeat iteratively until convergence.

$$u_{ij,t} = \frac{1}{\sum_{k=1}^{C} \frac{\|x_i - V_{j,t}\|^2}{\|x_i - V_{k,t}\|^2}^{\frac{2}{m-1}}} \tag{2}$$

$$V_{j,t+1} = \frac{\sum_{i=1}^{N} u_{ij,t}^m x_i}{\sum_{i=1}^{N} u_{ij,t}^m} \tag{3}$$

2.3 FeFCM Starting Centroid Initialization

To initialize the starting centroids, we construct a new eFCM procedure for the decentralized computing framework. The eFCM initialization scheme described in [8] is applied to the dataset of the client with the largest number of features or attributes, and the starting centroids are then projected into PCA space and inverse-projected back into the space of each individual client. The initialization is described in Figure 3.

2.4 FeFCM Global Model

Figure 4 describes how the global cluster model is constructed from each client. FeFCM begins with principal component analysis (PCA), a procedure that transforms data into a set of standardized, orthogonal principal components [1].

Projecting into PCA space allow datasets from different clients to be represented in a common space. The centroids derived from eFCM in a matrix format can be transformed into PCA space and projected onto an n-dimensional plane by dropping the principal components representing the least variance. Preserving influential features permits the aggregation of cluster results containing different numbers of time points in longitudinal data.

Let $V_{j,k}$ represent a final given centroid in PCA space for the client k after the convergence of eFCM and $V_{j,0}$ represent the starting centroid in PCA space. Equation 4 calculates final centroids V_j^* for the global cluster model by averaging the direction vector of the centroid j movement from start to end of eFCM across each

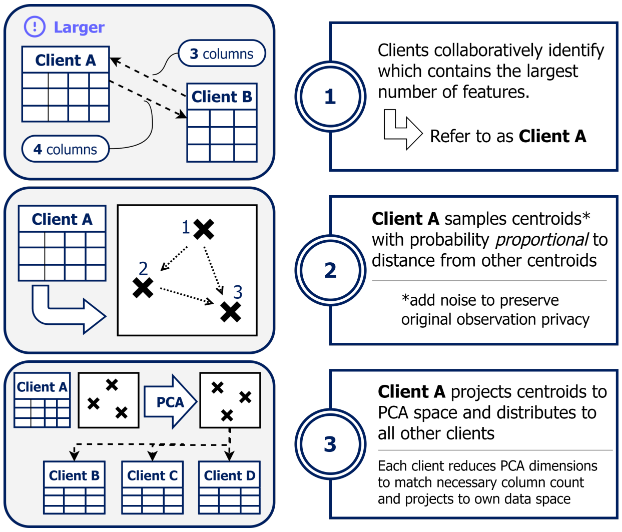

Figure 3: FeFCM starting centroid initialization algorithm.

client. The average is weighted by n_k, the number of observations contained in each client k, such that clients with more samples have greater influence. This produces the final global cluster model.

$$V_j^* = V_{j,0} + \sum_{k=1}^{N} \frac{1}{n_k}(V_{j,k} - V_{j,0}) \tag{4}$$

3 EVALUATION METHODS

FeFCM was implemented in Python 3.8 and evaluated on two clustering tasks using a desktop with an Intel Core i7-6700 CPU. Note that while in practice, FeFCM is intended to run on a CPU at each participating client, allowing the number of clients to be easily scaled, for ease of evaluation we simply run each client's portion separately on the same CPU and combine the results, averaging the time across trials. Below, we describe the empirical and simulated data for each task, the parameter settings, and the metrics reported.

3.1 Training Data

Dietary Study Data. First, this study applied FeFCM to harmonized data from four longitudinal dietary health studies in [6]. These studies recorded the types of foods which respondents consumed over 24 hours, which were processed into multiple overall diet quality scores. For this study, we used the 2010 Alternate Healthy Eating Index (AHEI-2010) section of the data [4].

Each study represents a separate client in the distributed algorithm. Client data sets contained between 2 and 4 features, each representing an AHEI-2010 score at a different time point in the longitudinal study. Note that in this work we sample only completed cases (with no missing values) for performance evaluation.

Simulated Data. Second, this study applied FeFCM to simulated clusters to test the algorithm in scenarios with different numbers of clients. To simulate the data, we used eFCM to identify five cluster centroids from each of the four clients and classify each of the client's observations into a cluster. To simulate new clusters, we

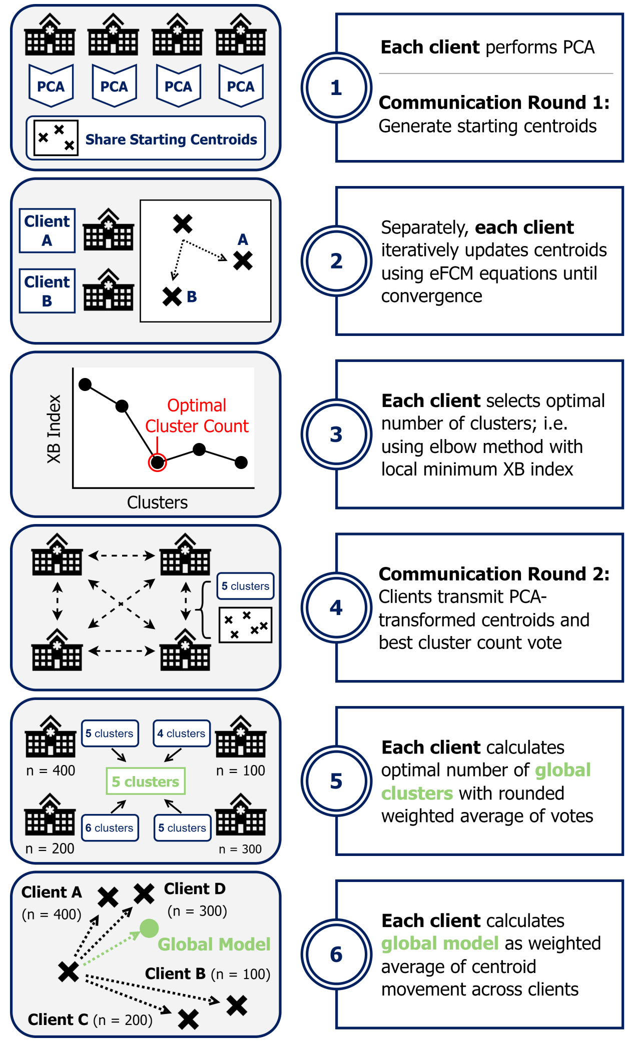

Figure 4: Steps in the full FeFCM algorithm.

randomly generated observations from a multivariate Gaussian distribution centered on each cluster centroid using the parameters learned from the four empirical dietary studies. The covariance matrix for a given distribution was calculated from the existing observations assigned to the cluster associated with its centroid.

We simulated datasets containing 4, 20, and 100 clients. With 4 clients, we simulated Gaussian clusters with 10 times the number of samples of the original data. With 20 clients, observations from the 4 simulated clients were randomly divided into 5 separate clients each. The same was done for the 100 client dataset, with a small number of observations dropped to ensure divisibility by 5. Table 1 displays the number of samples and time points for each dataset (where M represents the total number of clients).

3.2 Model Parameters

FeFCM used a fuzzifier of $m = 2.7$ (based on other empirical evaluation and visualization [5, 6]) using 300 maximum iterations and a tolerance of 10^{-4}. Clustering attempts were repeated 10 times with

Table 1: Number of empirical samples compared to number of samples per client across simulated datasets

Empirical Client ID	Time Points	Empirical Samples	Samples per client for $M =$ 4	20	100
1	2	263	2630	526	105
2	3	180	1800	360	72
3	3	129	1290	258	51
4	4	176	1760	352	70

different starting centroids, and the algorithm selected the set with the lowest total within-cluster sum of squares (inertia) as the final output.

3.3 Model Evaluation

FeFCM performance was evaluated using several metrics. The algorithm was repeated across 10 trials; the means and standard deviations of metrics are reported in Tables 2 and 3. Metrics are listed as follows:

- *Xie-Beni (XB) index*: ratio of cluster compactness to cluster separation [21] (Equation 5). Smaller values, indicating more compact and separate clusters, are preferred. This is standard for fuzzy clustering evaluation [8, 21].

$$I_{XB} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} ||V_i - X_j||^2}{n\left(\min_{i \neq j} ||V_i - V_j||^2\right)} \tag{5}$$

- *Inertia*: within-cluster sum of squares, which eFCM seeks to minimize [8]. Lower inertia is preferable.
- *Silhouette score*: how well entries fit their own cluster compared to others, on average [18]. A positive score indicates effective clustering; higher scores are preferable.

In addition, we include the total iterations for convergence and time taken in seconds across clients. These confirm the computational efficiency of the algorithm.

4 RESULTS

4.1 Dietary Study Data

On empirical dietary data across individual clients, as displayed in Table 2, FeFCM consistently achieved low XB indices of 0.11 to 0.14 and positive silhouette scores of 0.17 to 0.34 with low standard deviations, indicating consistent performance across trials. The algorithm also converged relatively quickly, averaging between 78 to 92 iterations per client and only requiring 0.14 to 0.49 seconds on average across clients to run in Python.

The global model also performed well, achieving a silhouette score of 0.3. While the global XB index of 0.85 was larger than any individual client, it was still small (less than 1), and likely grew simply due to the larger number of samples reducing cluster compactness. Finally, the global model inertia of 13988 was lower than the summed average inertia across individual clients, which totaled about 17534 - improving over individual cluster models.

4.2 Simulated Data

FeFCM obtains similar results when scaling up the number of clients. Table 3 depicts the performance of the FeFCM global model on 5 clusters simulated from multivariate Gaussian distributions. The global model achieved positive silhouette scores (0.24 to 0.35) and a reasonable XB index (0.75 to 1.01). However, increasing the number of clients decreased performance on these metrics.

FeFCM also successfully identified the correct number of clusters often. As shown in Table 3, across all client scenarios, the global model selected 5 to 6 clusters. The best result occurred for 100 clients, where FeFCM correctly found 5 clusters in every trial.

Furthermore, FeFCM leverages distributed computing to yield faster runtime and convergence. The 100 client scenario averaged 0.1 seconds and 78 iterations per client compared to the 4 client case, which averaged 2.8 seconds and 114 iterations per client, as shown in Table 3. This most likely occurred because the 100-client scenario featured fewer observations per client.

5 DISCUSSION

In this paper, we developed FeFCM for federated fuzzy clustering in a decentralized, distributed system of clients. In this way, multiple healthcare institutions can cooperate to identify cluster centroids while keeping patient data private. Using PCA, FeFCM also allows clustering of observations from varying client datasets that may not share the same number of features. Since longitudinal datasets from different institutions may collect different numbers of time points which may be represented as separate features, this method is highly useful for longitudinal health studies.

In experimental evaluation with 4 clients on simulated and empirical data, FeFCM achieves desirable clustering performance on the XB index and average silhouette score. Similar success was achieved in the 20 client and 100 client scenarios on simulated data. FeFCM also converges relatively quickly and requires only two rounds of communication with clients - one to initialize starting centroids, and one to aggregate the results into a global model - which is far less than traditional federated learning [10, 13].

To extend the model, further theoretical exploration under various statistical assumptions may be necessary. For example, federated learning assumes that data is identically and independently distributed (iid) among clients [10]. Though the dietary data used in this study is non-iid, more extreme data distributions, such as large imbalances in the sample sizes across clients, might result in performance differences. Hence, it may be worth adopting strategies for non-iid data from existing federated learning research [23, 24].

Future research in this area could also use alternate evaluation metrics or other techniques for selecting the optimal number of clusters. For example, existing techniques like the gap statistic [19] could be adapted to the federated paradigm. Alternative voting schemes like simple majority votes rather than a weighted average could be evaluated as well.

Additionally, to handle incomplete data which often occurs in longitudinal health studies, we can integrate imputation techniques such as MIFuzzy [5]. Federated learning can also be extended to neuro-fuzzy classification of clinical intervention outcomes [7]. These methods can yield new insights in medical research and improve individual intervention plans.

Table 2: FeFCM Performance on 4 Study Sites

Client	mean (SD) 1		2		3		4		**Global Model**	
XB Index	0.11	(0.00)	0.14	(0.01)	0.13	(0.03)	0.13	(0.00)	**0.85**	**(0.97)**
Silhouette	0.22	(0.00)	0.34	(0.05)	0.17	(0.11)	0.28	(0.00)	**0.30**	**(0.08)**
Inertia	3870	(0.01)	2313	(61)	3133	(202)	6147	(0.09)	**13988**	**(493)**
Time (s)	0.49	(0.26)	0.14	(0.07)	0.21	(0.10)	0.26	(0.16)		
Iterations	92	(41)	78	(38)	90	(44)	91	(56)		

Table 3: FeFCM Performance on 5 Simulated Clusters

Number of Clients	Global Model mean (SD) 4		20		100	
XB Index	0.75	(0.21)	0.81	(0.17)	1.01	(0.31)
Silhouette	0.35	(0.02)	0.25	(0.01)	0.24	(0.01)
Time (s)	2.85	(0.46)	0.64	(0.03)	0.10	(0.00)
Iterations	114	(13)	127	(5)	78	(1)
Clusters Found	5	(1)	6	(0)	5	(0)

6 CONCLUSION

Unlike existing federated learning, our proposed federated fuzzy clustering, called FeFCM, clusters longitudinal data with different numbers of time points, more efficiently identifies clusters distributed across several disparate healthcare institutions or other clients, and protects individual patient privacy. This method could even be deployed in a large-scale integrated pattern recognition platform for longitudinal health studies and clinical interventions, enabling collaboration between hundreds of different study sites. Such a platform would better improve health information communication and keep patients' personal data safe.

7 ACKNOWLEDGEMENT

This research was partly supported by NSF/IIS 2218596 to Dr. Hua Fang and Dr. Honggang Wang.

REFERENCES

[1] Hervé Abdi and Lynne J. Williams. 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* 2, 4 (June 2010), 433–459. https://doi.org/10.1002/wics.101

[2] James C. Bezdek, Robert Ehrlich, and William Full. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences* 10, 2-3 (Jan. 1984), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

[3] Min Chen, Yongfeng Qian, Jing Chen, et al. 2020. Privacy Protection and Intrusion Avoidance for Cloudlet-Based Medical Data Sharing. *IEEE Transactions on Cloud Computing* 8, 4 (Oct. 2020), 1274–1283. https://doi.org/10.1109/tcc.2016.2617382

[4] Stephanie E. Chiuve, Teresa T. Fung, Eric B. Rimm, et al. 2012. Alternative Dietary Indices Both Strongly Predict Risk of Chronic Disease. *The Journal of Nutrition* 142, 6 (April 2012), 1009–1018. https://doi.org/10.3945/jn.111.157222

[5] Hua Fang. 2017. MIFuzzy clustering for incomplete longitudinal data in smart health. *Smart Health* 1-2 (June 2017), 50–65. https://doi.org/10.1016/j.smhl.2017.04.002

[6] Venkata Sukumar Gurugubelli, Hua Fang, James M. Shikany, et al. 2022. A review of harmonization methods for studying dietary patterns. *Smart Health* 23 (mar 2022), 100263. https://doi.org/10.1016/j.smhl.2021.100263

[7] Venkata Sukumar Gurugubelli, Hua Fang, and Honggang Wang. 2019. Neuro-Fuzzy classifier for longitudinal behavioral intervention data. In *2019 International Conference on Computing, Networking and Communications (ICNC)*. IEEE. https://doi.org/10.1109/iccnc.2019.8685574

[8] Venkata Sukumar Gurugubelli, Zhouzhou Li, Honggang Wang, and Hua Fang. 2018. eFCM: An Enhanced Fuzzy C-Means Algorithm for Longitudinal Intervention Data. IEEE. https://doi.org/10.1109/iccnc.2018.8390419

[9] Health Insurance Portability and Accountability Act of 1996, 45 CFR § 164.502. 1996.

[10] Jakub Konecný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. 2016. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. *ArXiv* abs/1610.02527 (2016).

[11] Hemant H Kumar, Karthik V R, and Mydhili K Nair. 2020. Federated K-Means Clustering: A Novel Edge AI Based Approach for Privacy Preservation. IEEE. https://doi.org/10.1109/ccem50674.2020.00021

[12] Office for Civil Rights. 2013. Summary of the HIPAA Privacy Rule. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html Accessed January 3, 2022.

[13] Witold Pedrycz. 2021. Federated FCM: Clustering Under Privacy Requirements. (2021), 1–1. https://doi.org/10.1109/tfuzz.2021.3105193

[14] Nicola Rieke, Jonny Hancox, Wenqi Li, et al. 2020. The future of digital health with federated learning. 3, 1 (Sept. 2020). https://doi.org/10.1038/s41746-020-00323-1

[15] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications* 10, 1 (July 2019). https://doi.org/10.1038/s41467-019-10933-3

[16] Enrique H. Ruspini, James C. Bezdek, and James M. Keller. 2019. Fuzzy Clustering: A Historical Perspective. *IEEE Computational Intelligence Magazine* 14, 1 (Feb. 2019), 45–55. https://doi.org/10.1109/mci.2018.2881643

[17] Tobias Schulte. 2020. *The protection of personal data in health information systems – principles and processes for public health*. Technical Report 2021-1994-41749-57154. World Health Organization Regional Office for Europe, Copenhagen.

[18] Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster Quality Analysis Using Silhouette Score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE. https://doi.org/10.1109/dsaa49011.2020.00096

[19] Robert Tibshirani, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63, 2 (2001), 411–423. https://doi.org/10.1111/1467-9868.00293

[20] Volker Tresp, J. Marc Overhage, Markus Bundschus, et al. 2016. Going Digital: A Survey on Digitalization and Large-Scale Data Analytics in Healthcare. *Proc. IEEE* 104, 11 (Nov. 2016), 2180–2206. https://doi.org/10.1109/jproc.2016.2615052

[21] X.L. Xie and G. Beni. 1991. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13, 8 (1991), 841–847. https://doi.org/10.1109/34.85677

[22] Jie Xu, Benjamin S. Glicksberg, Chang Su, et al. 2020. Federated Learning for Healthcare Informatics. 5, 1 (Nov. 2020), 1–19. https://doi.org/10.1007/s41666-020-00082-4

[23] Yue Zhao, Meng Li, Liangzhen Lai, et al. 2018. Federated Learning with Non-IID Data. (2018). https://arxiv.org/pdf/1806.00582.pdf

[24] Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. 2021. Federated learning on non-IID data: A survey. *Neurocomputing* 465 (Nov. 2021), 371–390. https://doi.org/10.1016/j.neucom.2021.07.098