A Federated K-Means-Based Approach in eHealth Domains with Heterogeneous Data Distributions

Giovanni Paragliola[1][a], Patrizia Ribino[2][b] and Maria Mannone[2,3][c]
[1]National Research Council (CNR), Institute for High-Performance Computing and Networking (ICAR), Naples, Italy
[2]National Research Council (CNR), Institute for High-Performance Computing and Networking (ICAR), Palermo, Italy
[3]Institute of Physics and Astronomy, University of Potsdam, Potsdam, Germany
{giovanni.paragliola, patrizia.ribino, maria.mannone}@icar.cnr.it

Keywords: Federated Clustering, Healthcare, Heterogeneous Data Distribution.

Abstract: Healthcare organizations collect and store significant amounts of patient health information. However, sharing or accessing this information outside of their facilities is often hindered by factors such as privacy concerns. Federated Learning (FL) data systems are emerging to overcome the siloed nature of health data and the barriers to sharing it. While federated approaches have been extensively studied, especially in classification problems, clustering-oriented approaches are still relatively few and less widespread, both in formulating algorithms and in their application in eHealth domains. The primary objective of this paper is to introduce a federated K-means-based approach for clustering tasks within the healthcare domain and explore the impact of heterogeneous health data distributions. The evaluation of the proposed federated K-means approach has been conducted on several health-related datasets through comparison with the centralized version and by estimating the trade-off between privacy and performance. The preliminary findings suggest that in the case of heterogeneous health data distributions, the difference between the centralized and federated approach is marginal, with the federated approach outperforming the centralized one on some healthcare datasets.

1 INTRODUCTION

Healthcare organizations typically gather large amounts of patient health data. Multi-centre research plays a crucial role in developing machine learning (ML) algorithms for real-world scenarios. However, various factors hinder the dissemination or retrieval of this information outside the organization, such as privacy concerns (Bonawitz et al., 2021). The Health Insurance Portability and Accountability Act (HIPAA) (Act, 1996) and the General Data Protection Regulation (GDPR) (Voigt and Von dem Bussche, 2017) have established regulations that restrict the exchange of electronic health records (EHRs) between stakeholders and healthcare providers without patient consent (Sheller et al., 2020). Protecting confidential medical information while leveraging the collective knowledge of healthcare facilities presents a complex and demanding challenge (Dhade and Shirke, 2024).

Federated Learning (FL) systems are emerging as a promising solution to overcome the siloed nature of health data and the associated barriers to sharing them (Bharati et al., 2022). FL enables the decentralized training of ML models without transferring medical data to a central server. Each healthcare institution is a client node, which independently trains its model and transmits it to a central server for aggregation. A global model is formed by integrating local models from all nodes, which are then disseminated to the nodes by a centralized server responsible for coordinating and aggregating the models.

However, while federated approaches have been extensively studied, mainly in the context of classification problems (Marulli et al., 2021a; Marulli et al., 2021b), clustering-oriented approaches remain relatively scarce and less widespread. The formulation of algorithms tailored explicitly for federated clustering and their application in e-health domains are areas that are still in their early stages of exploration.

Despite the benefits of federated learning, such as preserving privacy and reducing communication overhead, a more comprehensive investigation is required to fully understand and leverage the advantages of federated clustering in healthcare domains. As the

[a] https://orcid.org/0000-0003-3580-9232
[b] https://orcid.org/0000-0003-3266-9617
[c] https://orcid.org/0000-0003-3606-3436

Paragliola, G., Ribino, P. and Mannone, M.
A Federated K-Means-Based Approach in eHealth Domains with Heterogeneous Data Distributions.
DOI: 10.5220/0012981200003837

In *Proceedings of the 16th International Joint Conference on Computational Intelligence (IJCCI 2024)*, pages 550-559
ISBN: 978-989-758-721-4; ISSN: 2184-3236

healthcare landscape evolves toward more collaborative and data-driven models, addressing these issues will become increasingly critical to realizing the full potential of federated learning to improve patient outcomes and advance medical research.

Furthermore, when investigating methodologies for distributed clustering, it is imperative to consider the distinct challenges associated with federated learning, such as data heterogeneity, meaning that data are not uniformly distributed among the participants. This problem, known as the non-independent and identically distributed (non-IID) data challenge (Wahab et al., 2021), occurs when the local data of individual clients does not accurately reflect the entire dataset due to heterogeneous class imbalances, distribution variations, and data size. These challenges can lead to significant performance degradation in federated learning models because traditional machine learning algorithms assume uniform data distribution across clients. In particular, local updates may diverge during model aggregation, resulting in slow convergence or even degraded performance in the global model. To address these issues, various strategies have been proposed, including methods for grouping clients with similar data distributions. These weighted aggregation schemes assign importance to clients based on data size or distribution and fine-tuning global models on local data.

This paper presents a study focused on evaluating a federated clustering approach in the e-health domain in non-IID scenarios. Specifically, the primary objective is to introduce a federated K-means-based (FKM) approach for clustering within the healthcare domain and explore the impact of heterogeneous data distributions. The evaluation of the proposed approach has been conducted on several health-related datasets through comparison with the centralized version and by estimating the trade-off between privacy and performance. The approach was evaluated mainly under three different data distribution scenarios to more accurately assess the impact of data heterogeneity across clients on the federated model. These scenarios range from independent and identically distributed data across clients to a non-independent and identically distributed scenario, within which we distinguish soft and hard heterogeneity.

The preliminary findings suggest that in the case of heterogeneous health data distributions, the difference between the centralized and federated approach is marginal, with the federated approach outperforming the centralized one on some healthcare datasets.

The paper's contribution can be summarized into two main points: 1) Establish a federated K-means approach to evaluate the effectiveness of clustering models in the federated eHealth domain; 2) Investigate how the hypothesis of heterogeneous health data distribution affects the convergence of local and global models.

2 RELATED WORKS

Centralized clustering approaches require storing and accessing all raw data from a single central node. The K-means clustering algorithm, introduced more than six decades ago, continues to be widely favored and utilized in contemporary research and practice.

Federated clustering is a framework within the field of federated learning whose objective is to cluster data distributed across multiple devices or locations while preserving privacy and data security. The process involves aggregating local data points that exhibit global similarity. The distributed nature of the data points allows for their clustering based on a global similarity measure, as they are distributed among multiple clients. Notably, the data remains local on client devices despite the clustering process. Based on our current understanding, a limited body of literature is devoted to investigating this issue.

Federated K-means clustering can be used to perform unsupervised learning (for a variable number of clusters between centers), also clustering on multiple datasets avoiding sharing the underlying data (Garst and Reinders, 2024). The idea of avoidance of local, sensitive data sharing is also used in the algorithm for Federated, Fair, and Fast K-means (F^3KM) (Zhu et al., 2023). With this approach, the K-means are efficiently solved in vertical FL. The overall problem is decomposed into multiple sub-problems that are solved at the level of single clients. Thus, the clients only transmit their results to the server rather than the original sensitive data. Clustering precision may be enhanced by exploiting local data stored on each client to help define better initial centroids (Yang et al., 2024). Defining the initial centroids directly at the server level results in the initial features being devoid of insights derived from the clients. Then, the server finds the initial global centroids, feeding a greedy algorithm with the client-based information. To improve the first step, the clients share the data point number for each centroid, allowing better partitioning. The problem of privacy and security in FL is addressed in (Pedrycz, 2022) using fuzzy logic, and in particular, a problem of unsupervised federating learning is solved via fuzzy sets based on federated clustering, named Fuzzy C-Means Federated (F-FCM). In (Pedrycz, 2022), an objective function is proposed. Clusters are formed via a partition function that assigns points to them, and a set of cluster prototypes is updated according to the exchanges between the server and clients. Another fuzzy approach is used for multistep federated clustering, validating clusters through the DB index (Stallmann and Wilbik, 2022).

An alternative methodology addressing the privacy concerns associated with FKM within proactive caching for cellular networks is put forth by (Liu et al., 2020). The base stations collect users' data, which may cause an information leakage. The solution of (Liu et al., 2020) uses privacy-preserving federated K-means and secret-sharing protocols. The passage of secret sharing is needed to avoid reconstructing users' data from the shared gradients uploaded to the macro-cell base station to update cluster centroids. Splitting the gradient into random shared, its reconstruction is only made possible when a minimal number of the shared is collected, thus guaranteeing the privacy of user data. Moreover, secure aggregation and homomorphic encryption strategies making K-means more robust concerning non-IID data (nonindependent and identically distributed) are addressed in (Brandão et al., 2021).

The current study introduces a novel K-means approach based on a global centroids similarity measure to improve the balance between maintaining privacy and achieving effective clustering. We start with the standard K-means clustering steps and add modifications to integrate them with federated learning. Additionally, our study explores the potential influence of the heterogeneous health data distribution on the convergence of local and global models using real-world healthcare datasets.

3 MATERIALS AND METHODS

3.1 Federated K-Means Approach

Beyond the aforementioned advantage concerning privacy issues, we focus on federated clusters for their multiple benefits. To name but a few, we mention their scalability efficiency (no need to store and process the whole dataset in the same location), data control (to remain in possession of individual clients), and communication efficiency (reduction of the amount of shared data). Other advantages include decreased large-scale data breaches, real-time latency reduction, and local pattern discovery. This last element is particularly relevant in medical applications. The investigation of heterogeneously-distributed medical datasets is, in fact, the primary motivation of our research.

The main steps of the federated K-means algorithm presented here are the following: In the first phase, (*Broadcast Parameters*), the central server provides the different clients with training parameters, including the number of clusters to set for each local model, the maximum number of iterations, and the centroid initialization method. After this phase, each client independently performs clustering by applying the K-means algorithm to its data. Each client transmits its set of centroids to the central server (*Collect Local Centroids*), randomly selecting a set of centroids from one of the n clients (*Random Selection of Centroid C^**), hereinafter referred to as C^*, indicating the initialization set of centroids to define the *global centroids*. Indeed, the *global centroids* definition needs an initialization of centroids' selections as the local K-means; the selection of C^* addresses this requirement. Then, for each centroid c belonging to C^*, the central server creates a set of centroids belonging to other clients closest to cosine similarity according to the following formula:

$$\forall c \in C^*, \quad S_c = \min_j(cosine(c, c_{ij})), \tag{1}$$

where C^* is the initial set of centroids for the definition of global centroids, c is the generic centroid belonging to C^*, and S_c is the set of the closer centroids to c belonging to C_j, with $C_j \neq C^*$.

After identifying for each centroid $c \in C^*$ the correspondence set S_c, the global centroids are computed as the average (*Calculate global centroids*). At this point, local centroids are evaluated and sent to local nodes (*Broadcast global centroids*), which repeat the local clustering until the rounds are completed.

3.2 Datasets

The following four publicly available datasets and a synthetic dataset are used to evaluate the FKM model:

- **OASIS2** (Battineni et al., 2019) is a dataset related to Alzheimer's disease (AD) containing patient sociodemographic characteristics and clinical variables. Size: 354 samples x 12 features.
- **Heart Disease** (Abid Ali Awan, 2021) is a dataset with features of heart disease patients. Size: 303 samples x 11 features.
- **Obesity** (Palechor and De la Hoz Manotas, 2019) is a dataset containing data to estimate obesity in patients according to their eating habits and physical condition. Size: 2111 samples x 17 features.
- **Breast Cancer** (W. H. Wolberg and Mangasarian, 2017) is a dataset of digitized images of fine needle aspirates from breast masses. Size: 569 samples x 30 features.

- **Custom** is a synthetic dataset. Size: 500 samples x 10 features.

3.3 Data Distribution Scenarios

The data contained in each dataset have been used to define three scenarios corresponding to three different data distributions. Specifically, we define a *uniform* distribution, where the data are uniformly distributed between clients; a *soft-heterogeneous* distribution, which exhibits a slight heterogeneity of data distribution; and finally, a *hard-heterogeneous* distribution, which is characterized by a pronounced heterogeneity of data distribution. Each scenario represents a different usage context for our model and will allow us to evaluate its performance under realistic and varied conditions. Figure 1 shows the architecture used for the Federated Clustering. It is delineated by two primary nodes: local nodes, also known as client nodes, and server nodes, which act as aggregators. Without loss of generality, we assume to work with three client nodes and one aggregator node.

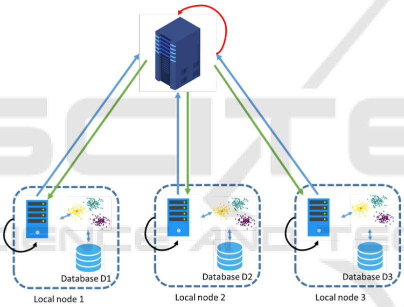

Figure 1: Federated architecture with 3 client nodes.

The clients emulate three entities where data are stored and are subject to mobility constraints to facilitate machine-learning training. The aggregator serves as the focal point for the aggregation process, receiving the local centroids generated by the clients during each round of communication.

In each of the three considered scenarios, each client stores a data portion according to a type of heterogeneous data distribution (HDD), as follows: *(i) Homogeneous* (H)- Each client has the same number of samples; *(ii) Soft-Heterogeneous* (SH)- Each sample in the dataset is randomly assigned to one of the 3 clients according to a uniform distribution; *(iii) Hard-Heterogeneous* (HH)- First, 80% of each dataset is equally distributed among all clients. Then, half of the clients are randomly selected to receive the remaining 20% of the data, again using a uniform distribution.

A sample allocation example is shown in Table 1.

Table 1: Client's samples for each dataset.

Client	Datatet	H	SH	HH
1	OASIS2	118	108	165
2	OASIS2	118	138	98
3	OASIS2	118	108	91
1	Heart Disease	101	105	146
2	Heart Disease	101	105	63
3	Heart Disease	101	93	94
1	Obesity	703	664	1003
2	Obesity	703	731	568
3	Obesity	703	716	540
1	Breast Cancer	189	176	238
2	Breast Cancer	189	166	135
3	Breast Cancer	189	158	127
1	Custom	166	173	257
2	Custom	166	201	173
3	Custom	166	195	139

It is worth introducing the hypothesis underlying our proposed approach. Since we are defining our use case in a federated environment with multiple clients, each client performs independent local clustering on its data. To compare clustering results across clients, we assume that the *probability distributions on the local clients are identical*. This assumption implies that the underlying probability distributions governing the data points are consistent across clients:

$$F_X(x) = F_Y(x) \quad \forall x \in X \quad \text{with} \quad X \neq Y, \tag{2}$$

where F is the probability distribution function, and X and Y are two generic clients. By assuming the same probability distribution, the comparison of the clusters produced by different clients becomes meaningful. Consequently, we can evaluate the effectiveness and consistency of the clustering algorithms employed across the distributed system.

4 COMPARISON BETWEEN CENTRALISED AND FEDERATED K-MEANS

We perform a comparative analysis from two perspectives to evaluate the proposed federated K-means approach. First, we compare the performance of the two approaches using widely known metrics for unsupervised techniques, assessing the quality and effectiveness of the clustering algorithms. Then, we compare the two approaches considering cluster composition and the similarity of items within clusters.

Experiments were conducted to evaluate different cluster configurations, with cluster numbers ranging from two to five. A random initialization was used alongside a max iteration parameter set at 100.

Table 2: Overview of the Performance Results - Hard Heterogeneous Distribution.

Dataset	Clusters	Approach	Silhouette	Calinski Harabasz	Davies Bouldin	Approach	Silhouette	Calinski Harabasz	Davies Bouldin
oasis2	2	Centralized	0.19	76.40	2.00	Federated	0.19	73.63	2.06
oasis2	3	Centralized	0.20	71.87	1.80	Federated	0.26	93.52	1.58
oasis2	4	Centralized	0.26	87.22	1.53	Federated	0.40	141.77	1.27
oasis2	5	Centralized	0.44	143.20	1.13	Federated	0.61	238.33	1.05
ObesityDataSet	2	Centralized	0.19	277.98	2.53	Federated	0.11	225.02	2.92
ObesityDataSet	3	Centralized	0.12	228.80	2.37	Federated	0.34	174.66	1.71
ObesityDataSet	4	Centralized	0.13	283.36	2.13	Federated	0.16	349.53	2.08
ObesityDataSet	5	Centralized	0.17	327.37	1.94	Federated	0.25	467.40	1.75
heart disease	2	Centralized	0.17	63.26	2.11	Federated	0.19	68.36	2.03
heart disease	3	Centralized	0.19	55.59	1.87	Federated	0.20	62.43	1.93
heart disease	4	Centralized	0.17	182.40	2.01	Federated	0.43	123.97	1.32
heart disease	5	Centralized	0.45	121.40	1.20	Federated	0.63	209.60	0.96
custom	2	Centralized	0.59	726.53	0.63	Federated	0.60	731.56	0.62
custom	3	Centralized	0.76	2992.18	0.35	Federated	0.56	1748.00	0.80
custom	4	Centralized	0.59	2343.33	0.84	Federated	0.66	2597.46	0.73
custom	5	Centralized	0.56	2344.11	0.78	Federated	0.68	2936.42	0.77
Breast Cancer	2	Centralized	0.34	267.69	1.32	Federated	0.30	260.51	1.33
Breast Cancer	3	Centralized	0.30	208.42	1.41	Federated	0.29	219.32	1.44
Breast Cancer	4	Centralized	0.32	212.07	1.20	Federated	0.32	220.81	1.29
Breast Cancer	5	Centralized	0.41	268.48	0.99	Federated	0.52	354.95	1.01

4.1 Performance Analysis

Tables 2, 3, and 4 provide an overview of the performance achieved with the centralized and federated approaches across three data distribution scenarios. Each row represents a separate run with different settings in terms of dataset and number of clusters, so each run was evaluated with both centralized and federated approaches. We consider here, as clustering-quality indexes, the Calinski-Harabasz (CH), the Silhouette, and the Davies-Bouldin ones, respectively. CH provides information on how well-separated and compact the clusters are. The lower the CH, the better. The silhouette index summarizes clusters' cohesion (proximity of points inside the same cluster) and separation (distance between clusters). Its values range between -1 and 1, where -1 indicates a misclassification. The higher the value, the better. Finally, the Davies-Bouldin (DB) index tells us how similar the clusters are, with intra-cluster dispersion and inter-cluster separation. In this case, the lower the value, the better. Focusing on hard-heterogeneous data distribution (Table 2), we notice an improvement of all the indices while considering the federated clustering method against the clustering performed classically. The improvement is still evident for data soft distribution (Table 4) and less noticeable for heterogeneous data distribution (Table 3).

The data reported in the tables represent averages calculated from multiple training runs. Performance was evaluated using three metrics: 1) the *Silhouette score* utilized to assess the cohesion and separation of clusters; it ranges from -1 to 1, with higher scores indicating better clustering performance. 2) the *Davies-Bouldin index* (DBI) that measures the ratio of within-cluster distances to between-cluster distances; a lower score indicates better clustering quality, and 3) the *Calinski-Harabasz index* (CHI) evaluates the ratio of between-cluster dispersion to within-cluster dispersion; a higher score indicates better-defined clusters.

From a high-level analysis, the results clearly show that the centralized approach does not consistently outperform the federated approach; on the contrary, federated approaches are proving to be highly competitive. Looking at the results in Table 2 for the hard distribution scenario, it is noticeable that federated approaches outperform centralized approaches in 14 out of 20 reported runs. Here, 'outperform' means that federated approaches give better results in at least 2 out of 3 metrics.

For H and SH scenarios (see tables 3 and 4), centralized approaches are more effective in 11 out of 20 runs and 13 out of 20 runs, respectively.

To further support the competitiveness of the approach, the average performance gap between the centralized and federated approaches is generally modest across all scenarios. Here, the gap is calculated as the average difference between the corresponding pairs of metrics for each run, considering

Table 3: Overview of the Performance Results - Homogeneous Distribution.

Dataset	Clusters	Approach	Silhouette	Calinski Harabasz	Davies Bouldin	Approach	Silhouette	Calinski Harabasz	Davies Bouldin
oasis2	2	Centralized	0.19	76.40	2.00	Federated	0.19	76.12	1.99
oasis2	3	Centralized	0.18	74.28	1.78	Federated	0.18	74.21	1.78
oasis2	4	Centralized	0.18	67.19	1.64	Federated	0.17	65.48	1.73
oasis2	5	Centralized	0.15	60.67	1.69	Federated	0.15	58.98	1.76
ObesityDataSet	2	Centralized	0.19	277.98	2.53	Federated	0.10	201.84	2.94
ObesityDataSet	3	Centralized	0.14	244.42	2.32	Federated	0.16	163.43	2.28
ObesityDataSet	4	Centralized	0.15	226.26	2.11	Federated	0.14	180.76	1.90
ObesityDataSet	5	Centralized	0.13	222.67	1.96	Federated	0.15	141.32	2.12
heart disease	2	Centralized	0.17	63.26	2.11	Federated	0.17	63.22	2.1
heart disease	3	Centralized	0.14	47.53	2.11	Federated	0.13	47.28	2.22
heart disease	4	Centralized	0.14	43.60	2.04	Federated	0.13	41.07	2.14
heart disease	5	Centralized	0.12	38.28	2.07	Federated	0.13	37.98	2.05
custom	2	Centralized	0.59	726.53	0.63	Federated	0.52	485.43	0.79
custom	3	Centralized	0.75	2918.89	0.36	Federated	0.67	2408.70	0.66
custom	4	Centralized	0.53	2048.49	1.39	Federated	0.54	2043.10	1.38
custom	5	Centralized	0.34	1618.97	1.95	Federated	0.45	1599.02	1.64
Breast Cancer	2	Centralized	0.34	267.69	1.32	Federated	0.35	267.44	1.30
Breast Cancer	3	Centralized	0.31	197.11	1.53	Federated	0.31	196.93	1.53
Breast Cancer	4	Centralized	0.27	158.68	1.51	Federated	0.20	156.09	1.76
Breast Cancer	5	Centralized	0.16	140.16	1.76	Federated	0.16	137.99	1.81

Table 4: Overview of the Performance Results - Soft Distribution.

Dataset	Clusters	Approach	Silhouette	Calinski Harabasz	Davies Bouldin	Approach	Silhouette	Calinski Harabasz	Davies Bouldin
oasis2	2	Centralized	0.19	76.40	2.00	Federated	0.19	76.37	2.00
oasis2	3	Centralized	0.18	74.28	1.78	Federated	0.19	74.22	1.77
oasis2	4	Centralized	0.18	67.19	1.64	Federated	0.17	65.00	1.70
oasis2	5	Centralized	0.15	60.67	1.69	Federated	0.16	59.16	1.72
ObesityDataSet	2	Centralized	0.19	277.98	2.53	Federated	0.19	277.76	2.53
ObesityDataSet	3	Centralized	0.14	244.42	2.32	Federated	0.15	228.32	2.33
ObesityDataSet	4	Centralized	0.1 5	226.26	2.11	Federated	0.12	223.13	2.21
ObesityDataSet	5	Centralized	0.13	222.67	1.96	Federated	0.15	213.95	2.07
heart disease	2	Centralized	0.17	63.26	2.11	Federated	0.17	63.24	2.11
heart disease	3	Centralized	0.14	47.53	2.11	Federated	0.14	47.18	2.25
heart disease	4	Centralized	0.14	43.60	2.04	Federated	0.14	42.63	2.04
heart disease	5	Centralized	0.12	38.28	2.07	Federated	0.13	38.35	2.03
custom	2	Centralized	0.59	726.53	0.63	Federated	0.56	625.57	0.69
custom	3	Centralized	0.75	2918.89	0.36	Federated	0.75	2918.89	0.36
custom	4	Centralized	0.53	2048.49	1.39	Federated	0.50	1597.72	1.45
custom	5	Centralized	0.34	1618.97	1.95	Federated	0.48	1594.48	1.54
Breast Cancer	2	Centralized	0.34	267.69	1.32	Federated	0.34	267.65	1.31
Breast Cancer	3	Centralized	0.31	197.11	1.53	Federated	0.32	196.47	1.54
Breast Cancer	4	Centralized	0.27	158.68	1.51	Federated	0.19	156.15	1.79
Breast Cancer	5	Centralized	0.16	140.16	1.76	Federated	0.16	137.18	1.79

all runs performed for each scenario. In fact, in Table 5, it is noticeable that the average gaps are small. For example, in the SH scenario, the average gap for the Silhouette is 0, indicating that even when the centralized approach performs better, the difference from the federated approach is minimal. From Table 5, we observe that, in the homogeneous case, the silhouette worst gap is circa 0, that is, the point is close to the decision boundary between the considered clusters. Moving from the homogeneous data distribution to the soft heterogeneous one, the silhouette worst gap is equal to 0 and slightly below 0 for the hard-

Table 5: Performance Gap Between Data Distribution Scenario.

Data Distribution	Silhouette Average Gap	Calinski-Harabasz Average Gap	Davies-Bouldin Average Gap	Silhouette Worst Gap	Calinski-Harabasz Worst Gap	Davies-Bouldin Worst Gap
Homogeneous	0.0085	53.63	-0.053	0.08	0.30	-2.58
Hard Heterogeneous	-0.04	-10.68	0.02	0.08	-6.84	-2.56
Soft Heterogeneous	0	30.78	-0.02	0	-0.06	-2.17

Table 6: Cluster similarity analysis.

	ARI	FMI
Alzheimer's Disease		
H-scenario	0.95	0.98
SH-scenario	1	1
HH-scenario	0.91	0.96
Heart Disease		
H-scenario	0.93	0.96
SH-scenario	1	1
HH-scenario	0.88	0.94
Obesity		
H-scenario	0.1	0.58
SH-scenario	0.96	0.98
HH-scenario	0.72	0.86
Breast Cancer		
H-scenario	0.94	0.97
SH-scenario	1	1
HH-scenario	0.94	0.97
Custom		
H-scenario	1	1
SH-scenario	1	1
HH-scenario	1	1

Table 7: Clusters composition of Alzheimer's patients obtained with the Centralised approach. EDUC- Education level; SES - Socio-Economic Status; eTIV - estimated Total Intracranial Volume; Normalized Whole Brain Volume - nWBV; MMSE - Mini-Mental State Examination; CDR - Clinical Dementia Rating.

	Cluster 1	Cluster 2
Age		
Mean (Std)	76.29 ± 7.3	77.46± 8.0
Range	[61, 98]	[60, 97]
Sex		
Female # (%)	51 (39.53%)	153 (68%)
Male # (%)	78 (60.47%)	72 (32%)
EDUC (years)		
Mean (Std)	13.79 ± 3.01	15.23 ± 2.69
Range	[6, 20]	[8, 23]
SES		
Mean (Std)	2.79 ± 1.19	2.27 ± 1.05
Range	[1, 5]	[1, 5]
eTIV		
Mean (Std)	1494.34 ± 172.84	1487.5 ± 176.99
Range	[1143.0, 1957.0]	[1106.0, 2004.0]
nWBV		
Mean (Std)	0.71 ± 0.03	0.74 ± 0.04
Range	[0.65, 0.81]	[0.64, 0.84]
MMSE		
Mean (Std)	24.39 ± 4.63)	29.14 (1.04)
Range	[4, 30]	[24, 30]
CDR		
Mean (Std)	0.67 ±0.31	0.04 ± 0.14
Range	[0, 2]	[0, 0.5]
Diagnosis		
CN	2 (1.6%)	188 (83.6%)
AD	127 (98.4%)	0 (0%)
Converted	0 (0%)	37 (16.4%)

heterogeneous case, thus showing a certain stability with respect to the passage toward the inhomogeneity. A higher instability is shown by the average gap in the Calinski-Harabasz index, from soft to hard homogeneous. Finally, the Davies-Bouldin index presents only a small clustering-quality diminution concerning the worst gap, even between homogeneous and hard-heterogeneous data distribution. It's also interesting to note that the average gap value for the Calinski-Harabasz metric ranges from -0.053 to 0.02, indicating that the clusters are well separated with low intra-cluster variance and high inter-cluster variance.

In addition, Table 5 also reports the gap between the worst-performing runs among all runs for each data distribution. Even in these cases, the gaps for the Silhouette and Calinski-Harabasz metrics remain small. However, the Davies-Bouldin gap is larger than the average gap.

These results aim to demonstrate the effectiveness of the proposed approach by comparing it with traditional centralized K-means. The performance from a more general perspective provides insight into the results in terms of the evaluation metrics. Thus, the evaluation of the average and worst gaps highlights the approach's consistency, showing that even in average and worst-case scenarios, the proposed federated approach converges to the centralized one.

4.2 Similarity Clusters Analysis

To determine the quality of the cluster composition generated by the federated approach, we compare the similarity between the clusters identified by the cen-

Table 8: Clusters of AD patients obtained with the federated approach under homogeneous distribution settings.

	Cluster 1	**Cluster 2**
Age		
Mean (Std)	76.53 ± 7.21	77.31 ± 89
Range	[61, 98]	[60, 97]
Sex		
Female # (%)	49 (39.2%)	155 (67.69%)
Male # (%)	76 (60.8) %	74 (32.31%)
EDUC		
Mean (Std)	13.69 ± 2.99	15.26 ± 2.68
Range	[6, 20]	[8, 23]
SES		
Mean (Std)	2.85 ± 1.17	2.25 ± 15
Range	[1, 5]	[1, 5]
eTIV		
Mean (Std)	1492.74 ± 170	1488.49 ± 178.44
Range	[1143 1957]	[1106 2004]
nWBV		
Mean (Std)	0.71 ± 03	0.74 ± 04
Range	[0.65, 0.81]	[0.64, 0.84]
MMSE		
Mean (Std)	24.25 ± 4.63	29.14 ± 14
Range	[4, 30]	[24, 30]
CDR		
Mean ± Std	0.67 ± 0.31	05 ± 0.15
Range	[0, 2]	[0, 0.5]
Diagnosis		
CN	2 (1.6%)	188 (82.1%)
AD	123 (98.4%)	4 (1.7%)
Converted	0 (0%)	37 (16.2%)

Table 9: Clusters of AD patients obtained with the federated approach under soft distribution settings.

	Cluster 1	**Cluster 2**
Age		
Mean (Std)	76.29 ± 7.3	77.46± 8.0
Range	[61, 98]	[60, 97]
Sex		
Female # (%)	51 (39.53%)	153 (68%)
Male # (%)	78 (60.47%)	72 (32%)
EDUC		
Mean (Std)	13.79 ± 3.01	15.23 ± 2.69
Range	[6, 20]	[8, 23]
SES		
Mean (Std)	2.79 ± 1.19)	2.27 ± 1.05
Range	[1, 5]	[1, 5]
eTIV		
Mean (Std)	1494.34 ± 172.84	1487.5 ± 176.99
Range	[1143.0, 1957.0]	[1106.0, 2004.0]
nWBV		
Mean (Std)	0.71 ± 0.03	0.74 ± 0.04
Range	[0.65, 0.81]	[0.64, 0.84]
MMSE		
Mean (Std)	24.39 ± 4.63)	29.14 (1.04)
Range	[4, 30]	[24, 30]
CDR		
Mean (Std)	0.67 ±0.31	0.04 ± 0.14
Range	[0, 2]	[0, 0.5]
Diagnosis		
CN	2 (1.6%)	188 (83.6%)
AD	127 (98.4%)	0 (0 %)
Converted	0 (0%)	37 (16.4%)

tralized approach and those identified by the federated one using two similarity measures: 1) *Adjusted Rand Index (ARI)*, while Rand Index (RI) measures cluster similarity through the percentage of consistent decisions between two clustering, the Adjusted Rand Index (ARI) corrects the RI by the chance grouping of elements, providing more robust statistics for comparing different clustering algorithms or methods; 2) *Fowlkes–Mallows index* (FMI) is a metric used to evaluate the clusters similarity obtained through various clustering algorithms. It is typically used to evaluate the clustering performance of a specific algorithm by assuming that the obtained cluster is compared to the ground truth–i.e., the perfect cluster.

To adopt such indexes, we assume that the results of the centralised clustering approach are the benchmark clusters to be compared.

The indexes obtained from the comparison between the centralised approach and the federated approach in three different scenarios and for each dataset are presented in Table 6. It can be noted that significant heterogeneity within the data may result in diminished cluster composition quality in federated k-means, analogously to the case of traditional supervised federated-learning methodologies. Nevertheless, a noteworthy outcome indicates that soft heterogeneity provides advantages in federated settings. This specific finding manifests with greater prominence in the Obesity dataset, which encompasses a fourfold greater number of samples than the Breast Cancer and Custom datasets, six times more than the Oasis2 and Heart Disease datasets. Indeed, the analysis reveals that the ARI for the Obesity dataset displays a notably low performance in the homogeneous scenario. At the same time, there is a notable increase in the SH scenario and a somewhat lesser increase in the HH scenario.

Finally, to further demonstrate the efficacy of the federated approach in generating comparative outcomes with the classical centralized K-means method, we present a descriptive analysis of the clusters identified through both approaches across three specified scenarios using the OASIS2 dataset related to Alzheimer's disease. Such a dataset was selected based on its unique provision of patients' diagnoses, which allows for a more practical demonstration of similarities and differences among clusters compared to other available datasets.

The clusters resulting from the centralized K-means approach and the federated one with homoge-

Table 10: Clusters of AD patients using the federated approach under Hard Heterogeneous distribution settings.

	Cluster 1	**Cluster 2**
Age		
Mean (Std)	76.94 ± 8.34	79.26 ± 8.21
Range	[66.0 , 98.0]	[60.0 , 97.0]
Sex		
Female # (%)	45 (40.18 %)	161 (66.53 %)
Male # (%)	67 (59.82 %)	81 (33.47 %)
EDUC		
Mean (Std)	13.16 ± 3.14	15.1 ± 2.76
Range	[6.0 , 20.0]	[8.0 , 23.0]
SES		
Mean (Std)	2.96 ± 1.16	2.34 ± 1.09
Range	[1.0 , 5.0]	[1.0 , 5.0]
eTIV		
Mean ± SD	1465.46 ± 154.55	1489.93 ± 199.58
Range	[1143.0 , 1911.0]	[1154.0 , 2004.0]
nWBV		
Mean (Std)	0.71 ± 0.04	0.73 ± 0.04
Range	[0.65 , 0.81]	[0.66 , 0.84]
MMSE		
Mean (Std)	23.04 ± 5.18	28.95 ± 1.17
Range	[4.0 , 30.0]	[24.0 , 30.0]
CDR		
Mean (Std)	0.67 ± 0.29	0.06 ± 0.16
Range	[0.5 , 2.0]	[0.0 , 0.5]
Diagnosis		
CN	0 (0%)	201 (83.1%)
AD	112 (100%)	3 (1.2%)
Converted	0 (0%)	38 (15.7%)

neous data distribution are shown in Tables 7 and 8.

The clusters identified by the proposed federated K-means closely resemble those identified by the classical K-means, as evidenced by the indexes presented in Table 6. The federated K-means algorithm successfully clustered 123 individuals with Alzheimer's disease, while the classical K-means algorithm clustered 127 individuals with the same condition. Both methods categorized an equal number of cognitively normal subjects and subjects likely to experience a conversion into cluster #2.

Moreover, Tables 9 and 10 show the clusters detected by the federated approach under heterogeneous data distribution, Soft-Heterogeneous and Hard-Heterogeneous distribution, respectively. As observed, in cases of a soft distribution condition, the federated approach demonstrates performance comparable to that of the traditional K-means algorithm. Any slight discrepancies identified in the preceding condition are mitigated, as evidenced by the indexes presented in Table 6. On the contrary, when the data distribution is heavily unbalanced under the previous conditions, the discriminatory power of the federated K-means decreases.

5 CONCLUSIONS

Federated Learning (FL) represents an emerging trend in machine learning. This approach allows for the development of a global model without sharing private data distributed among multiple data owners. Significant research in the field of FL has primarily focused on applying supervised learning techniques. On the contrary, the available literature about federated unsupervised learning is still limited.

In particular, among unsupervised techniques, clustering has shown numerous beneficial applications in the healthcare domain. Hence, the advancement of federated clustering has the potential to address certain limitations associated with data usage, particularly privacy concerns, which currently hinder the full realization of the vast potential of health data-clustering approaches. Moreover, the performance of Federated Learning is found to be satisfactory in scenarios where data are independent and identically distributed (IID). Conversely, in cases where data are non-independent and non-identically distributed (Non-IID), it becomes challenging to effectively train a machine learning algorithm that relies on global measures while ensuring that all data remain local.

The current study presents a new federated K-means clustering framework based on a global cosine-similarity measure to enhance the trade-off between privacy preservation and clustering effectiveness. Furthermore, our investigation explores the impact of the heterogeneous health data-distribution hypothesis on the convergence of both local and global models.

The efficacy of the proposed federated K-means algorithm has been evaluated across diverse healthcare datasets and under various experimental conditions, with a comparative analysis of its performance against the centralized K-means algorithm. According to the initial results, while dealing with heterogeneous distributions of health data, the difference between the centralized and federated approaches is minimal, and the federated approach demonstrates promising potential. The decentralized approach exhibits superior performance when evaluated on certain healthcare datasets compared to the centralised one. Finally, the results of this study provide empirical evidence that diversity within federated networks can have a beneficial effect on the overall quality of cluster composition.

Future developments of this research can also address the challenges associated with heterogeneous data distributions in federated learning, for instance, divergence of local models, and slow or unstable convergence, or class imbalance. The detailed exploration of solutions to undertake in these cases can strengthen the proposed clustering approach.

Future work will focus on evaluating the approach on larger datasets while removing the assumption of equal probability-distributions among clients.

ACKNOWLEDGEMENTS

This paper was developed within the project funded by Next Generation EU - "Age-It - Ageing well in an ageing society" project (PE0000015), National Recovery and Resilience Plan (NRRP) - PE8 - Mission 4, C2, Intervention 1.3. The views and opinions expressed are only of the authors and do not necessarily reflect those of the European Union or the European Commission.

REFERENCES

Abid Ali Awan (2021). Heart Disease patients - Targeting treatment for heart disease patients. URL: https://www.kaggle.com/datasets/kingabzpro/heart-disease-patients.

Act, A. (1996). Health insurance portability and accountability act of 1996. *Public law*, 104:191.

Battineni, G., Chintalapudi, N., and Amenta, F. (2019). Data for: *Machine learning in medicine: performance calculation of dementia prediction by support vector machines (SVM). Mendeley Data.*

Bharati, S., Mondal, M. R. H., Podder, P., and Prasath, V. S. (2022). Federated learning: Applications, challenges and future directions.

Bonawitz, K., Kairouz, P., McMahan, B., and Ramage, D. R. (2021). Federated learning and privacy. *ACM Queue*, 19.

Brandão, A., Mendes, R., and Vilela, J. P. (2021). Efficient Privacy Preserving Distributed K-Means for Non-IID Data. In Abreu, P. E., Pereira Rodrigues, P., Fernández, A., and Gama, J., editors, *Advances in Intelligent Data Analysis XIX, 19th International Symposium on Intelligent Data Analysis, IDA 2021*, volume 12695, pages 439–451. Springer.

Dhade, P. and Shirke, P. (2024). Federated learning for healthcare: A comprehensive review. *Engineering Proceedings*, 59(1):230.

Garst, S. and Reinders, M. (2024). Federated K-Means Clustering. *ArXiV preprint, arXiv:2310.01195v2*.

Liu, Y., Ma, Z., Yan, Z., Wang, Z., Liu, X., and a, J. (2020). Privacy-preserving federated k-means for proactive caching in next generation cellular networks. *Information Sciences*, 521:14–31.

Marulli, F., Balzanella, A., Campanile, L., Iacono, M., and Mastroianni, M. (2021a). Exploring a federated learning approach to enhance authorship attribution of misleading information from heterogeneous sources. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Marulli, F., Verde, L., Marrone, S., Barone, R., and De Biase, M. S. (2021b). Evaluating efficiency and effectiveness of federated learning approaches in knowledge extraction tasks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6.

Palechor, F. M. and De la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico. *Data in brief*, 25:104344.

Pedrycz, W. (2022). Federated FCM: Clustering Under Privacy Requirements. *IEEE Transactions on Fuzzy Systems*, 30(8):3384–3388.

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., et al. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):12598.

Stallmann, M. and Wilbik, A. (2022). On a Framework for Federated Cluster Analysis. *Applied Sciences*, 12(10455).

Voigt, P. and Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

W. H. Wolberg, W. N. S. and Mangasarian, O. L. (2017). Breast cancer Wisconsin (diagnostic) data set. URL: https://www.kaggle.com/datasets/nancyalaswad90/breast-cancer-dataset.

Wahab, O. A., Mourad, A., Otrok, H., and Taleb, T. (2021). Federated machine learning: Survey, multi-level classification, desirable criteria and future directions in communication and networking systems. *IEEE Communications Surveys Tutorials*, 23(2):1342–1397.

Yang, K., Mohammadi Amiri, M., and Kulkarni, S. R. (2024). Greedy centroid initialization for federated K -means. *Knowledge and Information Systems*.

Zhu, S., Xu, Q., Zeng, J., Wang, S., Sun, Y., Yang, Z., and Oeng, Z. (2023). F^{3}KM: Federated, Fair, and Fast *k*-means. *Proc. ACM Manag. Data*, 1(4):241.