JOURNAL OF DATA SCIENCE | Vol.2026, Issue 3, No.16
eISSN:2805-5160

Privacy-Preserving Data Mining Through Federated Clustering for Distributed Healthcare Data

Iwan Abadi[1]*, Nanang Fakhrur Rozi[2], Timothy Joshua M. Ventayen[3]

[1]Universitas Langlangbuana, Indonesia
[2]Informatics Engineering Department, Institut Teknologi Adhi Tama Surabaya, East Java, Indonesia
[3]ICT Management Office, Pangasinan State University, Philippines

Email: iwan.abadi69@gmail.com[1]*, nanang@itats.ac.id[2], tventayen@psu.edu.ph[3]

Abstract

Healthcare institutions often face difficulties in conducting collaborative data mining because patient-related records are sensitive and cannot be freely exchanged across organizations. Conventional centralized clustering requires raw data from multiple institutions to be pooled into a single repository, which may create privacy, ethical, and governance risks. This study proposes a privacy-preserving data mining framework based on federated clustering for distributed healthcare data analysis. In the proposed approach, each participating institution performs local preprocessing and clustering using its own data partition, while only cluster-level representations are shared for global aggregation. A benchmark-style structured healthcare dataset was partitioned across multiple institutional nodes to evaluate the proposed federated clustering framework under controlled distributed conditions. Clustering quality was evaluated using Silhouette Score, Calinski-Harabasz Index, and cluster consistency. The experimental results show that federated clustering achieved performance close to centralized clustering while avoiding the exchange of raw patient-level records. The proposed framework provides a practical approach for collaborative healthcare analytics under privacy-sensitive conditions and contributes to privacy-preserving data mining research in distributed medical systems.

Keywords

Privacy-preserving data mining; Federated learning; Federated clustering; Healthcare analytics; Public health

Introduction

Gathering, integrating, and mining diverse healthcare datasets is a powerful method to capture emerging patient trends and service patterns, and develop clinically relevant classifications. Data generated or collected by hospitals, outpatient clinics, and healthcare laboratories and

Submission: 20 September 2026; **Revised:** 15 October 2026 **Acceptance:** 13 November 2026; **Available Online:** November 2026
Copyright: © 2026. All the authors listed in this paper. The distribution, reproduction, and any other usage of the content of this paper is permitted, with credit given to all the author(s) and copyright owner(s) in accordance to common academic practice. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license, as stated in the website: https://creativecommons.org/licenses/by/4.0/

organizations contain a variety of structured or semi-structured data, such as demographic data, laboratory data, clinical data, treatment data, and administrative service data. These data can be leveraged to perform patient segmentation, analyze disease risk, plan health services, and make informed decisions on healthcare services. Among data mining techniques, clustering is applicable as it can recognize groups of records in the form of clusters with no prior groups assigned.

Healthcare data mining, especially for the purpose of collaboration, is particularly difficult due to the sensitive nature of patient data. Healthcare organizations are often bound by laws and regulations governing patient privacy, as well as their ethical obligations, internal governance policies, and the commitment to patient confidentiality. Centralized clustering, in its traditional sense, requires the data of disparate organizations to be moved to a single data repository for the purpose of conducting analyses. While this method may enhance the thoroughness of an analysis, it may also elevate the potential for a breach of privacy, unintentional exposure of data, and harm to patient data.

Despite advances in distributed data mining, many cluster analysis studies still assume data centralization. (Muñoz-Gama et al., 2022) Such an assumption can be unrealistic in actual healthcare environments, as data may be distributed across several institutions. Each institution has only partial views of the patient populations, and it may be illegal or unethical to share data. Therefore, to address the stated challenges, distributed clustering approaches that ensure privacy are needed. Recent advances in federated learning directed toward the detection of thoracic pathologies show that distributed analysis of medical data is feasible and significantly reduces the need for the data to be centralized (Mahmood et al., 2025).

The objective of this research is to investigate and implement a framework for the analysis of distributed healthcare data using privacy-preserving data mining techniques based on federated clustering. (Tao et al., 2026) The framework provides the capability to the healthcare institutions to perform clustering at their location and share only the clusters created for the purpose of aggregation at the global level. The analysis presents the following contributions: creating a collaborative healthcare data mining framework based on federated clustering; constructing a controlled distributed healthcare environment by partitioning the dataset across institutional nodes; providing an assessment of the clustering results based on the Silhouette Score and the Calinski-Harabasz Index as well as the consistency of the clusters; and providing an assessment of the privacy-utility trade-off based on the comparison of federated clustering and both centralized and local clustering. (Dwivedi et al., 2023)

The remainder of this document is structured as follows. Research design and experimental framework, including segmentation of data and central clustering benchmarks, algorithms and their metrics, and the associated methods are described in Section 2. The paper concludes by developing an agenda for future research in Section 2. (Yang et al., 2023) Section 3 focuses on results and discussions with respect to clustering and its limitations, centralized clustering, privacy valuation, and healthcare. (Murray et al., 2022)

Privacy-preserving clustering assumes even greater importance given the need for advanced healthcare collaborations. Many healthcare challenges cannot be addressed by relying on data from one healthcare institution. Patient characteristics, service patterns, and disease distributions can vary across different providers and across different regions. Depending on its location, specialization, and referral network, a hospital may only have a specific subset of patients. Healthcare service patterns may only be discerned with a unified analysis of multiple healthcare institutions. The analysis, however, must be conducted within the framework that protects the integrity and confidentiality of the data.

Given the requirements of this situation, federated clustering is particularly appropriate, as it combines the separation of data ownership and the process of knowledge discovery. Each healthcare institution is responsible for its own data, while the process of knowledge discovery involves the collaborative effort to interpret shared patterns in the outputs of locally trained models. This separation is particularly pertinent in healthcare datasets, as data governance is not only a technical issue, but also an ethical issue and an organizational burden. A collaborative research framework that permits participation without the transfer of raw data may foster the needed trust to remove the impediments to collaborative healthcare research. (Rehm et al., 2021)

The focus of this study is on the combination of healthcare analytics, distributed data mining, and privacy-preserving computation. This study differs from the typical approach of considering privacy as an add-on after the data analysis, in that privacy preservation is embedded in the design of the clustering workflow. It is believed that significant patterns of patient groups can be uncovered without the need to centralize the raw data. This assumption was evaluated by comparing centralized, local-only, and federated clustering under the same analytical conditions

Methodology

Research Design

An experimental data mining design was used to evaluate federated clustering in a controlled distributed healthcare setting. Federated learning was the privacy-preserving methodology, which entailed the aggregation of local model results without the transfer of the raw data. Three methodologies were compared: centralized clustering, local-only clustering, and federated clustering. Centralized clustering was the traditional methodology, which involved combining all institutional data partitions prior to performing the clustering analysis. Local-only clustering was the methodology that entailed independent clustering done by each institution (Zhang et al., 2024).

The comparison examined whether federated clustering could approximate centralized clustering in terms of clustering quality while maintaining data locality. The centralized model was utilized as the analytical benchmark, as it was the only model that had full access to the clustering data. The local-only model was utilized to demonstrate the shortcoming of independent institutional clustering. (Ali et al., 2023) The federated model was used to demonstrate the optimal compromise of clustering quality against privacy preservation. (Abbas et al., 2024)

Dataset and Data Partitioning

A benchmark-style structured healthcare dataset was used to evaluate distributed healthcare data mining under controlled federated conditions. The dataset contained numerical indicators related to patients' age, body mass index, blood pressure, glucose and cholesterol levels, and length of stay. Data privacy was maintained, as these variables cover general demographic, physiological, and metabolic data and service-related data. The dataset was partitioned into five institutional nodes representing heterogeneous healthcare data holders.

Each institution maintained its local dataset. Within the federated framework, unprocessed data were not transferred between the participating institutions, nor were they sent to the aggregation server. The partitioning strategy was designed to represent heterogeneous institutional

data distributions commonly encountered in distributed healthcare analytics. Table 1 outlines the simulated dataset design, partitioning strategy, and the features that characterize the dataset.

Table 1. Dataset and Distributed Partition Description

Dataset	Number of Records	Number of Features	Example Feature Types	Number of Simulated Institutions	Partitioning Strategy	Clustering Target or Analytical Purpose
Benchmark-style structured healthcare dataset	5,000	10	Age, BMI, blood pressure, glucose level, cholesterol level, length of stay, and other numerical health indicators	5	Heterogeneous institutional partitioning	Discover patient profile groups and distributed healthcare patterns

The partitioning strategy makes it possible to assess research questions related to the discovery of useful cluster structures in the absence of raw healthcare data and centralized data control. (Cuomo et al., 2022) The design seeks to address the more common situation that exists within the healthcare domain, where organizations are often compelled to analyze data within their own ecosystem to safeguard sensitive healthcare data.

The distributed partitioning strategy was designed to represent decentralized healthcare data ownership across multiple institutional nodes. In the healthcare context, most organizations build and manage their own databases, and there may/or may not be consistent data management systems. Even if similar data are collected pertaining to patients, institutional variation may lead to differences in distribution. For example, one hospital may have a larger patient volume of elderly patients, while another hospital may have a larger volume of patients with chronic conditions and diseases. This variation introduces even more complexity to distributed healthcare data mining than to centralized data mining.

The heterogeneous partitioning design enables evaluation of the effect of uneven data distributions on clustering stability. Having equal distribution across local partitions would make a federated environment simpler but more artificial. Using heterogeneous partitions, this study evaluates a setting in which institutional nodes contribute different local data patterns to the global clustering process. This approach is relevant because the federated clustering design should be applied to situations as close as possible to real-world, multi-organization healthcare scenarios.

Although the dataset is benchmark-style and used under controlled conditions, the experimental design reflects key characteristics of privacy-sensitive distributed healthcare analytics. The dataset is intended to support methodological evaluation of federated clustering rather than to represent the full diversity of real clinical populations. This method creates an opportunity to analyze the quality of clustering within the bounds of data locality and the privacy-utility spectrum to offer a foundation for potential real-world application.

Proposed Federated Clustering Framework

For an overview of the proposed privacy-preserving mechanism, the federated clustering process is illustrated as a sequential workflow. Local healthcare institutions are responsible for the

preprocessing and clustering of their data, and only the cluster-level representations are sent for global aggregation. This configuration shows the main idea of the framework that allows collaborative healthcare data mining without exchanging the raw data of individual patients. Figure 1 shows the workflow of the proposed federated clustering framework for distributed healthcare data.

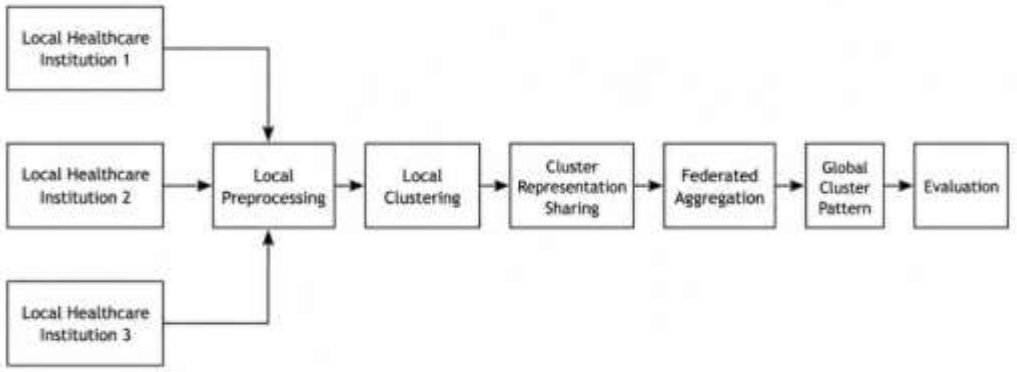


Figure 1. Proposed Federated Clustering Framework for Distributed Healthcare Data

In Figure 1, collaborative pattern discovery occurs using cluster-level representations, and the local institutions retain their raw data of individual patients. This design minimizes the privacy-related challenges of centralized data pooling and facilitates collaborative healthcare analytics within the framework of institutional regulations on data sharing.

Centralized Baseline

Centralized clustering was the main baseline for comparison. In the centralized approach, all institutional data segments were merged into a single, complete dataset prior to clustering. This scenario represents the reference condition with full access to all records, without data-locality constraints. However, centralized clustering requires raw data pooling and therefore introduces greater privacy and governance exposure than distributed alternatives.

Through the comparison of centralized and federated clustering, this study also aims to assess the trade-off in the quality of clustering when raw data are withheld. A small performance difference between the two approaches indicates that federated clustering can preserve much of the clustering utility while maintaining data locality. (Long et al., 2023)

Clustering Algorithm

K-Means clustering was the chosen algorithm for this study. It is the algorithm of choice for benchmark clustering studies and is distributed clustering-friendly, as it can use local cluster centroids as a compact representation for aggregation (Pang et al., 2024). It is simple to use, interpretable, and is computationally efficient.

K-Means was trained by each institution on its own local data. The local outputs of this clustering algorithm, the cluster centroids and the cluster sizes, were used for the aggregation step. The four clusters represent latent patient-profile groupings within the benchmark healthcare dataset.

The aggregation strategy used in this study was weighted centroid aggregation. After local clustering was completed, each institution sent only its local cluster centroid and the corresponding cluster size to the aggregation process. The global centroid was calculated by weighting each local centroid according to the number of records assigned to that cluster. In simple terms, a centroid produced from a larger local cluster contributed more strongly to the global centroid than a centroid produced from a smaller local cluster. The global centroid for cluster j can be expressed as

$$C_j^{global} = \frac{\sum_{k=1}^{K} n_{kj} C_{kj}}{\sum_{k=1}^{K} n_{kj}},$$

where C_{kj} represents the local centroid of cluster j from institution k, and nkj represents the number of records assigned to that cluster. This strategy allows the federated model to construct a global cluster representation while avoiding the transfer of raw patient-level records.

Considering the need for interpretability in healthcare analytics, K-Means clustering offers a significant advantage. Data mining outcomes need to be understood by both technical and non-technical personnel. Because cluster centroids signify the representative patient group and are less abstract than other representations, they show latent patient groups in more complex forms. This is advantageous for exploratory analysis, institutional reports, and early-stage decision support.

K-Means clustering also has the advantage of low communication needs in a federated setting. Because this clustering algorithm uses a centroid-based summarization, it requires each institution to only transmit a small number of mathematical representations. This is advantageous in a simple federated clustering design and is even more advantageous than other clustering techniques that require a high volume of data and model parameters.

Although the use of K-Means has its advantages, it demonstrates some limitations. (Dwivedi et al., 2023) K-Means relies on the concept of compact and well-separated clusters. Some healthcare datasets may have overlapping or irregular clusters that cannot be represented by distance-based or centroid-based clustering. For this reason, this research uses K-Means as the first interpretable baseline for federated clustering. (Δρίτσας & Τρίγκα, 2025) The future iterations of this research can expand the proposed framework to accommodate different clustering techniques. (Wang et al., 2024)

Evaluation Metrics

Clustering performance was evaluated using three primary metrics: Silhouette Score, Calinski-Harabasz Index, and cluster consistency. The Silhouette Score measures the compactness and separation of the clusters. A greater Silhouette Score suggests that a record was assigned to a cluster that is more compact and separate from the others (Ziakis & Vlachopoulou, 2023).

The Calinski-Harabasz Index is the comparison of the dispersion measured between the clusters and that measured within the clusters. Better cluster formation is indicated by higher values of this index. Cluster consistency measures the similarity or the stability of the cluster assignments between local/federated clustering and the centralized clustering. This metric is beneficial in assessing if the federated approach can come close to the centralized clustering in the absence of pooled raw data.

Experimental Procedure

The experimental methodology was constructed in such a way that centralized, local-only, and federated clustering approaches would be subjected to similar analytical conditions. After the preprocessing phase, the healthcare dataset was divided into five heterogeneous institutional partitions. This partitioning represented decentralized data ownership across multiple institutional nodes. Each partition performed clustering independently, and in the federated environment, clustering was aggregated at the illustration of the clusters to form a global clustering output. The centralized approach was implemented on the complete dataset. The local-only approach represented clustering as performed within an isolated institutional setting. Table 2 summarizes the main elements of the experimental setup by describing the configuration of each component, the setting of the implementation, and the purpose of the component.

Table 2. Experimental Configuration

Component	Configuration	Purpose
Dataset partitioning	Five simulated healthcare institutions	Represent distributed healthcare data ownership
Number of institutions	5 local clients	Simulate multi-institutional collaboration
Clustering algorithm	K-Means clustering	Provide simple and interpretable clustering
Aggregation strategy	Weighted centroid aggregation	Construct global cluster representation from local clusters
Evaluation metrics	Silhouette Score, Calinski-Harabasz Index, cluster consistency	Evaluate clustering quality and stability
Baseline model	Centralized clustering and local-only clustering	Compare privacy-utility trade-off

As shown in Table 2, the experiment was constructed so that centralized clustering, local-only clustering, and federated clustering would be subjected to the same analytical conditions. The same clustering methodology and assessment criteria were used for the comparison to be justified.

Results and Discussion

Federated Clustering Performance

The results in Table 3 compare the three clustering approaches under controlled distributed conditions. The same dataset structure, clustering algorithm, and evaluation metrics were applied to all approaches. Centralized clustering was used as the benchmark, as it had access to the complete dataset. It therefore represented the most thorough analytical environment in the absence of privacy constraints. The local-only and federated approaches, then, were benchmarked against it to evaluate the level of clustering quality that could be preserved when the raw data remained at local institutions.

Table 3 presents clustering quality comparisons for centralized, local-only, and federated clustering approaches to the application of unsupervised learning evaluation metrics.

Table 3. Clustering Performance Comparison

Method	Silhouette Score	Calinski-Harabasz Index	Cluster Consistency	Privacy Preservation
Centralized clustering	0.524	1876.42	1.000	Low
Local-only clustering	0.471	1512.36	0.781	High
Federated clustering	0.502	1764.88	0.918	High

The results in the table help to assess whether the federated approach can safeguard raw healthcare data while still achieving clustering of a quality that is deemed acceptable. (Joshi et al., 2022) Centralized clustering obtained the best results for the Silhouette Score and Calinski-Harabasz Index, as it had the complete dataset. Centralized clustering benefits from complete access to the records across all institutional partitions.

The federated approach to clustering obtained a Silhouette Score of 0.502 and a Calinski-Harabasz Index of 1764.88 compared to a score of 0.524 and an index of 1876.42 for the centralized approach. The results indicate that the federated clustering approach maintained a considerable amount of the global cluster structure in the absence of pooled data, and a cluster consistency score of 0.918 indicates that the federated clustering was very similar to the centralized clustering.

Local-only clustering performed the worst among the different clustering methods, yielding a Silhouette Score of 0.471 and a cluster consistency of 0.781. This indicates that closer clustering at the individual institutions is probably not adequate for examining the higher-order structures of the grouping of patients. (Luo et al., 2023) Each institution has only local access to its records. Consequently, local-only clustering may produce fragmentary structures that are meaningless in the context of the total distributed healthcare population.

Comparison with Centralized Clustering

When compared to centralized clustering, it can be stated that within the pure metric scope, federated clustering does not outperform centralized clustering (Zhao et al., 2026). Centralized clustering, given its access to the complete dataset, may score higher. This difference reflects the trade-off between centralized analytical access and preservation of data locality. In operational healthcare settings, collecting and consolidating raw patient data into a singular repository may be in violation of governing standards and of confidentiality and privacy concerns.

Within this context, federated clustering presents a privacy tradeoff. Its performance was only marginally worse than that of centralized clustering, meaning that important patterns may be preserved within local clusters. The federated approach, through the aggregation of local centroids, was able to define a global clustering space without the need to access the local datasets.

Within the field of healthcare analytics, a privacy tradeoff becomes even more evident. The clustering configuration with the strongest numerical performance may not be the most practical option under privacy, governance, and data-sharing constraints. A clustering solution that requires raw data sharing, even if its evaluation score is higher, may be practically unfeasible to implement. Within privacy-sensitive frameworks, federated clustering offers a more feasible solution as it allows for collaborative data analysis while preserving the data integrity of the participating institutions.

A minor decline in performance when comparing centralized to federated clustering should be understood as the analytic cost associated with maintaining data locality. (Liu et al., 2021) Centralized clustering allows the algorithm to interactively observe every record and therefore can manipulate cluster boundaries knowing the full data distribution. Federated clustering, on the other hand, works with a summarized and restricted local representation. During the summarization process, some of the local information is lost, which is the reason for the slight superiority of centralized clustering in the federated clustering metrics.

Nevertheless, the federated result remains practically meaningful because the performance difference relative to centralized clustering is small. (Khan et al., 2024) The Silhouette Score and Calinski-Harabasz Index confirm that the federated approach produces reasonably compact and well-separated clusters. Furthermore, the federated output is consistent with the centralized output, indicating that the overall configuration of the clusters within the federated approach is in agreement with the central clustering approach, even in the absence of direct data integration.

This finding bolsters the argument that privacy-preserving clustering techniques should be measured not only in terms of their ability to achieve an equivalent centralized clustering, but also in terms of the analytic utility they afford under more realistic constraints (Lim, 2024). In the case of healthcare, a small decrease in a clustering score may be warranted in order to decrease privacy concerns. Consequently, in such cases, federated clustering may be considered a suitable approach to cluster analysis.

Privacy-Preserving Value

The proposed framework preserves data locality by avoiding the exchange of raw patient-level records during the federated clustering process. Each health care organization performs preprocessing and clustering on its own. The centralized assembly process works on cluster-level information only, such as cluster centroids and cluster sizes. This inherently minimizes the amount of sensitive data shared among organizations.

The framework enables organizations to avoid total dataset centralization while still allowing collaboration. This is particularly useful for healthcare networks, multicenter studies, and regional healthcare systems where data are partitioned across several organizations. (Teramoto et al., 2024) In these cases, the framework supports collaborative knowledge discovery while reducing the need for raw data centralization.

Federated clustering alone should not be interpreted as providing complete privacy protection, because shared cluster representations may still reveal information about local data distributions. After federated clustering has been performed, knowledge of the cluster representations can be used to determine the local data distribution, especially in the case of very small clusters. Therefore, additional measures should be considered, such as secure aggregation, access control, setting a minimum cluster size, and differential privacy. Even without additional

cryptographic or differential-privacy mechanisms, the framework reduces exposure associated with centralized data pooling because raw patient-level records remain local. (Rieke et al., 2020)

Interpretation for Healthcare Analytics

The results indicate that federated clustering provides a practical trade-off between analytical utility and data locality (Alsamhi et al., 2024). Centralized clustering produced the best results, as it had access to all data, whereas local-only clustering generated privacy-preserving cluster data, but the cluster data were weaker, less reliable, and less consistent (Hollmann et al., 2025). Federated clustering yielded a compromise, as it permitted organizations to provide local clusters to a global model, while still protecting the raw, patient-level data. From the perspective of healthcare analytics, this is a noteworthy finding, as it is often necessary to analyze data in a collaborative manner, and the need for integrated datasets to perform interactive analytics often conflicts with privacy, ethical, and governance constraints (Kleczyk, 2023).

The main contribution of the proposed framework is the ability to perform exploratory healthcare analytics even with privacy-preserving constraints. The clusters generated by the framework represent groups of healthcare records that contain similar quantitative data. These data may include physiological data, metabolic data, and data pertaining to the healthcare services that the individuals may have utilized. These clusters should not be viewed as clinical diagnoses. However, they are useful in the healthcare domain for patient segmentation, population surveillance, and planning of healthcare services. The K-Means-based implementation also supports interpretability and relatively straightforward deployment in exploratory healthcare analytics settings.

The framework is also characterized by a degree of cooperative clustering among organizations. In order for the framework to be implemented successfully, the participating organizations must come to consensus on defining healthcare services, setting clustering rules, establishing how to aggregate the data, and determining the privacy protection measures that will be implemented. If such cooperative clustering is not accomplished, the local clusters that are produced will not be comparable. Therefore, healthcare federated clustering is a cooperative governance framework for privacy-protecting healthcare data analytics.

Limitations

This study has several limitations. First, the evaluation was conducted using a benchmark-style healthcare dataset under controlled distributed conditions rather than data from an operational multi-institutional healthcare network. While this offers an opportunity for a controlled evaluation, real-world healthcare data would likely have many issues related to missing values or heterogeneity, institutional bias, coding consistency, and regulations. Second, K-Means was used. This algorithm is fairly simple and efficient, but would also likely fail to identify clusters with complex or irregular shapes. Finally, privacy preservation was only examined with respect to the locality of the data. Future work should incorporate more in-depth privacy risk assessments, such as membership inference attacks or evaluations of the reconstruction risk and differential privacy.

Future research should evaluate the framework using independently sourced multi-institutional healthcare datasets and operational collaborative settings. (Singhal et al., 2023) It should also use additional privacy mechanisms and test larger, more complex distributed datasets

to increase the relevance and incorporate more privacy into the federated clustering system in healthcare.

Conclusion

Overall, these findings suggest that, in healthcare, federated clustering can reduce the gap between institution-specific analysis and centralized collaborative analytics (Alsamhi et al., 2024). Local clustering preserves privacy; however, it restricts data exploratory analysis. In contrast, central clustering fulfills the analytic need and exposes privacy risks. Federated clustering enables local institutions to contribute cluster-level information to a global representation without transferring raw patient-level records. This is the most significant benefit of this proposed system.

This study proposes a privacy-preserving federated clustering system for healthcare data that allows the analytic collaboration of healthcare institutions without the need to share patient data. In this system, each healthcare institution independently completes clustering and shares only the summary of the clusters for aggregation.

Results showed that the proposed federated clustering method achieved similar results to the centralized version. This was seen in the Silhouette Score, Calinski-Harabasz Index, and cluster consistency. Despite centralized clustering achieving the highest values, federated clustering was the better privacy-preserving choice as it kept the raw data in the respective organizations. The proposed method was also an improvement over local-only clustering, and indicated that even slight collaboration in the form of cluster representation sharing was beneficial to the discovery of global patterns.

The framework is suitable for healthcare analytics settings in which data locality and privacy governance are major considerations. Further evaluation using larger datasets, operational healthcare settings, and stronger privacy mechanisms such as secure aggregation and differential privacy is required.

Acknowledgements

The researcher did not receive any funding for this study, and the results have not been published in any other sources. Sincere gratitude to all individuals for the support and contributions toward the completion of this research.

References

Abbas, S. R., Abbas, Z., Zahir, A., Lee, S. W. (2024). Federated Learning in Smart Healthcare: A Comprehensive Review on Privacy, Security, and Predictive Analytics with IoT Integration. Healthcare. https://doi.org/10.3390/healthcare12242587

Ali, S., Abuhmed, T., El–Sappagh, S., Muhammad, K., Alonso, J. M., Confalonieri, R., Guidotti, R., Ser, J. D. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion. https://doi.org/10.1016/j.inffus.2023.101805

Alsamhi, S. H., Myrzashova, R., Hawbani, A., Kumar, S., Srivastava, S., Zhao, L., Wei, X., Guizan, M. (2024). Federated Learning Meets Blockchain in Decentralized Data Sharing: Healthcare Use Case. IEEE Internet of Things Journal. https://doi.org/10.1109/jiot.2024.3367249

Cuomo, S., Cola, V. S. D., Giampaolo, F., Rozza, G., Raissi, M., Piccialli, F. (2022). Scientific Machine Learning Through Physics–Informed Neural Networks: Where we are and What's Next. Journal of Scientific Computing. https://doi.org/10.1007/s10915-022-01939-z

Δρίτσας, Η., Τρίγκα, Μ. (2025). Federated Learning for IoT: A Survey of Techniques, Challenges, and Applications. Journal of Sensor and Actuator Networks. https://doi.org/10.3390/jsan14010009

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. Nature. https://doi.org/10.1038/s41586-024-08328-6

Joshi, M., Pal, A., Sankarasubbu, M. (2022). Federated Learning for Healthcare Domain—Pipeline, Applications and Challenges. ACM Transactions on Computing for Healthcare. https://doi.org/10.1145/3533708

Khan, S. U., García, C. E., Hwang, T., Koo, I. (2024). Radio Environment Map Construction Based on Privacy-Centric Federated Learning. IEEE Access. https://doi.org/10.1109/access.2024.3367589

Kleczyk, E. J. (2023). Importance of Patient Privacy in Healthcare Analytics Research. Ethics—Scientific Research, Ethical Issues, Artificial Intelligence and Education [Working Title]. https://doi.org/10.5772/intechopen.1001305

Lim, W. M. (2024). What Is Quantitative Research? An Overview and Guidelines. Australasian Marketing Journal (AMJ). https://doi.org/10.1177/14413582241264622

Liu, J. C., Goetz, J., Sen, S., Tewari, A. (2021). Learning From Others Without Sacrificing Privacy: Simulation Comparing Centralized and Federated Machine Learning on Mobile Health Data. JMIR mHealth and uHealth. https://doi.org/10.2196/23728

Long, Y., Ang, K. S., Li, M., Chong, K. L. K., Sethi, R., Zhong, C., Xu, H., Ong, Z. (2023). Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. Nature Communications. https://doi.org/10.1038/s41467-023-36796-3

Luo, Y., Abidian, M. R., Ahn, J., Akinwande, D., Andrews, A. M., Antonietti, M., Bao, Z., Berggren, M. (2023). Technology Roadmap for Flexible Sensors. ACS Nano. https://doi.org/10.1021/acsnano.2c12606

Mahmood, A., Sadique, T. K., Qureshi, A. N., Saleem, Y., & Lohithraman, S. (2025). Federated learning for thoracic disease detection: Comparative evaluation across chest x-ray classifications. Multidisciplinary Science Journal, 8(5), 2026336. https://doi.org/10.31893/multiscience.2026336

Muñoz-Gama, J., Martin, N., Fernández-Llatas, C., Johnson, O., Sepúlveda, M., Helm, E., Galvez-Yanjari, V., Rojas, E. (2022). Process mining for healthcare: Characteristics and challenges. Journal of Biomedical Informatics. https://doi.org/10.1016/j.jbi.2022.103994

Murray, C. J. L., Ikuta, K. S., Sharara, F., Swetschinski, L. R., Aguilar, G. R., Gray, A. P., Han, C., Bisignano, C. (2022). Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. The Lancet. https://doi.org/10.1016/S0140-6736(21)02724-0

Pang, Z., Lü, Y., Zhou, G., Hui, F., Xu, L., Viau, C., Spigelman, A. F., MacDonald, P. E. (2024). MetaboAnalyst 6.0: Towards a unified platform for metabolomics data processing, analysis and interpretation. Nucleic Acids Research. https://doi.org/10.1093/nar/gkae253

Rehm, H. L., Page, A., Smith, L., Adams, J., Alterovitz, G., Babb, L., Barkley, M. P., Baudis, M. (2021). GA4GH: International policies and standards for data sharing across genomic research and healthcare. Cell Genomics. https://doi.org/10.1016/j.xgen.2021.100029

Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N. (2020). The future of digital health with federated learning. npj Digital Medicine. https://doi.org/10.1038/s41746-020-00323-1

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Lee, J., Chung, H. W., Scales, N., Tanwani, A. K. (2023). Large language models encode clinical knowledge. Nature. https://doi.org/10.1038/s41586-023-06291-2

Tao, J., Liu, Z., Lyu, R. (2026). Research on a Privacy-Aware Data Processing Framework for Large-Scale Distributed Systems. https://doi.org/10.21203/rs.3.rs-8923292/v1

Teramoto, M., Roth, G. A., Aravkin, A. Y., Zheng, P., Abate, K. H., Abate, Y., Abbafati, C., Abbasgholizadeh, R. (2024). Global burden and strength of evidence for 88 risk factors in 204 countries and 811 subnational locations, 1990–2021: A systematic analysis for the Global Burden of Disease Study 2021. The Lancet. https://doi.org/10.1016/S0140-6736(24)00933-4

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science. https://doi.org/10.1007/s11704-024-40231-1

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B. (2023). Diffusion Models: A Comprehensive Survey of Methods and Applications. ACM Computing Surveys. https://doi.org/10.1145/3626235

Zhang, F., Kreuter, D., Chen, Y., Dittmer, S., Tull, S., Shadbahr, T., Schut, M. C., Asselbergs, F. W. (2024). Recent methodological advances in federated learning for healthcare. Patterns. https://doi.org/10.1016/j.patter.2024.101006

Zhao, W. X., Zhou, K., Li, J., Tang, T., Dong, Z., Hou, Y., Zhang, B., Min, Y. (2026). A Survey of Large Language Models. Frontiers of Computer Science. https://doi.org/10.1007/s11704-026-60308-3

Ziakis, C., Vlachopoulou, M. (2023). Artificial Intelligence in Digital Marketing: Insights from a Comprehensive Review. Information. https://doi.org/10.3390/info14120664