International Journal of Multiphysics
Volume 20, No. 1, 2026
ISSN: 1750-9548

Privacy-Preserving Federated Unsupervised Learning for Diabetes Risk Pattern Discovery Across Heterogeneous Tabular Health Datasets

Alireza Fathi, Reza Boostani, Rahil Hosseini, Mazin Abed Mohammed

Alireza Fathi — Dept. Computer Engineering, NT.C., Islamic Azad University, Tehran, Iran; alirezafathi84@gmail.com

Reza Boostani — CSE&IT Dept., ECE School, Shiraz University, Shiraz, Iran; rboostani87@gmail.com

Rahil Hosseini — Dept. Computer Engineering, Islamic Azad University, ShQ.C., Tehran, Iran; rahilhosseini@gmail.com

Mazin Abed Mohammed — Dept. Computer Science, University of Anbar, Ramadi, Iraq; mazinalshujeary@uoanbar.edu.iq

Abstract—Diabetes risk is influenced by interacting metabolic, anthropometric, cardiovascular, and lifestyle factors. Centralized machine-learning pipelines can discover latent patient subgroups, but they require pooling records from multiple data holders, which creates privacy and governance concerns. This paper presents a federated unsupervised learning framework for diabetes-risk pattern discovery using heterogeneous tabular health datasets. Public Kaggle datasets were used to simulate a multi-client healthcare environment under non-IID client partitioning. Federated standardization, Federated PCA, Federated KMeans, and federated centroid-distance anomaly scoring were compared with centralized PCA, centralized KMeans, and centralized Isolation Forest. The experiments covered 14 federated runs across Pima, BRFSS, and Health & Lifestyle data families, using five simulated clients per run. Federated PCA preserved the dominant variance direction with an average PC1 cosine similarity of 0.999. Federated KMeans achieved a mean ARI of 0.756 and a median ARI of 0.868 relative to centralized KMeans, while the median NMI was 0.781. Communication overhead was low, with an average estimated cost of only 0.072 MB per dataset. The findings indicate that federated unsupervised learning can recover major diabetes-risk structures without centralizing raw patient records, although highly non-IID partitions can degrade cluster agreement.

Index Terms—Federated learning, diabetes risk, unsupervised learning, privacy-preserving analytics, Federated PCA, Federated KMeans, anomaly detection, healthcare data mining.

I. INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder associated with hyperglycemia, insulin resistance, cardiovascular complications, kidney disease, and reduced quality of life. Modern diabetes analytics often combines laboratory biomarkers such as glucose and HbA1c with anthropometric indicators such as body mass index (BMI), lifestyle variables, cardiovascular history, and self-reported health indicators. In clinical and public-health settings, however, these variables are distributed across different institutions and populations. A model that performs well on a centralized benchmark may not be feasible in real healthcare networks because raw records cannot always be transferred to a central server.

Federated learning provides a mechanism for collaborative modeling without direct data centralization. Most federated healthcare studies focus on supervised classification, where the objective is to predict a known label such as diabetes or disease progression. In contrast, the current work focuses on federated unsupervised learning:

the goal is to discover latent risk structures, patient subgroups, dominant variance directions, and anomalous records without using diabetes labels for training. Labels are used only after clustering as post-hoc validation. This distinction is important because unsupervised discovery can reveal heterogeneous risk profiles even when labels are incomplete, inconsistent, or unavailable across clinical sites. This motivation is consistent with foundational and healthcare-oriented federated-learning literature [1], [2], [8]–[11].

The main research question is therefore: can federated unsupervised learning recover the same diabetes-risk structure that would be obtained through centralized analysis, while transmitting only aggregated summaries? To answer this question, a reproducible experimental pipeline was constructed using tabular diabetes-related datasets. The system simulates multiple healthcare clients through non-IID sharding and compares centralized and federated versions of PCA, KMeans clustering, and anomaly detection. The study is not intended to claim clinical diagnosis; rather, it evaluates privacy-preserving pattern discovery across heterogeneous diabetes datasets.

II. RELATED WORK AND MOTIVATION

Classical diabetes prediction studies often use Pima Indians Diabetes data and related tabular datasets to train supervised classifiers. These studies usually identify glucose, BMI, age, insulin, blood pressure, and family-history indicators as

$$R_i^* = w_g G_i + w_b B_i + w_a A_i + w_h H_i + w_l L_i, \qquad \sum_k w_k = 1$$

Grouped risk score used for conceptual interpretation of discovered clusters.(2)

relevant predictors. Public-health datasets such as BRFSS extend this view by including general health, physical activity, mobility limitation, income, education, high blood pressure, and cholesterol indicators. Lifestyle datasets add HbA1c, postprandial glucose, lipid profiles, waist-to-hip ratio, and physical activity. Across these sources, diabetes risk is multidimensional and heterogeneous.

Unsupervised learning is useful for this setting because it does not require a fully reliable label. PCA can reveal dominant latent dimensions in metabolic and lifestyle variables. KMeans and density-based clustering can identify lower-risk, intermediate-risk, and higher-risk patient subgroups. Anomaly detection can flag rare profiles or records that may require additional review. However, centralized unsupervised analysis requires access to all records. Federated unsupervised learning offers a privacy-preserving alternative: clients retain their local data and share only local sufficient statistics, such as sums, counts, covariance matrices, or cluster summaries.

This paper contributes a federated unsupervised framework tailored to diabetes-risk pattern discovery. Its contribution is methodological rather than clinical: it demonstrates how distributed health data can be analyzed while preserving raw-data locality, and it quantifies the trade-off between centralized and federated latent-structure discovery. The design is also related to privacy-preserving patient-grouping ideas explored in recent personalized and clustered federated-learning studies [12].

III. MATHEMATICAL FORMULATION

Let each patient or survey participant be represented by a feature vector x_i in R^d. Depending on the data family, the feature vector may contain glucose, BMI, blood pressure, insulin, age, HbA1c, lipid variables, general health, mobility limitation, or lifestyle indicators. A conceptual diabetes-risk score can be written as a logistic risk model:

$$R_i = \sigma\left(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij}\right), \qquad \sigma(z) = \frac{1}{1+e^{-z}}$$(1)

Logistic formulation of an interpretable diabetes-risk score.

A federated environment is represented as M clients, where each client m holds a local dataset D_m. The data are intentionally treated as non-IID, meaning that feature and label distributions can differ across clients:

$$\mathcal{D} = \bigcup_{m=1}^{M} \mathcal{D}_m, \quad \mathcal{D}_m \cap \mathcal{D}_{m'} = \emptyset, \quad P_m(X, Y) \neq P_{m'}(X, Y) \tag{3}$$

Non-IID federated data partition used to simulate heterogeneous healthcare clients.

Before federated modeling, each client computes local statistics for imputation and scaling. Global standardization can be obtained from client-level sums and squared sums rather than raw records:

$$\mu = \frac{1}{N} \sum_{m=1}^{M} \sum_{i=1}^{n_m} x_i^{(m)}, \qquad \sigma^2 = \frac{1}{N} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \left(x_i^{(m)} - \mu\right)^2 \tag{4}$$

Federated standardization using aggregated client-level sufficient statistics.

Federated PCA is performed through covariance aggregation. Each client sends only X_m^T X_m and its sample count. The server aggregates the covariance matrix and solves the eigenvalue problem:

$$C = \frac{1}{N-1} \sum_{m=1}^{M} (X_m^\top X_m), \qquad C v_j = \lambda_j v_j \tag{5}$$

Federated PCA using covariance aggregation and eigen-decomposition.

IV. FEDERATED CLUSTERING AND ANOMALY DETECTION

Federated KMeans minimizes the same within-cluster sum-of-squares objective as centralized KMeans, but the server never receives raw records. In each round, clients receive centroids, assign local samples to the nearest centroid, and send cluster sums and counts to the server. The objective is:

$$\min_{\{\mu_k\}_{k=1}^{K}} \sum_{m=1}^{M} \sum_{x_i \in \mathcal{D}_m} \min_k \|x_i - \mu_k\|_2^2 \tag{6}$$

In this expression, beta_j represents the contribution of each normalized feature. The model is not used as the primary training objective in this study; instead, it motivates why metabolic, anthropometric, and lifestyle variables are expected to form meaningful latent structures. For unsupervised interpretation, a normalized weighted risk score can also be defined by grouping features into clinically meaningful components: glucose status G_i, body-composition status B_i, age-pressure status A_i, general-health status H_i, and lifestyle status L_i:

Federated KMeans objective over distributed clients.

The server updates each centroid by aggregating local cluster sums and local cluster counts. This update is mathematically equivalent to centralized KMeans when all clients participate and use the same feature space:

$$\mu_k^{(t+1)} = \frac{\sum_{m=1}^{M} s_{mk}^{(t)}}{\sum_{m=1}^{M} n_{mk}^{(t)}}, \qquad s_{mk}^{(t)} = \sum_{x_i \in C_{mk}^{(t)}} x_i \tag{7}$$

Federated centroid update using client-level cluster summaries.

After the final federated centroids are obtained, a centroid-distance anomaly score is computed. Records with unusually large distance from their nearest centroid are treated as rare or atypical profiles:

$$a_i = \min_k \|x_i - \mu_k\|_2, \qquad \text{Anomaly}(x_i) = \mathbf{1}[a_i \geq Q_{1-\alpha}(a)] \tag{8}$$

Federated anomaly scoring using distance to the nearest global centroid.

Cluster quality is evaluated internally using Silhouette, Calinski-Harabasz, and Davies-Bouldin indices. The Silhouette coefficient is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \qquad S = \frac{1}{N} \sum_{i=1}^{N} s(i) \tag{9}$$

Silhouette coefficient for internal cluster-quality evaluation.

Agreement between centralized and federated clusters is evaluated using Adjusted Rand Index (ARI) and

Normalized Mutual Information (NMI). Post-hoc alignment with diabetes labels is also measured using ARI and NMI, but labels are not used to train the unsupervised models. Communication cost is estimated by counting the floating-point summaries transmitted per round:

$$C_{\text{round}} \approx M(2Kd + K) \cdot 8 \text{ bytes} \quad (10)$$

Approximate communication cost per federated KMeans round.

V. EXPERIMENTAL DESIGN

The experiment used a strict tabular-dataset selection strategy. Image-based diabetic retinopathy datasets and non-tabular sources were excluded to avoid mixing incompatible modalities. The final experiments focused on three data families: Pima-style clinical datasets, BRFSS health-indicator datasets, and a Health & Lifestyle diabetes dataset. Each dataset was split into five simulated clients using a non-IID sharding strategy. For BRFSS, features included high blood pressure, high cholesterol, BMI, smoking, stroke, heart disease, physical activity, diet variables, healthcare access, general health, mental health, physical health, difficulty walking, sex, age, education, and income. For Pima, features included pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. For Health & Lifestyle, features included glucose fasting, glucose postprandial, HbA1c, insulin level, BMI, waist-to-hip ratio, systolic and diastolic blood pressure, cholesterol fractions, triglycerides, physical activity, sleep, diet score, and screen time.

Table I. Dataset families and experimental scope.

Family	Runs	Total rows analyzed	Median features	Representative targets
BRFSS	9	928520	21	diabetes_012, diabetes_binary
Health_Lifestyle	1	97297	21	diabetes_stage
Pima	4	5072	8	outcome

The main comparison was Centralized KMeans versus Federated KMeans. Centralized KMeans was trained on the pooled standardized feature matrix, whereas Federated KMeans used only client-level sums and counts. Federated PCA was compared with centralized PCA using explained-variance differences and PC1 cosine similarity. Anomaly results were compared between federated centroid-distance scoring and centralized Isolation Forest through Jaccard overlap.

VI. RESULTS

A. Centralized versus Federated KMeans

Across the 14 federated runs, Federated KMeans recovered cluster structures that were broadly similar to centralized KMeans. The mean ARI was 0.756, the median ARI was 0.868, the mean NMI was 0.686, and the median NMI was 0.781. The average silhouette difference was 0.016, indicating that cluster compactness was largely preserved under federated optimization.

Table II. Centralized-vs-federated cluster agreement by dataset family.

Family	Mean ARI	Mean NMI	Mean Silhouette Difference
BRFSS	0.670	0.592	0.020
Health_Lifestyle	0.993	0.984	0.000
Pima	0.891	0.824	0.011

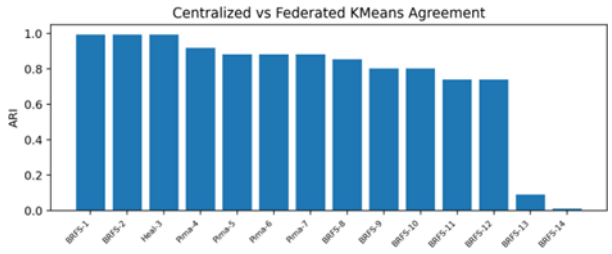

Fig. 1. ARI agreement between centralized and federated KMeans across datasets.

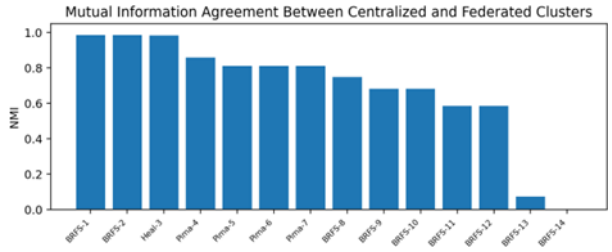

Fig. 2. NMI agreement between centralized and federated KMeans across datasets.

The strongest agreements were observed in the Health & Lifestyle dataset and balanced BRFSS subsets, where ARI and NMI approached or exceeded 0.98 in some runs. Pima datasets also showed stable agreement, with ARI around 0.88-0.92. The weakest agreement appeared in highly non-IID BRFSS 2021 partitions, indicating that client heterogeneity remains a key challenge for federated unsupervised learning.

B. Federated PCA

Federated PCA produced the strongest and most stable result. PC1 cosine similarity between centralized and federated PCA had a mean value of 0.999, a median of 1.000, and a minimum of 0.997. This indicates that covariance aggregation can reconstruct the global dominant variance direction almost exactly without transmitting raw patient records.

Table III. Federated PCA preserves centralized explained variance.

Component	Mean centralized variance	Mean federated variance	Mean absolute difference
PC1	0.200	0.206	6.04e-03
PC2	0.116	0.114	2.06e-03
PC3	0.088	0.087	4.81e-04

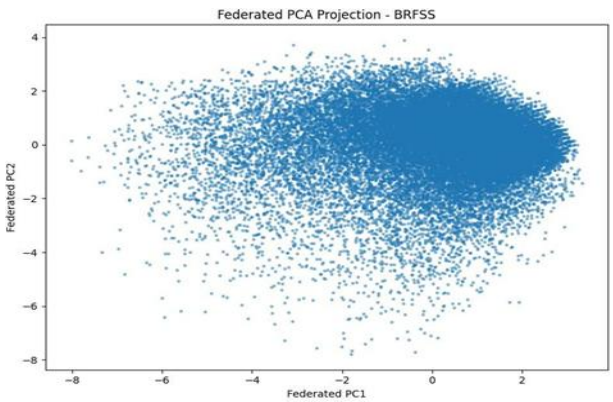

Fig. 3. Federated PCA projection for a BRFSS dataset.

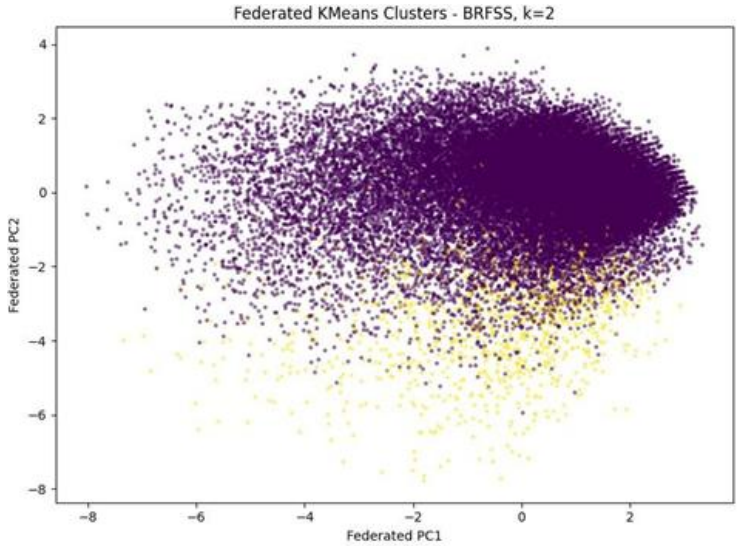

Fig. 4. Federated KMeans clusters for BRFSS.

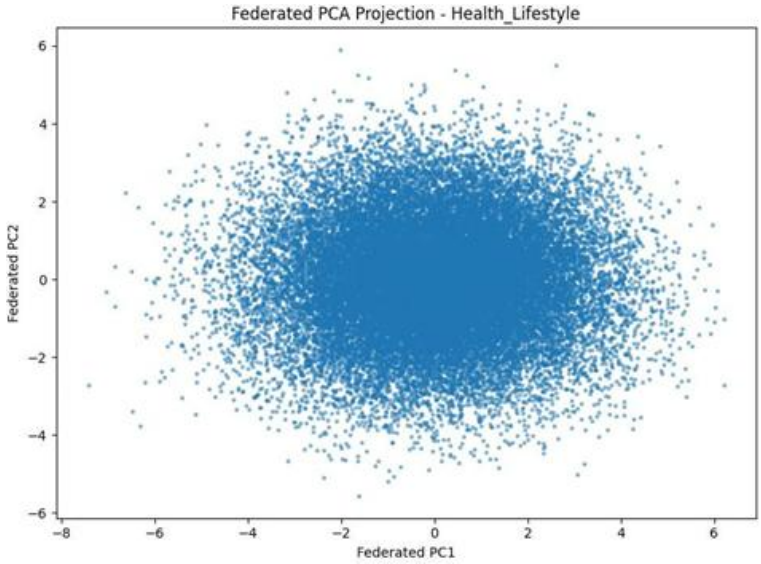

Fig. 5. Federated PCA projection for Health & Lifestyle data.

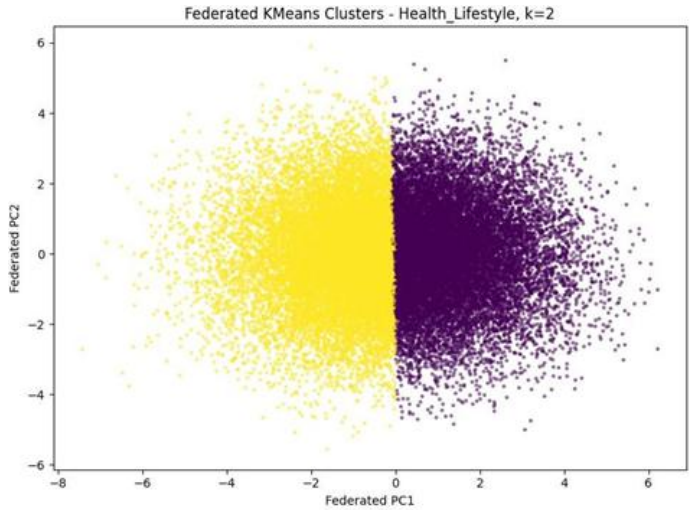

Fig. 6. Federated KMeans clusters for Health & Lifestyle data.

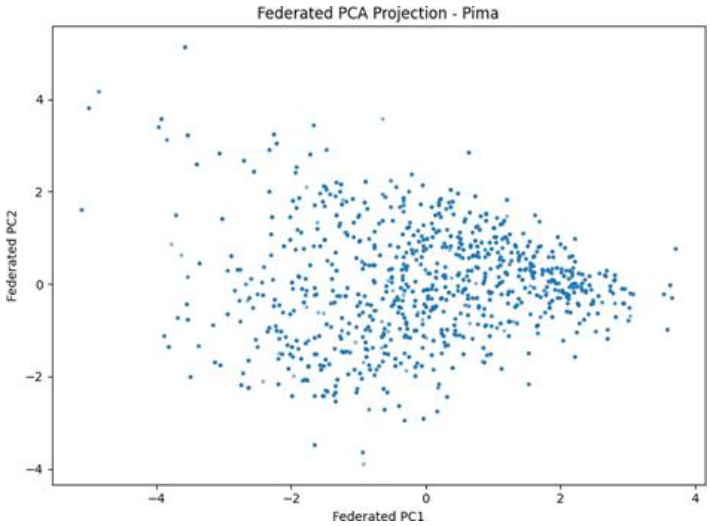

Fig. 7. Federated PCA projection for a Pima-style dataset.

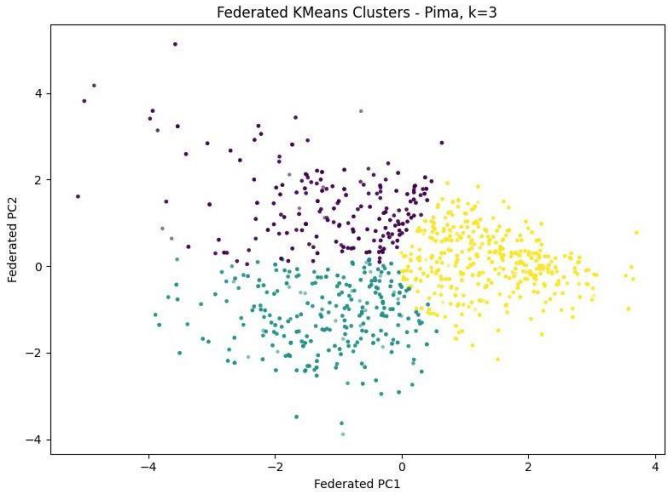

Fig. 8. Federated KMeans clusters for a Pima-style dataset.

C. Feature Importance and Risk Factors

Federated centroid-spread importance was used to identify the variables that most strongly separated latent clusters. The most consistent factors differed by dataset family. BRFSS clusters were primarily separated by healthcare access, cholesterol checks, stroke, difficulty walking, general health, physical health, income, and high blood pressure. Health & Lifestyle clusters were separated by BMI, LDL cholesterol, HbA1c, fasting glucose, waist-to-hip ratio, total cholesterol, postprandial glucose, and systolic blood pressure. Pima clusters were separated by age, pregnancies, BMI, skin thickness, glucose, insulin, and blood pressure.

Table IV. Dominant risk factors identified by federated centroid-spread importance.

Family	**Top federated cluster-separating features**
Pima	age (0.165), pregnancies (0.158), bmi (0.149), skinthickness (0.137), glucose (0.118)
BRFSS	cholcheck (0.129), stroke (0.105), anyhealthcare (0.070), diffwalk

	(0.069), genhlth (0.066)
Health_Lifestyle	bmi (0.092), ldl_cholesterol (0.085), hba1c (0.084), glucose_fasting (0.083), waist_to_hip_ratio (0.081)

D. Target Alignment and Anomaly Detection

Because the framework is unsupervised, target labels were used only for post-hoc interpretation. Across datasets, the average federated target NMI was 0.083 and the median was

0.095. These values indicate partial, not complete, alignment with diabetes labels. This is expected: clusters describe latent patient subgroups rather than directly optimizing diagnosis. Anomaly analysis identified approximately 3% of records as unusual using both federated centroid-distance scoring and centralized Isolation Forest. The mean federated anomaly rate was 0.030 and the mean centralized anomaly rate was

0.030. Mean Jaccard overlap was 0.385. This moderate overlap suggests that the federated method captures a privacy-preserving subset of rare profiles while maintaining a distinct sensitivity pattern.

Table V. Federated and centralized anomaly-detection comparison.

Metric	**Mean**	**Median**	**Min**	**Max**
Federated anomaly rate	0.030	0.030	0.030	0.031
Centralized anomaly rate	0.030	0.030	0.030	0.031
Anomaly Jaccard similarity	0.385	0.386	0.216	0.548

VII. DISCUSSION

The results support three major findings. First, Federated PCA is highly reliable for recovering global latent variance directions from distributed tabular diabetes data. Since PCA depends on covariance structure, client-level covariance aggregation is sufficient to reproduce centralized PCA almost exactly when the feature space is harmonized. This makes Federated PCA a strong candidate for privacy-preserving exploratory analysis across healthcare organizations. These observations align with broader medical federated-learning evidence on privacy-preserving collaboration under heterogeneous clinical data [9]–[12].

Second, Federated KMeans is effective but more sensitive to heterogeneity. In Pima and Health & Lifestyle data, the federated clusters were nearly identical to centralized clusters. In several BRFSS runs, agreement remained strong, but highly non-IID BRFSS 2021 partitions showed substantial degradation. This indicates that federated unsupervised learning is not only a question of privacy, but also a question of client distribution. If clients differ strongly in age, healthcare access, socioeconomic status, or disease prevalence, centroid aggregation may converge to a structure that differs from centralized clustering.

Third, the important cluster-separating variables are consistent with known diabetes-risk domains. Glucose, HbA1c, BMI, age, blood pressure, lipid variables, difficulty walking, general health, and physical activity repeatedly appeared as relevant variables. This supports the interpretation that federated clusters are not arbitrary mathematical artifacts; they reflect meaningful metabolic, functional, and lifestyle dimensions.

The low estimated communication overhead is also significant. The average communication cost was less than

0.1 MB per dataset because Federated KMeans transmits only centroid vectors, cluster sums, and counts. In a

real deployment, this would be far smaller than transferring raw health records. The privacy advantage is therefore accompanied by practical communication efficiency.

VIII. LIMITATIONS

This study has several limitations. First, the federated clients were simulated from public Kaggle datasets rather than deployed across real hospitals. Therefore, the results demonstrate methodological feasibility rather than operational clinical deployment. Second, some public datasets may contain duplicated benchmark sources or synthetic variables, so the results should not be interpreted as epidemiological estimates. Third, post-hoc target alignment was modest because the models were unsupervised and were not optimized for classification. Fourth, harmonized feature spaces are required for federated PCA and KMeans. In real multi-hospital deployment, additional work would be needed for feature mapping, missingness normalization, coding-system alignment, secure aggregation, differential privacy, and auditability.

Another limitation is that centroid-distance anomaly scoring is simple and interpretable but not equivalent to more flexible anomaly detectors such as Isolation Forest, One-Class SVM, or deep autoencoders. The moderate Jaccard overlap observed in the experiments suggests that federated anomaly detection should be interpreted as a complementary screening tool, not a replacement for full clinical review.

IX. CONCLUSION

This paper presented a privacy-preserving federated unsupervised learning framework for diabetes-risk pattern discovery. Using heterogeneous tabular diabetes datasets, the study compared centralized and federated PCA, KMeans clustering, feature-importance analysis, target alignment, and anomaly detection. Federated PCA almost perfectly preserved centralized principal components, while Federated KMeans recovered comparable clusters in most datasets, with high agreement in Health & Lifestyle and Pima datasets. The method transmitted only compact summaries and produced low communication overhead. These results suggest that federated unsupervised learning can support diabetes-risk subgroup discovery without centralizing raw patient-level records. Future work should extend the framework to real multi-institutional datasets, stronger non-IID correction methods, federated representation learning, secure aggregation, and privacy guarantees such as differential privacy.

REFERENCES

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," Proc. AISTATS, 2017.

[2] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," arXiv preprint arXiv:1610.05492, 2016.

[3] S. Lloyd, "Least squares quantization in PCM," IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129-137, 1982.

[4] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," Philosophical Transactions of the Royal Society A, vol. 374, 2016.

[5] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, 1987.

[6] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001.

[7] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," Proc. IEEE ICDM, pp. 413-422, 2008.

[8] P. Kairouz et al., "Advances and open problems in federated learning," Foundations and Trends in Machine Learning, vol. 14, no. 1-2, pp. 1-210, 2021.

[9] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, D. Pati, S. Kotrotsou, M. Milchenko, B. Xu, A. Marcus, R. R. Colen, and S. Bakas, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," Scientific Reports, vol. 10, art. no. 12598, 2020.

[10] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, and M. J. Cardoso et al., "The future of digital health with federated learning," npj Digital Medicine, vol. 3, art. no. 119, 2020.

[11] M. Ali, F. Naeem, M. Tariq, and G. Kaddoum, "Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 2, pp. 778-789, 2023.

[12] A. Elhussein and G. Gursoy, "Privacy-preserving patient clustering for personalized federated learnings," in Proceedings of the 8th Machine Learning for Healthcare Conference, vol. 219, pp. 150-166, 2023.