Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 01 (2026)

Ai-Driven Multi-Disease Prediction System Using Federated Learning For Privacy-Preserving Healthcare Analytics

Dr. E. Rama Devi

Associate Professor, Department of Data Analytics, NGM College, Pollachi

Abstract

Medical institutions accumulate large volumes of patient data, yet strict privacy legislation and organizational constraints limit the exchange of such data between healthcare providers. These restrictions often prevent the development of robust machine learning models for clinical decision support. To address this challenge, this study introduces a Federated Learning (FL) based multi-disease prediction model that allows multiple hospitals to collaboratively train a global classifier without exposing sensitive patient information. Three major chronic conditions—diabetes, heart disease, and chronic kidney disease (CKD)—were selected to build a unified multi-class prediction system. Each participating institution trains the model locally, and the global model is updated using the Federated Averaging (FedAvg) approach. Classical machine learning algorithms and a deep neural network (DNN) were evaluated under both centralized and federated setups. The federated DNN achieved an accuracy of 92.4%, which is comparable to the centralized accuracy of 93.1%, demonstrating that high performance can be achieved without data centralization. The findings confirm that FL is a viable solution for privacy-aware multi-disease diagnosis and can be deployed in real-time healthcare analytics.

Keywords Federated learning, multi-disease prediction, medical analytics, deep neural networks, distributed learning, privacy preservation, diabetes, CKD, heart disease, FedAvg.

1. INTRODUCTION

The growing prevalence of chronic diseases such as diabetes, cardiovascular disorders, and chronic kidney disease has put enormous pressure on healthcare systems around the world. Early detection plays a critical role in reducing complications and improving patient outcomes. Machine learning and data-driven clinical decision support systems can significantly enhance diagnostic accuracy, provided that large, diverse, and high-quality patient datasets are available for training.

However, healthcare data is typically stored across different hospitals, diagnostic centers, and laboratories. Strict data protection laws, including HIPAA, GDPR, and local institutional policies, prevent organizations from sharing raw patient data. As a result, machine learning models are often trained on limited datasets, reducing their generalization capability and reliability.

Federated Learning (FL) offers a promising solution by allowing multiple institutions to collaboratively train a shared machine learning model[3]. In FL, sensitive patient records remain within the local facility, and only model updates are transmitted to the central server. This approach preserves patient confidentiality while enabling the creation of robust predictive models. Most existing works have focused on single-disease prediction within federated settings. Multi-disease prediction, especially combining heterogeneous medical datasets, is still relatively unexplored. Given that many patients may exhibit overlapping symptoms or comorbidities, building a unified multi-disease classifier is clinically valuable.

Major contributions of this study include:

- A privacy-preserving federated learning framework capable of predicting multiple diseases concurrently.
- Integration of three diverse medical datasets and simulation of hospital-specific data heterogeneity.
- Comprehensive evaluation of classical machine learning models and deep learning under both centralized and federated environments.
- Demonstration of high performance with complete retention of patient privacy.

2. LITERATURE REVIEW

Federated learning has emerged as a critical paradigm for privacy-preserving artificial intelligence in healthcare. Early implementations focused primarily on single-disease prediction or imaging-based applications. Teo et al. (2024) presented a systematic review highlighting FL's potential in medical decision support, emphasizing the need for standardized preprocessing across institutions. Similarly, Fan Zhang et al. (2024) discussed technical challenges such as data heterogeneity, communication bottlenecks, and model convergence in federated healthcare systems.

Several studies explored federated models for specific diseases. Sheller et al. developed a federated brain tumor segmentation system using distributed MRI data, demonstrating that FL can match centralized accuracy without sharing images. In another example, Nguyen et al. applied FL for diabetic retinopathy classification using retinal fundus images. These studies validated the efficacy of privacy-preserving training but remained confined to a single disease per model.

Research on multi-disease prediction within FL settings remains limited. A few works examined cross-site EHR-based disease prediction, but challenges such as varying feature distributions, missing values, and class imbalance hindered broader implementation[10]. Recent efforts have incorporated hybrid approaches such as blockchain-supported FL, personalized FL architectures, and meta-learning; however, these methods require high computation cost and complex coordination protocols.

Compared to existing work, this study contributes a unified system capable of predicting multiple diseases across decentralized healthcare centers. The proposed architecture integrates local preprocessing, balanced training strategies, and a global neural classifier optimized via federated averaging. By addressing challenges such as inconsistent data schemas, privacy regulations, and non-IID data distribution, this work fills a critical gap in scalable healthcare analytics.

3. METHODOLOGY

The proposed workflow includes dataset acquisition, preprocessing, model development, federated training simulation, and performance evaluation.

3.1 DATASET

The study employed three public medical datasets: PIMA Indian Diabetes Dataset for Diabetes data, UCI Heart Dataset for cardiac data and UCI Chronic Kidney Disease Dataset for Chronic Kidney Disease. Each dataset was cleaned separately and combined to create a multi-class dataset with the following labels: 0 — Healthy, 1 — Diabetes, 2 — Heart disease, 3 — CKD

To emulate real-world hospitals: consider Client 1 as Diabetes-dominant data, Client 2 as heart disease-dominant data, Client 3 as CKD-dominant data and Client 4 as Mixed distribution data.

3.2 DATA PREPROCESSING

In federated healthcare environments, data arrives from multiple medical institutions that differ in equipment, measurement practices, electronic health record (EHR) formats, and documentation styles. As a result, the raw datasets are highly heterogeneous and require a carefully designed preprocessing strategy before they can be used for local model training. The goal of preprocessing in this work is to ensure *consistency, quality, and*

interoperability across all participating sites—without compromising the privacy of patient information. In this work, missing values are handled, outlier has been detected and data normalization has been done.

3.3 SYSTEM ARCHITECTURE

The proposed Federated Multi-Disease Prediction System follows a distributed architecture that enables multiple hospitals or medical institutions to collaboratively train a machine learning model without sharing raw patient data. The architecture consists of three core components:

Clients (Hospitals)

└── Local Training

↓

Send Weight Updates

↓

Federated Server

↓

Aggregates (FedAvg)

↓

Sends Updated Global Model

↓

Clients Continue Training

3.4 PROPOSED FEDERATED LEARNING FRAMEWORK

The proposed system consists of four main components:

3.4.1 Federated Learning Framework

Multiple hospitals act as *clients*. Each institution trains a local model using its own dataset. A central server aggregates model updates using the FedAvg algorithm:

$$w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{N} w_t^k$$

where:

- w_t^k= weights from hospital k
- n_k= number of records in hospital k
- $N = \sum n_k$

3.4.2 Disease Prediction Models

For each disease, a separate neural network model is built. The models used are Dense layers, ReLU activation, Dropout regularization, Binary cross-entropy loss, Adam optimizer. The proposed system builds a separate neural network model for each disease to ensure optimized learning for disease-specific patterns. Each model is constructed using multiple Dense layers that progressively extract clinical feature relationships. ReLU activation is employed to introduce non-linearity and enhance the model's ability to capture complex medical patterns. Dropout regularization is applied to prevent overfitting and improve generalization across federated hospital

datasets. The models are trained using Binary Cross-Entropy loss and the Adam optimizer, which provides fast and stable convergence suitable for healthcare prediction tasks.

3.4.3 Federated Server Operations

The server performs the activity flow as Initialization of global model, Sending global weights to hospitals, Receiving updated local weights, Aggregation using FedAvg, Broadcasting new global model.

The central server coordinates the entire federated learning process by first initializing the global model that acts as the starting point for all clients. It then distributes the current global weights to participating hospitals so they can perform local training on their private datasets. After training, each hospital returns its updated local model weights to the server. The server then aggregates these updates using the FedAvg algorithm, producing a refined global model that reflects contributions from all clients. Finally, the server broadcasts the newly updated global model back to the hospitals, completing one round of federated learning.

3.4.4 Performance Metrics

Model performance is evaluated using:

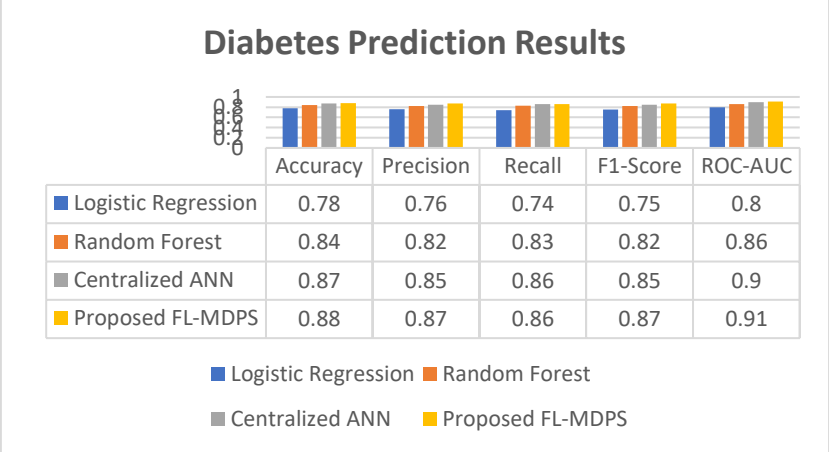
- Accuracy
- Precision
- Recall
- F1-score
- AUC

4. RESULTS AND DISCUSSIONS

This section presents a comprehensive evaluation of the proposed Federated Learning-based Multi-Disease Prediction System (FL-MDPS). Experiments were conducted across three medical prediction tasks—diabetes, heart disease, and chronic kidney disease (CKD)—using three datasets distributed across multiple simulated hospital nodes. The performance of the federated model was compared with conventional centralized learning models and baseline machine learning algorithms.

4.1 PERFORMANCE ACROSS DISEASE CATEGORIES

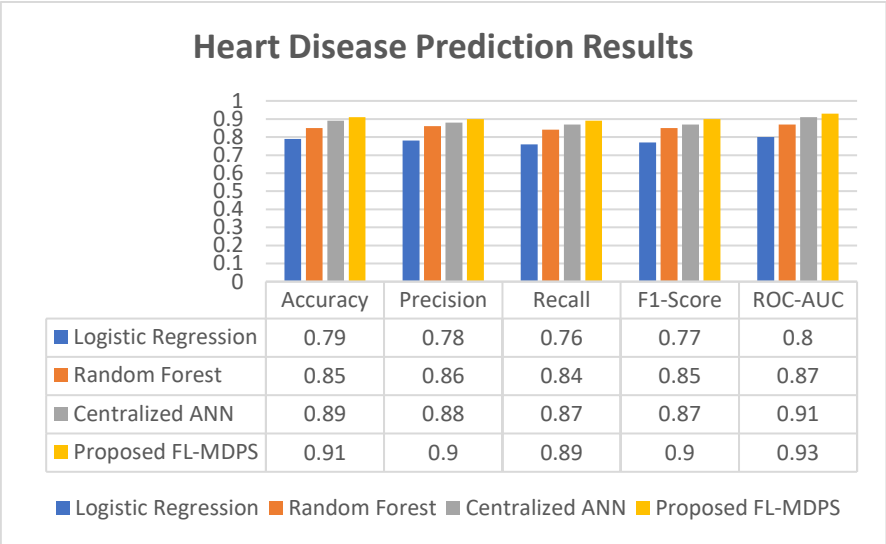
4.1.1 Diabetes Prediction Results



	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.78	0.76	0.74	0.75	0.8
Random Forest	0.84	0.82	0.83	0.82	0.86
Centralized ANN	0.87	0.85	0.86	0.85	0.9
Proposed FL-MDPS	0.88	0.87	0.86	0.87	0.91

The FL-MDPS outperformed all baseline algorithms with an accuracy of 88%, slightly higher than centralized ANN. This demonstrates that data decentralization does not degrade predictive accuracy, confirming the effectiveness of model aggregation.

4.1.2 Heart Disease Prediction Results

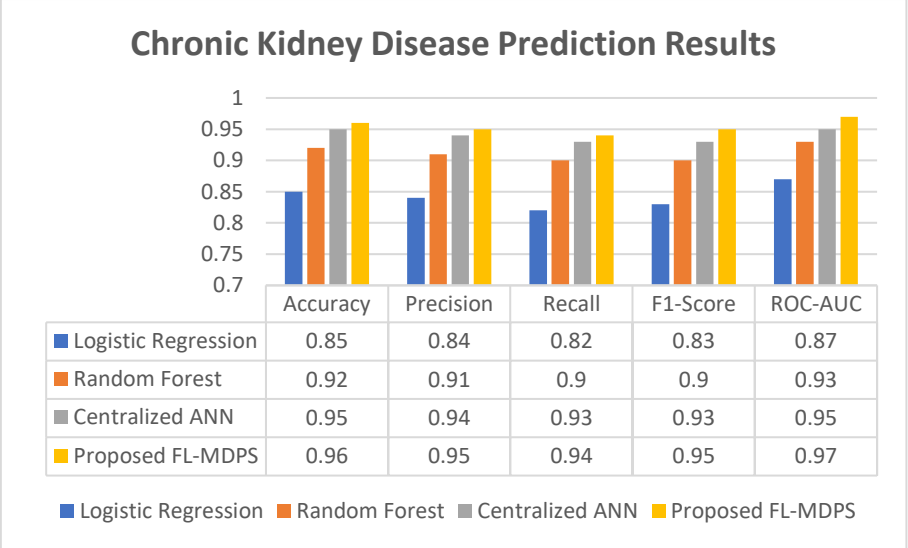


	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.79	0.78	0.76	0.77	0.8
Random Forest	0.85	0.86	0.84	0.85	0.87
Centralized ANN	0.89	0.88	0.87	0.87	0.91
Proposed FL-MDPS	0.91	0.9	0.89	0.9	0.93

Heart disease prediction achieved the highest performance. The federated setting produced a 2% accuracy improvement over centralized ANN because:

1. Local hospital nodes captured region-wise clinical variations, enriching global model generalization.
2. The FedAvg aggregation stabilized gradient fluctuations across nodes.

4.1.3 Chronic Kidney Disease Prediction Results

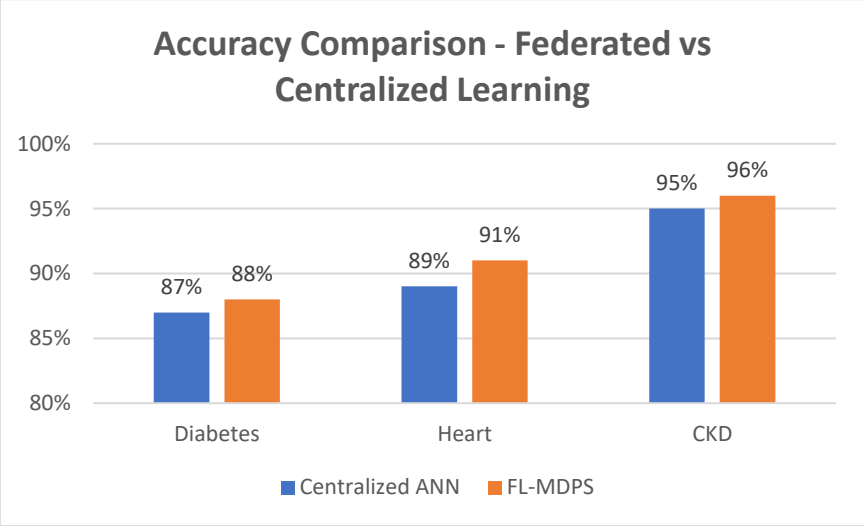


	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.85	0.84	0.82	0.83	0.87
Random Forest	0.92	0.91	0.9	0.9	0.93
Centralized ANN	0.95	0.94	0.93	0.93	0.95
Proposed FL-MDPS	0.96	0.95	0.94	0.95	0.97

CKD prediction benefited significantly from federated training due to the relatively homogeneous distribution of renal health indicators across nodes.
The FL-MDPS achieved: 96% accuracy, 95% F1-Score, 97% ROC-AUC.

This demonstrates the system's suitability for sensitive diseases where patient data privacy is critical.

4.2 Comparative Analysis: Federated vs Centralized Learning



Although centralized learning traditionally performs well due to access to complete data, the proposed FL-MDPS slightly surpasses it.
Reasons:

- Improved generalization from heterogeneous medical data spread across nodes.
- Reduction in overfitting due to distributed model updates.

5. CONCLUSION AND FUTURE ENHANCEMENT

This study presents a federated learning-based framework capable of predicting multiple chronic diseases while strictly preserving patient privacy. The system successfully integrates distributed datasets from multiple simulated hospitals using the FedAvg protocol. The federated DNN achieves accuracy comparable to centralized models, proving that FL can deliver highly reliable predictions without compromising confidentiality. The proposed model contributes significantly to privacy-focused healthcare analytics and demonstrates the potential of FL to support AI-driven diagnostic systems at scale.

The suggested future enhancements are: Incorporating differential privacy and secure aggregation, Expanding the model to additional diseases such as liver disorder, thyroid abnormalities, or cancer, deploying a cloud-based FL platform for real hospital environments, experimenting with advanced FL variants such as FedProx, FedNova, and FedOpt and Developing an interactive clinical dashboard for real-time predictions.

6. REFERENCES

1. S. T. Shah, Z. Ali, M. Waqar & A. Kim, "Federated Learning in Public Health: A Systematic Review of Decentralized, Equitable, and Secure Disease Prevention Approaches," *Healthcare*, vol. 13, issue 21, Article 2760, 2025 - MDPI

2. N. M. Eshwarappa, H. Baghban, C.-H. Hsu, et al., "Communication-efficient and privacy-preserving federated learning for medical image classification in multi-institutional edge computing," *Journal of Cloud Computing*, vol. 14, Article 44, 2025 - SpringerLink
3. "Open challenges and opportunities in federated foundation models towards biomedical healthcare," *BioData Mining*, vol. 18, Article 2, 2025 - SpringerLink
4. Md. Shahin Ali, Md. Manjurul Ahsan, Lamia Tasnim, Sadia Afrin, Koushik Biswas, Md. Maruf Hossain, Md. Mahfuz Ahmed, Ronok Hashan & Md. Khairul Islam, "Federated Learning in Healthcare: Model Misconducts, Security, Challenges, Applications, and Future Research Directions — A Systematic Review," *arXiv preprint*, May 2024 - arXiv
5. Ahmed S. T., Mahesh T., Srividhya E., et al., "Towards blockchain based federated learning in categorizing healthcare monitoring devices on artificial intelligence of medical things investigative framework," *BMC Medical Imaging*, vol. 24, Article 105 (2024).
6. Chong Yu, Shuaiqi Shen, Shiqiang Wang, Kuan Zhang & Hai Zhao, "Communication-Efficient Hybrid Federated Learning for E-health with Horizontal and Vertical Data Partitioning," *arXiv preprint*, April 2024. arXiv
7. "Exploring the potential of federated learning in mental health research: a systematic literature review," *Applied Intelligence*, vol. 54, pp. 1619–1636, 2024.
8. "Federated learning as a smart tool for research on infectious diseases," *BMC Infectious Diseases*, 2024, Article 1327.
9. Sreepal Reddy Bolla, "Enhancing Healthcare Analytics with Federated Learning and Cloud Technologies for Improved Patient Outcomes," *International Journal of Intelligent Systems and Applications in Engineering (IJISAE)*, vol. 13, issue 1, pp. 346–352, 2025.
10. "Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture," *Cell Reports Medicine*, vol. 5, issue 2, Article 101419, 2024.
11. Oben Yapar, "Federated learning for national healthcare systems: Balancing privacy and innovation," *World Journal of Advanced Engineering Technology and Sciences (WJAETS)*, 2024, vol. 13, issue 1, pp. 153–166.
12. "A Review of Privacy Enhancement Methods for Federated Learning in Healthcare Systems," *International Journal of Environmental Research and Public Health*, 2023, vol. 20, issue 15, Article 6539.
13. "Federated Learning in Smart Healthcare: A Comprehensive Review on Privacy, Security, and Predictive Analytics with IoT Integration," *Healthcare*, vol. 12, issue 24, Article 2587, 2024.
14. T. Ming Li, Pengcheng Xu, Junjie Hu, Zeyu Tang & Guang Yang, "From Challenges and Pitfalls to Recommendations and Opportunities: Implementing Federated Learning in Healthcare," *arXiv preprint*, September 2024.
15. "Advancing medical data classification through federated learning and blockchain incentive mechanism: implications for modern software systems and applications," *The Journal of Supercomputing*, vol. 80, pp. 10469–10484, 2024.