IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# DIABETES PREDICTION USING FEDERATED LEARNING

**Manoj Kumar C[1], Manjunath Doddmani[2] , Prajwal P K[3], Prajwal Hukkeri [4]**

**Prof. Arathi P[5]**

Student, Computer Science and Engineering, Dr Ambedkar Institute of Technology Bengaluru, India [1-4]

Assistant Professor, Computer Science and Engineering, Dr Ambedkar Institute of Technology, Bengaluru, India [5]

**Abstract:** Federated learning is an innovative approach used in medical field to resolve the issues like centralization, privacy and confidentiality. It gathers diverse data from several local models and aggregates it in a global model where only results are shared instead of data. Its a collaborative model training method to achieve best performance. We build a frame work for diabetics prediction, which consists of local models like Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks. These models are trained independently on local data distributed across multiple hospitals, ensuring privacy and data security. To improve the dataset and address class imbalance, both Exploratory Data Analysis (EDA) techniques and Synthetic Minority Over-sampling Technique (SMOTE) are used. EDA helps in understanding the underlying patterns and characteristics of the data, while SMOTE generates synthetic data points to balance the classes. And at the global model we aggregate all local models weights and check for the best model among the existing local models based on their predictive accuracy. In our frame work ANN performs with 89% accuracy. Hence those values are considered for prediction. After training different models , we achieved 89.00% accuracy with an RNN,89.99% accuracy with an ANN, and 89.08% accuracy. With an LSTM model .therefore, we proceed with the ANN model to predict diabetes .After successfully submitting all the weights, we obtained these accuracy levels through the best performer strategy in the global model. This approach ensures that the highest-performing model is utilized for predications,enhancing the overall effectiveness and realiability of the diabetics predication system in a collaborative healthcare environment.

## I. Introduction

Federated Learning (FL) is a decentralized machine learning technique that facilitates model training across numerous edge devices or servers while maintaining data privacy[1]. Unlike traditional centralized methods, FL allows model training to occur locally on devices where data resides, avoiding the necessity to transfer raw data to a central server. The key components of FL include client devices, a central server, and a secure communication protocol. The workflow starts with the central server initializing a global model[2].

Local models performs training using their local data, compute model updates, and securely transmit the weight updates to the central server[4]. The server consolidates the updates to enhance the global model and then transmits it back to the clients for additional training. Privacy preservation techniques, such as differential privacy and secure aggregation, Phrase to ensure the confidentiality of individual data points. FL can scale to a large number of client devices, making it suitable for applications with vast decentralized data sources. Challenges include handling heterogeneous devices, adversarial attacks, and balancing privacy with model utility[3]. FL can efficiently scale to a large number of client devices, making it suitable for applications with vast decentralized data sources, such as mobile phones or IoT devices. Despite its advantages, FL faces challenges such as dealing with heterogeneous client devices with varying computational capabilities and data distributions, vulnerability to adversarial attacks targeting model updates during transmission or aggregation, and balancing the trade off between data privacy and model utility. FL empowers the advancement of robust and accurate models without compromising the confidentiality of individual data points, making it an ideal choice for sectors that handle sensitive information, such as healthcare, finance, and smart city applications[5]. The significance of federated learning becomes even more pronounced in light of the growing emphasis on personalized medicine. The ability to tailor predictive models to individual characteristics, lifestyles, and genetic predispositions holds the potential to revolutionize diabetes management[6]. Federated learning not only enables more precise predictions by leveraging diverse data sources but also empowers individuals to actively contribute to medical research without compromising their privacy.

## II. OBJECTIVES

- Implement Federated Learning (FL) to enable collaborative model training across decentralized edge devices.
- Develop and deploy machine learning models, including ANNs, LSTMs, and RNNs, for various healthcare prediction tasks.
- Ensure data privacy and security through techniques like differential privacy and secure aggregation.
- Conduct comprehensive exploratory data analysis (EDA) to understand dataset characteristics and address data imbalances.
- Design a scalable and efficient distributed system architecture for model training and deployment.
- Assess the effectiveness of the proposed system in terms of accuracy, privacy preservation, and scalability.

## III. LITERATURE REVIEW

An academic publication that covers current understanding, noteworthy findings, and conceptual and methodological developments in a particular field is known as a survey of literature or review. It takes into account the dimensions and scope of the project and gives a variety of analyses and research done in the subject of interest, in addition to results that have already been published.
Through automation and improved decision-making in the fields of financial services, healthcare, and transportation, machine learning (ML) has enhanced daily routines and increased productivity. However, its wider acceptance has been hampered by worries about data privacy. Federated learning (FL) shares training parameters rather than raw data, which allays these worries. FL is becoming more popular; according to projections, 32% of businesses want to use it during the next 12 to 24 months, and investment will rise from USD 107 million in 2020 to USD 538 million in 2025. This article examines FL's technological features, how it differs from other technologies, the aggregation methods that are currently in use, and how it is used in the transportation, financial, and healthcare industries. It also discusses the difficulties and potential paths for advancement in FL's diagnosis of cancer, diabetes, and cardiovascular disease [1].
Every year, heart disease takes millions of lives since it is one of the world's leading causes of mortality. Health service organizations (HSPs) collect biomedical data, which is sensitive patient records and is governed by international privacy standards, limiting data exchange. Furthermore, there are delays in the prediction of cardiac illness due to the transmission and gathering of biological data, which result in considerable network communication costs. We propose a hybrid architecture that is applied at the client's end of HSPs to reduce the danger of one point of failure, training delay, and communication costs. The framework uses support vector machines and modified artificial bee colonies (MABC-SVM) to find the best features and classify cardiac disease. To address privacy issues, we suggest federated matched averaging for the HSP server. suggestion for the HSP server to tackle privacy issues is to implement federated matched averaging. tested and assessed our method using a pooled dataset on cardiovascular illness, comparing it to conventional federated learning approaches. Our findings show that the suggested hybrid strategy takes 17.7% fewer cycles to reach maximum accuracy while improving prediction accuracy by 1.5% and lowering classification error by 1.6%[2].
Large datasets are essential for the success of investigations based on machine learning (ML). Nevertheless, gathering such information inside a single hospital or health system is frequently impracticable. Although

accessing massive datasets is made easier by multicenter research, sharing data between institutions presents financial, legal, and technological challenges. Federated learning (FL), which addresses data sharing issues across participating institutions, has emerged as a unique machine learning (ML) framework designed for multicenter investigations. By maintaining data access control and facilitating the collation of locally trained models from the participating institutions to create a global model, FL streamlines multicenter investigations. This study discusses the difficulties and unresolved problems in applying FL to clinical experiments using organized medical data, as well as evaluates current studies using FL in clinical trials with structured medical data[3].

Heart disease is among the worst diseases, taking millions of lives every year. Early illness identification and diagnosis through the use of Internet of Medical Things (IoMT)-based healthcare has been shown to be useful in lowering death rates. However, highly individualized patient information is present in biological data obtained by IoMT, which raises serious privacy concerns. Conventional machine learning methods now face difficulties from planned international data protection rules intended to address these problems. We provide a methodology that combines federated matched averaging with a modified Artificial Bee Colony (M-ABC) optimization algorithm to address these privacy concerns and improve heart disease prediction. Compared against state-of-the-art federated learning algorithms on real-world heart disease datasets, our proposed method shows improvements in prediction accuracy, classification error, and communication efficiency[4].

Diabetes mellitus is a prevalent global health issue, often leading to various complications and significant health, economic, and mortality burdens. Timely diagnosis and prediction are crucial for implementing appropriate preventive and treatment measures. In our study, we aim to enhance understanding of the risk factors for type 2 diabetes among Pima Indian women. Employing a logistic regression model and decision tree machine learning algorithm, we identify five key predictors: glucose levels, pregnancy, body mass index (BMI), diabetes pedigree function, and age. Additionally, we employ a classification tree to validate our findings. The six-fold classification tree highlights glucose, BMI, and age as vital factors, while the ten-node tree underscores the importance of glucose, BMI, pregnancy, diabetes pedigree function, and age. Our preferred model achieves a prediction accuracy of 78.26% with a cross-validation error rate of 21.74%. We assert that our model offers valuable insights for predicting type 2 diabetes and could serve as a useful tool to complement existing preventive strategies, thereby potentially reducing the incidence and associated costs of diabetes[5].

Diabetes is a prevalent condition worldwide with no known cure, leading to significant healthcare costs annually in our country due to associated disabilities. Accurately predicting patients' conditions is crucial for effective treatment planning, emphasizing the need for highly precise and reliable forecasting methods. Artificial intelligence, particularly neural networks, presents a promising avenue for achieving this goal. This study aims to enhance prediction accuracy and reliability by combining statistical models such as logistic regression with neural networks. By integrating these approaches and conducting comparative analyses, the study evaluates the efficacy of the proposed hybrid model. Results indicate that the combined model yields superior performance compared to standalone neural network or logistic regression methods, with a notable reduction in error function from 0.1 in neural network training to 0.0002 in the hybrid model[6].

Diabetes Mellitus poses a significant health risk, often leading to complications such as heart disease, kidney issues, stroke, vision problems, and nerve damage. Currently, hospitals rely on various tests to diagnose diabetes, followed by tailored treatments based on the diagnosis. Big Data Analytics has emerged as a vital tool in healthcare, given the vast databases available. By analyzing extensive datasets, hidden information and patterns can be unearthed to enhance understanding and predict outcomes. However, existing methods often fall short in terms of classification and prediction accuracy. In this paper, we propose a diabetes prediction model that incorporates additional external factors alongside traditional metrics like Glucose, BMI, Age, and Insulin. The inclusion of these factors improves classification accuracy compared to previous datasets. Additionally, we introduce a pipeline model designed specifically for diabetes prediction, with the aim of further enhancing classification accuracy [7].

Diabetes stands as one of the most debilitating illnesses worldwide, often paving the way for other serious conditions such as kidney disease, blindness, and heart failure. Patients facing this challenge typically undergo diagnostic procedures, investing both time and money in obtaining their reports after consultations. However, advancements in machine learning techniques offer a promising avenue for finding effective solutions. With sophisticated information processing systems now at our disposal, we can predict whether a patient is at risk of diabetes or not, thus enabling early intervention before the situation escalates. The primary aim of this study is to develop a predictive system for assessing the risk of diabetes in patients. Experimental results demonstrate high accuracy in diabetes prediction through the utilization of support vector machines [8].

## IV. MODEL DESGIN

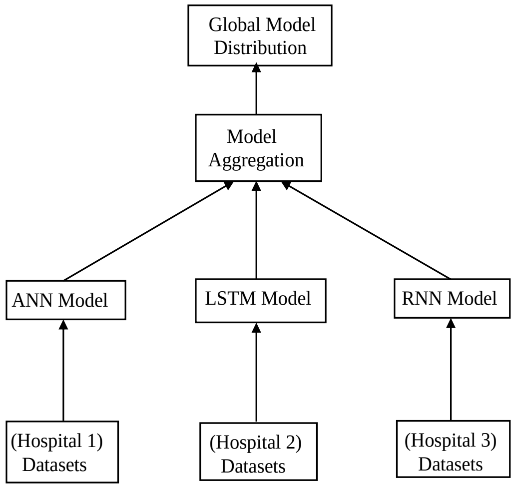


**Fig 1: Architecture Diagram**

The diagram illustrates a Federated Learning (FL) system architecture involving three hospitals, each with a unique machine learning model: an Artificial Neural Network (ANN) at Hospital 1, a Long Short-Term Memory (LSTM) network at Hospital 2, and a Recurrent Neural Network (RNN) at Hospital 3. Each hospital uses its local dataset to train its respective model. After training, the local models' parameters are sent to the central server for aggregation, ensuring that no raw data leaves the hospitals. The central server integrates these parameters to update the global model, which is then redistributed to all hospitals. This cycle continues iteratively, allowing collaborative model improvement while maintaining data privacy. The global model is refined based on the aggregated updates, and hospitals utilize the global model for diabetes prediction, benefiting from the collective learning. This system architecture addresses data privacy, security, and scalability concerns by keeping data decentralized and leveraging local computation power. The process ensures robust and accurate predictive models in a healthcare setting while protecting sensitive patient data.

## V.METHODS AND RESULTS

- Data Collection and Preprocessing :Acquire patient information from three hospitals, each providing a dataset (e.g., "hospital1.csv","hospital2.csv", "hospital3.csv"). Preprocess the data by encoding categorical variables and splitting it into features (X) and target variables (y).
- Local Model Training : Train different machine learning models at each hospital using the local data. For example, train an Artificial Neural Network (ANN) at Hospital 1, a Long Short-Term Memory (LSTM) network at Hospital 2, and a Recurrent Neural Network (RNN) at Hospital 3.
- Model Saving: After training, save the model weights for each trained model locally. This ensures that the trained model's parameters can be shared without exposing raw data.
- Secure Model Transmission : Implement a secure communication protocol to send the model weights from each hospital to a central server. Use encryption to ensure the security and privacy of the transmitted model weights.
- Model Aggregation : At the central server, aggregate the received model weights to update the global model. This aggregation can be done using techniques like Federated Averaging.
- Global Model Distribution : Distribute the updated global model back to the hospitals. Each hospital will use this global model as a new starting point for the next round of local training.

- Prediction Interface : Develop a frontend interface using Streamlit to allow healthcare providers to input patient data and request diabetes predictions. This interface will communicate with the backend to use the aggregated global model for making predictions
- Model Accuracy Evaluation : Continuously evaluate the accuracy of the global model by comparing predictions with actual outcomes. Use metrics such as accuracy, precision, recall, and F1 score to monitor performance.
- User Authentication : Implement user authentication to ensure that only the right people can access the prediction interface and manage model updates.
- Continuous Improvement: Implement a feedback loop where prediction results and new data can be fed back into the system for continuous model improvement. This involves periodic retraining of local models, re-aggregation at the central server, and updating the global model accordingly.

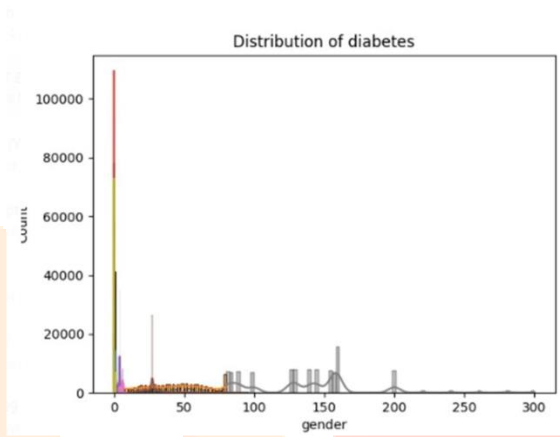


**Fig 2: Distribution of diabetes**

The depiction of how diabetes-related data is spread or distributed across a population or sample. It helps to visualize patterns, trends, and variations in the distribution of diabetes-related data.

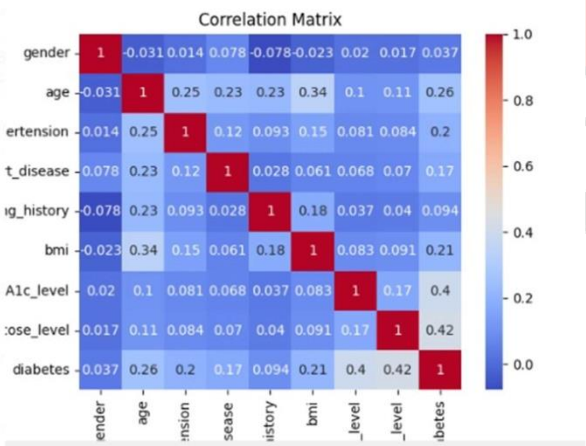


**Fig 3: Correlation Matrix**

It is particularly useful in data analysis and statistics for understanding the relationships between multiple variables simultaneously. It helps identify potential patterns, associations, and dependencies within the data.

**5.1 SMOTE :** (Synthetic Minority Over-sampling Technique) is an algorithm that balances imbalanced datasets by generating synthetic samples of the minority class. It's extensively utilized in machine learning for tasks like fraud detection and medical diagnosis to prevent bias towards the majority class during model training.


**Fig 4: Before SMOTE**

Before applying SMOTE, the dataset typically contains class imbalance, where one class is underrepresented compared to the other classes. trained on such data may have a bias towards the majority class, leading to poor performance in predicting the minority class.

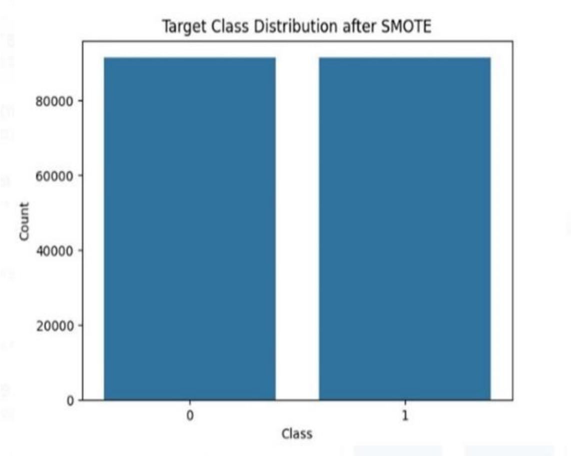


**Fig 5: After SMOTE**

After applying SMOTE, the dataset is rebalanced by generating synthetic samples for the minority class, thus addressing the initial class imbalance.

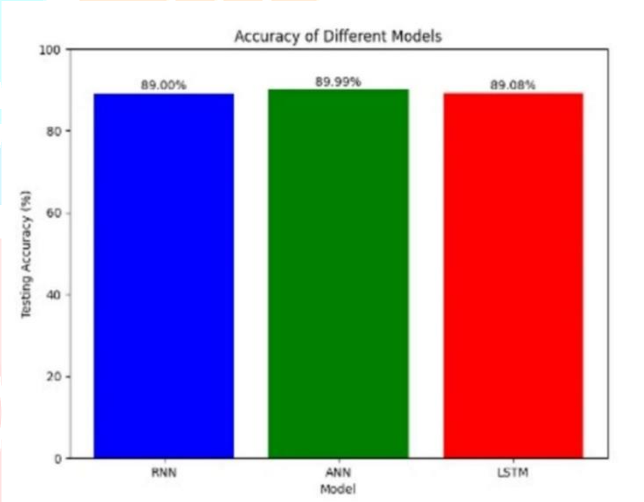


**Fig 6: Accuracy Of Different Models**

The accuracy of three different models is displayed in the figure above. Finally, diabetes is predicted based on the best performance of the model.

## VI. CONCLUSION

The implementation of Federated Learning (FL) revolutionizes healthcare analytics by addressing data privacy, security, and scalability concerns inherent in centralized approaches. FL enables collaborative model training across distributed healthcare institutions while preserving patient data confidentiality. Secure communication protocols and encryption mechanisms safeguard model updates during transmission, mitigating privacy and tampering risks. FL's scalability accommodates diverse healthcare institutions, crucial for managing increasing data volume efficiently. Practical implementation demonstrates promising results in predictive healthcare analytics, particularly in diabetes prediction. Aggregating locally trained models into a global one yields competitive performance while maintaining data privacy and security. The user-friendly interface streamlines user experiences, enabling healthcare professionals to input patient data and access predictive analytics results seamlessly. FL in healthcare signifies a paradigm shift toward decentralized, privacy-preserving machine learning, offering a robust framework for collaborative model training and predictive analytics in real-world healthcare settings. After training different models, we achieved 89.00% accuracy with an RNN, 89.99%

accuracy with an ANN, and 89.08% accuracy with an LSTM model; therefore, we proceed with the ANN model to predict diabetes. After successfully submitting all the weights, we obtained these accuracy levels through the best performer strategy in the global model/central server. This approach ensures that the highest-performing model is utilized for predictions, enhancing the overall effectiveness and reliability of the diabetes prediction system in a collaborative healthcare environment.

## REFERENCES

[1] Mohammad Moshawrab, Mehdi Adda, Abdenour Bouzouane, Hussein Ibrahim, Ali Raad, "ReviewingFederated Machine Learning and Its Use in Diseases Prediction", Sensors, vol.23, no.4, pp.2112, 2023.

[2] Yaqoob, M.M.; Nazir, M.; Khan, M.A.; Qureshi, S.; Al-Rasheed, A. Hybrid Classifier Based Federated Learning in Health Service Providers for Cardiovascular Disease Prediction. Appl. Sci. 2023, 13, 1911. https://doi.org/10.3390/app13031911

[3] Wonsuk Oh1 , Girish N Nadkarni1,2,3, "Federated learning in healthcare using structured medical data" https://doi.org/10.1053/j.akdh.2022

[4] Yaqoob, M.M.; Nazir, M.; Yousafzai, A.; Khan, M.A.; Shaikh, A.A.; Algarni, A.D.; Elmannai, H. Modified Artificial Bee Colony Based Feature Optimized Federated Learning for Heart Disease Diagnosis in Healthcare. Appl. Sci. 2022, 12, 12080.
https://doi.org/10.3390/ app122312080

[5] Joshi, R.D.; Dhakal, C.K. Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. Int. J. Environ. Res. Public Health 2021, 18, 7346. https://doi.org/10.3390/ ijerph18147

[6] RAHIMLOO1 , Ahmad JAFARIAN2, "Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them" Bulletin de la Société Royale des Sciences de Liège, Vol. 85, 2016, p. 1148 – 1164.

[7] Mujumdara , Dr. Vaidehi Vb " Diabetes Prediction using Machine Learning Algorithms " international conference on recent trends in advanced computing 2019, icrtac 2019.

[8] N. Srividhya1 , K. Divya1 , N. Sanjana1, K. Krishna Kumari1 , M. Rambhupal1, "DIABETES PREDICTION USING SUPPORT VECTOR MACHINES ", 1Department of Information Technology, Vasireddy Venkatadri Institute of Technology Guntur.https://doi.org/10.36713/epra14769

[9] Rahman, Karimon Nesha, Mariam Akter, Md. Sheikh Giash Uddin. Application of Artificial Neural Network and Binary Logistic Regression in Detection of Diabetes Status, Science Journal of Public Health, Vol. 1, No. 1, 2013, pp. 39-43. doi: 10.11648/j.sjph.20130101.16

[10] Mujumdara , Dr. Vaidehi Vb " Diabetes Prediction using Machine Learning Algorithms " international conference on recent trends in advanced computing 2019, icrtac 2019

[11] N. Srividhya1 , K. Divya1 , N. Sanjana1, K. Krishna Kumari1 , M. Rambhupal1, "DIABETES PREDICTION USING SUPPORT VECTOR MACHINES ", 1Department of Information Technology, Vasireddy Venkatadri Institute of Technology Guntur, https://doi.org/10.36713/epra14769