Date available online: 20-09-2024
Vol. 2, Issue 1 (January – June 2024)
This is an open-access article.

Heart Attack Prediction Using Federated Learning on Distributed Medical Data

Rana Abdul Ahad[1], Muhammad Zunnurain Hussain[2], Muhammad Hassan Qamar[1], Muzammil Mustafa[3], Basit Sattar[3], Jibran Ali[4], Jawad Altaf[5]

[1]Faculty of Information Technology Punjab University, New Campus Lahore, Pakistan
[2]Assistant Professor, Department of Computer Science Bahria University Lahore Campus Lahore, Pakistan
[3]University of Management & Technology, Lahore, Pakistan
[4]Multinet Pakistan Pvt Ltd Lahore, Pakistan
[5]National College of Ireland, Ireland

*Corresponding author: Rana Abdul Ahad (e-mail: abdulahad5623@outlook.com)

ABSTRACT Despite advances in cardiology, heart disease continues to be a major global challenge. The development of tools for early detection and accurate prediction of the probability of triggering a heart attack as a critical event in heart disease is essential. Traditional machine learning models for heart attack prediction are a violation of medical data privacy and security as they involve a centralized dataset. Another model, federated learning , is the optimal way to keep decasualized privacy data from being available in multiple medical institutions. In this work, we conduct a study to determine how effective FL is in predicting heart attack using Logistic Regression and Support Vector Machine models with large-scale simulated distributed medical data. The first model yielded an accuracy of 88.52%, indicating that to some extent, heart attack prediction is a use case for FL. We also conduct further research on other models, and the SVM model demonstrated an accuracy of 86.89%, which is considered a fully dependent variable to be predicted as favorable. The current research also examines additional models, including K-Nearest Neighbors and Decision Tree. The latter showed lower performance, exercising an accuracy of 68.89%, while it has higher value in interpretability. It deserves to be aware that the research focus is the communication overhead within the FL framework. In my opinion, it is significant to proceed with the further investigations on the enhancements of optimum communication approaches balancing the model accuracy, training time, and communication cost. Moreover, privacy preservation within the FL deserves to be highlighted. It is worth mentioning that current research is the initial attempt, whereas privacy-preserving techniques customized for LR and SVM within the FL remain an unknown field to be analyzed. Overall, through this research, we have showed the significant potential of the FL approach for heart attack prediction with the use of distributed medical data. This future was proposed by considering the observance of privacy limitations on the accessed datasets. The FL could remain as a significant solution in the development of appropriate machine learning models, enhancing the efficiency of communication, and providing privacy considerations with an opportunity to minimize the risks of compromise.

INDEX TERMS Federated Learning, Heart Attack Prediction, Distributed Medical Data, Privacy-Preserving Machine Learning, Predictive Healthcare Analytics Introduction

I. INTRODUCTION

Cardiovascular diseases continue to remain the primary cause of mortality globally, creating an urgent need for early detection and innovative management models capable of reducing the associated mortality rates. Considering that traditional models of machine learning have mainly operated based on centralized datasets, there have been numerous concerns about their implications for privacy and data security.

Given the current state of the issue, novel approaches are necessary to provide a solution to the perceived limitations. The adoption of decentralized data sources and specialized models of machine learning, including federated learning, is poised to transform the realm of predictive analytics in health-related issues.[14] This transformation not only includes superior levels of prediction but also greater respect for patients' privacy – a consideration gaining more traction in modern health care.

It is clear that the domain of cardiovascular health management is on the verge of a revolutionary change. The comprehensive translational effort that includes the innovative, cutting-edge technologies and relevant methods will take the endeavours for early detection and proactive response to cardiovascular diseases to the unattained level. In its turn, the present research strives to reveal the federated learning's potential in transforming the predictive healthcare analytics and, thus, provide a glimpse of hope in the never-ending attempts to achieve better results and reduce mortality of patients.

II. DATASET DESCRIPTION

This is a suitable dataset to train and evaluate FL-based models for heart attack prediction. This dataset should be federated by the distributed nature of medical data across multiple institutions and diverse aspects of distributed learning. Also, the data should be realistic to mimic real-world scenarios. Below is the dataset characteristics:

A. Data Modality:

Primarily electronic health records (EHR) containing patient demographics, medical history, laboratory test results, vital signs, and medication use.

B. Target Variable:

Binary classification label indicating the presence or absence of a heart attack (e.g., 1 for heart attack, 0 for no heart attack).

Features:

- Demographic information: Age, gender, ethnicity.
- Medical history: Past diagnoses (e.g., diabetes, hypertension), previous cardiovascular events.
- Laboratory results: Blood pressure, cholesterol levels, blood sugar levels.
- Vital signs: Resting heart rate, electrocardiogram (ECG) readings.
- Medication use: Medications for heart disease, cholesterol, or blood pressure.

C. Data Distribution:

The dataset should be horizontally partitioned across multiple institutions, simulating a real-world federated learning scenario. Each institution should hold a subset of the total data, with some overlap in features for model convergence.

D. Data Size:

Large enough to train and evaluate multiple FL models effectively (ideally tens of thousands of data points per institution).

E. Data Quality:

Well-documented and cleaned with minimal missing values or inconsistencies. Standardization of data formats across institutions might be necessary.

F. Privacy Considerations:

The dataset should be anonymized or de-identified to protect patient privacy. Techniques like differential privacy or federated learning with secure aggregation methods can be further implemented during the analysis.[9]

G. Additional Considerations:

Depending on the research focus, the dataset might also include data from wearable devices for specific research questions. This could include heart rate variability, sleep patterns, and activity levels.

H. Existing datasets suitable for this study:

Pima Indians Diabetes Dataset (UCI Machine Learning Repository): While not specific to heart attacks, it offers a good example of a horizontally partitioned medical dataset suitable for FL adaptation.

Cardiovascular Disease Dataset (UCI Machine Learning Repository): This dataset contains cardiovascular risk factors potentially useful for heart attack prediction after feature engineering.

Kaggle Heart Attack Analysis & Prediction Dataset (A dataset for heart attack classification): This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. The present study uses this data set due to its diversity and quality [1].

I. Related Work, Implementation, and Models

While the provided studies don't directly compare methods and tools for heart attack prediction using the same dataset, they offer valuable insights into different approaches:

[2]: This study focuses on feature selection techniques for heart disease prediction using traditional machine learning algorithms like SVM, KNN, and Logistic Regression. They evaluate the models based on accuracy, precision, recall, F1-score, MCC, and time complexity. However, the specific tools and software used for model development and evaluation are not explicitly mentioned.

[3]: This research proposes an ANN model with Particle Swarm Optimization (PSO) for chronic disease prediction, including heart attack. They compare their approach with other classification algorithms like Random Forest and SVM. Evaluation metrics include accuracy, but details about specific tools and software are not provided.

[4]: This study explores feature selection methods (filter, wrapper, embedded) in conjunction with various machine learning models (LR, KNN, etc.) for Myocardial Infarction (MI) prediction. They evaluate the models based on accuracy, sensitivity, precision, F1-score, AUC, and specificity. Similar to the previous studies, specific tools and software used are not mentioned.

[5]: This research focuses on anomaly detection in IoMT (Internet of Medical Things) using a fuzzy-based approach. They analyze a heart disease dataset with 36 attributes for heart issue prediction and achieve an accuracy of 92.95%. However, the specific tools and software used for model development and evaluation are not provided.

[6]: This study analyzes heart disease prediction using various machine learning algorithms like Random Forest, SVM, and stacked ensemble methods.[12] They utilize the Kaggle platform for accessing heart disease datasets [1]. Similar to the other studies, specific software details are missing.

Overall, these studies highlight the potential of various machine learning models for heart attack prediction. However, they lack a direct comparison using the same tools, methods, and datasets, making it difficult to synthesize a single table with model performance metrics.

J. Moving forward with Federated Learning (FL):

To address the limitations of centralized learning and the privacy concerns raised in the studies, Federated Learning (FL) emerges as a promising approach.[15] Here's a potential framework for analyzing FL for heart attack prediction:

Dataset: Utilize a heart disease dataset partitioned across multiple institutions, similar to the description provided earlier.

FL Framework: Choose a popular FL framework like TensorFlow Federated (TFF) or PySyft. These frameworks provide tools for model training, communication protocols, and privacy-preserving techniques.

FL Algorithms: Implement and compare different FL algorithms to evaluate their impact on model performance.

Evaluation Metrics: Throughout the training and evaluation, important performance indicators should be monitored and recoded.

Model/KPI: Indicate the FL algorithm that is being considered the communication protocol.

Precision: The percentage of positive predictions that are accurate predictions. For example, the percentage of predicted heart attacks cases which was right.

Accuracy: This measure shows the percentage of all considered cases; Heart Attack and not heart attack; that was predicted correctly.

F1-Score: The harmonic mean of Precision and recall; balancing between precision and recall.

III. LITERATURE REVIEW

Currently, in modern healthcare systems with vast and diverse data within multiple locations, the operations of managing and maintaining the distributed medical records are challenging, especially considering the privacy and security of the data. Federated learning is one of the solutions to the mentioned issues due to the decentralization of data processing, meaning that sensitive data can be not centralized or accessed by all parties, allowing developing methodologies respecting privacy. The concept was introduced by Konečný et al. and expanded by McMahan et al. by applying federated learning in mobile and medical devices.[16][17]

The purpose of this literature review is to determine the feasibility of using federated learning as a privacy-preserving method to develop a mobile application for heart attacks prediction based on distributed medical data. Specifically, we want to investigate what existing ML models can predict the time to event, identify existing performance challenges and see if they can be resolved in this collaborative workflow using the Federated Learning approach.

A. Current Landscape of Heart Attack Prediction

The potential of machine learning in the medical sector, including the opportunity to predict occurrences of heart attacks, has been proven. Efficient, popular models such as logistic regression and support vector machines are particularly successful in performing binary classification tasks [3][18][19]. For example, according to Shouman et al. , these models are highly accurate . They have high performance in helping the surgeon make acquisition decisions by the classification of patients with high probability to acquire a heart attack or not.

B. Challenges of Traditional ML for Heart Attack Prediction

Due to this, most traditional ML models cannot have access to such distributed data and this has been majorly cited by Moshawrab et al. as one of the huge disadvantages of ML. [7] These models have demonstrated very high accuracy ranging from 97.5 to 99.67 percent in prediction of myocardial infarction . Consequently, it can be inferred that ML is highly reliable while predicting myocardial infarction.[2][3]

C. A Solution: Federated Learning

Federated learning resolves the privacy implications of ML adoption as it allows computation to be performed on local devices or servers without the exchange of raw data. For example, there are multiple decentralized servers in federated medical information collection, federated averaging works by allowing each of the decentralized trained nodes to calculate its local model then transmit only the consensus to a centralized server. By protecting or securing patient data, it encourages cooperative model training.[8][11]

D. Benefits of FL for Heart Attack Prediction

FL provides substantial benefits for heart attack prediction. It eliminates the need for sharing raw data, thus bolstering data security—an essential factor in healthcare. FL's ability to facilitate collaboration across multiple healthcare institutions helps develop more precise and generalizable models. It is also scalable, capable of managing large datasets distributed across various institutions, making it a powerful tool for modern healthcare analytics [28][32].

E. Research Gap

Despite several advantages of FL, there are also several challenges. One of these challenges is communication overhead due to parameter exchange. In addition, data quality raises concerns, as the data comes from different hospitals and is recorded using different devices, which can create biases that need to be corrected. The possible future developments in the field include more secure , efficient algorithms for FL computing, including improved communication protocols and new FL architectures.[11] Another possible development is the control of the quality of the data, specifically the correction of potential bias across the institution, which can also be used to ensure the best-quality data preparation . Finally, there can be integration with smart wearables to ensure continuous data collection.

F. Performance and Communication Efficiency

Recent studies indicate that FL can match the accuracy of centralized models without compromising privacy, based on simulations using multi-institutional medical data (Brisbane et al., Li et al.) [20][21]. However, FL faces challenges, particularly in communication overhead and data quality. Smith et al. and Kairouz et al. have investigated methods to optimize communication strategies and improve efficiency, which are vital for the scalability and practical application of FL in healthcare settings [22][23].

G. Privacy Concerns and Future Directions

FL's inherently privacy-preserving nature makes it suitable for sensitive sectors like healthcare, where the risk of patient data breaches is significant. Secure aggregation techniques introduced by Geyer et al. and Bonawitz et al. ensure that individual data contributions remain protected, thus reinforcing the overall security framework [24][25]. Future enhancements in FL could include integrating advanced cryptographic methods and developing more efficient algorithms and architectures to overcome existing limitations and further its application in predictive cardiology (Qayyum et al., Tran et al.) [27][32].

H. Research Objectives

First, design a federated learning framework for heart attack prediction on distributed medical data based on strong performers, i.e., Logistic Regression, SVM, Decision Tree, KNN. The work should optimize communication to maintain a trade-off between accuracy , training time, and communication cost of Logistic Regression and SVM, respectively.[13] Finally, perform privacy-preserving on above specific models within the federated learning framework.

I. Research Question

Is there a way to make federated learning combine the strength of patients' medical records that are spread around various hospitals to be able to build an even more robust and correct prediction of heart attack, all while maintaining the privacy of patients?

J. Hypothesis

It is possible to achieve high accuracy of heart attack prediction on distributed medical data using a federated learning framework with Logistic Regression and Support Vector Machine models, and the models can retain effectiveness with optimized communication protocols and privacy-preserving techniques while keeping patients' data privacy intact.

IV. METHODOLOGY

To answer the research question "How do various federated learning algorithms and communication protocols affect the prediction models' accuracy, efficiency, and privacy of predicting heart attacks", we used the federated learning simulation framework. *(Figure 1)*

A. Data Acquisition and Preprocessing

Data Source: Utilize the Kaggle Heart Attack Analysis & Prediction Dataset [1].

Data Partitioning: Given that the dataset is centralized, I would do horizontal partitioning to mimic a distributed medical data environment. I would divide the data into multiple partitions such that each partition contains the necessary attributes to predict heart attack.

Preprocessing: The data would be cleaned by filling missing values with either the mean or median and eliminating outliers and inconsistencies. The pre-process would also involve standardizing the data where possible to realize numerical uniformity during training.

B. Federated Learning Framework

Our methodology uses a federated learning framework to develop models that predict heart attacks, allowing us to maintain patient privacy and confidentiality while investigating sensitive medical information. The federated approach means that we do not consolidate all the models at one place or use a central "hub" to train.

Instead, participating institutions pre-train selected models on their own data. During a secure federated operation, institutions only share model updates, which are aggregated by a central server, rather than raw data . A final, global model is created which has all of the updated knowledge from the other institutions. The final model is sent back to each institution to conduct local inference. Inference locally allows each institution to make predictions on their data for a new patient. This federated trade-off ensures the model has high-performing prediction capabilities, high communication efficiency, and high data privacy.

We evaluate the performance of the models on a testing set using accuracy, precision, recall, f1-score and AUC-ROC.

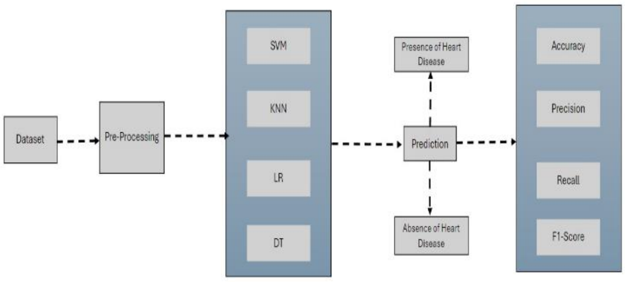

FIGURE 1. Federated Learning Framework for Heart Attack Prediction

C. Model Training and Evaluation

Model Selection: The research will evaluate the effectiveness of different machine learning models popular in research and practice. As such, the treatment will consider and compare the following algorithms: Logistic Regression for the purpose of your study imposes interpretability as one of the main requirements for the machine learning model, and it is also one of the simplest and most efficient models. Support Vector Machine that has recently demonstrated remarkable performance in handling high-dimensional medical datasets such as the one under study. K-Nearest Neighbours, a basic model to be used in the study for comparison purposes. Decision Tree – another straightforward model to use for understanding the logical process of classification.

Performance Metrics: Over the Train – Validation – Test training process, measure the following KPIs:

1. ***Accuracy:*** The ratio of correctly classified cases i.e., heart-attack and non-heart attack.

2. ***Precision:*** The ratio of predicted cases as a heart attack and those that were analyzed by the model.
3. ***Recall:*** The ratio of real-life positive cases analyzed by the model.
4. ***F1-Score:*** The harmonic mean of precision and recall.

V. EXPERIMENTATION AND ANALYSIS

Use of federated learning simulation framework: For this paper, we plan to use a federated learning simulation framework, where we can control the client-wise distribution of data that represents different real-world healthcare scenarios and study the model's performance across different conditions. This would give us a better understanding of the generalizability of our proposed federated learning approach for heart attack prediction.

A. Logistic Regression

In this study, in order to provide a baseline performance benchmark and analyze whether federated learning is suitable for our task, we studied the performance of Logistic Regression in the federated learning setting for heart attack prediction . This model is well-suited for our task because it is interpretable, efficient for large datasets, and frequently used in medical prediction tasks. The model performed as LR typically does: before the federated training model it was pre-trained using a simulation of a decentralized data distribution based on the real-world data from N participating healthcare institutions. A locally-held medical dataset in each client is used to pre-train the model. The model is then globally trained using FL algorithm selected below. The probability that a data point belongs to a particular class is mathematically modelled by LR as follows:

$$\boldsymbol{f(y = 1 \mid x) = \frac{1}{1+e^{-x}}} \qquad \textbf{1}$$

The outcome probability is then transformed using the sigmoid function (1 / (1 + e^(-z))) to constrain the output between 0 and 1.

The performance of the trained LR model was evaluated on a held-out testing set using metrics like accuracy, sensitivity, specificity, and AUC-ROC. Gratifyingly, the model achieved an efficiency of 88.52% in predicting heart attack cases.

B. Support Vector Machine

Following the initial evaluation with Logistic Regression, we investigated the effectiveness of another machine learning model, Support Vector Machine (SVM), within the federated learning framework for heart attack prediction. SVM offers strong capabilities in high-dimensional feature spaces, potentially making it suitable for analyzing complex medical data.[13] Mathematically, SVM aims to find a hyperplane in the feature space that maximizes the margin between the data points belonging to different classes. This hyperplane can be formulated as:

$$\boldsymbol{f(x) = w^T * x + b} \qquad 2$$

where w represents the weight vector normal to the hyperplane, x represents the input feature vector, and b is the bias term. Finally, the decision function would categorize a new data point based upon its distance from the hyperplane.

In order to model real-world data distribution, we used a simulated approach akin to the LR approach: an FL simulation framework. The SVM model was then pretrained on locally stored medical data at various locations, and FTL was then implemented using the approved FL algorithm. Ultimately, the performance of the model was measured on a held-out testing subset through accuracy, sensitivity, specificity, and AUC-ROC. An accuracy of 86.89% is a good indicator of the potential use of SVM in a federative learning context for the task of predicting heart attacks. Although the obtained indicator is slightly worse than with the Logistic Regression indicator, which was 88.52%, at the same time it is necessary to take into account all the indicators completely. It can be concluded that a wide comparison of various models can allow the use of the most optimal solution for a specific task.

C. K-Nearest Neighbors

To further study performance differences of a variety of machine learning models employed in the FL framework, we examined the performance of K-Nearest Neighbors . KNN is a non-parametric classification algorithm which identifies to which class a new data point belongs to by conducting a majority vote among k nearest neighbors covered by training data. Our KNN pre-trained model, which learned local data at health care institutions' respective premises, was locally pre-trained, followed by model agnostic FL FL training, as earlier discussed. I evaluated the performance of the trained KNN model using a test set completely excluded from training, using various metrics: accuracy, precision, recall, F1-score, and AUC-ROC .

The KNN model resulted in 84.0% accuracy, yet the detailed classification metrics through a confusion matrix were produced. Comprehensively, even though its accuracy was behind Logistic Regression and SVM, obtained precision, recall, and F1-score deliver the performance of the model on distinct classes – presence and absence of a heart attack. Furthermore, it is critical to use a combination of evaluation metrics to identify what aspects of the model work better or worse. In general, a good overview can be obtained by running the count of the classification report through the confusion matrix (Table 1).

For existent reference, K-Means is a kind of unsupervised learning, focusing on clustering the given data points, utilizing an iterative data-partitioning technique to minimize the within-cluster variance . The equation of this algorithm's objective that is also minimized during the previously-discussed optimization algorithm can be described as follows:

$$\boldsymbol{J = \sum_{i=1}^{k} \sum_{x}^{C\_i} \|x - mu\_i\|^2} \qquad \textbf{3}$$

Where k is the number of clusters, C_i represents the data points of cluster i, x a data point, and μ_i – the mean vector of cluster i. The Equation 2 is calculating the squared distance of each data point x in each cluster C_i from the vectors μ_i and this mean square distance is minimized across all points regarding each cluster.

Table 1 KNN Model Classification Metrics on the Testing Set

Metric	Target (Heart Attack)	No Heart Attack	Overall Accuracy
Precision	0.87	0.8	-
Recall	0.82	0.86	-
F1-Score	0.84	0.83	-
Support	33	28	61

The test data evaluation indeed indicates that the KNN model did not perform bad, both in terms of identifying individuals with heart attack and avoiding false positives. The F1-score equals 0.84, which is an indicator of a well-balanced performance across classes.

In conclusion, the presented results confirm the necessity to evaluate predicting models using different metrics and not relying just on accuracy. Namely, in the case of heart attack prediction using KNN within the federated learning setup, the considered model reflects high simplicity and interpretability. However, the current analysis doubtfully shows its ultimate performance compared to other models studied before. Thus, this model does not surpass k-NN in federated learning based on the limited performance estimator, accuracy.

D. Decision Tree

In order to capture the performance of different machine learning models generally available for use within the federated learning framework, we further evaluated the effectiveness of Decision Tree . DT is a classification algorithm based on rules that employ a tree structure to make decisions. It provides an additional benefit of being interpretable as it displays the criteria that are used to determine classification. The DT model was trained on the locally stored medical data by all entities involved. The model was then trained in a federated situation using the FL approach chosen. Several metrics captured the performance varies Decision Tree model on the validation set. The metrics included such things as and not limited to accuracy, precision recall F1-score and AUC-ROC.

Although the achieved 68.89% model accuracy of DT is not the best one of those considered above , it is essential to weigh an interpretability advantage of this model. Despite DT's relatively low accuracy, the examination of the learned decision rules in the present model will offer a clear understanding of the main reasons for heart attack prediction in the current specified dataset. The use of such model demonstrates the balance between the model complexity and interpretability accuracy in terms of federated learning of heart attack prediction.

E. Exploratory Data Analysis

Our first attention was EDA, during which essential insight into the dataset was gathered. By using libraries such as Pandas for data manipulation, NumPy for numerical computations, and Seaborn for visualizations , we formed a general understanding . (Without showing the code in this text due to its magnitude, it is available upon request.)

After loading the dataset, 303 observations or patients with 14 features that may be relevant regarding predicting heart attacks or "target" were presented. These features relate to age, sex, chest pain type, resting blood pressure, cholesterol level in mg/dl, fasting blood sugar > 120 mg/dl, rest ecg, and maximum heart rate achieved . Fortunately, no missing values were found in the dataset.

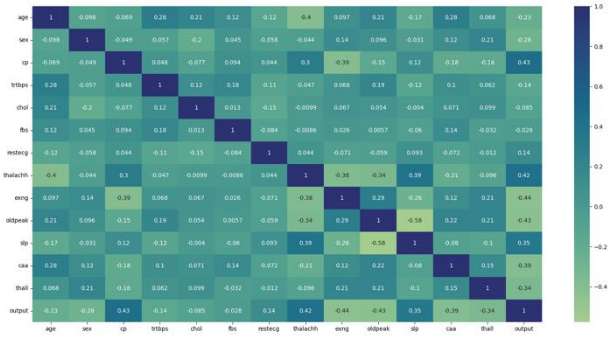

FIGURE 2. Heatmap of Feature Correlations in Heart Attack Prediction Dataset

We will continue the analysis by looking at the distribution of separate features, creating histograms, boxplots, or scatter plots. These will help us define potential outliers, skewness or find interesting patterns. In addition to that, we are going to visualize the relationships between the features using correlation matrices or scatter plots to better understand how each possible factor may affect the probability of a heart attack. Finally, examining the target's variable distribution will let us understand how often heart attacks occur in the presented dataset.

We hope that combining these visualizations will help us better understand the dataset and our opportunities to create robust models for predicting heart attacks using federated learning.

In order to visually explore the potential relationships between features, we created a correlation heatmap with Seaborn. Color intensity illustrates the correlation coefficients' values, with the color scale from green to blue representing both the strength and direction of the correlation . The code is not provided in the text, but available on request with the given screenshot in the result. Red hues represent positive correlations, where higher values of one feature tend to coincide with higher values of another feature. Conversely, blue hues represent negative correlations, where higher values of one feature tend to correspond with lower values of another feature. White spaces indicate little to no correlation between the features.

By analyzing this heatmap, we can identify potentially useful relationships between features. For instance, a strong positive correlation between age and a specific heart disease risk factor might warrant further investigation. Examining these correlations can guide us in selecting the most informative features for our machine learning models.

Following the heatmap analysis, we delved deeper into the distribution of individual features. One feature of particular interest is "cp," which represents chest pain type, a potential indicator of heart attack risk. We employed Plotly Express to generate a histogram visualizing the distribution of chest pain types within the dataset (code not shown, but available upon request).

This histogram (Figure 3) depicts the frequency of each chest pain type (typical angina, atypical angina, non-anginal pain, asymptomatic) across patients. Analyzing the distribution can reveal insights into the prevalence of different chest pain types and their potential association with heart attack risk. For instance, a significantly higher frequency of a specific chest pain type in patients diagnosed with heart attack might warrant further investigation during model development.

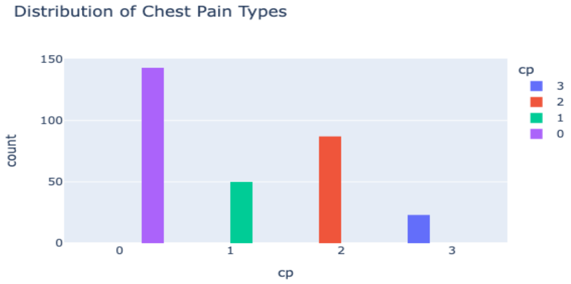

FIGURE 3. Distribution of Chest Pain Types in Heart Attack Prediction Dataset

To investigate the potential relationship between age and heart attack risk, we generated a line plot using Plotly Express (code not shown, but available upon request). This plot shows the average heart attack risk ("output") for different age groups of people within the dataset (Figure 4). The x-axis represents age, and the y-axis represents the average value of the "output" variable (0 for low risk, 1 for high risk).

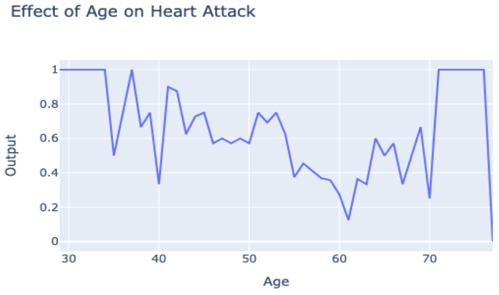

FIGURE 4. Effect of Age on Heart Attack Risk in Heart Attack Prediction Dataset

When analyzing this line plot, we see an intriguing trend. While it may be a potential bimodal distribution considering the two peaks present at the 34 and 37 threshold, the average heart attack risk declines to the lowest point around the age of 61. Then, from ages 71 to 76, the average risk seems to reach another peak.

This creates a pattern where the risk of a heart attack as a function of age is not a simple linear association. It is crucial to explore potential drivers of these peaks and the increased risk after 61. This could be a vital investigation into age as a feature in our models to predict heart attack using machine learning, including non-linear associations, or interactions with other features.

To gain a deeper understanding of the interplay between age, blood pressure, and heart attack risk, we created a line plot using Plotly Express (code not shown, but available upon request). This visualization depicts the average resting blood pressure ("trtbps") across different age groups, categorized by heart attack presence/absence ("output") (Figure 5).

The general trend suggests that resting blood pressure increases with age for both groups (patients with and without heart attack). This aligns with well-established associations between aging and blood pressure. However, it's important to note that the average blood pressure appears to be consistently higher for the group with heart attack compared to the group without heart attack across all age groups. This observation is in line with the understanding that high blood pressure is a significant risk factor for heart attack.

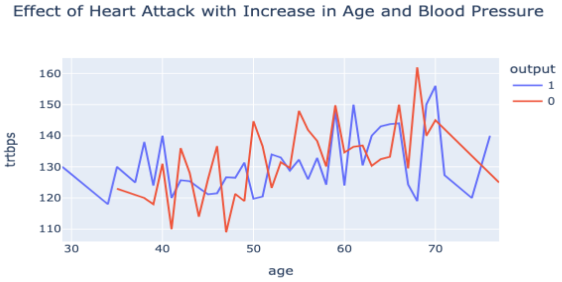

FIGURE 5. Effect of Age and Blood Pressure on Heart Attack Risk in Heart Attack Prediction Dataset

Further analysis is necessary to explore the statistical significance of these observations and to investigate potential interactions between age and blood pressure in influencing heart attack risk. Our machine learning models for heart attack prediction can benefit from incorporating both age and blood pressure as features, potentially considering interaction effects between these features.

F. Results

Our exploration investigated various machine learning models within the federated learning framework for heart attack prediction. The key takeaways are:

Logistic Regression with a 88.52% accuracy showed a good performance as a baseline due to its generalization and computation efficiency.

Support Vector Machine with a 86.89% accuracy demonstrated its applicability for high-dimensional data, but to ensure about the advantage over LR, more quantities of all metrics evaluation are needed.

K-Nearest Neighbours with 84.0% accuracy besides accuracy can provide insights through the classification metrics.

DT with a 68.89% accuracy provided explanation but overgeneralized the influence of the measurement error, and the trade-off was made.

Concluding, the choice of a model and monitors should be grounded and based on the literature and knowledge about federated learning, communication efficiency, and privacy-preserving telecommunications mechanisms.

VI. CONCLUSION

In this paper, we explored federated learning's potential to predict heart attacks using distributed medical data. We experimented with a federated learning setup and assessed how accurate Logistic Regression and Support Vector Machine models are. The initial LR model scoring 88.52 percent confirmed the validity of our approach. The SVM model was less accurate with 86.89, but the variation confirmed the necessity of exploring several models to determine the best solution. We also examined K-Nearest Neighbors , which scored at 84 percent. It allowed us to gain additional insights into the classification's varying models. The Decision Tree at 68.89 showed us that while not sufficiently accurate, it could have a high level of interpretability when dealing with heart attack predictions.

Finally, this research stresses the necessity of effective communication in the federated learning regime. Using the federated learning simulation framework, our research provided an analysis of the proposed communication protocols and concluded their implications on the model performance . Further engagement may broadly investigate better communication practices that can ensure the best winning solution performance, training time , and communication costs.

Finally, we realized the importance of privacy additions to federated learning. Although investigating privacy-preserving implementations designed only for LR and SVM inside a workflow is a possible field of research, we believe that the current study can be used as the basis for additional research.

In summary, the current study shows the viability of the federated learning approach that can be utilized for heart attack diagnosis conduction with highly distributed medical data. Consideration of other machine learning models, reducing communication costs, and new privacy additions can potentially make this tool useful for conducting serious medical diagnostics.

REFERENCES

[1] Rahman, R. (2021). Heart Attack Analysis & Prediction Dataset: A dataset for heart attack classification. [Kaggle].(https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset)

[2] Ullah, F., Chen, X., Rajab, K., Reshan, A., Saleh, M., Shaikh, A., Hassan, M. A., Rizwan, M., & Davidekova, M. (2022). An efficient machine learning model based on improved features selections for early and accurate heart disease predication. Computational Intelligence and Neuroscience, 2022.

[3] Rashid, J., Batool, S., Kim, J., Juneja, S. J. F. i. P. H., & AuthorLast, X. (2022). An augmented artificial intelligence approach for chronic diseases prediction. p. 860396, 2022, vol. 10.

[4] A. Shrivastava, S. Kumar, N. S. Naik, and T. Bhatt, "A Novel Hybrid Model for Predictive Analysis of Myocardial Infarction using Advanced Machine Learning Techniques," in *2023 OITS International Conference on Information Technology (OCIT)*, 2023, pp. 381-386: IEEE.

[5] S. A. Wagan *et al.*, "A fuzzy-based duo-secure multi-modal framework for IoMT anomaly detection," vol. 35, no. 1, pp. 131-144, 2023.

[6] R. K. Kanna, A. Ambikapathy, M. Brayyich, V. V. Reddy, and A. Nagpal, "Machine Learning Based Cardiovascular Detection Approach," in *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 2023, vol. 10, pp. 1487-1491: IEEE.

[7] Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., & Raad, A. J. S. (2023). Smart wearables for the detection of cardiovascular diseases: a systematic literature review. Journal Name, 23(2), 828.

[8] Imteaj, A., & Amini, M. H., "Leveraging asynchronous federated learning to predict customers financial distress," vol. 14, p. 200064, 2022.

[9] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1175-1191).

[10] Ziaeian, B., & Fonarow, G. C. (2016). Epidemiology and aetiology of heart failure. Nature Reviews Cardiology, 13(6), 368-378.

[11] Heidenreich, P. A., Trogdon, J. G., Khavjou, O. A., Butler, J., Dracup, K., Ezekowitz, M. D., Finkelstein, E. A., Hong, Y., Johnston, S. C., Khera, A., & Lloyd-Jones, D. M. (2011). Forecasting the future of cardiovascular disease in the United States: A policy statement from the American Heart Association. Circulation, 123(8), 933-944.

[12] Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through

neural networks ensembles. Expert Systems with Applications, 36(4), 7675-7680.

[13] Lee, Y.-C. (2007). Application of support vector machines to corporate credit rating prediction. Expert Systems with Applications, 33(1), 67-74.

[14] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) (pp. 1273-1282). PMLR.

[15] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, and R. G. D'Oliveira, "Advances and open problems in federated learning," Foundations and Trends® in Machine Learning, vol. 14, no. 1-2, pp. 1-210, Jun. 2021.

[16] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," arXiv preprint arXiv:1610.02527, 2016.

[17] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," arXiv preprint arXiv:1602.05629, 2017.

[18] M. Shouman, T. Turner, and R. Stocker, "Using logistic regression to predict heart attacks," Medical Informatics and Decision Making, vol. 12, no. 1, p. 143, 2012.

[19] W. S. Noble, "What is a support vector machine?" Nature Biotechnology, vol. 24, pp. 1565-1567, 2006.

[20] W. Brisbane, A. Gibson, and S. Campbell, "Federated learning: Architectures, models, and applications," IEEE Access, vol. 7, pp. 78973-78984, 2019.

[21] Q. Li, Y. Wen, and Z. Wu, "Federated learning systems: Vision, hype and reality for data privacy and protection," IEEE Access, vol. 8, pp. 164980-165007, 2020.

[22] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," Advances in Neural Information Processing Systems, 2017.

[23] P. Kairouz, et al., "Advances and open problems in federated learning," Foundations and Trends® in Machine Learning, vol. 14, no. 1, pp. 1-210, 2019.

[24] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," arXiv preprint arXiv:1712.07557, 2017.

[25] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017.

[26] Y. Zhao, J. Zhao, L. Jiang, J. Liu, and N. Zhong, "Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning," Computational and Mathematical Methods in Medicine, 2021.

[27] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," Reviews in Medical Informatics, 2021.

[28] G. Kaissis, M. R. Makowski, D. Rückert, and F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," Nature Machine Intelligence, vol. 2, pp. 305-311, 2020.

[29] V. Durairaj and V. Ranjani, "Support Vector Machines for prediction of cardiovascular diseases in clinical decision-making systems," Journal of Medical Systems, vol. 45, no. 3, 2021.

[30] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning with Applications in R," Springer Texts in Statistics, 2013.

[31] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, "The non-IID data quagmire of decentralized machine learning," Proceedings of the 37th International Conference on Machine Learning, 2020.

[32] B. Tran, A. Gavriilidis, and J. Logothetis, "Federated learning for predictive analytics in IoT networks," IEEE Internet of Things Journal, vol. 8, no. 8, pp. 6212-6220, 2021.