Precision at Heart: An IoT-based Vertical Federated Learning Approach for Heterogeneous Data-Driven Cardiovascular Disease Risk Prediction

Sulfikar Shajimon [1], Raj Mani Shukla [2], and Amar Nath Patra [1]

[1]Affiliation not available
[2]Anglia Ruskin University

December 02, 2023

Precision at Heart: An IoT-based Vertical Federated Learning Approach for Heterogeneous Data-Driven Cardiovascular Disease Risk Prediction

Sulfikar Shajimon
Computing and Information Science
Anglia Ruskin University, Cambridge
ss2894@student.aru.ac.uk

Raj Mani Shukla
Computing and Information Science
Anglia Ruskin University, Cambridge
raj.shukla@aru.ac.uk

Amar Nath Patra
Radford University
Radford, USA
apatra@radford.edu

***Abstract*—Cardiovascular disease (CVD) poses a serious threat to individual health, highlighting the importance of early detection and proactive mitigation. With advancements in consumer electronics such as wearables and IoT, there exists an opportunity for enhanced CVD prediction for users. Machine Learning (ML) has been widely used to predict CVD risk (high/low) based on various factors and is a critical area of healthcare research. However, sharing data needed to predict CVD with machine learning models is challenging due to privacy concerns. Federated Learning (FL) enables distributed training of ML models without sharing raw data. However, it requires all training features to be available to all clients. To address this problem, we propose a Vertical Federated Learning (VFL) based method designed for use with consumer electronics platforms. The proposed method trains Neural Network (NN) model in a distributed manner where different data features are held by different parties. In this work, each party maintains a portion of separate data features, performs calculations on them locally, and then transfers only the necessary information to jointly train an NN model. We employ the proposed method for different use cases where the dataset features are distributed between: i) the patient and the hospital (2-splits); ii) the patient, the doctor, and the laboratory (3-splits); and iii) the patient, the doctor, the Electrocardiogram (ECG) center, and the laboratory (4-splits). Using a realistic dataset publicly available, we test the proposed methodology.**

***Index Terms*—Vertical Federated Learning (VFL), Machine Learning (ML), Internet of Things (IoT), Cardiovascular Disease (CVD), Privacy-preservation.**

I. Introduction

In recent years, cardiovascular disease (CVD) has been ranked as the leading cause of death worldwide, accounting for almost one-third of all deaths [1]. There is a pressing need to develop solutions for the early detection and diagnosis of the disease. In the realm of healthcare, the integration of the Internet of Things (IoT) has brought about transformative changes, particularly by interconnecting medical professionals, patients, and diagnostic laboratories. This integration aims to improve patient care, diagnostic procedures, and treatment outcomes. Furthermore, the advancement of consumer electronics, such as wearable devices and interconnected devices, has promised to enhance the early detection of CVD risk through techniques like Artificial Intelligence (AI) [2].

Research has been conducted on the use of AI to predict cardiovascular disease (CVD) [3]. To achieve the goal of automatic CVD risk prediction, Machine Learning (ML) algorithms are used to meticulously analyze extensive datasets, often collected from devices such as wearable sensors and hospitals, to identify patterns that may not be obvious to humans. However, one of the concerns with using ML algorithms is that they require a large amount of data for training, which is often inaccessible due to user privacy concerns, especially in healthcare applications.

Federated Learning (FL), occasionally known as distributed collaborative learning, is a popular ML methodology. By strategically utilizing algorithms, FL employs multiple distinct and autonomous clients, each with its unique collection of related datasets. FL differs from traditional centralized ML methods, which typically condense localized data into a single training session. Furthermore, FL contrasts with algorithms that rely solely on datasets with identical distributions. FL enables many entities to collaboratively construct an impervious ML model while negating the need to disclose sensitive data. Consequently, this approach successfully mitigates the pressing predicaments related to data privacy, security, and access rights, as well as enabling unfettered access to heterogeneous datasets [4].

However, a drawback of traditional FL methods is their assumption that every client contains the necessary features and information for training an ML model. This assumption often fails, as different clients may have dissimilar features from the same dataset required for ML algorithm training. In this scenario, a Vertical Federated Learning (VFL) approach involves training a model using sample features stored at disparate locations [5]. This unique approach enables multiple entities to collaborate without collecting disparate sample features at a centralized location. Each entity retains a partial feature set of the data, processes it independently, and shares gradients to build an AI model.

This paper introduces novel techniques for Vertical Federated Learning (VFL) tailored specifically for healthcare-based Internet of Things (IoT) devices. The goal is to train AI models across a network of heterogeneous devices and servers

to predict the risk (high/low) of CVD automatically. Unlike traditional FL methods, where all devices have access to the same data features and labels, in our approach, each device holds distinct features and is unaware of the target labels. The clients in our system, which are healthcare IoT devices, exhibit diversity in the information (features) they possess. For instance, one client could be a hospital managing patient records based on diagnoses from IoT medical instruments, while another might be a laboratory conducting tests like ECG using IoT devices. As a result, these clients maintain different sets of features about the same patients. Our proposed method stands out by addressing this feature disparity among clients. Additionally, our approach assumes that clients lack access to the target labels, unlike in traditional FL setups where clients possess such labels for training the model. The major contributions of this work are as follows:

- We present a novel VFL framework for predicting the risk of CVD (high/low) using machine learning (ML). Our framework prioritizes strict user privacy and acknowledges the diversity of IoT devices (clients), considering the varying features they possess. Additionally, it functions under the realistic assumption that clients lack access to the labels required for training the ML model.
- This paper presents a novel algorithm for training an ML model using VFL, where clients and servers coordinate with each other while simultaneously holding disparate feature sets to maintain confidentiality as well as feature set separation.
- We implement and assess the proposed VFL framework for various realistic case studies on sample feature distribution.
- Finally, we test the proposed framework on a real-world dataset and assess its performance compared to the state-of-the-art.

To the best of our knowledge, this article represents the first attempt to address user privacy, feature separation, and the absence of training labels for clients in predicting CVD risk using ML. The remainder of the paper is organized as follows: In Section II, we discuss recent works pertaining to the utilization of IoT, ML, and FL for CVD risk prediction. Section III elucidates our proposed system model and presents the problem statement. Section IV presents the proposed algorithms for CVD risk prediction. Section V analyzes implementation details, describes datasets, examines various case studies, and presents a comparison with state-of-the-art methods. Finally, in Section VI, we conclude our work and highlight future research directions.

II. Literature Review

CVD remains a significant threat to public health, contributing significantly to global mortality rates. Scholars are increasingly turning to technologies such as Machine Learning (ML) and the Internet of Things (IoT) to develop strategies for mitigating and predicting CVD as technological capabilities expand. This section offers an in-depth literature review on CVD diseases and their prediction using ML, IoT-based frameworks, and Federated Learning (FL) to safeguard privacy.

A. Cardiovascular Disease

CVD is defined as a complex and multifaceted array of destructive disorders that impact the intricately sophisticated cardiovascular system, which comprises an interwoven network of cardiac muscles, arteries, and veins [6], [7]. This encompasses various conditions, including coronary artery diseases, cerebrovascular diseases, congenital heart diseases, rheumatic heart diseases, peripheral arterial diseases, deep vein thrombosis (DVT), and pulmonary embolism, contributing to its numerous and intricate characteristics, posing formidable challenges for modern medical sciences to address [7]. Among these ailments, CVD affects masses with alarming frequency due to its high incidence rate.

B. CVD Prediction Using Machine Learning

Recent studies have explored ML techniques for predicting CVD. Anuar et al. [8] conducted a study to predict CVD from Electrocardiogram (ECG) data using ML. Their research involved a prospective population-based case-control study with sixty participants from the Malaysian cohort. They focused on five variables statistically significant in predicting CVD, including the R-R interval, root mean square of sequential differences recovered from the ECG, systolic and diastolic blood pressures, and total cholesterol levels. Comparing the performance of six ML techniques, including k-nearest neighbor (KNN), linear discriminant analysis (LDA), decision trees, linear and quadratic support vector machines, and artificial neural network (ANN), they found that ANN achieved the highest prediction performance, with 90% specificity, 90% sensitivity, and 90% accuracy. Marbaniang et al. [9] also investigated six similar ML algorithms, except for ANN, where they used Naïve Bayes. Similar to Anuar et al., they found that introducing feature selection facilitated the identification of important risk factors. They observed increased accuracy with the inclusion of 'Blood pressure' and 'Body Mass Index (BMI)' factors, with KNN outperforming other ML techniques and achieving an accuracy of approximately 73%. Mai et al. proposed a non-contact-based method using ballistocardiogram (BCG) signals for CVD prediction, employing UNet coupled with bidirectional long short-term memory (Bi-LSTM) [10]. Their focus was on the robustness of noise in BCG signals to provide reliable predictions. Zarkogianni et al. [11] discussed a comparison of ML-based approaches for CVD risk prediction. They proposed a novel method using ensemble learning to combine multiple models for handling unbalanced datasets. However, their analysis focused on specific classes of patients rather than all patients.

Mishra et al. [12] developed a JAVA application system named the *Heart Disease Risk Predictor*, providing an online platform to forecast disease occurrences based on various symptoms. Users can select from a range of symptoms to identify diseases along with their probability percentages. They used sophisticated systems that implemented data mining

techniques such as Naïve Bayes and Decision Tree. Despite slight differences in performance, the authors claim that the Naïve Bayes algorithm outperformed the Decision Tree. The *Heart Disease Risk Predictor* system maintains all patient data in a single database, which physicians utilize for patient counseling and record maintenance.

The *HeartCare+* mobile application, developed by Elsayed et al. [13], helps assess the risk of coronary heart disease over 10 years using clinical and nonclinical data, categorizing patients' risk as low, moderate, or high. In addition, *HeartCare+* provides alerts for additional treatment suggestions. Its primary objective is to provide assistance to rural residents. One of the scoring methods utilized to estimate a person's risk of CVD is the Framingham Risk Score, developed to calculate the 10-year risk of coronary heart disease using data from the Framingham Heart Study. A gender-specific method based on this score is used to calculate the 10-year cardiovascular risk of an individual [14].

In [15], a CVD prediction technique was developed using multiple ML techniques, including logistic regression, random forest, Naïve Bayes, SVM, KNN, decision tree classifiers, and ANN. KNN exhibited the lowest accuracy, around 68.65%, while most other algorithms, except the decision tree classifier, achieved accuracy rates greater than 85%. According to [15], Naïve Bayes performed the best, with an accuracy rate of 90.16

In [16], the use of a trained recurrent fuzzy neural network (RFNN) based on a genetic algorithm (GA) was investigated to diagnose cardiac diseases. The performance of the proposed method was evaluated using the Cleveland heart disease dataset from the University of California, Irvine (UCI) as a benchmark, comprising 297 patient data samples, 45 for testing and 252 for training. The experiment yielded a remarkable 97.78% accuracy for the test set. Additionally, measures including root mean square error, F score, sensitivity, specificity, precision, and misclassification error were assessed alongside accuracy. Compared to related studies, the findings of the study [16] were considered satisfactory.

C. IoT-based Frameworks in Healthcare and CVD Prediction

IoT finds applications in various fields, including healthcare. The study by Al-Makhadmeh et al. [17] introduces an IoT-based medical device to collect heart details from patients both before and after the occurrence of heart disease. Subsequently, these data are processed using a method known as the higher-order Boltzmann deep belief neural network (HOBDBNN). Bardia et al. proposed a cloud-based ECG monitoring system for IoT devices in [18]. The system comprises hardware, firmware, and AI-based analytics. Additionally, the authors introduced a novel encoding method to enhance performance, with a primary focus on hardware-based solutions. Golec et al. introduced a Function as a Service (FaaS) named HealthFaaS for CVD risk prediction using AI, IoT, and serverless computing in [19]. They compared the performance of serverless and non-serverless platforms for CVD risk prediction, evaluating various ML models to achieve the highest F-score of 92.06. Khan et al. [20] also proposed an IoT-based framework similar to [17], utilizing a Modified Deep Convolutional Neural Network (MDCNN) instead of a deep belief neural network. Their study involved connecting a smartwatch and a heart monitor device to the patient via IoT technology to collect sensor data for the diagnosis and prognosis of heart disease. The acquired data was processed using the Modified Deep Convolutional Neural Network (MDCNN) to classify it into normal and abnormal categories.

D. Federated Learning for CVD Prediction

Research has explored the use of Federated Learning (FL) to safeguard privacy in applications related to CVD prediction. Linardos et al. [21] utilized cardiovascular magnetic resonance (CMR) data from four distinct centers employing FL to diagnose hypertrophic cardiomyopathy (HCM). Their findings illustrate that FL exhibits greater robustness and sensitivity to domain-shift effects, yielding promising results despite limited data. The efficacy of FL models for CMR diagnosis was compared to conventional centralized learning models while ensuring patient privacy. Results indicate that FL offers prospective results comparable to collective data sharing, even with a modest sample size of 180 patients from four centers.

Yaqoob et al. [22] developed a hybrid FL-based technique with MABC-RB-SVM architecture that uses federated matched averaging at the cloud end of health service providers (HSPs) to address data privacy concerns for heart disease prediction in HSPs systems. This method enables HSPs to protect patient privacy while sharing only the necessary information for heart disease prediction. Enhancing the privacy of patient data is the modified artificial bee colony optimization with support vector machine (MABC-SVM) technique, employed at the client end of HSPs for optimal feature selection and classification of heart disease. Compared to conventional FL techniques, the study [22] suggests that the hybrid FL-based method with MABC-RB-SVM architecture increases the prediction accuracy by 1.5%, achieves 1.6% less classification error, and requires 17.7% fewer rounds to reach maximum accuracy. The proposed framework outperforms current FedAvg-SVM, FedMA-SVM, and FedMA algorithms with GA-SVM by achieving 93.8% accuracy after 4500 rounds of communication.

Research Gap: The literature review highlights the importance of developing prognosis applications for CVD risk prediction, along with research on using machine learning (ML) and Internet of Things (IoT) frameworks for CVD predictions. Additionally, there has been work on utilizing Federated Learning (FL) to protect user privacy while developing ML models for CVD diagnosis. However, there is a gap in the literature regarding an integrated framework that combines IoT and FL techniques to improve CVD prognosis.

Furthermore, most FL research assumes homogeneous participation in the distributed learning process, presuming that FL clients are homogeneous because they all possess the same set of data features and are aware of the target labels. However, this assumption is unrealistic as clients could be

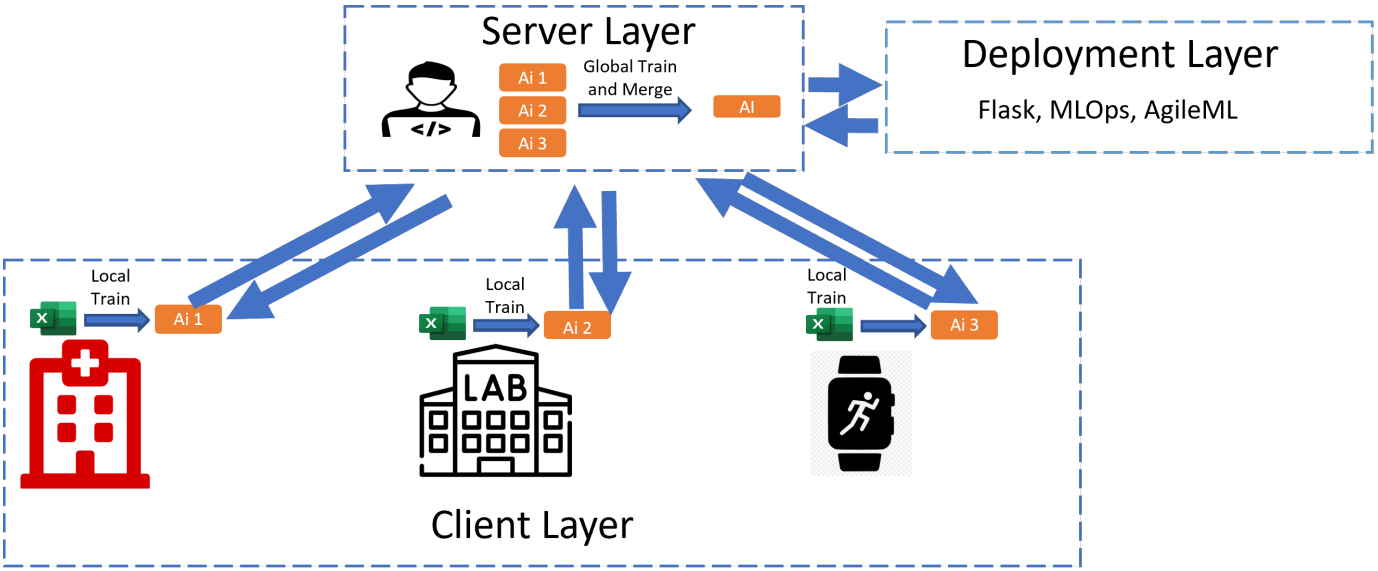

Fig. 1: Proposed System Model

heterogeneous entities with different feature sets. Moreover, in realistic scenarios, clients may not be aware of the labels for training ML models. Therefore, we propose an integrated IoT-based framework that utilizes VFL for CVD prediction, where distributed nodes hold and maintain different types of data features and are unaware of the target class labels.

To the best of our knowledge, this is the first attempt to address the scenario in which different sample features are present with different clients, and they are not aware of the target labels for the distributed training of ML models for CVD prediction applications.

III. System Model

This section presents the proposed IoT-based system model for predicting CVD based on distributed features at different locations. Additionally, it provides the problem statement and the proposed VFL-based algorithms for CVD prediction.

A. Proposed System Model

Figure 2 shows the basic overview of the proposed system that has three layers, heterogeneous clients, server, and application deployment layer.

In the proposed system model, IoT-based heterogeneous clients do not share the raw data. Instead, clients transform their raw data into a low-dimensional vector representation using ML models, which are then shared with the central server. Additionally, the different clients have disparate feature sets. For example, one client could be a hospital where doctors examine symptoms physically or using IoT-based medical instruments, another client could be a lab where IoT-based medical instruments have taken patients' vitals, and the third one could be IoT devices holding patients' demographic information. Thus, the three clients not only have distributed storage of the data but also have distinct features. Furthermore, individual clients do not have labels to train the ML models and therefore cannot be used independently to develop ML applications for CVD prognosis.

The central server may hold some additional features as well as the label of the dataset. Its role is to synchronize the various Artificial Intelligence (AI) models placed at the disparate clients, collect information from the clients in the form of low-dimensional data representation, and merge them using the gradients and true labels it maintains to obtain the converged global AI model. The AI model can be deployed on the server itself or on external cloud-based platforms as an application or Software-as-a-Service (SaaS) platform using deployment tools such as Flask, MLOps, or AgileML [23], [24], [25]. The globally deployed model is used by the stakeholders involved in the VFL process, as well as by any party that has similar input data.

The proposed framework can be utilized to develop mobile or Software-as-a-Service (SaaS) applications, wherein user data is dispersed across various locations. The VFL-based ML model would be deployed on users' devices/wearables, in hospitals on doctors' devices or measuring instruments, and at testing centers on IoT-based devices like ECG instruments. Based on available patient data at disparate locations, an AI-based mobile application can be developed to alert users if they are at risk of CVD. Moreover, such applications can be beneficial for telehealth care providers. Often, telehealth providers lack access to patients' data, such as blood reports, as it is provided by a third party. This issue is particularly common in regions with strict privacy regulations. Therefore, by employing the proposed framework, an application for telehealth care providers can be developed where access to raw data is not necessary for diagnostic purposes.

B. Problem Statement

We consider N clients, each represented by the variables n, where $n \in \{1, 2, ..., N\}$. Spatially distributed clients contain a unique set of features represented as x^m such that $\underset{m}{\cap} x^m = \emptyset$. It should be noted that although the different clients have disparate feature sets, they have the same number of samples P for synchronization purposes. Consequently, for a specific patient, disparate data and features are available to all the clients. Furthermore, during the training process, the clients and servers communicate with each other to get a globally trained ML model.

To ensure the preservation of user privacy, distinct clients do not share their feature space or their data with the server. Instead, they transform the higher-dimensional data x^m into a lower-dimensional vector h^m parameterized by θ^m in the form of smashed information that cannot be deciphered. In addition, we assume that the clients do not have access to the true categories or labels of the training data set. Thus, they only have their own partial set of features x^m. The true category or label y is maintained and stored on the server, which is not shared with clients.

The objective of the proposed method is to solve the equation 1 subject to the fact that the feature set x^m does not leave the client for every client, clients do not have labels y, and the labels do not leave the server.

$$F(\theta) := \frac{1}{|y|} \Sigma L(\theta^0, h^1, h^2, ..., h^n) \tag{1}$$

In equation 1, θ_0 represents a global model, while $\theta = [\theta'_1, \theta'_2, \ldots, \theta'_n]'$ represents a set of variables. The variable L denotes the loss function, and $|y|$ indicates the cardinality of the set y. Given that the optimization function utilizes the first derivative, the optimization algorithm employed here is of the first-order [26]. To address the optimization problem, we utilize the algorithm outlined in the subsequent section, which leverages the first derivatives of the cost with respect to the weights to iteratively minimize the optimization function.

C. Implementation Details

Fig. 2 illustrates the workflow for the proposed system. The proposed process is divided into three distinct steps. In the initial step, an open source dataset is used [27]. The data undergoes processing, normalization, and is then partitioned among different clients based on the feature set. Subsequently, in the next step, a Neural Network model is initiated. The model is trained by sharing the loss and updates between the clients and the server to obtain the global model. Once the model is trained, in the final step, its performance is evaluated using classification accuracy metrics and other relevant metrics such as precision, recall, and F-score. Performance calculation is done using test data that were separated before training the model. Once we achieve satisfactory performance, the trained model can be deployed for users.

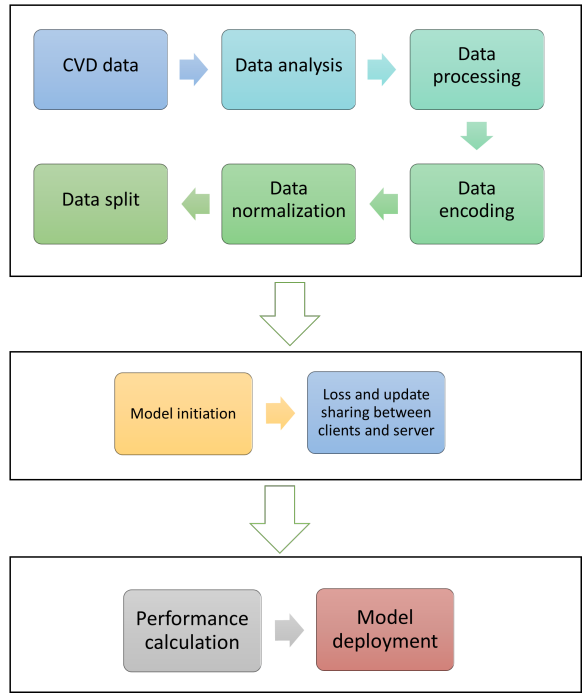

Fig. 2: Overall research framework

IV. Proposed Vertical Federated Learning (VFL) Algorithm

To solve the optimization model mentioned above, we propose VFL for CVD prediction that minimizes loss function using a gradient-based optimizer in a distributed manner, instead of traditional centralized servers, to protect client privacy while applying the fundamental VFL system in the context of classification problems. The proposed approach is aimed at improving the confidentiality of client information because data attributes, as well as feature space, are not gathered from multiple clients situated at spatially separate locations. In contrast to centralized ML, our proposed parameter-based learning provides advanced privacy measures for each client, ensuring that only the parameters of local models are shared with the global model for aggregation and that none of the actual data, features, and labels are shared.

The complete data set can be represented by the variable x, which is $\underset{m}{\cup} x^m$, where as mentioned before each client has only the partial feature set x^m. The data set can be represented in the form of a matrix of size $P \times Q$. x_i is one row, that is, the set of all the features of the dataset, and $x_{i,j}$ is a particular feature of the row. Each client x^m holds and maintains features of size $|x^m|$, where $|x^m|$ is the cardinality of the set x^m.

The algorithms 1 and 2 present the proposed VFL algorithm which is run by the client and server in a synchronized manner for reducing the error based on the y and y', where y and y' respectively are the true prediction and predictions made by

the globally converged model $F(\theta)$ after the training process.

Algorithm 1 Client pseudocode

Require: Shared row ids i
Batch size B
Gradients $\nabla(h^s)$
Ensure: Updated parameters θ^m
Gradients Δh^m
if $e == 0$ **then**
  $\theta^m = rand()$
end if
$Get\ x^m\ from\ i's$
$h^m = f(x^m, \theta^m, \zeta)$
if $e! = 0$ **then**
  Calculate $\nabla(h^m)$ based on $\nabla(h^s)$
  Update θ^m using equation 4
end if
Transfer $h^m\ and\ \nabla(h^m)$ to the server

Algorithm 2 Server pseudocode

Require: h^m
∇h^m
$\nabla L^m(y^{e'}, y^e)$
Ensure: Updated parameters θ^s
(Get a set of IDs i)
if $\zeta! = 0$ **then**
  Convert to a low dimensional representation $h^s = f(x^s, \theta^s, \zeta)$
end if
for all $e\ in\ E$ **do**
  if $\zeta == 0$ **then**
    Calculate the prediction confidence score $p^e = argmax((\frac{\Sigma h^m + h^s}{(N+1)*B})$
  else
    Calculate the prediction confidence scores $p^e = argmax(\zeta * (\frac{\Sigma h^m}{N*B})$
    Perform mapping $p^e \rightarrow y^{e'}$
    Find overall loss $l^e = L(y^{e'}, y^e)$
    Update parameters θ^s based on the equation 4
  end if
  Send updated gradients to all the clients
end for
Deploy the global model

1) Client Process: Algorithm 1 describes in detail the pseudocode run by the clients to train the local model. The client does not share the information of the sample features with each other as well as with the server. They only share the local gradients via different communication rounds. In each such communication round, it first receives a unique set of IDs i from the server for synchronization purposes. For this unique set of IDs, it extracts samples from the feature set x^m of batch size B. In every round of communication, the client converts the high-dimensional vector x^m into a lower-dimensional representation h^m, using a highly nonlinear function $f(x^m, \theta^m, \zeta)$, parameterized by $\theta^m \in \mathbb{R}^{|\theta^m|}$ and a nonlinear function ζ. During the initial iterations, the set of θ^m is initialized to a low random value. The h^m, as it is a lower dimensional representation of the feature ser maintained by a client, cannot be deciphered by the server. This ensures that the server does not have access to the raw sample features and thus client's privacy is ensured. After the initial communication round, the clients also receive the server gradients. In addition, clients calculate their gradients $\nabla(h^m)$ using the one received from the server. Subsequently, it updates the set of θ^m based on $\nabla(h^m)$ using Equation 4. It should be noted that the Equation 4 is also used by the server to update its gradients. The calculated gradients from the client $\nabla(h^m)$ are sent to the server for further aggregation and parameter updates.

2) Server Process: During different communication rounds, the server collects a list of sample IDs i for a batch of data sets B. The list of row IDs is sent to the client for synchronization purposes so that the same sample is used for the gradient update by all the clients and the server. The server communicates with clients, and one of these communication rounds is considered an epoch e and the total number of such epochs is E. The server merges the partial information predictions from the disparate clients according to Equation 2.

$$p^e = argmax(\zeta * (\frac{\Sigma h^m}{N * B}) + (1 - \zeta) * (\frac{\Sigma h^m + h^s}{(N+1) * B}) \quad (2)$$

In equation 2, p^e is the confidence score of the prediction in rounds e for a batch of the dataset of size B, h^s is the low dimensional representation of the server features and the binary variable ζ represents if the server contains the portion of the feature set. Based on the probability of prediction, a mapping $p^e \rightarrow y^e$ is performed to the corresponding target categories where $y^{e'} \in y$ for the communication round e. Furthermore, using $y^{e'}$ and known target values, the loss is calculated using Equation 3, where l^e is the loss in the communication round e.

$$l^e = L(y^{e'}, y^e) \quad (3)$$

For every round of communication e, the server also collects the gradients $\nabla L^m(y^{e'}, y^e)$ from the clients. It should be noted that $\nabla L^m(y^{e'}, y^e)$ is indirectly calculated using server loss and chain rule and the $y^{e'}$ and y^e are not directly utimilized to maintain privacy. The loss is used by the server to update its gradient based on the Adam optimization algorithm, as explained in equation 4 which has been customized for the proposed VFL setup [28]. In the equation, η is the learning rate, g_e is the gradient in round e, μ_e is the exponential average of the gradients, s_e is the exponential average of the square of the gradients and β_1 and β_2 are the hyperparameters used for optimization.

$$
\begin{aligned}
\theta_{e+1} &= \theta_e + \Delta\theta_e \\
\Delta\theta_e &= -\eta * g_e * \frac{\mu_e}{\sqrt{s_e + \kappa}} \\
\mu_e &= \beta_1 * \mu_{e-1} - (1 - \beta_1) * g_e \\
s_e &= \beta_2 * s_{e-1} - (1 - \beta_2) * g_e
\end{aligned} \tag{4}
$$

Both the client and the server update their gradients in different communication rounds, resulting in a converged global model that minimizes prediction loss. Subsequently, the global model can be deployed for real-time use. It should be noted that as new patterns in the data become available in the future, the global model serves as a base model for retraining. Since clients involve IoT devices like ECG machines that might lack sufficient power, they do not need to train the model every time. Instead, they can continue collecting data, and when a sufficient amount is available, they could train the model for a few epochs. Furthermore, to conserve power, they might opt to train it when plugged into a power source.

V. Results and Discussion

This section presents the implementation and simulation details of the proposed VFL-based CVD prediction mechanism. Firstly, we discuss the implementation details and the dataset used in this study. Later, we analyze the performance of the proposed method for various case studies. We also compare the proposed method with the state-of-the-art.

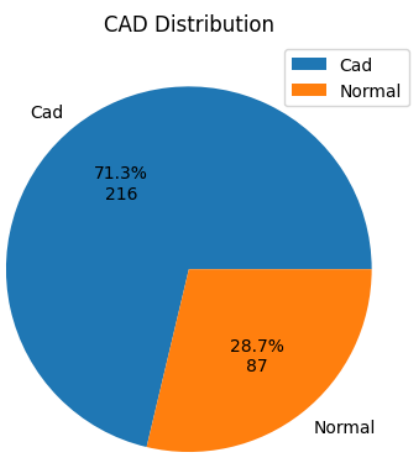

Fig. 3: CVD distribution in the dataset

A. Dataset Discription

We utilize the *Z-Alizadeh Sani Data Set'* [27], available in the public repository *UCI Machine Learning Repository'*. The dataset comprises 55 characteristics and includes information on user demographics, symptoms, and examination, ECG measures, and laboratory analysis. The distribution of the dataset is illustrated in Fig. 3. During data analysis, we discovered that the dataset was devoid of duplicates and null values. The 'Label' column contains two categories, namely 'normal' or 'CVD'. In the dataset, approximately 71.3% are affected by CVD, while around 28.7% are normal (Fig. 3).

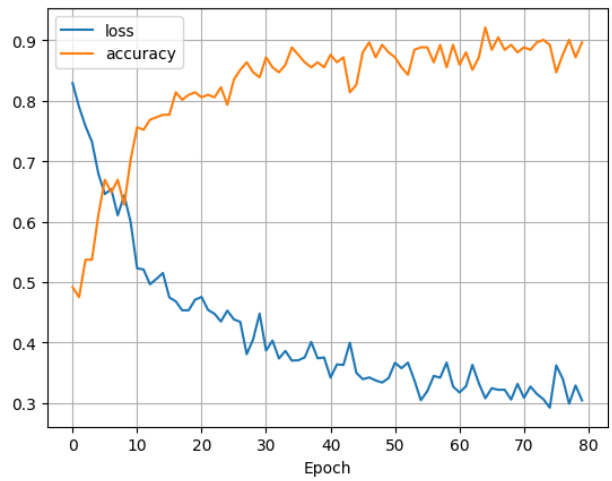

Fig. 4: 2-split VFL (Patient and hospital)

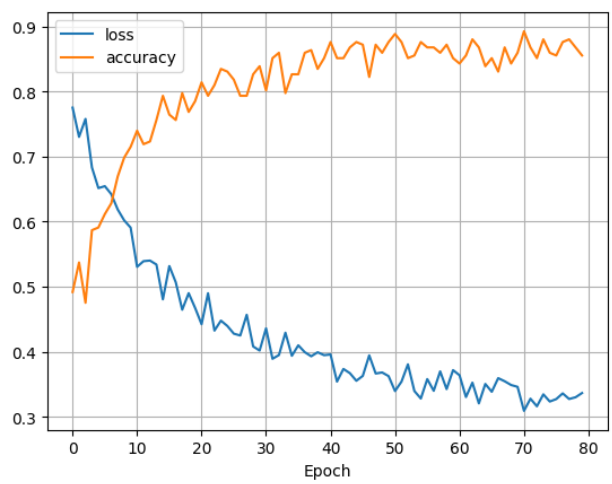

Fig. 5: 3-split VFL (Patient, doctor, and laboratory)

B. Case Studies

We conduct three different case studies based on how features are distributed among the clients in this study. The 55 features of the dataset can be categorized into four different types: 1) demographic information (e.g., age, gender), symptoms and examination such as blood pressure and pulse rates, 3) specific characteristics of the ECG, and 4) laboratory measurements such as hemoglobin level. Based on the type of features, we considered the possible combinations of features that could be available to different clients. These include a 2-split where data features are considered to be divided between the patient and hospital, such that the demographic features are available with the patient and the others are with the hospital. Similarly, in 3-split scenarios, data features are considered to be separated among the patient, hospital, and laboratory. For 4-split, features are considered to be divided among the patient,

the hospital, the ECG center, and the laboratory. These splits present a realistic scenario, as many hospitals may not have a laboratory and generally refer the patient for testing at a pathological center. On the other hand, some hospitals may have a testing facility within their premises. Consequently, we tested against different possible combinations of feature splits based on realistic scenarios.

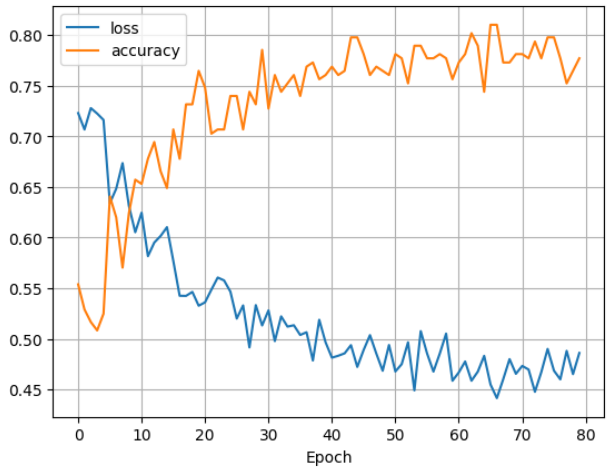

Fig. 6: 4-split VFL (patient, doctor, ECG, and laboratory)

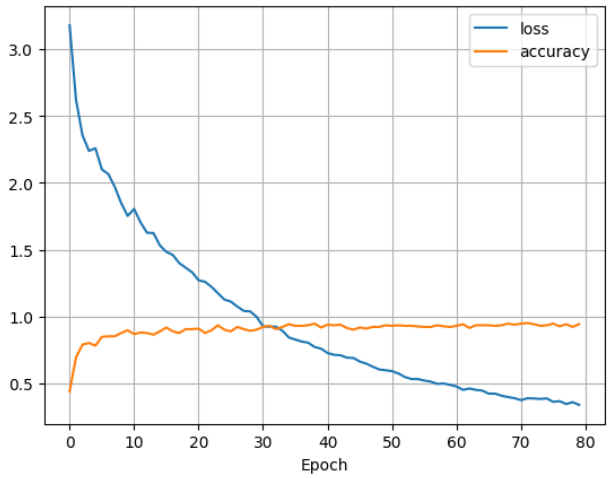

Fig. 7: Conventional DNN

C. Performance Measure

We compare the proposed VFL-based CVD prediction across different case studies. Additionally, we compare the performance of the traditional DNN-based approach for CVD prediction. To ensure consistency across all cases, we train and test using the same hyperparameters.

Figures 4, 5, 6, 7 present the performance of loss and accuracy with the number of communication rounds for the test data for the different case studies and the conventional DNN-based approach. As depicted in the figures, compared to the conventional DNN model, the novel implementation of proposed VFL algorithms in the prediction of CVD provides comparable accuracy. The results indicate that the convergence of the proposed algorithms is achieved in around 60 communication rounds, after which there is minimal improvement in accuracy.

We also compute and compare other relevant classification metrics such as Precision, Recall, F-score, and AUC. Table I presents the classification metrics of the mentioned case studies and the conventional DNN model.

The results indicate that for different case studies, the performance is comparable to that of the traditional DNN method. For instance, when the sample features are situated at two different locations (patient and hospital), the F1 score is notably higher, surpassing that of the conventional DNN methods. Similarly, for 3-splits (patient, lab, and hospitals), the metrics are comparable to those of the traditional methods. Hence, it can be inferred that the proposed VFL-based method demonstrates comparable performance to traditional centralized DNN methods, while also offering the added benefits of data privacy for heterogeneous clients.

D. Comparison With the State-of-the-art

We compare our proposed system with the state-of-the-art using Table II and III. Table II illustrates the performance of our proposed method compared to the centralized ML-based training approach. The table also displays the percentage performance difference between the proposed approach and the highest-performing state-of-the-art method in brackets for all three case studies. As depicted in the table, the advantage of our proposed method lies in user privacy as well as feature splitting. Our 2-split VFL method achieves an F-score of 91.92.

It is important to note that the state-of-the-art methods are capable of achieving better performance compared to our proposed method. However, the state-of-the-art techniques utilize feature engineering approaches. For instance, [30] utilized wrapper methods and Recursive Feature Elimination (RFE) to enhance accuracy. Feature engineering helps improve the accuracy of the model by leveraging the relationship between features and labels in the entire dataset and employing statistical methods to select the optimal number of features.

In the proposed scenario of VFL, each client possesses different features, while the server holds the labels. Similarly, clients hold only a portion of the dataset. Consequently, a client does not have knowledge of other clients' features, data distribution, or the labels stored on the server. Therefore, feature engineering cannot be applied in our scenario as it would violate privacy, which is an essential component of VFL. Thus, our method is able to achieve good accuracy despite not employing feature engineering techniques, which we also cannot apply to guarantee clients' privacy.

Furthermore, it should be noted that in [29] and [30], although they achieve higher accuracy, they employ ensemble learning. In ensemble learning, multiple classifiers are trained, and predictions are made by each classifier. Subsequently, these predictions are combined using methods such as voting.

TABLE I: Classification accuracy metrics of the models

Model	Accuracy	Precision	Recall	F1-Score	AUC
2-split VFL	88.53%	93.02%	90.91%	91.95%	91.85%
3-split VFL	85.25%	88.89%	90.91%	89.89%	92.51%
4-split VFL	80.33%	86.36%	86.36%	86.36%	83.56%
DNN	85.25%	87.23%	93.18%	90.11%	92.65%

TABLE II: Comparison of the proposed VFL-based system with state-of-the-art centralized systmes

Benchmark	Method	Privacy	Feature split	Accuracy	Precision	Recall	F-score
Qin et al. [29]	Ensemble-based	✗	✗	93.70	95.65	97.63	95.53
Wang et al. [30]	Stacking-based model	✗	✗	95.43	97.71	95.84	96.77
Kolukisa et al. [31]	Ensemble-based model	✗	✗	83.48	83.83	82.77	83.3
Shahid et al. [32]	Emotional Neural Network	✗	✗	88.34	92.37	91.87	92.12
Proposed VFL	2-split	✓	✓	88.53 (6.9)	93.02 (4.69)	90.91 (6.72)	91.95 (4.82)
Proposed VFL	3-split	✓	✓	85.25 (10.18)	88.89 (8.82)	90.91 (6.72)	89.89 (6.88)
Proposed VFL	4-split	✓	✓	80.33 (15.1)	86.36 (11.35)	86.36 (11.27)	86.36 (10.41)

TABLE III: Comparison of the proposed VFL-based system with state-of-the-art FL systems

Benchmark	Method	Federated Privacy	Feature split	Label storage
Mahalingam et al. [33]	Edge FL	✓	✗	Client
Linardos et al. [21]	Pre-trained CNN based FL	✓	✗	Client
Wang et al. [34]	Noise and incentive-based FL	✓	✗	Client
Zhou et al. [35]	Hierachical FL	✓	✗	Client
Proposed VFL	2-split	✓	✓	Server
Proposed VFL	3-split	✓	✓	Server
Proposed VFL	4-split	✓	✓	Server

As ensemble learning involves training multiple classifiers, it is computationally complex and not feasible for IoT devices. In contrast, our method employs a single classifier and is computationally less complex.

We also compare our proposed method with other state-of-the-art FL methods in Table III. In contrast to the other FL methods, our proposed method assumes that the sample features of the data are split among different clients, meaning they have different kinds of features. Our scenario is more realistic as it not only guarantees that data privacy is maintained but also ensures that different clients may have different information. Furthermore, the benchmark methods assume that the labels are stored on each client. This is again an unrealistic scenario for training an ML model, as clients (for example, patients) are not aware of the data labels. Labels are generally available at the server. Thus, our method considers a more realistic scenario where clients are not aware of the training labels.

VI. Conclusions and Future work

This paper introduces an IoT-based framework utilizing Vertical Federated Learning (VFL) for automated cardiovascular disease prediction using Machine Learning (ML). Unlike traditional federated learning (FL), our proposed method considers data set features to be separated across different clients. Through comparison with various case studies, the results demonstrate that our proposed methods offer comparable performance to traditional DNN-based methods while also providing the advantages of privacy, feature separation, and clients being agnostic of training labels. Furthermore, our framework preserves user privacy and feature separation among different client devices.

In the future, we plan to collaborate with the local hospital to evaluate the proposed method using a recent dataset that includes a large number of features. Additionally, we aim to integrate transfer learning into the proposed VFL framework to optimize the energy requirements for training the model on IoT devices. Moreover, one of the challenges in VFL is that clients are unaware of each other's data distribution and feature sets. Therefore, data balancing and feature engineering are challenging in the case of VFL. In the future, suitable techniques for performing data balancing and feature engineering need to be investigated as open research problems.

References

[1] World Health Organization, "Cardiovascular diseases (CVDs)," 2021.

[2] T.-M. Chen, Y.-H. Tsai, H.-H. Tseng, K.-C. Liu, J.-Y. Chen, C.-H. Huang, G.-Y. Li, C.-Y. Shen, and Y. Tsao, "Srecg: Ecg signal super-resolution framework for portable/wearable devices in cardiac arrhythmias classification," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 3, pp. 250–260, 2023.

[3] D. Zhang, X. Liu, J. Xia, Z. Gao, H. Zhang, and V. H. C. de Albuquerque, "A physics-guided deep learning approach for functional assessment of cardiovascular disease in iot-based smart health," *IEEE Internet of Things Journal*, vol. 10, no. 21, pp. 18505–18516, 2023.

[4] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3347–3366, 2023.

[5] J. Li, Y. Zhang, Q. Wang, and K. Liu, "Application of vertical federated learning in predicting cad using iot devices," *arXiv preprint arXiv:2303.09531v1*, 2023.

[6] A. Molloy, K. Beaumont, A. Alyami, M. Kirimi, D. Hoare, N. Mirzai, H. Heidari, S. Mitra, S. L. Neale, and J. R. Mercer, "Challenges to the development of the next generation of self-reporting cardiovascular implantable medical devices," *IEEE Reviews in Biomedical Engineering*, vol. 15, pp. 260–272, 2022.

[7] World Health Organization, "Cardiovascular diseases (cvds)," *World Health Organization*, Jun 2021.

[8] N. Anuar, H. A. Hamid, M. Z. Suboh, A. Noraidatulakma, R. Jaafar, M. Y. N. Ain, H. M. Akma, Z. N. Farawahida, K. A. A. Shawani, M. A. D. Syakila, K. M. Arman, and A. J. Rahman, "Cardiovascular disease prediction from electrocardiogram by using machine learning," *Int. J. Online Biomed. Eng.*, vol. 16, pp. 34–48, 2020.

[9] I. A. Marbaniang, N. A. Choudhury, and S. Moulik, "Cardiovascular disease (cvd) prediction using machine learning algorithms," *2020 IEEE 17th India Council International Conference (INDICON)*, pp. 1–6, 2020.

[10] Y. Mai, Z. Chen, B. Yu, Y. Li, Z. Pang, and Z. Han, "Non-contact heartbeat detection based on ballistocardiogram using unet and bidirectional long short-term memory," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3720–3730, 2022.

[11] K. Zarkogianni, M. Athanasiou, A. C. Thanopoulou, and K. S. Nikita, "Comparison of machine learning approaches toward assessing the risk of developing cardiovascular disease as a long-term diabetes complication," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1637–1647, 2018.

[12] R. Mishra, P. Saharan, and A. Jyoti, "Heart disease risk predictor," Aug 2019.

[13] H. A. Elsayed, M. A. Galal, and L. Syed, "Heartcare+: A smart heart care mobile application for framingham-based early risk prediction of hard coronary heart diseases in middle east," *Mobile Information Systems*, vol. 2017, pp. 1—11, Sep 2017.

[14] P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation*, no. 18, pp. 1837—-1847, 1998.

[15] O. Voloshynskyi, V. Vysotska, and M. Bublyk, "Cardiovascular disease prediction based on machine learning technology," pp. 69–75, 2021.

[16] K. Uyar and A. İlhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," *Procedia Computer Science*, vol. 120, pp. 588—593, 2017.

[17] Z. Al-Makhadmeh and A. Tolba, "Utilizing iot wearable medical device for heart disease prediction using higher order boltzmann model: A classification approach," *Measurement*, vol. 147, p. 106815, Dec 2019.

[18] B. Baraeinejad, M. F. Shayan, A. R. Vazifeh, D. Rashidi, M. S. Hamedani, H. Tavolinejad, P. Gorji, P. Razmara, K. Vaziri, D. Vashaee, and M. Fakharzadeh, "Design and implementation of an ultralow-power ecg patch and smart cloud-based platform," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[19] M. Golec, S. S. Gill, A. K. Parlikad, and S. Uhlig, "Healthfaas: Ai-based smart healthcare system for heart patients using serverless computing," *IEEE Internet of Things Journal*, vol. 10, no. 21, pp. 18469–18476, 2023.

[20] M. A. Khan, "An iot framework for heart disease prediction based on mdcnn classifier," *IEEE Access*, vol. 8, pp. 34717–34727, 2020.

[21] A. Linardos, K. Kushibar, S. Walsh, P. Gkontra, and K. Lekadir, "Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease," *Scientific Reports*, vol. 12, no. 1, p. 3551, 2022.

[22] M. M. Yaqoob, M. Nazir, M. A. Khan, S. Qureshi, and A. Al-Rasheed, "Hybrid classifier-based federated learning in health service providers for cardiovascular disease prediction," *Applied Sciences*, vol. 13, no. 3, p. 1911, 2023.

[23] G. Dwyer, *Flask By Example*. Packt Publishing Ltd, 2016.

[24] GoogleCloud, "MLOps: Continuous delivery and automation pipelines in machine learning," 2021.

[25] R. M. Shukla and J. Cartlidge, "AgileML: A machine learning project development pipeline incorporating active consumer engagement," in *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2021.

[26] G. Lan, *First-order and stochastic optimization methods for machine learning*, vol. 1. Springer, 2020.

[27] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, and Z. A. Sani, "A data mining approach for diagnosis of coronary artery disease," *Computer methods and programs in biomedicine*, vol. 111, no. 1, pp. 52–61, 2013.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] C.-J. Qin, Q. Guan, and X.-P. Wang, "Application of ensemble algorithm integrating multiple criteria feature selection in coronary heart disease detection," *Biomedical Engineering: Applications, Basis and Communications*, vol. 29, no. 06, p. 1750043, 2017.

[30] J. Wang, C. Liu, L. Li, W. Li, L. Yao, H. Li, and H. Zhang, "A stacking-based model for non-invasive detection of coronary heart disease," *IEEE Access*, vol. 8, pp. 37124–37133, 2020.

[31] B. Kolukisa and B. Bakir-Gungor, "Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis," *Computer Standards & Interfaces*, vol. 84, p. 103706, 2023.

[32] A. H. Shahid and M. Singh, "A novel approach for coronary artery disease diagnosis using hybrid particle swarm optimization based emotional neural network," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 4, pp. 1568–1585, 2020.

[33] P. Mahalingam and J. Dheeba, "A heart disease prognosis pipeline for the edge using federated learning," *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, p. 100490, 2024.

[34] X. Wang, J. Hu, H. Lin, W. Liu, H. Moon, and M. J. Piran, "Federated learning-empowered disease diagnosis mechanism in the internet of medical things: From the privacy-preservation perspective," *IEEE Transactions on Industrial Informatics*, 2022.

[35] X. Zhou, X. Ye, I. Kevin, K. Wang, W. Liang, N. K. C. Nair, S. Shimizu, Z. Yan, and Q. Jin, "Hierarchical federated learning with social context clustering-based participant selection for internet of medical things applications," *IEEE Transactions on Computational Social Systems*, 2023.