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) encompasses a wide range of diseases that affect the heart and blood vessels, including coronary artery disease, heart failure, arrhythmia, and stroke. Machine Learning (ML) has been widely used to predict CVD risk based on various factors and is a critical area of healthcare research. However, due to privacy concerns, sharing the data needed to predict CVD with ML is challenging. Even though Federated Learning (FL) enables distributed training of ML models without sharing raw data, it assumes that all training features are available to all clients. To address the problem, we propose a Vertical Federated Learning (VFL) based method that trains ML models in a distributed manner and has different features available to several spatial locations. 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 ML model. We employ the proposed method for different use cases where the data characteristics are split 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). We test the proposed methodology on the realistic publicly available dataset.**

***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, causing almost one-third of deaths [1]. This startling revelation stresses the pressing need for heightened awareness and deterrent measures against this perilous ailment. CVD has different types, such as coronary artery disease (CAD), heart failure, arrhythmia, and stroke, among others, that affect heart health along with blood vessels. Medical professionals strive to predict risk factors such as high cholesterol levels and smoking habits, but it remains difficult to detect early signs of CVD. In addition to prompt identification protocols, proper management strategies are necessary, including reducing hypertension and regulating body weight, among other determinants [2].

There has been research on the use of Machine Learning (ML) to predict CVD [3], [4]. To achieve the goal of automatic CVD prediction, ML algorithms are utilized to meticulously analyze vast data sets and identify patterns that may not be obvious to humans. However, one of the concerns of using ML algorithms is that it requires a large amount of data for training that is often inaccessible due to user privacy concerns, especially for healthcare applications.

Federated learning (FL), occasionally known as distributed collaborative learning, is a popular ML methodology. Using an algorithm strategically, it employs several distinct and autonomous clients, each employing its unique collection of related data sets. FL offers a unique departure from traditional centralized ML methods, where localized data compilations are often condensed into one training session. Additionally, it stands opposed to algorithms that solely rely on datasets exhibiting identical distributions across the board. 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 [5].

However, one of the problems with traditional FL methods is that they assume that every client is identical so that they contain the features and information of the target class needed to train an ML model. Often, the assumption does not hold, as different clients may have dissimilar features of the same dataset required to train ML algorithms. In this scenario, a Vertical Federated Learning (VFL) approach involves training a model using sample features stored at disparate locations [6]. 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 ML model.

This paper proposes the use of VFL techniques and presents novel algorithms for the distributed training of ML models between multiple heterogeneous clients and servers for the automatic prediction of CVD. In contrast to the traditional FL techniques, the disparate clients hold and maintain a different type of sample data (features) and are not aware of the target class or labels. Clients may be Internet of Things (IoT) devices or edge servers and are heterogeneous so they hold a variety of information (features). For example, one of the clients could be a hospital that maintains patient records based on the doctor's diagnosis, while another client could be a laboratory where patient tests (for example, blood tests) are performed. These two clients maintain different types of information about the same patient. Consequently, the novelty of the proposed method is that it also considers feature separation, in contrast to the traditional FL where each client has all the features and information about the target class. The main contributions of this research are as follows.

- We propose an innovative framework based on IoT to predict CVD risk using ML. It maintains user privacy and considers the heterogeneity of disparate clients in terms of the features they hold and maintain.
- This paper provides a novel algorithm for training an ML model using VFL where clients and servers coordinate with each other to maintain confidentiality.
- We implement and assess the proposed VFL framework for various possible realistic case studies on sample feature distribution.
- Finally, we test the proposed framework in a real-world dataset and assess its performance compared to the state-of-the-art.

To the best of our knowledge, this article is the first attempt to consider both user privacy and sample feature separation for CVD risk prediction using ML. The rest of the paper is organized as follows. In Section II, we discuss recently published work on the use of IoT, ML, and FL for CVD prediction. Section III elucidates our proposed system model and presents the problem statement. Section IV analyzes the implementation details, describes the datasets, considers different case studies, and presents the comparison with the state-of-the-art. Finally, in Section V, we conclude our work and highlight future research directions.

II. Literature Review

CVD continues to pose a considerable risk to public health as it is a major factor in mortality worldwide. Scholars are increasingly using technologies such as Machine Learning (ML) and the Internet of Things (IoT) to devise pathways toward alleviating and prognosticating CVD as technology boundaries expand. This section provides a detailed literature survey on CVD diseases and their prediction using ML, IoT-based frameworks, and FL to preserve privacy.

A. Cardiovascular Disease

Around 7.6 million people in the UK suffer from heart and circulatory disorders, one person dies every three minutes as a result of heart and circulatory disease, one person is hospitalized every five minutes due to a stroke, and 13 babies are diagnosed with congenital heart defects each day [7]. CVD is characterized as a complex and multifaceted array of destructive disorders that affect the exquisitely sophisticated cardiovascular system, encompassing an interwoven network of cardiac muscles, arteries, and veins [8].

Multiple conditions, including coronary artery diseases, cerebrovascular diseases, congenital heart diseases, rheumatic heart diseases, peripheral arterial diseases, deep vein thrombosis (DVT), and pulmonary embolism, define its numerous and complicated characteristics, developing intricate obstacles for modern medical sciences to overcome [8]. Among these ailments, CVD affects masses with alarming frequency due to its high incidence rate. Cerebrovascular diseases are another group of diseases caused by problems in blood vessels that carry blood toward brain tissues. These pathologies are characterized by an interruption in cerebral blood flow stemming from obstructive paths leading to grave consequences on cognitive functions and may include conditions like carotid artery disease or cerebral hemorrhages which can result in fatal consequences like stroke compromising mental agility [9]. Congenital heart disorders position themselves unequivocally at birth because they originate from abnormalities present within different fetal aspects involved in the formation of valves or vascular structures characterizing human cardiac life. Rheumatic heart disease establishes itself after infection produced by Streptococcus bacteria followed by distressing damage to the heart muscle and heart valves due to rheumatic fever [8]. Peripheral arterial diseases show up when abnormal lumps hinder healthy bloodstream flow inside veins that are tasked with transporting oxygen-rich nutrients critical for meeting body requirements primarily geared toward lower extremities regulating muscle movement [10]. Finally, DVT entails a disorder characterized by a collection of coagulated fluid inside veins lodged primarily within the lower limb areas and pulmonary embolism materializes when a discrete segment of the thromboembolic disengages and navigates its way toward the lungs, serving as a potentially hazardous threat to one's life [11].

B. CVD Prediction Using Machine Learning

Studies have been conducted recently on ML techniques for predicting cardiovascular diseases (CVD). Anuar et al. [12] conducted a study to predict CVD from an Electrocardiogram (ECG) by using ML. Their research includes a prospective population-based case-control study with sixty people taking part from the Malaysian cohort. They concentrated on five variables because they were statistically significant in predicting CVD, including the R-R interval, the root mean square of sequential differences recovered from the ECG, systolic and diastolic blood pressures, and total cholesterol levels. The best accurate classifier to predict CVD risk was determined by 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). The authors observed that among the six algorithms, ANN had the highest prediction performance, achieving 90% specificity, 90% sensitivity, and 90% accuracy. Marbaniang et al. [13] also conducted a study using six similar ML algorithms except for ANN. Instead, they used Naïve Bayes and found, as did Anuar et al. [12], that introducing feature selection made it easier to identify important risk factors. They discovered an increase in accuracy when ‘Blood pressure’ and ‘Body Mass Index (BMI)’ factors were introduced to the data set. Here, KNN outperformed all other ML techniques and provided an accuracy of around 73%.

Mishra et al. [14] developed a JAVA application system called the *Heart Disease Risk Predictor* that offers an online platform to forecast disease occurrences based on a variety of symptoms. The user can choose from a range of symptoms to locate diseases along with their probability percentages. They used sophisticated systems implementing data mining techniques such as Naïve Bayes and Decision Tree. Despite the slight difference in performance, the authors claim that the Naïve Bayes algorithm outperformed the Decision Tree. The Heart Disease Risk Predictor system under consideration maintains all patient data in a single database. To counsel patients and maintain records, physicians utilize the same database.

The *HeartCare+* mobile application, created by Elsayed et al. [15], helped adequately assess the risk of coronary heart disease over a 10-year period using clinical and non-clinical data and categorizes the risk for patients as low, moderate, or high. *HeartCare+* also alerts patients for additional treatment suggestions. Its main objective is to offer help to rural residents. One of the scoring methods used to estimate a person’s risk of acquiring CVD is the Framingham Risk Score. This score was created to calculate the 10-year risk of coronary heart disease using information from the Framingham Heart Study. A gender-specific method based on this score is used to calculate the 10-year cardiovascular risk of a person [16].

In [17], a CVD prediction technique was created using several ML techniques, including logistic regression, random forest, nave Bayes, SVM, KNN, decision tree classifiers, and ANN. KNN offered the lowest accuracy, at about 68.65%, whereas the majority of the other algorithms, with the exception of the decision tree classifier, were able to attain accuracy rates of more than 85%. According to [17], Naïve Bayes performed the task the best, with an accuracy rate of 90.16%. Similar to the system created by [14], Naïve Bayes performed better in this particular system than that created by [17]

The use of trained recurrent fuzzy neural network (RFNN) based on a genetic algorithm (GA) for the diagnosis of cardiac diseases has been explored in [18]. The performance of the suggested method is evaluated using the Cleveland heart disease dataset from the University of California, Irvine (UCI) as a benchmark. The dataset consists of 297 samples of patient data, of which 45 were chosen for testing and 252 were used for training. The results of the experiment indicate that 97.78% accuracy was achieved for the test set. Notably, additional measures including root mean square error, F-score, sensitivity, specificity, precision, and misclassification error are assessed in addition to accuracy. Compared to the results of related studies, the study findings [18] were considered adequate.

C. IoT-based Frameworks in Healthcare and CVD Prediction

IoT is used in many fields, including healthcare. The study by Al-Makhadmeh et al. [19] introduces an IoT-based medical device to collect patient heart details before and after heart disease. These data are later processed using a method called the higher-order Boltzmann deep belief neural network (HOBDBNN). The effectiveness of this heart disease recognition system, using HOBDBNN as its basis, is thoroughly examined, evaluating a number of fundamental performance indicators including f-measure, sensitivity, specificity, loss function, and receiver operating characteristic (ROC) curve. These metrics provide insight into the precision with which the system recognizes instances of cardiac disease and promotes an evaluation of both the speed complexity and the false positive rate. On the basis of the conclusive findings, the proposed system manifested a recognition accuracy that reached 99. 03% with a minimal time complexity of 8.5s at maximum. This compelling evidence shows its potential to effectively reduce heart disease mortality and facilitate diagnostic procedures by reducing their inherent complexity. Khan et al. [20] also proposed an IoT-based framework similar to [19]. Instead of a deep belief neural network, the authors in [20] used a Modified Deep Convolutional Neural Network (MDCNN). In this study, IoT technology is employed by connecting a smartwatch and a heart monitor device to the patient in order to collect sensor data for the diagnosis and prognosis of heart disease. In order to classify the acquired data into normal and abnormal, the acquired data is subsequently processed using the Modified Deep Convolutional Neural Network (MDCNN). With the help of the given IoT framework, early detection and treatment of heart disease are made possible by accurate prediction of the condition. MDCNN was able to outperform existing classifiers with an accuracy of 98.2%. In his performance research, Khan et al. [20] also contrasted the proposed MDCNN with current deep learning neural networks and logistic regression and discovered that MDCNN surpassed other approaches.

D. Federated Learning for CVD Prediction

Research has been conducted on employing FL to safeguard privacy in applications related to CVD prediction. The work by Linardos et al. [21] uses cardiovascular magnetic resonance (CMR) data from four distinct centers using FL to diagnose hypertrophic cardiomyopathy (HCM). The findings demonstrate that FL demonstrates greater robustness and sensitivity to domain-shift effects and produces promising outcomes despite the limited quantity of data. The efficacy of FL models for CMR diagnosis is compared in the research to conventional centralized learning models while also protecting patient privacy. Their results show that FL offers prospective results comparable to collective data sharing, even with a modest sample size of 180 patients drawn from four centers.

Using federated matched averaging at the cloud end of health service providers (HSPs), Yaqoob et al. [22] developed a hybrid FL-based technique with MABC-RB-SVM architecture that solves the issue of data privacy for heart disease prediction in HSPs systems. Using this method, HSPs can protect patient privacy while sharing only the information required for the prediction of heart disease. Further improving the privacy of patient data is the modified artificial bee colony optimization

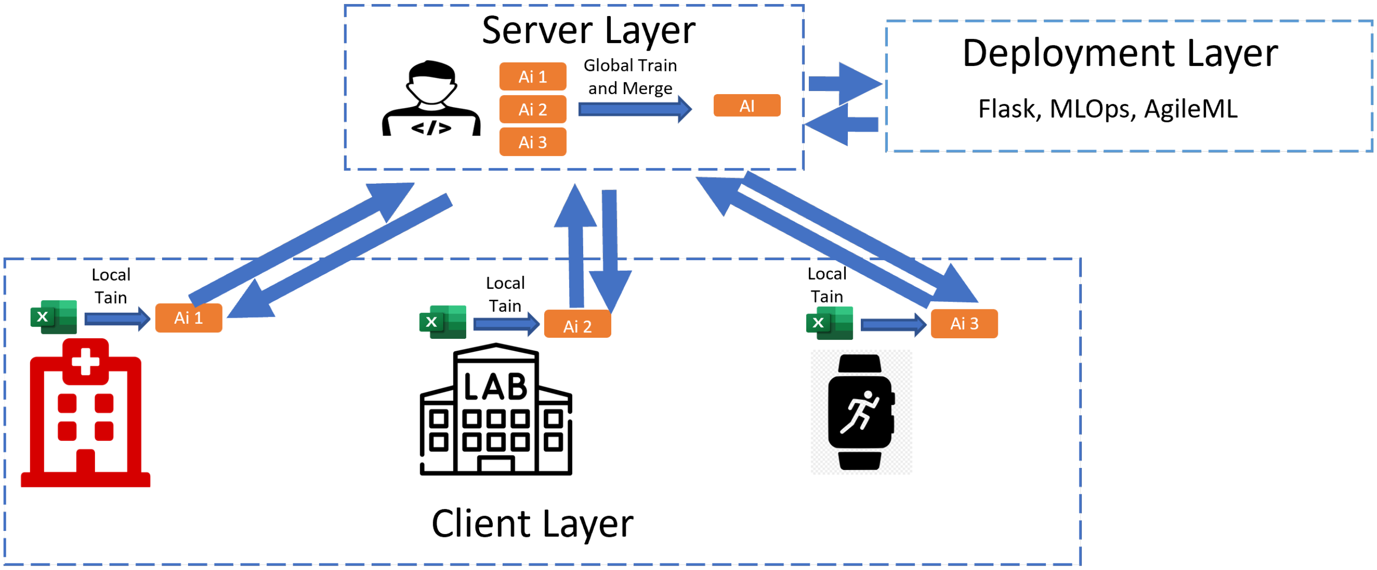

Fig. 1: Proposed System Model

with support vector machine (MABC-SVM) technique, which is employed at the client end of HSPs for the best selection of features and classification of heart disease. Compared to conventional FL techniques, the study [22] suggested that the hybrid FL-based method with the MABC-RB-SVM architecture increases prediction accuracy by 1.5%, achieves 1.6% less classification error, and uses 17.7% fewer rounds to reach maximum accuracy. The proposed framework outperforms the 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 study reveals the importance of developing prognosis applications for CVD prediction, and the work done on the use of ML and IoT frameworks for CVD predictions. Additionally, there has been work done on the use of FL for maintaining user privacy while developing ML models for CVD presentation. However, there is a gap for an integrated framework that uses IoT as well as FL techniques to improve the prognosis of CVD. Furthermore, most of the work done on FL assumes homogeneous parties participating in the distributed learning process. Thus, it is assumed that the FL clients are homogeneous since they all have the same feature set of data. It is also assumed that the clients participating in the FL process are aware of the target labels. The assumption is not a realistic scenario as clients could be heterogeneous entities having disparate feature sets. Therefore, we propose an integrated IoT-based framework that uses FL for CVD prediction where distributed nodes hold and maintain different types of data features. To the best of our knowledge, this is the first attempt to consider the scenario in which distinct sample features are present with different nodes for the distributed training of the ML models for CVD prediction application.

III. Methodology

This section presents the proposed IoT-based system model to predict the CVD based on the distributed features at different locations. It further 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 having three layers consisting of heterogeneous clients, server, and application deployment layer.

In the proposed system model, the heterogeneous clients do not share the raw data with each other. Instead, clients transform their raw data into a low-dimensional vector representation using ML models, which, in turn, are shared with the central server. In addition, the different clients have disparate feature sets. For example, one client could be a hospital holding patients' medical records, another client could be a lab where patients' lab reports are generated, and the third one could be IoT devices having patients' demographic information. Thus, the three clients do not have only distributed storage of the data but they also have distinct features. Additionally, 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 of the additional features as well as the label of the dataset. The role of the central server is to synchronize the various Artificial Intelligence (AI) models placed at the disparate clients, collect information from the clients which are in the form of the low-dimensional data representation together, and merge them using the gradients and true labels that it maintains to get the converged global AI model. The global AI model can be deployed on the central server itself or on external cloud-based platforms as an application or Softaware-as-a-Service (SaaS) platform using deployment tools such as Flask, MLOPs or AgileML [23], [24], [25]. It is not imperative that the globally deployed model be used only by the stakeholders involved in the VFL process. However, it could be used by any party having similar input data data.

B. Problem Statement

We consider N clients and each client is represented by the variables n, where $n \in \{1, 2, ..., N\}$. Spatially distributed client contains a unique set of features represented as x^m such that $\bigcap_{\text{m}} x^m = \emptyset$. 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 user privacy preservation, the distinct clients do not share their feature space with each other as well as 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 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 the 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 is a global model and θ is the $[\theta_1', \theta_2' ....., \theta_n']'$, variale L represents the loss function, and $|y|$ is the cardinality of the set y.

C. Proposed Vertical Federated Learning (VFL) Algorithm

To solve the optimization model mentioned above, we proposed VFL for the prediction of CVD that minimizes the loss function using the gradient-based optimizer in a distributed manner as opposed to the traditional centralized server, 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, 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 $\bigcup_{\text{m}} x^m$, 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 features of the dataset, and $x_{i,j}$ is a particular feature of the row. Each client holds and maintains x^m length features $|x^m|$, 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 $\theta^m$
    Gradients $\Delta h^m$
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 $\theta^m$ using equation 4
end if
Transfer $h^m\ and\ \nabla(h^m)$ to the server
```

Algorithm 2 Server pseudocode

```
Require: $h^m$
    $\nabla h^m$
    $\nabla L^m(y^{e'}, y^e)$
Ensure: Updated parameters $\theta^s$
(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 $\theta^s$ 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 the 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. Furthermore, clients calculate their own 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 client and 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 information partial 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 communicatio 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 [26]. 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.

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

Both the client and server thus update their gradients in different communication rounds resulting in a converged global model that minimizes the 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 is used as a base model to retrain the model again.

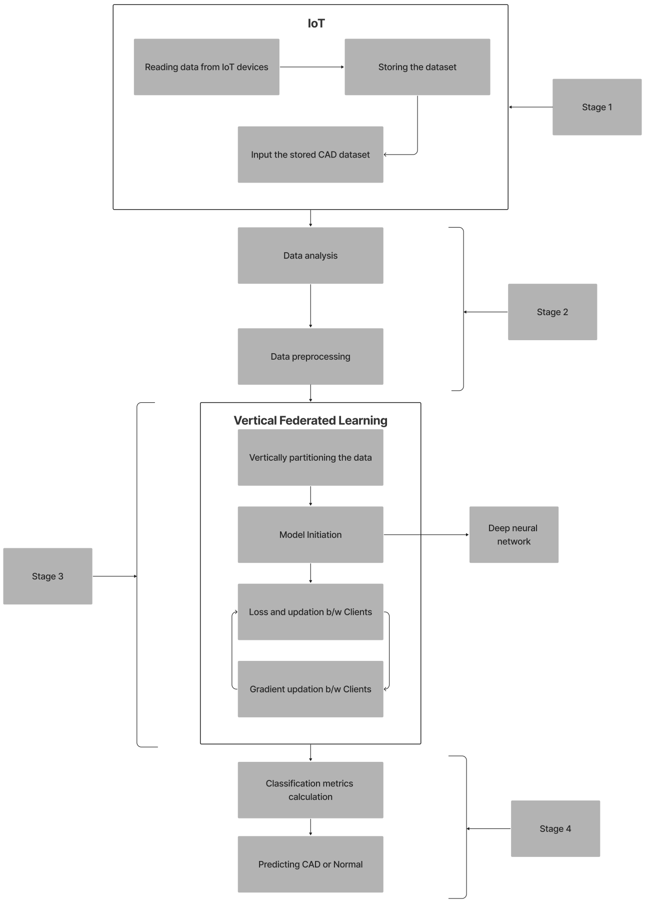
IoT
Reading data from IoT devices
Storing the dataset
Stage 1
Input the stored CAD dataset
Data analysis
Stage 2
Data preprocessing
Vertical Federated Learning
Vertically partitioning the data
Model Initiation
Deep neural network
Stage 3
Loss and updation b/w Clients
Gradient updation b/w Clients
Classification metrics calculation
Stage 4
Predicting CAD or Normal

Fig. 2: Overall research framework

IV. Results and Discussion

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

A. Implementation Details

Fig. 2 presents the workflow for the proposed system is shown below in Fig.2. The proposed process is divided into four different steps. The initial step involves collecting data from IoT devices. Subsequently, we analyze and process the data in the second step. Then follows, as is where VFL is put into effect. It starts with the vertical partitioning of the data feature with numerous clients based on different case studies. Subsequently, the model is also created and initiated along with parameters, the loss and gradients of the clients' models are calculated, and afterward, the model weights are updated using them. In the final step, the model's performance is evaluated using classification accuracy metrics alongside prediction is done using the test data that we have separated before training the model.

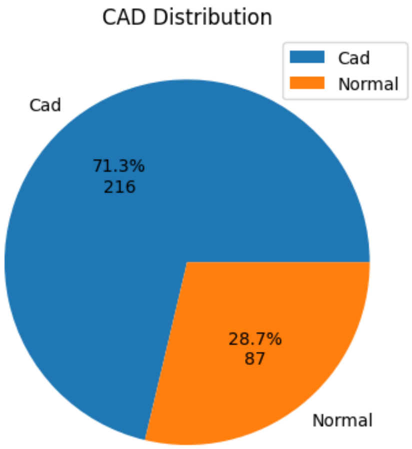

Fig. 3: CVD distribution in the dataset

B. Dataset Discription

We employ *'Z-Alizadeh Sani Data Set'* [27] available in the public repository *'UCI Machine Learning Repository'*. The dataset consists of 55 characteristics and information on the user demographics, symptoms and examination, ECG measures, and laboratory analysis. The distribution of the data set is shown in Fig. 3. During data analysis, we found that the data set was free of duplicates and null values. There are two categories for the 'Label' column called *Cath*: 'Normal' or 'CVD'. In the data set, approximately 71.3% are affected by CVD, while around 28.7% are normal (Fig. 3).

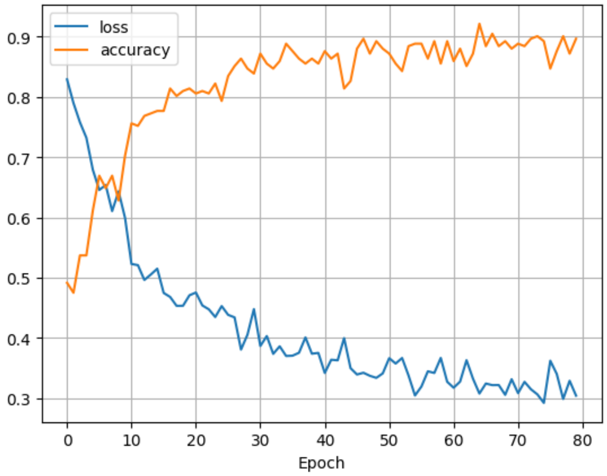

Fig. 4: 2-split VFL (Patient and hospital)

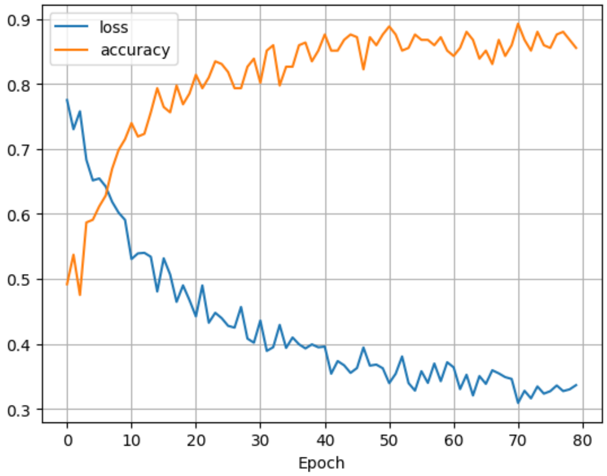

Fig. 5: 3-split VFL (Patient, doctor, and laboratory)

C. Case Studies

We conduct three different case studies based on how features are split between the clients in this study. The 55 features of the data set can be categorized into four different types: 1) demographic information (for example, 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 split between 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 between the patient, hospital, and laboratory. For 4-split data, features are considered to be divided between the patient, the hospital, the ECG center, and the laboratory. The 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.

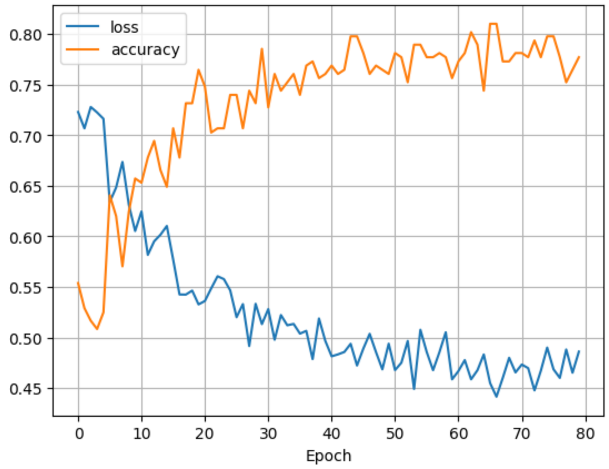

Fig. 6: 4-split VFL (patient, doctor, ECG, and laboratory)

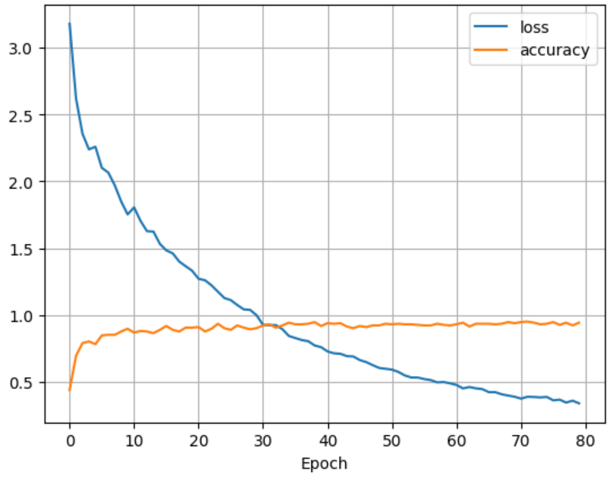

Fig. 7: Conventional DNN

D. Performance Measure

We compare the proposed VFL-based CVD prediction for different case studies. We also compare the performance of the traditional DNN-based approach for CVD prediction. To guarantee consistency for all cases, we trained and tested all cases with the same hyperparameters.

Fig 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 shown in the figure, compared to the conventional DNN model, the novel implementation of proposed VFL algorithms in the prediction of CVD provides comparable accuracy. The results reveal that the convergence of the proposed algorithms is obtained in around 60 communication rounds, after which there is less improvement in accuracy.

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

The results reveal that for different case studies, the performance is comparable to the traditional DNN method. For example, if the sample features are located at two different locations (patient and hospital), then the F1 score is the highest, that is, better than the conventional DNN methods. Similarly, for 3-splits (patient, lab, and hospitals) the metrics are comparable to that of the traditional methods. Thus, it can be concluded that the proposed VFL-based method shows comparable performance with respect to the traditional centralized DNN methods in addition to providing the added benefits of data privacy for heterogeneous clients.

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

We compare the proposed system with the state-of-the-art as presented in Table II. The system we propose provides a lot of advantages over the state-of-the-art. Our proposed system, mainly the 2-split, performs significantly better than the work in [13], [12], [28], [29] in terms of overall accuracy. Although the work in [20], [17] has slightly higher accuracy, it fails to address the issue of privacy preservation. In [22], FL has been utilized to preserve privacy, but it assumes that all clients are homogeneous, all having the same features and the target class. This scenario is unrealistic, as distinct clients may have different feature spaces in the data set and do not necessarily have the data label. The proposed method addresses this research gap. To the best of our knowledge, our study is the first to use a VFL-based system for the prognosis of CVD. In general, the proposed study addresses the research gap and also provides a performance comparable to the traditional DNN-based model and the FL-based approach.

V. Conclusion and Future work

This paper proposed an IoT-based framework that uses Vertical Federated Learning (VFL) for the automated prediction of cardiovascular disease using Machine Learning (ML). In contrast to the traditional federated learning (FL) mechanism, our proposed method considers the features of the data set to be separated for different clients. The proposed technique is compared with different case studies. Our results reveal that the proposed methods provide comparable performance as compared to the traditional DNN-based method in addition to providing the advantages of privacy as well as feature separation among clients. The presented results are equivalent in terms of performance as compared with the state-of-the-art. Additionally, our proposed framework preserves user privacy and feature separation between different client devices. In the future, we plan to use encryption methods in a shared model to provide robust privacy measures between the client and the server. In addition, differential privacy could be integrated into the proposed framework. Furthermore, a combination of

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

System Author's	Method	Privacy	Feature separation	Accuracy
[13]	K-Nearest Neighbor (KNN) model	✗	✗	72.91%
[12]	Artificial Neural Network (ANN) model	✗	✗	87.23%
[17]	Naïve Bayes model	✗	✗	90.16%
[28]	PTM (Problem Transformation Method)	✗	✗	80.89%
[20]	IoT-based Framework MDCNN (Modified Deep Convolutional Neural Network)	✗	✗	98.2%
[22]	Hybrid FL-based technique with MABC- RB-SVM	✓	✗	93.8%
Proposed system	Two-split vertical federated learning	✓	✓	88.53%
Proposed system	Three-split vertical federated learning	✓	✓	85.25%
Proposed system	Four-split vertical federated learning	✓	✓	80.33%

traditional FL and the proposed VFL could be integrated to reap the benefits of both technologies where there is a cluster of heterogeneous clients participating in the distributed model training process.

References

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

[2] T. Read, "Acsm cpt chapter 11: Preparticipation physical activity screening guidelines," *Personal Trainer Pioneer*, Mar 2023.

[3] M. Swathy and K. Saruladha, "A comparative study of classification and prediction of cardio-vascular diseases (cvd) using machine learning and deep learning techniques," *ICT Express*, no. 1, pp. 109–116, 2022.

[4] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. J. M. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boer, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," *IEEE Access*, vol. 9, 2021.

[5] L. T. Rajesh, T. Das, R. M. Shukla, and S. Sengupta, "Give and take: Federated transfer learning for industrial iot network intrusion detection," 2023.

[6] 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.

[7] BHF, "Heart statistics," Feb 2023.

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

[9] "Cerebrovascular disease," Sep 2022.

[10] CDC, "Peripheral arterial disease (pad)," Dec 2022.

[11] CDC, "Deep vein thrombosis & pulmonary embolism - chapter 8 - 2020 yellow book," Jun 2019.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

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

[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] A. Jamthikar, D. Gupta, A. M. Johri, L. E. Mantella, L. Saba, and J. S. Suri, "A machine learning framework for risk prediction of multi-label cardiovascular events based on focused carotid plaque b-mode ultrasound: A canadian study," *Computers in Biology and Medicine*, vol. 140, p. 105102, Jan 2022.

[29] X. Zhu, "A vertical federated learning algorithm for classfication problems with gradient-based optimization," 2021.