

Privacy-preserving Federated Learning for Echocardiographic Ejection Fraction Stratification under Non-IID Clinical Distributions

Abhigna Jalli, Kodukula Subrahmanyam, and Pachipala Yellamma *

Department of Computer Science and Engineering,

Koneru Lakshmaiah Education Foundation (Deemed to be University), Guntur, India

Email: jalliabhigna26@gmail.com (A.J.); smkodukula@kluniversity.in (K.S.); Pachipala.yamuna@gmail.com (P.Y.)

*Corresponding author

Abstract—Echocardiography plays a vital role in assessing cardiac function, with left ventricular Ejection Fraction (EF) serving as a key indicator of systolic performance. However, privacy regulations and restrictions on sharing patient data across institutions make it difficult to develop robust machine learning models for EF assessment. Federated Learning (FL) offers a promising solution by enabling collaborative model training without transferring raw clinical data. This study presents a privacy-preserving FL framework for echocardiographic EF classification and evaluates its performance using the Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) dataset under both Independent and Identically Distributed (IID) and Non-Independent and Identically Distributed (Non-IID) client data settings. In addition, the effect of injecting Gaussian noise into client updates is investigated as a lightweight privacy-enhancing strategy. For the clinically realistic EF classification task ($EF < 50\%$ versus $EF \geq 50\%$), centralised learning achieved an AUC of 0.765, whereas FL under IID and Non-IID settings achieved AUCs of 0.746 and 0.702, respectively. For the extreme EF cohort classification task ($EF \leq 35\%$ versus $EF \geq 60\%$), centralised learning and IID-FL achieved an AUC of 0.966, while Non-IID FL maintained the same Area under Curve (AUC) with a modest reduction in classification accuracy. The proposed framework also demonstrated strong resilience to moderate update perturbations, maintaining test AUC values above 0.96 across all evaluated noise levels. These findings highlight the potential of federated learning for privacy-aware, multi-institutional echocardiographic analysis and identify data heterogeneity as the primary challenge affecting federated model performance.

Keywords—federated learning, echocardiography, ejection fraction classification, privacy preservation, Non-Independent and Identically Distributed (Non-IID) data, medical image analysis

I. INTRODUCTION

Echocardiography is a cornerstone of clinical cardiology for imaging the heart and assessing cardiac function. Echocardiography allows us to see the heart's structure in real time and assess how well it is functioning. One important variable used to measure how well the heart is pumping (systolic function) is the Left Ventricular Ejection Fraction (LVEF). Having accurate LVEF estimates is important for diagnosing heart failure, determining treatment, and monitoring treatment success. In recent years, various techniques from Artificial Intelligence (AI), such as machine learning and deep learning, have shown great promise for automating different aspects of echocardiography, including view classification, segmentation, and LVEF estimation. However, developing models that generalise effectively and are suitable for clinical deployment typically requires access to large, diverse datasets collected from multiple institutions. Unfortunately, privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) impose strict restrictions on data sharing, making it difficult to aggregate sufficient high-quality patient data for training robust AI models [1].

Federated Learning (FL) is an attractive alternative to centralised training because it enables all parties to collaborate on building a model across multiple datasets without exposing the original patient data from individual organisations [2, 3]. Each organisation (client) creates its own model using its own data, but only sends its model updates (improvements) to a central server to combine all the improvements across organisations, thereby improving the privacy of each organisation while leveraging the different data types from all organisations. FL is gaining increasing attention in medical imaging research due to its ability to comply with privacy regulations while enabling

the development of more generalizable models through collaborative learning. It has been successfully applied to various medical imaging tasks, including brain tumour detection, Magnetic Resonance Imaging (MRI) segmentation, and histopathological diagnosis. These studies have demonstrated that federated models can achieve performance comparable to centralised models while preserving patient privacy and data confidentiality.

While there have been improvements made to the current methodology for clinical echocardiography with respect to federated learning, numerous obstacles remain before FL can become commonplace for this application. The first challenge is that real-world clinical data is highly heterogeneous (not only do hospitals serve patients with differing disease prevalences, but they also use differing imaging protocols and even different equipment characteristics), which contributes to the non-independent and non-identically distributed (Non-IID) nature of data across clients [4]. This heterogeneity increases the likelihood that federated models using standard aggregation techniques, such as Federated Averaging (FedAvg), will underperform because they assume similar data distributions across clients. A second consideration concerns the various methods used to preserve patient privacy (i.e., model isolation, noise injection, and differential privacy), which can help minimise the risk of information leakage when sharing updated parameters; however, these techniques can also adversely affect diagnostic quality if not calibrated appropriately. The existing body of research on federated learning in medicine has begun to investigate privacy-enhancing mechanisms (e.g., differential privacy and secure aggregation), but their empirical impact on medical imaging has yet to be thoroughly examined [5].

There has been relatively little research on the impact of federated learning on echocardiographic EF classification, despite its relevance to clinical practice amid real-world heterogeneity and privacy constraints. To date, most federated learning studies in medical imaging have focused on segmentation/classification tasks for MRI/Computed Tomography (CT); we have been unable to find any that focus on cardiac and ultrasound imaging. In addition, most of the existing literature does not capture how Non-IID clinical data distributions affect federated performance, nor does it evaluate lightweight privacy perturbations in conjunction with corresponding clinical tasks [2, 3, 5]. It is therefore imperative that we address these gaps to establish the practical applicability of Federated Learning within the specific workflows comprised of echocardiographic analyses.

Our aim in this paper is to propose a privacy-preserving federated learning method for identifying the echocardiographic EF class label using the Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) dataset [6]. We systematically compare centralised and federated training on two classification tasks using two types of clinical data distributions: IID and Non-IID. Since both IID and Non-IID distributions exist in clinical practice, we consider two relevant clinical tasks for our federated learning model. First, we use a threshold

of EF below 50% as low EF vs EF 50% and above as good EF. Second, we use EF 35% or below as very low EF, and EF 60% or above as normal.

EF to create clinically separated cohorts with low ambiguity. Next, we show that when we use a temporal feature extraction process to produce segmentation, we can achieve high-quality outcomes (i.e., over 95% classification accuracy). Additionally, within parameter-set periods, we will investigate whether adding a small amount of Gaussian noise to each client model update has any measurable effect on accuracy [7]. Our results suggest that federated learning with IID client distributions can achieve performance comparable to conventional centralised training. In contrast, Non-IID data distributions lead to a noticeable reduction in federated model performance due to increased data heterogeneity across clients. Nevertheless, the proposed framework maintained strong predictive capability, and clinically meaningful distinctions between EF categories remained evident even when Gaussian noise was incorporated into client model updates.

The main contributions of this work are threefold: (1) the development of a privacy-preserving federated learning framework for echocardiographic EF classification; (2) a systematic evaluation of model performance under IID and Non-IID client distributions with Gaussian update perturbation; and (3) a comparative analysis of federated optimisation strategies, including FedAvg, FedProx, and SCAFFOLD, for clinically relevant EF stratification tasks.

The rest of the paper will be organised as follows. Section II will review prior research on the use of federated learning methods in medical imaging and the challenges of handling Non-IID data. Section III details the dataset, feature extraction, and the federated framework used in this research. Section IV provides experimental results and analysis. Section V discuss their findings, limitations, and future directions for this research. Finally, Section VI presents the conclusion, with key insights and implications regarding collaborative echocardiographic analysis that protects patient privacy.

II. LITERATURE REVIEW

Federated Learning (FL) is a distributed optimisation framework intended to allow clients to train models collaboratively without having to share their original data with other clients [8]. Unlike the conventional centralised methodology, where data from multiple clients are pooled, under FL, clients train on their respective local datasets but send the resulting model parameters to a central server for aggregation. Since patient privacy regulations, institutional governance policies, and ethical obligations often prohibit direct data sharing between hospitals, the FL process presents an attractive alternative in healthcare.

A. Federated Learning in Medical Imaging

Medical image analysis has received significant attention from researchers in the field of federated learning over the last several years. Comprehensive literature reviews by Guan *et al.* [2] and da Silva *et al.* [3] discuss

aspects including architectural frameworks, communication protocols, privacy mechanisms, and domain-specific challenges associated with medical federated learning. Both surveys consistently report that FL can achieve nearly identical performance to centralised FL systems when data are both uniformly partitioned and statistically similar across all data sources; however, they also note that real-world hospital data typically does not conform to either property due to inherent differences between hospitals, making robustness to data heterogeneity the primary research focus in this area.

Choi *et al.* [9] conducted research to determine whether Fuzzy Logic has been adopted in health care and identified several barriers to its adoption: system interoperability, communication latency, regulatory compliance, and validation across diverse populations. They concluded that having the technical capability to use Fuzzy Logic is much more than technical feasibility; robust performance, at a minimum, requires an understanding of how distribution levels of data change from one population to another (i.e., Real World Settings).

Empirical feasibility has been established for large, collaborative federated studies in medical domains through Sheller *et al.*'s [10] work, which indicates that federated multi-site collaborations can produce competitive results with federated models compared to centralised baselines. Federated models can also be useful in privacy-sensitive imaging applications. Thus, other multi-site studies in radiology/pathology support the idea that federated collaboration improves model generalisation performance while maintaining proper data governance.

B. Impact of Data Heterogeneity

There are many issues with federated learning due to the differing distributions of clients' data. Applications of federated learning for medical imaging yield Non-IID data due to variations in medical imaging acquisition protocols, scanner vendors, ultrasound settings, the demographics of patients included in the images and the prevalence of disease. Such heterogeneous data can disrupt the optimisation process of federated learning, resulting in poor global performance of the model.

Federated convergence has been empirically evidenced to worsen when there is a major imbalance in distribution [11]. As a means of addressing this issue, Li *et al.* [12] presented FedProx, which uses a proximal regularisation term to constrain local updates, thereby increasing stability in the presence of heterogeneity. Beyond regulation, some personalised federated learning methods attempt to fit the global model to clients' individual characteristics [13]. These strategies also attempt to strike a balance between global and client-specific knowledge, thereby minimising degradation due to heterogeneity.

Despite this, there have been several studies that primarily rely on large-scale image segmentation or classification tasks, without systematically analysing clinically relevant stratification thresholds. In particular, the relationship between the selection of clinical threshold

values and federated heterogeneity has not been thoroughly investigated in echocardiography data analysis.

C. Privacy Enhancement in Federated Learning

Even without sharing raw patient data directly via federated learning, model updates themselves can leak data through various adversarial scenarios. As a result, there has been an active effort to develop formal privacy guarantees for federated learning. Differential Privacy (DP) algorithms add calibrated random noise to your model's parameters (model updates or gradients), thereby providing mathematically bounded privacy guarantees.

Adnan *et al.* [14] have shown that using differential privacy in Federated Learning (FL) enables an acceptable level of diagnostic accuracy in a medical image classification task when privacy budgets are appropriately set. Sensitivity-aware DP mechanisms have recently been developed to dynamically ensure a balance between privacy and utility [15]. Concurrently, secure aggregation protocols have been developed to prevent the server from viewing individual client updates during aggregation [16]. Although these solutions provide increased privacy, they also incur computational or communication costs and can alter the model's performance.

Very few researchers have explicitly evaluated the empirical robustness of medical imaging Federated Models using controlled perturbations of update methods in conjunction with heterogeneous data distributions to date; understanding the robustness of these models is necessary for evaluating real-world comparisons of privacy and utility.

D. Deep Learning in Echocardiography

Research for deep learning techniques in echocardiography has been primarily focused on segmentation (detecting the border of an object or region) and quantifying the left and right chambers of the heart, as well as how well the heart is pumping blood based on its size (the ejection fraction), with the use of centralised training. The CAMUS dataset has been established as a standard benchmark for segmenting and performing functional assessments, and many centralised models trained to segment images and regress ejection fraction have proven effective.

Recent advances in deep learning have significantly improved automated echocardiographic analysis. Ghorbani *et al.* [17] demonstrated that deep learning models can automatically identify cardiac structures and estimate clinically relevant cardiac parameters from echocardiographic images. Ouyang *et al.* [18] proposed EchoNet-Dynamic, a video-based framework for automated left ventricular ejection fraction estimation and cardiac segmentation with performance comparable to expert assessment. Other studies have further demonstrated the effectiveness of deep learning in assessing cardiac function and diagnosing disease from echocardiographic data [19, 20]. These advances highlight the growing role of AI in echocardiography; however, most existing approaches rely on centralised training and do not address privacy-preserving multi-institutional collaboration.

As a result of operator dependence, probe settings, and device heterogeneity, echocardiographic information varies significantly across facilities; however, very few published papers have examined how federated evaluation will help classify patients' EFs according to clinical classifications. There has been a recent increase in interest in utilising privacy-aware AI to develop federated foundation models of ultrasound [21]; however, Non-IID robustness and lightweight privacy perturbation analysis have not been thoroughly investigated for use in EF-based stratifications with federated evaluation.

E. Positioning of the Present Study

This research is the first to focus on federated echocardiographic EF classification across two practice-relevant stratification scenarios. These stratification scenarios are as follows: (1) a realistic threshold representing the assessment of systolic dysfunction; and (2) a relatively extreme cohort of EF values with minimal ambiguity associated with the labels. In addition, this research simulates the heterogeneous distribution of clients across institutions, capturing variability between institutions, while evaluating the robustness of the findings to controlled Gaussian perturbations arising from updates.

A thorough examination of centralised training, federated IID/Non-IID settings, and privacy-oriented noisy updates will provide structured insights into the interactions among clinical threshold design, data heterogeneity, and lightweight privacy mechanisms in echocardiographic federated learning. Ultimately, a systematic evaluation of the findings will help reduce the disconnect between theoretical federated frameworks and considerations for practical deployment in clinical settings.

Although prior studies have demonstrated the utility of deep learning and federated learning for cardiac image analysis, several limitations remain. Existing echocardiography studies predominantly rely on centralised learning frameworks, while many federated learning investigations focus on privacy preservation without specifically addressing clinically relevant Ejection Fraction (EF) classification tasks. Moreover, limited attention has been paid to the systematic evaluation of the effects of data heterogeneity, privacy-oriented update perturbations, and different federated optimisation strategies in echocardiographic analysis. To address these gaps, the present study proposes a privacy-preserving federated learning framework for EF classification, evaluates performance under both IID and Non-IID client distributions, investigates robustness to Gaussian update perturbation, and compares multiple federated optimisation strategies, including FedAvg, FedProx, and SCAFFOLD.

III. MATERIALS AND METHODS

The following outlines the methods utilised in this research. First, we introduce the dataset of echocardiographic studies that will serve as the basis for the analysis that follows. Then, we will provide an

overview of how patient data has been partitioned for analysis. Next, the EF Classification problem will be clearly defined and will contain a description of both the clinical EF threshold and the "extreme" cohort criteria. Following this will be a detailed description of how we extract temporal features from LV segmentation masks. In addition, we will provide a description of the baseline centralised model and the federated learning methods used, including simulations of IID and Non-IID distributions. Finally, we will define our Privacy Perturbation method based on Gaussian noise injection, as well as the evaluation metrics used to assess classifier performance.

A. Dataset

This study's experimental data was obtained from the CAMUS dataset, an open-access echocardiography dataset used to further research on the automated segmentation and functional evaluation of cardiac structures [6]. The CAMUS dataset consists of data collected from 500 echocardiographic patients and supports standardised 2D echocardiographic images.

For each individual participating in the study, 2CH and 4CH views are available through the apical window. For each of these views, sequence images are assessed and provided with expert-annotated segmentation masks for two important clinical phases: to determine Ejection Fractions (EF) at the End of Diastole (ED) and End-Systole (ES). Each patient has a record containing ground-truth EF information, which will be the primary clinical measure used in the study to evaluate ventricular performance.

Echocardiography researchers have widely adopted the CAMUS dataset because of its standardised acquisition protocol and high-quality manual annotations [6]. We will focus on the 4CH view throughout our work because it is a standard clinical functional assessment view that clearly shows the left ventricle. The full cardiac cycle sequence, along with its segmentation mask, was used to derive EF-based temporal contraction features.

The dataset was partitioned at the patient level rather than the frame level to enable a fair, even evaluation and prevent data leakage. The patients were randomly split into three groups (70%/15%/15%) to ensure that frames from an individual patient were not included across multi-splits; this will help maintain independence between the training and test data.

B. Problem Definition

This study will explore whether it is feasible to use federated learning to automatically compute left ventricular Ejection Fraction (EF) from echocardiograms. EF is frequently utilised as an objective measure of cardiac systolic function and provides important evidence for both the diagnosis of heart failure and the monitoring of response to treatment. Two classification problems that utilise binary-coded definitions of EF were designed to evaluate classifier performance across varying levels of clinical complexity.

1) Task 1: Realistic clinical threshold

The first task presents a clinically relevant scenario commonly encountered during a typical cardiac evaluation. The two groups of patients could be divided into the following based on clinically defined systolic function criteria:

- Class 0: EF of 50% or greater (your systolic function is either normal or slightly decreased)
- Class 1: EF of less than 50% (your systolic function is decreased)

The 50% threshold is commonly utilised in clinical cardiology to delineate the boundary between preserved and failing ventricular function. In real diagnostic practice, EF values will fall within the grey area or borderline range, thereby creating overlap of functional characteristics between these two classes and a moderate degree of difficulty for this classification problem, while enabling the evaluation of federated learning performance under conditions of clinically relevant ambiguity.

2) Task 2: Extreme EF cohort

The second activity was meant to provide a clearer example of a distinctly different type of clinical situation. We also selected patients with different EF values than in the first study to create an additional classification based on EF extremes, rather than a single broad category that included a Threshold category between the two classes.

- Class 0 (Normal Systolic Function): EF 60% and above
- Class 1 (Severe Systolic Dysfunction): EF 35% and below

By removing intermediate EF values, the task reduces uncertainty in labelling and increases the physiological separation between classes. While patients in the low EF group have significant ventricular dysfunction, those in the high EF group exhibit strong systolic contraction. These two datasets enable evaluation of the performance of federated learning algorithms under functional differentiation, with clear distinctions between classes. Furthermore, it enables evaluation of model robustness to client variability at high levels of class differentiability.

The two tasks are evaluated together and serve as two distinct evaluation scenarios for our model. Task 1 is a clinical threshold that may have overlapping classes, whereas Task 2 allows evaluation of model performance at a very high level of confidence and with little ambiguity.

The performance of both classification tasks is quantitatively evaluated using Accuracy, F1-score, and Area under the Receiver Operating Characteristic (ROC) Curve (AUC), with AUC serving as the primary metric for assessing class discrimination. Detailed performance comparisons for centralised learning, federated learning under IID settings, and federated learning under Non-IID settings are presented in Section IV.

C. Feature Extraction

In this study, a feature-based approach was implemented rather than directly training a model based on the raw pixel values of ultrasound images. Temporal contraction features for the Left Ventricle (LV) were derived from the 4-Chamber (4CH) half-sequence

segmentation masks of LV in the CAMUS dataset, thereby reducing the model's complexity but still maintaining functionally relevant, clinically meaningful information.

To obtain a temporal sequence of LV areas as the LV expanded and contracted throughout the cardiac cycle, each patient was segmented at each time frame using the segmentation mask to determine the LV area. The LV area for each time frame was determined by counting the number of pixels corresponding to the LV region.

The feature extraction procedure begins by identifying the Left Ventricular (LV) region from the segmentation mask at each frame of the cardiac cycle. The LV area is then calculated by counting the pixels in the segmented region, thereby generating a temporal area signal that captures ventricular contraction and relaxation dynamics. From this signal, a set of physiologically meaningful descriptors, including maximum area, minimum area, area difference, pseudo-ejection fraction, mean area, and standard deviation, is computed and used as input to the classification model. From this area-time sequence, the following quantitative features were extracted.

- Maximum LV area:

$$A_{max} = \max_t A_t$$

The maximum LV area approximately corresponds to the end-diastolic phase, when the ventricle reaches its maximum filling.

- Minimum LV area:

$$A_{min} = \min_t A_t$$

The minimum LV area approximately corresponds to the end-systolic phase, representing the point of maximum ventricular contraction.

- Area difference:

$$\Delta A = A_{max} - A_{min}$$

The area difference represents the absolute magnitude of ventricular contraction during the cardiac cycle.

- Pseudo-ejection fraction:

$$pEF = \frac{A_{max} - A_{min}}{A_{max}}$$

The pseudo-ejection fraction approximates the conventional ejection fraction by using area measurements instead of volumetric estimates.

- Mean LV area across the sequence:

$$\mu_A = \frac{1}{T} \sum_{t=1}^T A_t$$

The mean LV area represents the average ventricular size throughout the cardiac cycle.

- Standard deviation of LV area:

$$\sigma_A = \sqrt{\frac{1}{T} \sum_{t=1}^T (A_t - \mu_A)^2}$$

The standard deviation of the LV area reflects the variability in ventricular motion throughout the cardiac cycle.

These temporal descriptors together define the dynamics of ventricular contraction (systolic function), and they provide a low-dimensional functional biomarker of cardiac function. These attributes are physiologically interpretable, rather than high-dimensional pixel-based representations of the data, and they all include physiological information directly related to systolic function. The use of these features should decrease the potential for overfitting, improve computational efficiency, and increase the ease of interpreting results.

Before training the model, we standardised all predictors using z-scores to make them comparable on a common scale and to stabilise optimisation by removing potential biases. The standardisation was done using only the statistics from the training set. We then applied the same statistics to the validation and testing datasets, preventing data leakage.

The overall image analysis and classifier pipeline is shown in the following Fig. 1.

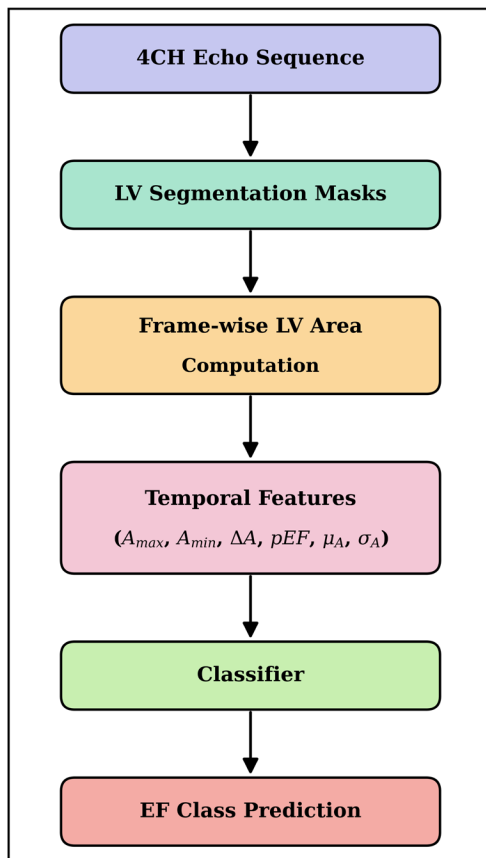


Fig. 1. Feature-based EF classification pipeline using temporal contraction descriptors derived from segmentation masks.

The starting point of the pipeline is a 4-chamber echocardiogram sequence, which is processed to obtain Left Ventricular (LV) segmentation masks (i.e., tracing the contour of the left ventricle through each frame) for each frame of the echocardiographic sequence. Next, LV area measurements are computed for each frame throughout the

entire cardiac cycle. Then, from these area measurements, a time series of LV area measurements across the cardiac cycle is created (temporal contraction signal). Finally, a few LV area-based physiologically relevant characteristic measurements (e.g., maximum area, minimum area, pseudo-ejection fraction, 1st–3rd statistical moments) from each temporal contraction signal are extracted. Each of the LV area-based characteristics is then input into a lightweight classifier to predict the LF category. Fig. 1 illustrates the organised progression from raw imaging data through image analysis to interpretable functional biomarkers.

D. Centralized Learning Baseline

To derive a reference performance level, the first step was to create a centralised training framework that pooled training data from each individual patient in the selected cohort to train a single global model on the pooled data. This serves as a reference benchmark in which all training data are pooled for model development.

The two distinct modelling frameworks used for the two distinct classification tasks were:

- Image-based classification (Task 1): A Convolutional Neural Network (CNN) based on the “ResNet18” architecture was used. This is a very popular deep learning architecture that uses residual connections (skip connections) to increase gradient stability and improve convergence speed in moderately deep neural networks. CNN was trained on echocardiographic images to produce binary EF classifications.
- Feature-based classification (Task 2): For the extreme EF cohort case, a logistic regression classifier was trained on low-dimensional temporal contraction features derived from segmentation masks. Logistic regression was used because of the models’ high interpretability, their robustness to small datasets, and the fact that the cohorts were linearly separable.

To assess the model’s performance, three metrics were utilised as complementary indices. Accuracy indicates how correct the model is, while the F1-score incorporates prior knowledge of the class distribution when evaluating the model’s performance, considering both precision and recall. The Area Under the ROC Curve (AUC) will be used as the primary measure of discrimination because it evaluates the classifier’s ranking ability without relying on a specific cut-off value.

Upon evaluating the probability outputs on the validation set, based on accuracy or F1 score, the optimal classification threshold for the model was determined. This determined threshold was then applied to the holdout test set. This process ensures that no bias is present in performance reporting and prevents overfitting to the test set.

E. Federated Learning Framework

The decentralised, multi-institutional settings that preserve privacy were simulated using a Federated Averaging (FedAvg) model based on federated learning [8]. In this decentralised environment, clients retain their data locally; however, they share only their model parameters (or updates) during training.

In round t of communication, the current global model is denoted by its global parameters (w_t), and the client began to create their local model using that information, followed by training on their private data set for several epochs. Once complete, the clients send their local updates to the server, where they are averaged with weights.

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

The set of model parameters, denoted by $w_t^{(k)}$, is client-specific, local model parameters created at round t by client k ; n_k is the total number of training samples for client k ; and $N = \sum_{k=1}^K n_k$ identifies the total number of training samples across all clients included in the design. Each client's total number of training samples contributes to the global model parameters, such that clients with larger numbers of training samples will have a greater impact on this global update than clients with smaller numbers of training samples.

This research involved the creation of simulated client data from 3 independent health care organisations, designed to represent a realistic situation of small-scale collaboration across multiple centres.

The overall architecture for federated learning used in the present study is depicted in Fig. 2 below. Each participating institution stores its own echocardiography data and then performs independent local model training relative to the existing global model. The local model parameters at the client are optionally modified by adding Gaussian noise before being sent to the central server. The central server receives prior updates from clients and aggregates them using the FedAvg method (Eq. (1)) to create a new global model. The new global model will be sent back to all the clients for the next round of communication. In this type of decentralised architecture, no raw patient data would leave the control of the individual institutions involved in the collaborative effort, while allowing cooperation among the institutions to improve the models created.

Fig. 2 illustrates the complete federated learning workflow, including local feature extraction and model training at each hospital, Gaussian noise injection for privacy preservation, FedAvg-based aggregation at the central server, and redistribution of the updated global model for subsequent training rounds. Each participating hospital performs feature extraction and local model training on private echocardiography data. The resulting model updates are perturbed through Gaussian Noise Injection (GNI) before being transmitted to a central server for FedAvg aggregation. The updated global model is then redistributed to participating hospitals for subsequent training rounds.

It is important to note that the Gaussian noise injection strategy adopted in this study differs from formal Differential Privacy (DP) mechanisms. While Gaussian perturbation obscures client model updates and provides n empirical assessment of privacy robustness, it does not provide mathematically quantifiable privacy guarantees, such as those offered by differential privacy frameworks.

Therefore, the proposed approach should be interpreted as a privacy-enhancing perturbation technique rather than a formal differential privacy implementation.

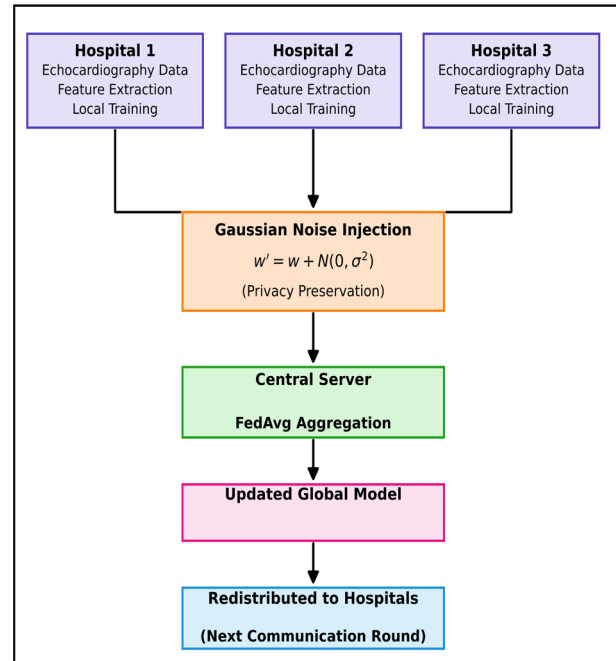


Fig. 2. Privacy-preserving federated learning workflow for echocardiographic EF classification.

Two ways of making client distribution configurations were investigated:

IID Setting: All patients are distributed evenly among the clients, so that each institution has approximately the same proportions of classes and similar characteristics (i.e., data) in general. Thus, these conditions would represent the closest to an ideal collaborative environment, or in other words, statistically matching hospital populations to each other.

Non-IID Setting: To simulate realistic clinical heterogeneity, the patient cohorts were split unevenly to introduce variation in class prevalence across clients. For example, one institution may have mostly reduced EF patients while another institution may have many preserved EF patients. This will reflect what occurs in a true clinical environment, including patterns of clinic referrals, community-level demographics, and the burden of disease as reported across different hospitals [13].

Fig. 3 shows how Federated Learning can work with either IID or Non-IID distributions of clients. In the case of the IID distribution, we can see that all institutions have equivalent proportions of EF classes, so that statistically similar training data is available at all sites. On the other hand, the nonIID distribution emulates a distribution of clinical cases with different prevalences, for example, one site primarily treats patients with reduced EF while another site treats patients with preserved EF only. Therefore, this distributional imbalance presents challenges for optimisation and reflects real-world variations between multiple centres.

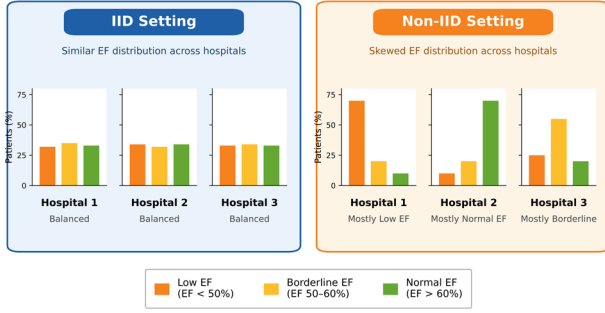


Fig. 3. Illustration of IID versus Non-IID client distributions in federated learning.

Based on the above client distribution settings, federated model training proceeded through a sequence of communication rounds. Each communication round consisted of the following steps:

- Local training at each client for a predefined number of epochs.
- Transmission of updated model parameters to the server.
- Weighted aggregation of client updates to form a new global model.

Throughout several rounds of communication, the final performance of the Federation was assessed using a test-set protocol that provided an opportunity to systematically compare the Centralised/Federated approaches on a fixed testing basis.

The experiments were implemented in Python using Scikit-learn and NumPy. Logistic regression was employed as the base classifier with a learning rate of 0.05 and two local training epochs per communication round. Federated training was conducted using three simulated clients representing participating institutions. For Non-IID experiments, client data distributions were intentionally varied to emulate institutional heterogeneity. The FedAvg algorithm was used as the primary aggregation strategy, while FedProx and SCAFFOLD were additionally evaluated for comparative analysis. Gaussian noise perturbations with standard deviations ranging from 0.01 to 0.20 were applied to client model updates to assess privacy robustness. All experiments were performed on a standard workstation environment.

F. Privacy-preserving Update Perturbation

Federated learning protects against the direct transmission of unaltered patient data by using model updates exchanged between participating clients and the central server. However, even if an update does not explicitly include information about local data distributions, it can still implicitly convey some information. Therefore, we added a lightweight perturbation technique that utilised noise to the federated training mechanism, providing an extra layer of privacy.

Specifically, Gaussian noise was added to the locally trained model parameters prior to transmission to the central server. The locally trained model parameters at the client after completing local optimisation are denoted by w . Let the perturbed update transmitted to the server be denoted by W' and defined as follows:

$$W' = w + N(0, \sigma^2) \quad (2)$$

where $N(0, \sigma^2)$ represents zero-mean Gaussian noise with variance σ^2 . The parameter σ controls the magnitude of perturbation and therefore determines the strength of privacy masking applied to the model update.

The impact of perturbation level was analysed by examining the different noise levels used to randomly change the model weights, based on the defined equation for each noise level.

$$\sigma \in \{0.0005, 0.001, 0.005, 0.01, 0.02, 0.05\} \quad (3)$$

The noisy update mechanism defined in Eq. (2) was systematically evaluated across the noise levels shown in Eq. (3). For each value of σ , federated training was repeated, and performance was measured on the centralized test set.

The strategy outlined in this paper approximates several common privacy-enhancing mechanisms used in federated learning research, particularly those based on Differential Privacy (DP). That is, we will use an approach similar to those that inject noise into a model's updates to reduce the likelihood of leaking information [16, 17]. However, unlike formal implementations of DP, where privacy budgets and mathematically defined privacy leakage are calculated, this implementation does neither. Rather, it serves as an empirical robustness analysis method to assess whether the federated model is sensitive to disturbances arising from client updates.

The investigation of performance stability as a function of increased noise levels will yield insights into both the stability of the learned feature representation and the practical trade-off between update perturbation and prediction accuracy.

G. Evaluation Metrics

The application of three separate yet complementary measures of quantitative performance (Accuracy, F1-score, Area Under the Receiver Operating Characteristic Curve (AUC)) to classify different performances in a balanced manner when the clinical classes represented in the sample may be imbalanced.

1) Accuracy

Accuracy is a measure of the number of correctly classified samples among all samples tested. It is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Whereas accuracy gives a simple measure of overall correctness, it can be misleading in skewed datasets, such as those where one group prevails over the other.

2) F1-score

It refers to the harmonic mean of recall and precision and can be defined as:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Precision is the fraction of predicted positive results that are positive, and recall is the fraction of actual positive results that are predicted. F1-score is especially applicable in medical classification, where false negatives and false positives have varying clinical consequences.

3) Area under the ROC Curve (AUC)

AUC measures the model's ability to discriminate between classes across all possible decision thresholds. It evaluates the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity). An AUC value of 1.0 indicates perfect discrimination, whereas 0.5 corresponds to random guessing.

Among these metrics, AUC was selected as the primary metric due to its threshold independence and robustness to class imbalance. It is often more useful in medical decision-making to rank patients correctly across many operating points than to evaluate at only one fixed threshold. Consequently, AUC also provides an overall evaluation of how well the model separates patients.

The criterion for model selection was the optimal model, defined as the one with the highest validation AUC on the dataset during the project's training phase. Once the optimal model was identified, it was fitted to the dataset and then evaluated against the dataset using the test dataset, i.e., the models were evaluated to ensure that they performed equally well regardless of whether the data was trained using a centralised or non-centralised (federated) training paradigm in order to minimise the effects of overfitting on a given metric or decision threshold.

IV. RESULTS

In this section, we present the experimental results for EF classification tasks in both centralised and federated learning settings. The performance results were generated for both the IID and Non-IID client distributions, enabling a clear assessment of the effect of data heterogeneity on the model. Furthermore, we will analyse each result based on the model's ability to discriminate, its robustness (or lack thereof), and, finally, its consistency across all evaluation metrics used to assess the model's performance.

A. Task 1: Realistic EF Classification

Table I summarises the performance comparison between centralised training, federated IID training, and federated NonIID training for the clinically realistic EF threshold (EF below 50% versus EF 50% and above). This task reflects a practical clinical scenario where borderline EF values may reduce class separability.

To further assess the impact of federated optimisation strategies under heterogeneous client distributions, two advanced methods, FedProx [14] and SCAFFOLD [22], were evaluated alongside the standard FedAvg algorithm. FedProx incorporates a proximal regularisation term to mitigate client drift, while SCAFFOLD employs control variates to reduce optimisation bias arising from Non-IID data distributions. Their performance is compared with

centralised training and FedAvg-based federated learning in Table I and Table II.

TABLE I. PERFORMANCE COMPARISON FOR REALISTIC EF CLASSIFICATION (TASK 1)

Method	Test Accuracy	Test F1	Test AUC
Centralized	0.660	0.794	0.765
FL-IID	0.693	0.798	0.746
FL-Non-IID	0.660	0.794	0.702
FedProx	0.649	0.774	0.689
SCAFFOLD [22]	0.601	0.732	0.643

Under centralised training, the model achieved a test accuracy of 0.660 and an AUC of 0.765. This represents the performance obtained when all training data are pooled together without institutional separation.

The federated IID model achieved a slightly higher accuracy (0.693) and a similar F1-score (0.798), while the AUC dropped to a moderate level (0.746) in the IID federated environment, where patients were equally distributed among the clients. This implies that federated learning can be used to model centralised performance when client data have statistically similar distributions.

On the other hand, using a federated setup with Non-IID data resulted in a decrease in discriminative power (AUC = 0.702) compared with a centralised approach, where the AUC was 0.765. While the accuracies of both settings were similar (66%), lower AUCs indicate that the classes were not as clearly separated when the client populations were heterogeneous. These results are consistent with past research showing that data heterogeneity adversely impacts federated optimisation and convergence stability [13].

Among the advanced federated optimisation methods, FedProx achieved a test accuracy of 64.9%, F1-score of 0.774, and AUC of 0.689, while SCAFFOLD achieved 0.601, 0.732, and 0.643, respectively. Both methods underperformed compared with FedAvg under Non-IID conditions. This suggests that the additional optimisation constraints introduced by FedProx and SCAFFOLD did not yield measurable benefits for the proposed feature-based EF classification framework and may have limited its adaptation to the underlying data characteristics.

These results indicate that, when there is a balanced distribution of clients, federated learning is just as capable as traditional training using centralised repositories. However, when simulated conditions reflect true clinical variability between institutions, a drop in performance from the federated learning process will result.

B. Task 2: Extreme EF Classification

To further evaluate how well the federated learning approach handles clearer physiological differences among patients, extreme thresholds for Ejection Fraction (EF) were determined. For this analysis, patients with an EF value of 35% or lower were considered to have severely depressed systolic function, while patients with an EF value of 60% or higher were considered to have preserved systolic function. By removing all EF values that fall between these two thresholds, label uncertainty is reduced,

and the separability between the two class levels is increased.

Table II shows the performance of both centralised and federated training methods in this extreme EF classification situation.

TABLE II. PERFORMANCE COMPARISON FOR EXTREME EF CLASSIFICATION (TASK 2)

Method	Test Accuracy	Test F1	Test AUC
Centralized	0.917	0.944	0.966
FL-IID	0.917	0.941	0.966
FL-Non-IID	0.875	0.914	0.966
FedProx [14]	0.856	0.892	0.923
SCAFFOLD [22]	0.812	0.854	0.884

Data from the centralised training indicate that the model achieves a test accuracy of 0.917 and an AUC of 0.966. These two values show that this model has a strong ability to differentiate and classify patients into two distinct groups. Because of the model's excellent ranking performance across a range of classification thresholds, the AUC is extremely high.

With the IID federated architecture, the performance of centralised and federated training was fairly similar; both achieved approximately the same accuracy (0.917) and a similar AUC (0.966). Based on this outcome, it appears that federated optimisation methods can adequately replicate centralised learning results when class distributions across institutions are similar, regardless of whether the learning environment is centralised or decentralised.

Test accuracy showed a moderate decline under the NonIID client partitioning approach (0.875), and AUC held steady at 0.966. The fact that AUC remained constant despite the drop in accuracy demonstrates the ability to rank patients by EF severity, but, due to the use of a fixed classification cut-off, slightly more misclassifications occurred. This behaviour indicates that distribution shifts among Clients primarily affect calibration at a single decision point rather than overall discriminative ability.

A comparison with FedProx and SCAFFOLD further confirmed the effectiveness of the FedAvg-based framework. FedProx achieved a test accuracy of 0.856, F1-score of 0.892, and AUC of 0.923, whereas SCAFFOLD achieved 0.812, 0.854, and 0.884, respectively. Although both methods are designed to address client heterogeneity, neither outperformed FedAvg in the current experimental setting. These findings indicate that the proposed feature representation and moderate client heterogeneity can be effectively handled using the standard FedAvg optimisation strategy.

Compared with the realistic EF threshold task, the performance in the extreme EF condition is superior across all training paradigms. The improved performance is likely due to the increased physiological separation between extremely reduced and well-preserved systolic function. The enhanced signal-to-noise ratio within the feature space also likely contributes to the robustness of both central and federated systems.

C. Privacy Perturbation Analysis

Gaussian noise was added to each client's parameters before they were aggregated at the server to test the robustness of the federated framework in response to noise being added to the model updates as defined in Eq. (2). This experiment investigated how increasing the magnitude of noise added to the model update would affect the model's predictive performance, thereby demonstrating the stability of the learned representation under privacy-based noise addition.

Table III summarizes the performance of the proposed federated framework under different Gaussian noise levels, reporting the validation AUC together with the test accuracy, F1-score, and AUC for each standard deviation (σ).

TABLE III. EFFECT OF GAUSSIAN NOISE ON FEDERATED EXTREME EF CLASSIFICATION

σ	Val AUC	Test Acc	Test F1	Test AUC
0.0005	0.765	0.875	0.914	0.966
0.001	0.765	0.875	0.914	0.966
0.005	0.765	0.875	0.914	0.966
0.01	0.775	0.917	0.944	0.966
0.02	0.775	0.875	0.914	0.966
0.05	0.794	0.875	0.914	0.975

For each level of noise added (all tested levels), classification performance remained fairly constant, all samples at each level produced a test AUC greater than 0.96, indicating that discriminative ability was preserved as the level of noise increased. Of particular interest is that the performance at a noise level of $\sigma = 0.01$ is equal to or slightly higher than that at the other lower levels, indicating that small stochastic changes in the data did not negatively impact convergence in the model.

Discrimination performance didn't degrade significantly, even at the maximum tested noise level ($\sigma = 0.05$). Performance showed consistent stability in this testing, indicating a robust feature representation that can withstand moderate parameter perturbations due to noise. The low-dimensional, physiologically relevant temporal features were also likely a factor in producing more smoothly defined decision boundaries than would be found with higher-dimensional pixels.

In this experiment, no formal differential privacy is provided; only that sensitivity has been empirically established as acceptable for training this model with moderate perturbations to the updates. This aligns with the literature on federated learning for classifying images from different hospitals within a privacy-enabling framework [16, 17].

Overall, the results demonstrate that the proposed federated framework maintains good predictive stability after adding very little noise designed for privacy-preserving purposes, suggesting that this approach to federated training can be useful when multiple hospitals wish to jointly train a shared predictive model using their combined patient repositories.

D. Summary of Observations

The findings of our experiments show distinct patterns between centralised and federated configurations in privacy-preserving collaborative learning for Echocardiography's EF classification. We summarise the main insights from this analysis below.

Federated learning performs similarly to centralised learning in terms of achieving high accuracy when using IID patient distributions across different sites. When patient data are uniformly distributed across the various health care sites, the federated models achieve performance metrics (e.g., accuracy, F1-score, and AUC) comparable to those obtained using centralised data. This suggests that an approach that uses the FedAvg algorithm to optimally train across multiple sites with private patient records can be just as effective when there is a common statistical pattern across all sites.

The introduction of non-identically distributed clients in a distributed learning model resulted in reduced performance on the EF classification task, particularly when EFs fell below 50% or above 49%. Although total accuracy remained unchanged, AUC decreased, indicating that the ability to discriminate was adversely affected by the data distribution. These results demonstrate how sensitive federation-based optimisation is to class imbalance and institutional variability, supporting existing observations about the problems that arise in heterogeneous medical data systems.

The clinical utility of separating subjects into two distinct categories: subjects with an EF 35% and below and subjects with an EF 60% and above. Diagnostic performance was assessed from the EF cohort ($n = 400$) at each level of separation ($n = 400$). We found that diagnostic classification accuracy consistently exceeded 90% in both federated and centralised approaches across all scenarios, with Non-IID partitioning and moderate Gaussian perturbation applied to the model updates remaining present throughout all tests. In summary, our findings indicate that subjects with more pronounced physiological separability exhibited greater robustness in the federated model and were less affected by heterogeneity and noise in the patient population than subjects in the same cohort with less pronounced separability.

These findings support the idea that echocardiographic EF classification can be performed using federated learning in a clinically relevant way for EF class definitions that are clearly separated from one another. However, sensitivity to potential distributional shifts in real-world scenarios highlights the need to address data heterogeneity when conducting multi-institutional collaborations on this topic.

E. Graphical Comparison of Performance

The visual representation of model accuracy is supported by the tabular results shown below. Figs. 4–6 present a way to visually assess how the models performed when trained centrally and via federated learning, including both IID and Non-IID client distributions, as

well as the effect of adding Gaussian noise to the updates to account for privacy considerations.

1) Task 1: Realistic EF classification

Fig. 4 summarises performance for the clinically realistic EF threshold (EF below 50% versus EF 50% and above). Fig. 4(a) reports test AUC, while Fig. 4(b) contrasts test accuracy and F1-score.

In Fig. 4(a), the AUC is highest when all training data is centralised, indicating better discrimination when all training data is pooled. When evaluating the IID federated configuration, there is only a small decrease in AUC, indicating that FedAvg can serve as a baseline for centralised clinical discrimination when the clients' distributions are highly overlapping. The Non-IID federated configuration shows a larger drop in AUC, indicating lower case separability due to wide clinical-level variation. Fig. 4(b) further shows that accuracy and F1-score were similar between centralised and IID federated settings, whereas the scores for the Non-IID arrangement did not appear stable, consistent with the AUC results.

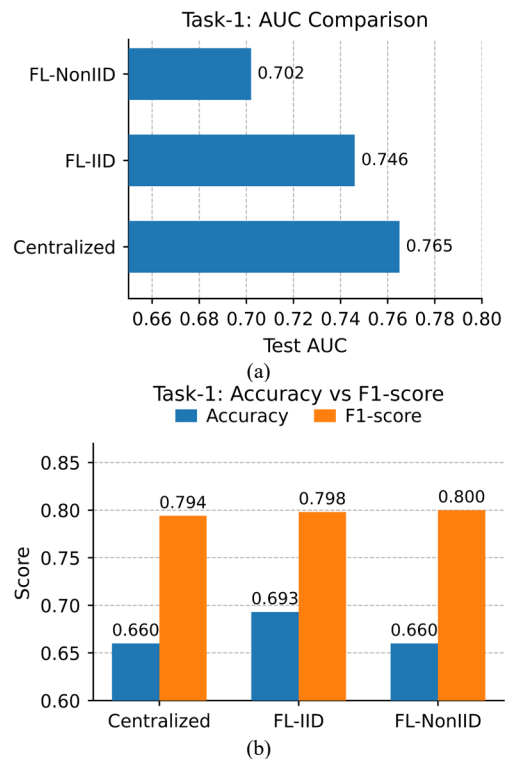


Fig. 4. Performance comparison for Task-1 (realistic EF threshold) across centralized training, FL under IID partitioning, and FL under Non-IID partitioning. (a) Task-1: Test AUC comparison. (b) Task-1: Test accuracy and F1-score comparison.

2) Task 2: Extreme EF classification

Fig. 5 presents results for the extreme EF cohort task (EF 35% and below versus EF 60% and above), which yields clinically well-separated groups. Fig. 5(a) shows the comparison of accuracy and F1-score, while Fig. 5(b) reports test AUC.

In Fig. 5(a), federated IID training matches centralised performance in both accuracy and F1-score, demonstrating

that decentralised learning can replicate pooled-data results under balanced client distributions. Under Non-IID partitioning, a moderate reduction in accuracy and F1-score is observed, but the overall decline is smaller than that seen in Task 1. Fig. 5(b) indicates consistently high AUC across all settings, suggesting that rank-based discrimination is preserved even when client data distributions differ. This pattern implies that heterogeneity affects threshold-dependent classification decisions more than overall separability in the extreme EF setting.

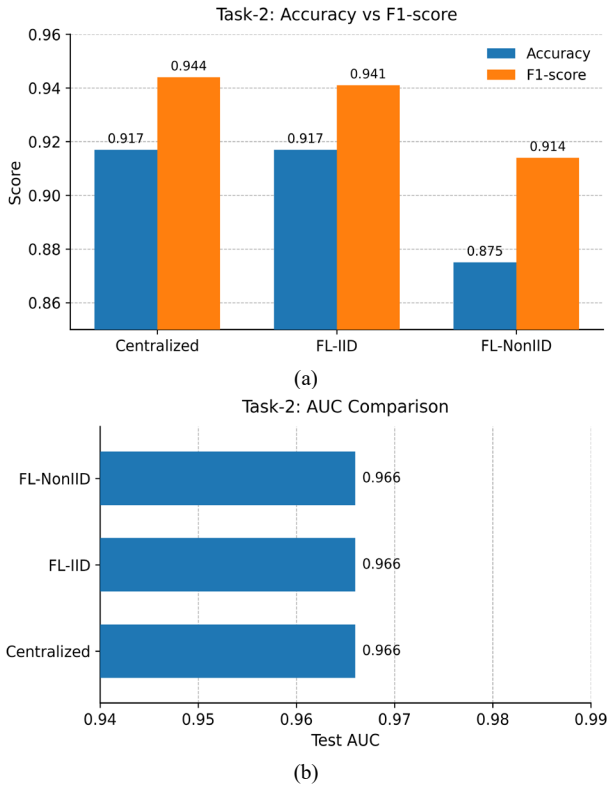


Fig. 5. Performance comparison for Task-2 (extreme EF cohorts) across centralized training, FL-IID, and FL-NonIID. (a) Task-2: Test accuracy and F1-score comparison. (b) Task-2: Test AUC comparison.

3) Privacy perturbation analysis

To evaluate robustness to noisy client updates, Fig. 6 illustrates the effect of Gaussian perturbation magnitude σ on performance for the Non-IID federated extreme EF setting. Fig. 6(a) reports test accuracy as σ increases, and Fig. 6(b) reports test AUC.

For every noise magnitude tested, performance remained almost unchanged, with only a negligible amount of change. More importantly, AUC was consistently high across tests, indicating that moderate amounts of disturbance in the updates do not significantly affect discriminative ability. The consistent performance also means that the low-dimensional temporal characteristics extracted from the segmentation masks yield a relatively flat decision boundary that is less susceptible to moderate noise.

The findings illustrated in the figures provided are: (1) federated learning can achieve performance outcomes similar to those of centralized learning for Independent Identical Data (IID) partitions; (2) the presence of data

heterogeneity from Non-IID clinical populations will provide measurable degradation in performance but to a predictable degree so long as the EF stratification is more realistic; and (3) Cohorts with extreme EFs allow for high-quality diagnostics while also having a degree of robustness against both distributional changes and moderate levels of perturbation in the data.

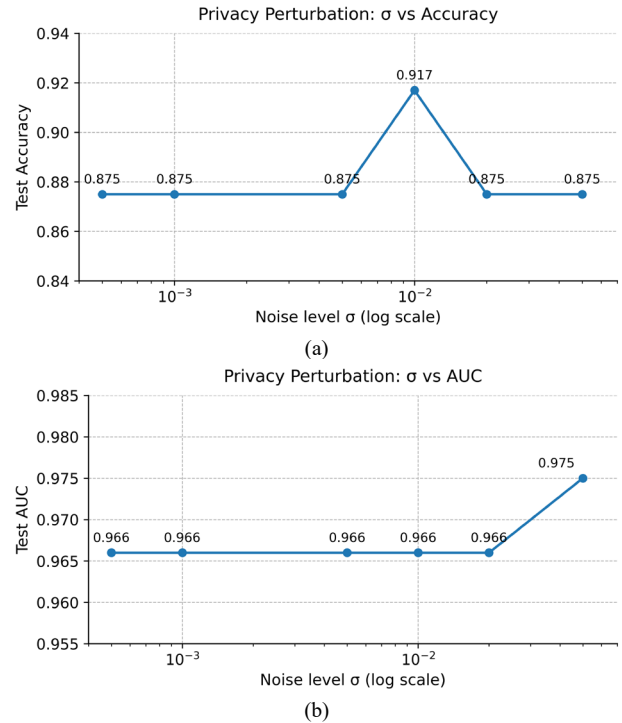


Fig. 6. Privacy perturbation analysis on Non-IID federated learning for the extreme EF task: Effect of Gaussian noise magnitude. (a) Noise level σ vs test accuracy (log scale). (b) Noise level σ vs test AUC (log scale).

V. DISCUSSION

This study evaluated the feasibility of using privacy-preserving Federated Learning (FL) to classify echocardiographic Ejection Fractions (EFs). The results provide important insights regarding model robustness, clinical separability, and the impact of privacy perturbations under both IID and heterogeneous Non-IID clinical distributions.

First, in Federated Learning (FL), when dealing with the realistic EF stratification task (EF of 49% or less versus 50% or greater), the performance of FL under IID partitioning was like that of centralised training. This is consistent with previously published evidence that FL can approximate centralised model performance when the class distributions of the client data are IID [9, 10]. When client data distributions simulating client distributions in a Non-IID manner were created, substantial decreases in the Area Under the Curve (AUC) were observed. These decreases in AUC corroborate results from previous publications demonstrating how Non-IID data affects the ability of FL to converge and perform well in clinical practice, as different hospitals generally have patients with different disease distributions and clinical characteristics. The use of Non-IID data is a common scenario in clinical

practice [13]. The large decrease in AUC supports the need to evaluate FL under heterogeneous distributions rather than idealised IID data.

Second, the very extreme EF classification assignment (EF 35% and below, EF 60% and above) exhibited much improved performance once again, with accuracy exceeding 90% with centralised and IID federated training. This can be attributed to reduced label ambiguity and clearer physiological separation between classes. The contraction features derived from temporal masks provided a sufficiently strong, discriminative signal for disease classification despite the low dimensionality of the feature space. Importantly, although Non-IID partitioning decreased overall accuracy relative to IID partitioning, the AUC remained stable between the two conditions, indicating that the relative ranking of test participants was uniformly preserved across heterogeneous clients. This provides strong evidence that federated learning may be a very effective platform for screening situations involving distinct, clinically separable patient cohorts.

Third, the privacy perturbation analysis showed that adding a moderate amount of Gaussian noise to client model updates did not affect performance. The accuracy and AUC were relatively constant across various levels of noise. These results are consistent with previous studies of privacy-enhanced federated learning in medical imaging, which frequently show a lower performance drop at moderate perturbation levels [16, 17]. The high strength in this research can be explained by the high signal-to-noise ratio of the contraction features of temporal organisation and the stability of the logistic regression decision boundary. Although the implementation does not obtain formal guarantees of differential privacy, it provides empirical evidence that moderate update perturbations can be added without affecting diagnostic utility.

The results provide clinical evidence that federated learning increases collaboration among institutions in echo analysis while maintaining patient confidentiality by not requiring data to be shared centrally. In health care settings with specific laws governing patients' privacy, the results on how federated performance was affected by whether data was shared between sites indicate a need for differential weighting strategies for aggregating site-specific performance results in future studies. The excellent performance of federated techniques applied to extreme EF groups may also provide opportunities to develop automated screening workflows or triage systems.

In a real-world deployment, the proposed federated learning framework could be integrated into existing hospital imaging workflows by allowing each institution to train the model locally on echocardiographic data stored within its own infrastructure. Only model updates would be shared with a central server for aggregation, ensuring compliance with privacy regulations and institutional data-governance policies. The resulting global model could support clinicians by providing automated EF classification during routine echocardiographic assessment, while continuously improving through collaboration among participating hospitals.

Although the findings of this study offer new potential for implementation, numerous limitations should be noted. First, all data used in this study are from a single published dataset. Second, all participants were simulated rather than drawn from actual sites or physical institutions. Third, although differential privacy was approximated by applying Gaussian noise, formal differential privacy accounting was never applied. Fourth, additional research should conduct an analysis of multi-centric datasets, explore the implementation of adaptive federated optimisation strategies for Non-IID robustness, and implement formal privacy guarantees.

Another limitation relates to potential inter-observer variability in echocardiographic assessment. Although the CAMUS dataset provides expert-annotated reference measurements and segmentation masks, EF estimation in routine clinical practice may vary among clinicians due to differences in image interpretation and measurement techniques. Such variability may influence label consistency and model generalizability when deployed across institutions with different clinical practices.

In addition, the present study primarily focused on classification performance and privacy robustness. Computational cost, communication overhead, and training efficiency across federated clients were not explicitly analysed. Although the proposed framework employs lightweight feature-based representations that are expected to reduce computational and communication requirements compared to image-based federated learning approaches, a detailed evaluation of resource consumption and scalability remains an important direction for future work, particularly in large-scale multi-institutional deployments.

Furthermore, the privacy evaluation in this study was limited to the Gaussian perturbation of client updates. Advanced privacy threats, such as membership inference, model inversion, and gradient reconstruction attacks, were not explicitly investigated. Evaluating the framework under these attack models remains an important direction for future research.

Finally, the present study focused on two clinically relevant binary EF classification tasks and did not investigate a finer-grained multiclass stratification of cardiac function. While the selected thresholds enabled a systematic evaluation of federated learning under varying levels of class separability, future work may explore multiclass EF categorisation and broader cardiac function assessment to support more detailed clinical decision-making.

Overall, the findings indicate that federated learning offers a plausible architecture for implementing privacy-aware echocardiographic EF classification, in which data heterogeneity and cohort separability have a significant impact on performance metrics.

VI. CONCLUSION

In this research, we evaluated a federated learning framework that preserves privacy for classifying echocardiographic Ejection Fraction (EF) under realistic clinical heterogeneity. The CAMUS dataset provided the

basis for a systematic comparison of centralised and federated training under both IID and Non-IID client distributions, and for an assessment of the robustness of these models to perturbations in the form of noise from client updates.

In the clinically realistic EF threshold (EF below 50% versus EF 50% and above), federated learning demonstrated performance equal to centralised training in IID settings, whereas Non-IID distributions showed an appreciable drop in discrimination performance. These results underscore the fragility of federated models to client heterogeneity, an important factor in practice.

Compared to patients with a Left Ventricular Ejection Fraction (LVEF) of 35% or below (low) and LVEF of 60% or greater (high), Federated Learning has achieved very high diagnostic performance (greater than 90% accuracy) for both clinically discrete LVEF cohorts when distributed data were classified as IID. Although diagnostic performance decreased modestly under Non-IID partitioning, adding Gaussian noise to client updates showed considerable robustness to perturbations across all tested sizes, demonstrating that low-dimensional temporally contracted features can maintain discriminative power even in the presence of moderate noise in client updates.

Altogether, the findings suggest that federated learning is a viable and privacy-conscious approach to collaborative echocardiographic EF classification across institutions. But data heterogeneity remains the primary factor in federated performance. Future work will focus on heterogeneity-aware aggregation methods, formal combination with differential privacy, and testing on multi-centre clinical data to enhance the validity of the application in the real world.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

A.J. (Abhigna Jalli) conceived the study and designed the experiments. A.J. and K.S. (Kodukula Subrahmanyam) conducted the experiments and collected the data. K.S. and P.Y. (Pachipala Yellamma) performed the data analysis and interpreted the results. P.Y. contributed to the development of the methodology and provided supervision. A.J. drafted the manuscript, and K.S. and P.Y. critically reviewed and revised it. All authors read and approved the final manuscript.

ACKNOWLEDGMENT

The authors would like to thank the Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Guntur, India, for providing the facilities and research environment that supported this work.

REFERENCES

- [1] S. T. Shah *et al.*, "Federated learning in public health: A systematic review of decentralized, equitable, and secure disease prevention approaches," *Healthcare (Basel)*, vol. 30, no. 13, 2760, 2025.
- [2] H. Guan *et al.*, "Federated learning for medical image analysis: A survey," *Pattern Recognition*, vol. 151, 110424, 2024.
- [3] F. R. da Silva *et al.*, "Federated learning in medical image analysis: A systematic survey," *Electronics*, vol. 13, no. 1, 47, 2024.
- [4] F. S. Alhafiz and A. A. Basuhail, "Non-IID medical imaging data on COVID-19 in the federated learning framework: Impact and directions," *COVID*, vol. 4, no. 12, pp. 1985–2016, 2024.
- [5] N. Hernandez-Cruz *et al.*, "Review of federated learning and machine learning-based methods for medical image analysis," *Big Data Cogn. Comput.*, vol. 8, no. 9, 99, 2024.
- [6] S. Leclerc, E. Smistad, J. Pedrosa *et al.*, "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.
- [7] S. Wassan *et al.*, "Federated learning and differential privacy: Machine learning and deep learning for biomedical image data classification," *Digit Health*, vol. 11, 2025.
- [8] B. McMahan, E. Moore, D. Ramage *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th International Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.
- [9] G. Choi *et al.*, "Survey of medical applications of federated learning," *Healthcare Informatics Research*, vol. 30, no. 1, pp. 3–15, 2024.
- [10] M. J. Sheller *et al.*, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, 12598, 2020.
- [11] M. Babar *et al.*, "Investigating the impact of data heterogeneity on the performance of federated learning algorithms in medical imaging," *PLOS ONE*, vol. 19, no. 5, e0302539, 2024.
- [12] T. Li *et al.*, "Federated optimization in heterogeneous networks," in *Proc. Machine Learning and Systems 2 (MLSys 2020)*, 2020.
- [13] A. Z. Tan *et al.*, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 9587–9603, 2023.
- [14] M. Adnan *et al.*, "Federated learning and differential privacy for medical image analysis," *Scientific Reports*, vol. 12, 1953, 2022.
- [15] L. Zheng *et al.*, "Sensitivity-aware differential privacy for federated medical imaging," *Sensors*, vol. 25, no. 9, 2847, 2025.
- [16] K. Bonawitz *et al.*, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. on Computer and Communications Security*, 2017, pp. 1175–1191.
- [17] A. Ghorbani *et al.*, "Deep learning interpretation of echocardiograms," *Nature Digital Medicine*, vol. 3, 10, 2020.
- [18] D. Ouyang *et al.*, "Video-based AI for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, pp. 252–256, 2020.
- [19] E. S. Lau *et al.*, "Deep Learning-Enabled assessment of left heart structure and function predicts cardiovascular outcomes," *Journal of the American College of Cardiology*, vol. 82, no. 20, pp. 1936–1948, 2023.
- [20] T. Barry, "The role of artificial intelligence in echocardiography," *J. Imaging*, vol. 9, no. 2, 50, 2023.
- [21] Y. Jiang *et al.*, "From pretraining to privacy: Federated ultrasound foundation model with self-supervised learning," *NPJ Digital Medicine*, vol. 8, 714, 2025.
- [22] S. P. Karimireddy *et al.*, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. 37th International Conf. on Machine Learning (ICML)*, 2020, pp. 5132–5143.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).