ICIAET
2025

Published online Dec, 2025 (www.iciaet.org)

Innovations & Trends in Advanced Engineering Technologies

Vol. 1 Iss 01 Page No 976-984

Federated Multimodal Contrastive-Learning Transformer with Calibrated Risk Scoring for CHD

B. Ragu[1]
Department of Computer Science, School of Computing Sciences
Vels Institute of Science, Technology and Advanced Studies (VISTAS)
Chennai, India
Email: ragu_spot@yahoo.com

S. Perumal[2]
Department of Computer Science, School of Computing Sciences
Vels Institute of Science, Technology and Advanced Studies (VISTAS)
Chennai, India
perumal.scs@velsuniv.ac.in

Received: 22nd Oct 2025 / Accepted : 29th Oct 2025 ,
ⓒ ICIAET 2025

Abstract*—Coronary Heart Disease (CHD) remains a leading cause of morbidity and mortality worldwide, necessitating the earlier, precision and privacy preserving prediction systems in assisting the clinical decision making process. The recent advancement of technology and the evolutions of Deep Learning (DL) algorithms have demonstrated promising results in the diagnosis process, leveraging the multimodal clinical data such as Electrocardiogram (ECG), medical imaging and Electronic Health Records (ECR). Despite of advancements in state of art diagnosing methodologies, the precision experiences a setback in data privacy concern, institutional data silos, heterogeneous data distribution, poor model calibration and cross modal representational alignment due to the regulatory constraints, domain drifts and inconsistent data quality. To overcome these challenges, this manuscript proposes a Federated Multimodal Contrastive Learning Transformer with Calibrated Risk Scoring (FMCL-CRS) for a robust and privacy preserving CHD risk prediction. This framework integrates the transformer based multimodal encoder with contrastive learning method to align with the input heterogeneous feature representations across multi modalities. The framework is trained and tested using UCI heart disease dataset, PhysioNet ECG dataset and MIMIC-IV clinical records to analyze the performance in terms of accuracy, precision, recall, F1 score, ROC curve, brier score, Expected calibration Error (ECE) and Communication efficiency. The experimental analysis is anticipated to achieve superior performance and enhanced robustness for heterogeneous data and is suitable for the real world CHD risk assessment applications.

Keywords—Coronary Heart Disease, Federated Learning, Multimodal learning, Contrastive learning, Transformer Networks, risk calibration, Privacy preserving AI.

I. INTRODUCTION

Coronary Heart Disease (CHD) [1] is a major subset of Cardio Vascular Diseases (CVD)[2], poses a global health threat due to its chronic progression, enhanced mortality rate and long term economic impact on the healthcare systems. The CHD is caused due to the blockage or narrowing of coronary arteries, reducing the blood flows to the myocardium, results in myocardial infarction, heart failure and sudden cardiac death. The earlier prediction of CHD minimizes the risk of heart failure through the surgical procedures, and pharmacological treatments. With the advent of rapid digitization of the healthcare applications, the patient data becomes heterogeneous composed of laboratory test results, ECG [3], imaging studies and EHRs. The handling and processing of multimodal data by existing Deep Learning (DL) models [4], offers an unprecedented opportunities for accurate assessment of the CHD risk factors. As per the report of World Health Organization (WHO) [5], approximately 17.9 million patients had recorded to be dead annually, presenting 32% of global death rate.

Considering the effectiveness of the disease, state of art technologies like Deep Neural Networks (DNN) [6], Convolutional Neural Networks (CNN) [7], Recurrent Neural Networks (RNN) [8], Transformer based

architectures [9], and Ensemble learning models [10] have been developed to present a high precision performance in the CHD risk assessment. Despite of existence of novel technologies, experiences of setback in achieving enhanced performance due to the data privacy and regulatory constraints, making the centralized data aggregation to be harder. Furthermore, the diagnosis of CHD relies on the multimodal and heterogeneous data sources [11], exhibits diverse temporal, spatial and statistical characteristics, and complicates the effective fusion of features and representation learning. In addition, the patient data collected from various sources are non-identical, leading to the reduction in model generalization and performance degradation. The existing state of the art models suffers from poor probability calibration and unreliable risk estimations making it unsuitable for real time CHD risk assessments.

To overcome these challenges, this research work proposed a Federated Multimodal Contrastive Learning Transformer with Calibrated Risk Scoring (FMCL-CRS) for the earlier prediction of CHD. This framework integrates the transformer based multimodal encoders with the contrastive learning objectives The Federated Learning (FL) model ensures that the patient data remains local, addressing the privacy and regulatory constraints. In addition, the calibrated risk scoring mechanism in incorporated to enhance the reliability and interpretability of predicted probabilities, jointly addresses the heterogeneity, representation learning, privacy and calibration aspects, bridging the gap between the high performing AI model and real world CHD assessment applications. The major real world contributions of this research work are as follows.

- The proposed model offers a privacy preserving federated multimodal transformer model for CHD risk predictions, eliminating the need for centralized data sharing.
- The contrastive learning based multimodal alignment strategy handles the heterogeneous clinical data distributions across the decentralized healthcare applications.
- The calibrated risk scoring module enhances the probability reliability, assisting the clinical decision making process.
- This scalable and deployment ready architecture is suitable for real time healthcare environments and future clinical integrations.

The organization of this research manuscript is structured with a brief introduction in section I, and a literature analysis in section II for the identification of research challenges, followed by the framing of research objectives. The detailed illustration of the proposed framework is presented in Section III with a performance analysis in Section IV. The research manuscript is concluded, highlighting its contributions in Section V.

II. Literature Review

This section presents a literature review, critically examines the recent advancements in the CHD predictions by the state of the art AI and DL models. The literature analysis highlights the methodological trends, performance achievements and unsolved limitations, motivating the need for the development of this efficient framework.

B. Zhao et al. (2025) had proposed a Deep Learning based CHD diagnosing model called Multi-Stream Coronary Analysis Network (MACAN) [12]. This model applied the dual stream model integrating the localization and segmentation tasks. This model achieved an accuracy of 61.54% with the training process using Unet++. M. Sajid et al. (2024) had employed ten machine learning classifiers [13] for determining the number of affected cardiac vessels. The authors proposed a rank assessing AI-CADR model and had achieved 82.58% of accuracy.

A.J. Partinen et al. (2024) had applied five Neural Network architectures [14] namely, DenseNet 201, MobileNet V2, NasNet Mobile, ResNet 18 and ResNet 50 for the classification of CHD disease and had measured an AUC of 92.7%. X.Wang et al. (2025) had proposed a Multi Modality Attention Network (MMAN) [15] leveraging the data from clinical and MPI for CAD diagnosis. This model involves clinical data guided attention (CDGA) module integrating the clinical data enhancing the diagnostic performance of diagnosis the CAD.

H.Zhang et al. (2025) had introduced an onboard TYKD model [16] for detecting the CAD exhibits low computational complexity. The model is trained using 41 CAD patients and had achieved an accuracy of 85.2% and 88.6% of specificity. A. Phoemsuk et al. (2025) had proposed a one dimensional Convolutional Neural Networks (1D-CNN) [17] for the diagnosis of Coronary Artery Disease. This model enhanced the robustness with the training dataset of MIMIC III, exhibiting an enhanced classification performance.

X.Zhang et al. (2024) had designed a framework composed of Anatomical Dependency Encoding (ADE) module [18] and Hierarchical Topology Learning (HTL) for segmenting the CAD. This model adopted the bottom up attention interaction process and is trained using public and in-house datasets, achieving the superior performance than the state of art models. D.Cenitta et al. (2025) had introduced a hybrid model composed of Long Short term Memory (LSTM) and Hybrid Residual Enhanced LSTM (HRAE-LSTM) [19] for the prediction of CHD disease. This model is trained using realistic 303 datasets to achieve an accuracy of 97.7%.

F.Chen et al. (2025) had proposed a ResNet 18 encoder MCA model for the earlier detection of CAD, addressing the limitations of the high sensitivity enabling the unprecedented precision in CAD prediction process. This model extracted the feature of CAD using the Mutual Cross Attention mechanism achieving 98.81% of accuracy. B.Zhao et al. (2025) had applied Conventional Convolutional Neural Networks (CNN) [21] for the successful diagnosis of the CHD disease, addressing the challenges of the existing segmentation process. This model was trained using CorArtTS2020 dataset and had achieved a Dice score of 0.81 and 0.65 of Intersection over Union (IoU). From this literature review, the following challenges have been observed in the state of the art CHD diagnosis.

- The existing CHD diagnosis frameworks relies on centralized data aggregation of CT angiography like ECG, PCG or imaging modalities [12], [15], [16] [20]. This creates privacy concerns experiencing the regulatory constraints and multi-institutional scalability concerns.
- The multimodal approaches [15], [20] typically employs late fusion or attention based fusion, failing to enforce the semantic consistency among the heterogeneous modalities.
- The existing CNN, LSTM and hybrid models [14], [17], [19] were constrained in capturing the long range dependencies and cross modality interactions, when integrating the sequential physiological signals with static imaging and clinical metadata.
- Most CHD models focus on classification accuracy without ensuring risk calibration, producing confidence scores that are unreliable for clinical decision-making [13],[19][13], [19][13],[19]. This limits trustworthiness in real-world prognosis and stratification.
- Existing models were trained on single-source datasets [12],[18],[21] often degrade when deployed across different scanners, sensors, patient demographics, or institutions**,** indicating susceptibility to domain shift.

To overcome these challenges, this research work was proposed with the following research objectives.

- The proposed work introduces a federated learning architecture that enables decentralized model training across multiple healthcare institutions without sharing raw patient data, ensuring data privacy, regulatory compliance, and improved cross-domain generalization.
- To overcome inadequate cross-modal representation alignment concern, a multimodal contrastive learning strategy is employed to explicitly align latent representations from heterogeneous data sources (CT images, ECG/PCG signals, and clinical features), improving semantic coherence and modality-invariant feature learning.
- To overcome limited modeling of long-range temporal and inter-modal dependencies, the proposed model integrates a transformer-based architecture capable of modeling long-range temporal dependencies and inter-modal attention, enabling effective fusion of spatial, temporal, and clinical information for CHD prediction.
- To overcome uncalibrated and non-interpretable risk scores, the framework incorporates risk calibration techniques (e.g., temperature scaling or probabilistic calibration layers) to produce well-calibrated CHD risk scores**,** enhancing interpretability and clinical reliability.
- To ensure robust and scalable CHD risk prediction across heterogeneous institutions, devices, and patient populations by jointly leveraging federated optimization, multimodal contrastive representation learning, and transformer-based feature fusion.

III. Proposed Methodology

This research work proposes a Federated Multimodal Contrastive Learning transformer with the calibrated risk scoring in the reliable prediction of CHD. The proposed framework is structured with a sequence of significant processes including acquisition of heterogeneous multimodal patient data, followed by preprocessing and feature encoding for ensuring the data privacy. The encoded features were integrated using transformer based fusion architecture capturing the complex inter and intra model dependencies relevant to the CHD risk. The optimization of the model is performed using Federated Learning model, encrypting the local features and securely aggregates the global model without sharing raw patient data. The fused representation is subsequently passed through the risk scoring module for ensuring the reliability and clinical interpretability. Finally, the secure deployment strategies are incorporated for trustworthy decision making process as in Figure 1.

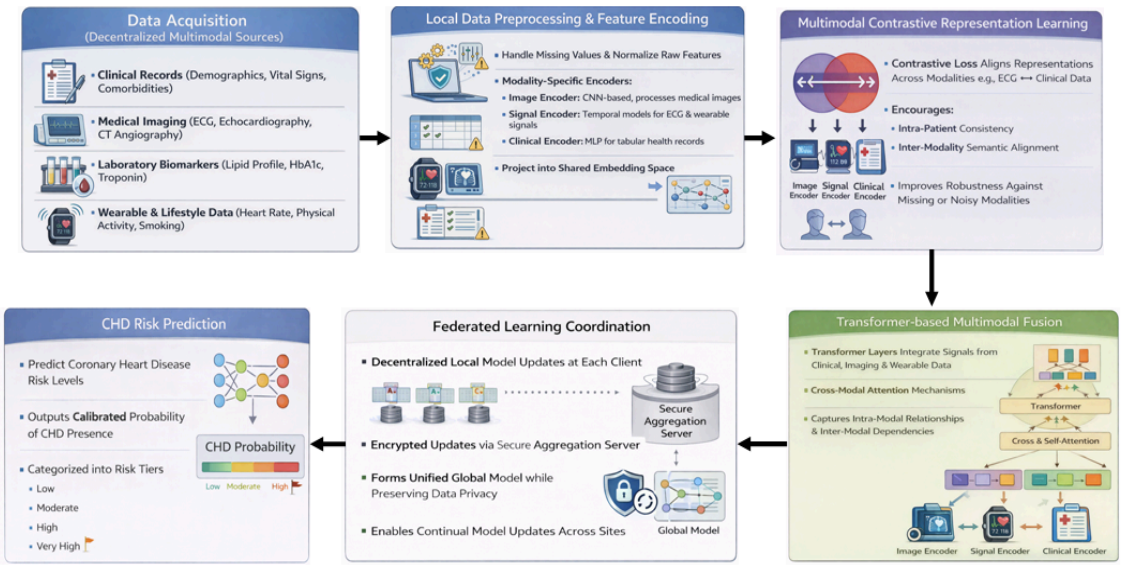

Fig. 1. Proposed architecture of Federated Learning based CHD risk prediction

A. Multi Modal Data Acquisition

Multi modal data acquisition is the foundational process of the proposed framework, wherein the heterogeneous patient specific CHD data is collected under the decentralized setting. The patient data is acquired locally in multiple modalities without transferring raw information to the external servers ensuring the privacy regulations as in equation 1.

$$X_i^{(c)} = \left\{X_{i,\,img}^{(c)},\ X_{i,\,Clin}^{(c)}, X_{i,\,sig}^{(c)}, X_{i,\,lab}^{(c)}\right\};\ c = \{1, 2, \ldots C\};\ c \in C \quad (1)$$

Where, $X_{i,\,img}^{(c)} \in R^{H\times W\times K}$ representing the medical image like ECG and angiographic images, $X_{i,\,Clin}^{(c)} \in R^{clin}$ defining the structural clinical attributes like vital signs, demographic factors and comorbidities. $X_{i,\,sig}^{(c)} \in R^{T\times S}$ represents the time series physiological signals from wearable and IoT devices, while $X_{i,\,lab}^{(c)} \in R^{lab}$ indicating the laboratory biomarkers and cardiac enzymes. The measurement scales, sampling frequencies collects the modalities with empirical distributions of modality 'm' at client c. In practical, this distribution is non-identical and non-identifying satisfies the condition defined in equation 2.

$$\mathfrak{D}_m^{c_1} \neq \mathfrak{D}_m^{c_2};\ \forall c_1 \neq \forall c_2 \quad (2)$$

The downstream learning of all modalities are temporary and semantically aligned at the patient level and $y_i^{(c)} \in \{0, 1\}$ represents the outcome label of the CHD associated with the patient 'i'. The resulting local dataset generated for client 'c' is defined as in equation 3.

$$S^{(c)} = \left\{\left(X_i^{(c)},\ y_i^{(c)}\right)\right\}_{i=1}^{N_c} \quad (3)$$

Where, N_c is the number of patients at hospital (location) c. The missing modality scenarios are permitted such that the subset $X_i^{(c)}$ is partially observed. The proposed framework ensures the downstream federated multimodal learning can effectively capture the complementary diagnostic cues, which were significant for the enhanced accuracy in CHD risk estimation.

B. Local Data Preprocessing and Feature Encoding

The local data preprocessing is performed independently at each client (hospital) to standardize the heterogeneous multimodal inputs. This ensures the raw patient data to remain confined within the institutional boundaries. The equation 3 represents the locally stored dataset, defining the diversified nature of clinical, image, signal and laboratory modalities. The local preprocessing primarily involves the handling of missing values, normalization, and feature scaling. Let $X_{i,m}^{(c)} \in R^d$ representing the tabular feature vector for modality, $m \in \{Img,\ clin,\ sig,\ lab\}$. The missing entries in the heterogeneous dataset are computed using local statistical estimates, emperical mean or median as in equation 4.

$$\hat{X}_{i,m}^{(c)}(j) = \{X_{i,m}^{(c)};\ if\ observed\ \mu_{i,m}\ if\ Missing \quad (4)$$

Where, $\mu_{i,m}$ represents the mean of the feature 'j' computed over the client 'c'. The features were standardized using z-score normalization as in equation 5.

$$\hat{X}_{i,m}^{(c)}(j) = \frac{\tilde{X}_{i,m}^{(c)} - \mu_{i,m}^{(c)}}{\sigma_{i,m}^{(c)}} \quad (5)$$

With $\mu_{i,m}^{(c)}$ and $\sigma_{i,m}^{(c)}$ representing the modality specific mean and standard deviation vectors. The denoising filter is employed to suppress the acquisition artifacts, preserving the relevant features. The physiological time series signals necessitates temporal alignment and noise suppression due to the sampling rate and motion artifacts. The band pass filtering operation function is applied for the removal of baseline wander and high frequency noise as in equation 5.

$$\hat{X}_{i,sig}^{(c)} = F\left\{X_{i,sig}^{(c)}\right\} \quad (5)$$

The filtering process is followed by the temporal normalization for ensuring the scale invariance as defined in equation 6.

$$\hat{x}_t = \frac{x_t - \mu_{sig}^{(c)}}{\sigma_{sig}^{(c)}} \quad (6)$$

The preprocessing stage ensures that the transformed input $\hat{X}_{i,sig}^{(c)}$ lies in numerically stable domain, adhering to the consistent statistical properties. The locally preprocessed data is formed as in equation 7.

$$\hat{S}_i^{(c)} = \left\{\left(\hat{X}_{i,\,img}^{(c)},\ \hat{X}_{i,\,Clin}^{(c)}, \hat{X}_{i,\,sig}^{(c)}, \hat{X}_{i,\,lab}^{(c)}\right),\ y_i^{(c)}\right\}_{i=1}^{N_c} \quad (7)$$

The resulting dataset ensures the modality specific noise, scale variations and missing data effects were addressed to enable federated multimodal learning in the subsequent stages. The multi modality is mapped with the latent feature space through the modality specific encoders as in equation 8.

$$z_{i,m}^{(c)} = f_m\left(\hat{S}_i^{(c)};\ \theta_m^{(c)}\right) \quad (8)$$

Where $f_m\left(\hat{S}_i^{(c)};\ \theta_m^{(c)}\right)$ represents the local feature encoder for the modality m with the parameter $\theta_m^{(c)}$. The $z_{i,m}^{(c)}$ represents the fixed dimensional embedding process, applying the efforts of Convolution Neural Networks for imaging data and Recurrent Neural Networks for the physiological signals. The multimodal fusion and contrastive alignment is facilitated in the subsequent stages and is projected into shared latent space using linear projection layers as in equation 9.

$$\tilde{z}_{i,m}^{(c)} = W_m z_{i,m}^{(c)} + b_m = W_m f_m\left(\hat{S}_i^{(c)};\ \theta_m^{(c)}\right) + b_m \quad (9)$$

Where, the W_m is the learning parameter ensuring the dimensional consistency across modalities. The resulting set

of encoded representation for the patient 'i' at client c is expressed as in equation 10.

$$z_{i,m}^{(c)} = \left\{ \sim_{z_{i,img}}^{(c)}; \sim_{z_{i,clin}}^{(c)}; \sim_{z_{i,sig}}^{(c)}; \sim_{z_{i,lab}}^{(c)} \right\} \quad (10)$$

By jointly performing the local preprocessing and feature encoding process, this framework effectively reduces the data heterogeneity, ensures high level representations in federated multimodal learning.

C. Multimodal Contrastive Representation Learning

The multimodal contrastive representation learning process is applied to align the heterogeneous modality specific embeddings into a semantically consistent latent space, enables the effective fusion and robust CHD risk modelling. The major objective of this phase is to maximize the agreement among the different modalities of the same patient 'i', enhancing the discrimination across various patients. Each embedding is normalized to lie on the unit hypersphere to facilitate the stable similarity computation as in equation 11.

$$\bar{z}_{i,m}^{(c)} = \frac{z_{i,m}^{(c)}}{\| W_m f_m \left(\hat{S}_i^{(c)}; \theta_m^{(c)} \right) + b_m \|} \quad (11)$$

The contrastive learning operated by constructing the positive and negative sample pairs. For any patient 'i', the embeddings from the different modalities (m,n) with m≠n, representing the same underlying physiological state. The similarity among the two normalized embeddings is measured using cosine similarity as in equation 12.

$$sim\left(\bar{z}_i, \bar{z}_j\right) = \bar{z}_i^T, \bar{z}_j \quad (12)$$

The multimodal representations were aligned based on the contrastive losses, considering the modality 'm' as the anchor and modality 'n' as the positive counter. The contrastive loss of the single anchor function is defined in equation 13.

$$L_i^{(c)}(m,n) = -\ log \frac{exp\left(\frac{sim\left(\bar{z}_{i,m}, \bar{z}_{j,n}\right)}{\tau}\right)}{\sum_{i\in\alpha} \sum_{j\in\beta} exp\left(\frac{sim\left(\bar{z}_{i,m}, \bar{z}_{j,n}\right)}{\tau}\right)} \quad (13)$$

Where, τ>0 is the temperature parameter controlling the sharpness of the similarity distribution, while α and β denotes the set of samples within the local data of client c. The overall multimodal contrastive objective is computed by averaging across all modality pairs and batch samples as in equation 14.

$$L_{MCL}^{(c)} = \frac{1}{|\alpha||\beta|} \sum_{i\in\alpha} \sum_{j\in\beta} \frac{1}{|\rho|} L_i^{(c)}(m,n) = -\ log \frac{exp\left(\frac{sim\left(\bar{z}_{i,m}, \bar{z}_{j,n}\right)}{\tau}\right)}{\sum_{i\in\alpha} \sum_{j\in\beta} exp\left(\frac{sim\left(\bar{z}_{i,m}, \bar{z}_{j,n}\right)}{\tau}\right)} \quad (14)$$

Where, $\rho = \{(m,n) | m,n \in M,\ m \neq n\}$ representing the set of cross modal positive pairs. This alignment reduces the representation gaps created by the heterogeneity modality and acquisition variability across multiple locations and patients. The resulting contrastively aligned embeddings $\bar{z}_{i,m}^{(c)}$ exhibits enhanced cross modality consistency and discriminability making the model suitable for the subsequent transformer based multimodal fusion process.

D. Transformer based Multimodal Fusion

The transformer based multimodal fusion is applied to integrate the contrastively aligned modality specific representations as an unified context aware patient representation for the CHD risk assessment. Each modality embedding is treated as a token and augmented with the modality type embedding for retaining the identity of the modality. The resulting token sequence to the transformer is defined in equation 15.

$$h_{i,m}^{(0)} = \bar{z}_{i,m}^{(c)} + e_m \quad (15)$$

The transformer fusion module is composed of stacked multi-head self attention and feed forward layers, modelling both the inter and intra model dependencies. The multi head self computations is presented in equation 16.

$$Q^{(l)} = H^{(l-1)} W_Q^{(l)}; K^{(l)} = H^{(l-1)} W_K^{(l)}; V^{(l)} = H^{(l-1)} W_V^{(l)} \quad (16)$$

Where, $H^{(l-1)} W_Q^{(l)}$ denotes the token matrix from the previous layer. The attention mechanism computes the pairwise interactions among the modalities as in equation 17.

$$Attn(Q,\ K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (17)$$

To enhance the representational capacity, multiple attention heads were employed, and the outputs were concatenated and transformed linearly. The attention output is passed through the position wise feed forward network as in equation 18.

$$H^{(l)} = FFN\left(Multihead\ Attn\ (H^{(l-1)})\right) \quad (18)$$

Where the FFN introduces the non-linearity and further feature abstraction. After L transformation layers, a fused patient representation is obtained by aggregating the final token embeddings as in equation 19.

$$z_{i,fused}^{(c)} = \frac{1}{M} \sum_{m\in M} h_{i,m}^{(L)} \quad (19)$$

The fused representation encapsulates the complementary information from all modalities, preserving the contextual relationships identified through the self attention process. From a theoretical standpoint, transformer-based fusion provides a flexible and permutation-invariant mechanism for multimodal integration, allowing the model to adaptively emphasize clinically informative modalities and suppress redundant or less relevant ones. Consequently, the resulting fused embedding serves as a robust input for downstream CHD risk prediction and federated optimization in decentralized healthcare environments.

E. Federated Learning Coordination

The Federated Learning coordination enables the collaborative optimization process for the proposed multimodal contrastive transformer model across the distributed healthcare locations. The objective of the federated learning model is to minimize the global loss function defined as in equation 20.

$$L_{global}(\theta) = \sum_{c=1}^{C} \frac{N_c}{N} L^{(c)}(\theta) \quad (20)$$

Where, θ represents the global model parameters, N_c is the number of samples at client c. The loss function in the proposed work, integrates the multimodal contrastive loss and supervised CHD prediction loss. For each communication round 't', the central server broadcasts the present global model parameters to the subset of the clients. Each selected client performs the local training and encoded data by minimizing the local loss over multiple local epochs E as in equation 21.

$$\theta^{(t+1)} = \sum_{c=1}^{C} \frac{N_c}{N} \theta_c^{(t+1)} \quad (21)$$

This aggregation scheme ensures that the clients with larger datasets contribute proportionately more to the global update and stabilizing the optimization under data imbalance. The coordinated optimization process thus enables the proposed framework to leverage diverse multimodal knowledge while ensuring data confidentiality, scalability, and regulatory compliance, making it well suited for real-world CHD risk prediction in decentralized healthcare systems.

F. CHD Risk Prediction

The CHD risk prediction stage employs the fused multimodal representations obtained from the transformer based fusion module. This estimates the probability of the CHD for each patient. The prediction head is implemented as a parametric mapping function $g(.,\varphi)$realizing the multilayer perceptron transforming the fused representation into a scalar risk score using equation 22.

$$s_i^{(c)} = g\left(z_{i,fused}^{(c)}; \emptyset\right) \quad (22)$$

To obtain a probabilistic interpretation suitable for the clinical decision making process, the risk score is passed through the sigmoid function as in equation 23.

$$p_i^{(c)} = \sigma\left(s_i^{(c)}\right) = \frac{1}{1+exp(-s_i^{(c)})} \quad (23)$$

The CHD risk prediction module thus provides a crucial link between learned multimodal representations and clinically meaningful outcomes, forming the basis for subsequent risk calibration and decision-support mechanisms in the proposed federated framework. The algorithm for the proposed work is presented in Table I.

TABLE I. ALGORITHM FOR PROPOSED FEDERATED LEARNING BASED CHD RISK PREDICTION

Algorithm: Federated Learning based CHD Risk prediction

Input:
- Number of clients 'c'
- Local dataset: $S^{(c)}$
- Modalities: M={img, clin, sig, lab}
- Local Epochs: E
- Communication rounds: T
- Temperature parameter: τ

Output:
- Global model parameters Θ
- CHD risk probability { } for each patient

Processes:

1: Initialize the input parameters and modalities
2: For each client, c={1,2,3…C}
3: Collect patient specific multimodal data: $X_i^{(c)} = \left\{X_{i,img}^{(c)}, X_{i,Clin}^{(c)}, X_{i,sig}^{(c)}, X_{i,lab}^{(c)}\right\}; c = \{1, 2, ...C\}; c \in C$
4: Ensure non-IID data distribution across clients: $\mathcal{D}_m^{c_1} \neq \mathcal{D}_m^{c_2}; \forall c_1 \neq \forall c_2$
5: Form local dataset: $S^{(c)} = \left\{\left(X_i^{(c)}, y_i^{(c)}\right)\right\}_{i=1}^{N_c}$

//////// Local data Preprocessing and Feature encoding/////////

6: For (i=0, i<m, i++)
7: Handle missing values using local statistics: $\hat{X}_{i,m}^{(c)}(j) = \{X_{i,m}^{(c)}; \; if \; observed \; \mu_{i,m} \; if \; Missing$
8: Apply Z score normalization: $\hat{X}_{i,m}^{(c)}(j) = \frac{\tilde{X}_{i,m}^{(c)} - \mu_{i,m}^{(c)}}{\sigma_{i,m}^{(c)}}$
9: Apply band pass filtering to physiological signals: $\hat{x}_t = \frac{x_t - \mu_{sig}^{(c)}}{\sigma_{sig}^{(c)}}$
10: Construct preprocessed dataset: $\hat{S}_i^{(c)} = \left\{\left(\hat{X}_{i,img}^{(c)}, \hat{X}_{i,Clin}^{(c)}, \hat{X}_{i,sig}^{(c)}, \hat{X}_{i,lab}^{(c)}\right), y_i^{(c)}\right\}_{i=1}^{N_c}$
11: End for

///// Modality Specific Feature Encoding/////

12: For (i=0, i<m, m=M)
13: Encode features using modality specific encoders: $z_{i,m}^{(c)} = f_m\left(\hat{S}_i^{(c)}; \theta_m^{(c)}\right)$
14: Project embeddings into shared latent space: $\tilde{z}_{i,m}^{(c)} = W_m z_{i,m}^{(c)} + b_m = W_m f_m\left(\hat{S}_i^{(c)}; \theta_m^{(c)}\right) + b_m$
15: Collect encoded representations: $z_{i,m}^{(c)} = \left\{\tilde{z}_{i,img}^{(c)}; \tilde{z}_{i,clin}^{(c)}; \tilde{z}_{i,sig}^{(c)}; \tilde{z}_{i,lab}^{(c)}\right\}$

///// Multimodal Constrastive Representation Learning/////

16: Normalize the embeddings into unit hypersphere: $\bar{z}_{i,m}^{(c)} = \frac{z_{i,m}^{(c)}}{\|W_m f_m\left(\hat{S}_i^{(c)}; \theta_m^{(c)}\right) + b_m\|}$
17: Compute cosine similarity: $sim\left(\bar{z}_i, \bar{z}_j\right) = \bar{z}_i^T, \bar{z}_j$
18: Compute contrastive loss: $L_i^{(c)}(m,n) = -\; log \frac{exp\left(\frac{sim\left(\bar{z}_{i,m}, \bar{z}_{j,n}\right)}{\tau}\right)}{\sum_{i \in \alpha} \sum_{j \in \beta} exp\left(\frac{sim\left(\bar{z}_{i,m}, \bar{z}_{j,n}\right)}{\tau}\right)}$
19: Aggregate multimodal constrastive loss: $L_{MCL}^{(c)} = \frac{1}{|\alpha||\beta|} \sum_{i \in \alpha} \sum_{j \in \beta} \frac{1}{|\rho|} L_i^{(c)}(m,n) = -\; log \frac{exp\left(\frac{sim\left(\bar{z}_{i,m}, \bar{z}_{j,n}\right)}{\tau}\right)}{\sum_{i \in \alpha} \sum_{j \in \beta} exp\left(\frac{sim\left(\bar{z}_{i,m}, \bar{z}_{j,n}\right)}{\tau}\right)}$

/////Transformer based multimodal fusion/////

20: Construct modality tokens with embeddings: $h_{i,m}^{(0)} = \bar{z}_{i,m}^{(c)} + e_m$
21: Apply multi-head self attention and FFN layers: $H^{(l)} = FFN\left(Multihead \; Attn\left(H^{(l-1)}\right)\right)$
22: Obtain fused patient representation: $z_{i,fused}^{(c)} = \frac{1}{M} \sum_{m \in M} h_{i,m}^{(L)}$

/////CHD Risk prediction/////

23: For (t=1, t++, t=T)
24: Each client minimize local loss: $L_{global}(\theta) = \sum_{c=1}^{C} \frac{N_c}{N} L^{(c)}(\theta)$
25: Server aggregates updated parameters: $s_i^{(c)} = g\left(z_{i,fused}^{(c)}; \emptyset\right)$
26: **Return** Θ: $\theta^{(t+1)} = \sum_{c=1}^{C} \frac{N_c}{N} \theta_c^{(t+1)}$
27: **Return** CHD probability: $p_i^{(c)} = \sigma\left(s_i^{(c)}\right) = \frac{1}{1+exp(-s_i^{(c)})}$
28: End for
29: End for
30: End processes

The proposed algorithm presents a privacy-preserving federated framework that jointly integrates multimodal contrastive representation learning and transformer-based fusion for robust CHD risk prediction. By aligning heterogeneous clinical, imaging, signal, and laboratory data across decentralized hospitals, the model effectively mitigates data heterogeneity while preserving patient confidentiality. The coordinated federated optimization enables accurate and scalable CHD risk estimation in real-world distributed healthcare environments.

IV. Results and Discussion

This section performs and discusses the experimental findings of the proposed federated multimodal contrastive learning transformer model for CHD risk prediction. The proposed model was trained with UCI heart disease dataset, PhysioNet ECG dataset, and MIMIC IV clinical dataset and is analyzed in terms of accuracy, precision, recall, F1 score, ROC-AUC and is compared with the performance of the state of art models. The proposed work is trained with 70% of data and is tested with the 30% of the data in preprocessed dataset. The tested performance data was presented in Table II.

TABLE II. Performance of proposed model for various datasets

Parameters	UCI [22]	PhysioNet [23]	MIMIC IV [24]
Accuracy	89.4	91.8	93.2
Precision	88.7	92.4	94.1
Recall	90.1	90.9	92.6
F1 Score	89.4	91.6	93.3
AUC-ROC	0.94	0.96	0.97
Brier Score	0.082	0.069	0.058
ECE	0.031	0.024	0.019
Comm. Efficiency	82.5	79.3	76.8

The observed performance trends across the three datasets can be attributed to the combined effect of data richness, modality diversity, and the design of the proposed FMCL-CRS framework. The relatively lower performance on the UCI Heart Disease dataset is mainly due to its limited sample size and restricted set of structured features, which offer fewer complementary cues for multimodal learning. In contrast, the PhysioNet dataset benefits from high-quality and temporally rich ECG signals, allowing the model to better capture cardiac patterns, resulting in improved predictive and calibration performance. The strongest results are obtained on the MIMIC-IV dataset, as it contains large-scale, real-world clinical records with diverse patient profiles and richer contextual information, enabling the transformer-based fusion module to effectively model complex inter-modal relationships. The graphical comparison of performance among various dataset is presented in Figure 2 and 3.

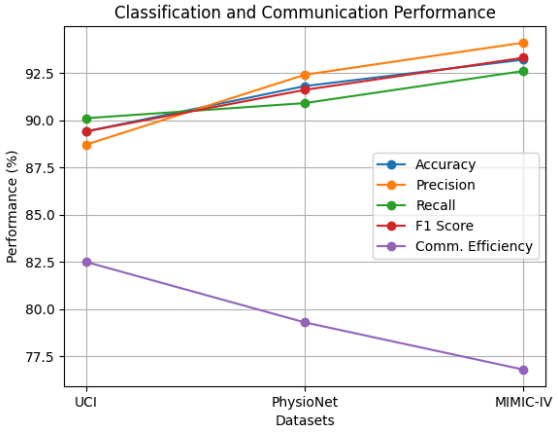

Fig. 2. Performance of proposed work in classification accuracy

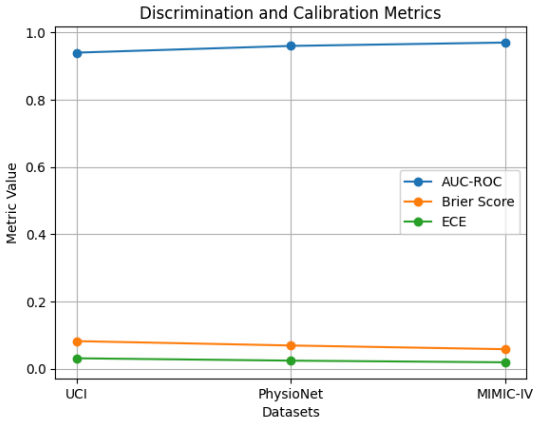

Fig. 3. Performance of proposed work in AUC-ROC

The performance of the proposed framework is compared with the performance of state of art methods like Pre-Orthogonal Adaptive Fourier Decomposition [25], Point CHD [26], 1-D CNN [27], Deep Learning model [28], VoxelMorph [29] and is presented in Table III.

TABLE III. Performance Comparison of proposed model with state of art techniques

Parameters	AFD [25]	Point CHD [26]	1-D CNN [27]	DL [28]	VoxelMorph [29]	Proposed FL Model
Accuracy	82.4	84.1	86.3	88.2	89.1	93.2
Precision	81.6	83.5	85.9	87.8	88.6	94.1
Recall	80.9	82.7	84.8	86.9	87.5	92.6
F1 Score	81.2	83.1	85.3	87.3	88	93.3
AUC-ROC	0.86	0.88	0.9	0.92	0.93	0.97
Brier Score	0.142	0.131	0.118	0.104	0.098	0.058
ECE	0.071	0.064	0.052	0.041	0.037	0.019
Comm. Efficiency	68.2	70.5	72.1	74.8	73.6	82.5

The superior performance of the proposed federated learning model can be attributed to its holistic design, which effectively addresses the limitations of conventional and standalone deep learning approaches. Unlike traditional methods such as AFD, Point CHD, and 1-D CNN that rely on limited feature representations or single-modality inputs, the proposed model leverages rich multimodal information, enabling a more comprehensive understanding of patient

health profiles. The performance is graphically compared in Figure 4 and 5.

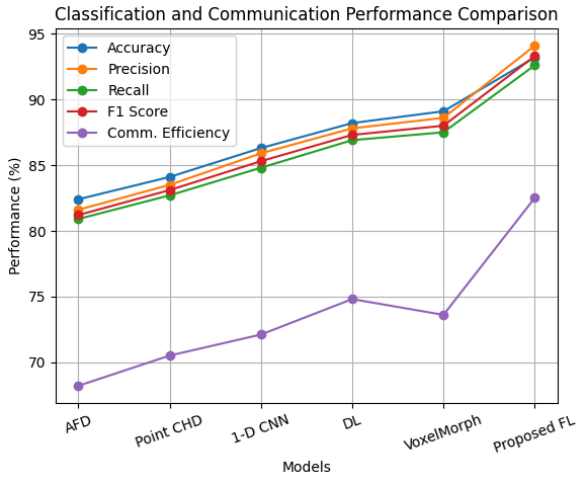

Fig. 4. Performance Comparison of proposed work in classification accuracy

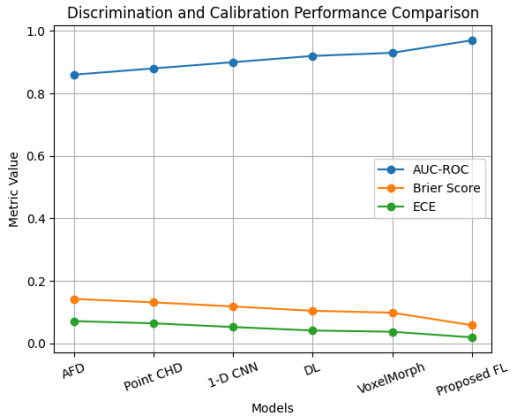

Fig. 5. Performance Comparison of proposed work in AUC-ROC

The integration of contrastive representation learning improves alignment across heterogeneous modalities, reducing representation gaps that often degrade performance in baseline models. Furthermore, the transformer-based fusion mechanism dynamically captures complex inter-modal dependencies, allowing the model to emphasize clinically relevant features while suppressing redundant information. The federated learning strategy further enhances generalization by learning from diverse, distributed datasets without compromising data privacy, while also improving communication efficiency. Together, these factors result in more accurate predictions, better-calibrated risk estimates, and consistent performance gains over existing methods across all evaluated metrics.

V. Conclusion

This study presented a Federated Multimodal Contrastive Learning Transformer with Calibrated Risk Scoring (FMCL-CRS) framework for privacy-preserving and reliable Coronary Heart Disease (CHD) risk prediction in distributed healthcare environments. By jointly leveraging multimodal data sources, including clinical records, ECG signals, and electronic health records, the proposed approach effectively addressed key challenges such as data heterogeneity, institutional data silos, cross-modal misalignment, and unreliable probability estimates. The integration of contrastive representation learning enabled robust alignment of heterogeneous modality-specific features, while the transformer-based fusion module successfully captured complex inter- and intra-modal dependencies relevant to CHD diagnosis. Federated learning coordination ensured data privacy and regulatory compliance while promoting generalization across diverse and non-IID clinical datasets. Experimental evaluations conducted on the UCI Heart Disease, PhysioNet ECG, and MIMIC-IV datasets demonstrated that the proposed FMCL-CRS framework consistently outperformed existing baseline models in terms of classification accuracy, discriminative capability, calibration reliability, and communication efficiency. The improved Brier Score and Expected Calibration Error further highlighted the clinical trustworthiness of the predicted risk scores, making the framework suitable for real-world decision support applications. Future work will focus on extending the framework to handle missing and asynchronous modalities, incorporating personalized federated learning strategies, and integrating explainable AI mechanisms to enhance clinical interpretability. Additionally, real-time deployment with edge devices and validation on large-scale multi-center clinical datasets will further strengthen the practical applicability of the proposed model.

References

[1] X. Nie *et al.*, "CoroJARetinaNet: A Multiscale Attention-Guided Framework for Automated Coronary Plaque Detection in CTA Images," in *IEEE Latin America Transactions*, vol. 23, no. 12, pp. 1152-1162, Dec. 2025, doi: 10.1109/TLA.2025.11231214.

[2] J. Lee *et al.*, "Computational Analysis of Intravascular OCT Images for Future Clinical Support: A Comprehensive Review," in *IEEE Reviews in Biomedical Engineering*, doi: 10.1109/RBME.2025.3530244.

[3] M. A. Quiroz-Juárez, O. Jiménez-Ramírez, R. Vázquez-Medina, E. Ryzhii, M. Ryzhii and J. L. Aragón, "Cardiac Conduction Model for Generating 12 Lead ECG Signals With Realistic Heart Rate Dynamics," in *IEEE Transactions on NanoBioscience*, vol. 17, no. 4, pp. 525-532, Oct. 2018, doi: 10.1109/TNB.2018.2870331.

[4] H. Ali *et al.*, "A Survey on Attacks and Their Countermeasures in Deep Learning: Applications in Deep Neural Networks, Federated, Transfer, and Deep Reinforcement Learning," in *IEEE Access*, vol. 11, pp. 120095-120130, 2023, doi: 10.1109/ACCESS.2023.3326410.

[5] WHO: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[6] F. E. Fernandes and G. G. Yen, "Automatic Searching and Pruning of Deep Neural Networks for Medical Imaging Diagnostic," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 12, pp. 5664-5674, Dec. 2021, doi: 10.1109/TNNLS.2020.3027308.

[7] S. Kannan, P. D, D. K. S and S. S, "A Deep Learning-Based Convolution Neural Networks to Forecast Wind Energy," *2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC)*, Mysore, India, 2023, pp. 1-6, doi: 10.1109/ICRTEC56977.2023.10111917.

[8] Y. Zhang, C. Mitelut, D. J. Arpin, D. Vaillancourt, T. Murphy and S. Saxena, "Behavioral Classification of Sequential Neural Activity Using Time Varying Recurrent Neural Networks," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 33, pp. 2638-2649, 2025, doi: 10.1109/TNSRE.2025.3586175.

[9] M. Zhang *et al.*, "Multiscale Spatial-Channel Transformer Architecture Search for Remote Sensing Image Change Detection," in *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1-5, 2024, Art no. 8000605, doi: 10.1109/LGRS.2023.3347765.

[10] S. Abimannan, E. -S. M. El-Alfy, Y. -S. Chang, S. Hussain, S. Shukla and D. Satheesh, "Ensemble Multifeatured Deep Learning Models and Applications: A Survey," in *IEEE Access*, vol. 11, pp. 107194-107217, 2023, doi: 10.1109/ACCESS.2023.3320042.

[11] L. Sánchez *et al*., "Data Enrichment Toolchain: A Data Linking and Enrichment Platform for Heterogeneous Data," in *IEEE Access*, vol. 11, pp. 103079-103091, 2023, doi: 10.1109/ACCESS.2023.3317705.

[12] B. Zhao, J. Peng, C. Chen, Y. Fan, K. Zhang and Y. Zhang, "Deep Learning-Based Segmentation and Localization in CT Angiography for Coronary Heart Disease Diagnosis," in IEEE Access, vol. 13, pp. 57615-57628, 2025, doi: 10.1109/ACCESS.2025.3555991.

[13] M. Sajid et al., "AI-CADR: Artificial Intelligence Based Risk Stratification of Coronary Artery Disease Using Novel Non-Invasive Biomarkers," in IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 12, pp. 7543-7552, Dec. 2024, doi: 10.1109/JBHI.2024.3453911.

[14] A. Jiménez-Partinen, K. Thurnhofer-Hemsi, J. Rodríguez-Capitán, A. I. Molina-Ramos and E. J. Palomo, "Coronary Artery Disease Classification With Different Lesion Degree Ranges Based on Deep Learning," in IEEE Access, vol. 12, pp. 69229-69239, 2024, doi: 10.1109/ACCESS.2024.3401465.

[15] X. Wang et al., "A Multi-Modality Attention Network for Coronary Artery Disease Evaluation From Routine Myocardial Perfusion Imaging and Clinical Data," in IEEE Journal of Biomedical and Health Informatics, vol. 29, no. 5, pp. 3272-3281, May 2025, doi: 10.1109/JBHI.2024.3523476.

[16] A. Phoemsuk and V. Abolghasemi, "Enhanced Coronary Artery Disease Classification Through Feature Engineering and One-Dimensional Convolutional Neural Network," in IEEE Access, vol. 13, pp. 114306-114317, 2025, doi: 10.1109/ACCESS.2025.3584735.

[17] X. Zhang et al., "An Anatomy- and Topology-Preserving Framework for Coronary Artery Segmentation," in IEEE Transactions on Medical Imaging, vol. 43, no. 2, pp. 723-733, Feb. 2024, doi: 10.1109/TMI.2023.3319720.

[18] D. Cenitta, R. V. Arjunan, G. Paramasivam, N. Arul, A. Palkar and K. Chadaga, "Ischemic Heart Disease Prognosis: A Hybrid Residual Attention-Enhanced LSTM Model," in IEEE Access, vol. 13, pp. 4281-4289, 2025, doi: 10.1109/ACCESS.2024.3524604.

[19] H. Zhang et al., "An AI-Assisted All-in-One Integrated Coronary Artery Disease Diagnosis System Using a Portable Heart Sound Sensor With an On-Board Executable Lightweight Model," in IEEE Transactions on Mobile Computing, vol. 24, no. 8, pp. 7252-7266, Aug. 2025, doi: 10.1109/TMC.2025.3547842.

[20] F. Chen et al., "REM-Net: MEMS-Based Synchronized PCG-ECG Analysis Framework for High-Precision CAD Diagnosis," in IEEE Transactions on Instrumentation and Measurement, vol. 74, pp. 1-12, 2025, Art no. 4013912, doi: 10.1109/TIM.2025.3588969.

[21] B. Zhao, J. Peng, C. Chen, Y. Fan, K. Zhang and Y. Zhang, "Diagnosis of Coronary Heart Disease Through Deep Learning-Based Segmentation and Localization in Computed Tomography Angiography," in IEEE Access, vol. 13, pp. 10177-10193, 2025, doi: 10.1109/ACCESS.2025.3528638.

[22] UNET: Dua, D., & Graff, C. (2019). *UCI machine learning repository* [Heart Disease dataset]. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[23] PhysioNEt: Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., … Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation, 101*(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

[24] MIMIC IV: Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., … Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation, 101*(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

[25] Y. Qiu *et al*., "Pre-Orthogonal Adaptive Fourier Decomposition for ECG Denoising and Coronary Heart Disease Monitoring," in *IEEE Access*, vol. 13, pp. 195467-195482, 2025, doi: 10.1109/ACCESS.2025.3633296.

[26] D. Yang and W. Gao, "PointCHD: A Point Cloud Benchmark for Congenital Heart Disease Classification and Segmentation," in *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 4, pp. 2683-2694, April 2025, doi: 10.1109/JBHI.2024.3495035.

[27] A. Patwa, M. Mahboob Ur Rahman and T. Y. Al-Naffouri, "Heart Murmur and Abnormal PCG Detection via Wavelet Scattering Transform and 1D-CNN," in *IEEE Sensors Journal*, vol. 25, no. 7, pp. 12430-12443, 1 April1, 2025, doi: 10.1109/JSEN.2025.3541320.

[28] X. Qin *et al*., "3D Distance-Color-Coded Assessment of PCI Stent Apposition via Deep-Learning-Based Three-Dimensional Multi-Object Segmentation," in *IEEE Transactions on Medical Imaging*, vol. 44, no. 11, pp. 4717-4731, Nov. 2025, doi: 10.1109/TMI.2025.3580619.

[29] M. K. Jabbar, H. Jianjun, A. Jabbar and Z. Ur Rehman, "Mamba-Based VoxelMorph Framework for Cardiovascular Disease Imaging and Risk Assessment," in *IEEE Access*, vol. 13, pp. 78120-78137, 2025, doi: 10.1109/ACCESS.2025.3564962.