Received: 2 April 2026 | Revised: 15 June 2026 | Accepted: 24 June 2026 | Published online: 8 July 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–18
DOI: 10.47852/bonviewAIA62029883

BON VIEW PUBLISHING

Beat-Level Echocardiographic Clip Analysis with Geometry-Video Fusion and Cross-Attention for Accurate Ejection Fraction Estimation and Heart Failure Stratification

Chaithra C. S.[1,*], Siddesha S.[1], Vinay Kumar N.[2] and V. N. Manjunath Aradhya[1]

[1] *Department of Computer Applications, JSS Science and Technology University, India*

[2] *NTT Data Inc., India*

Abstract: Left ventricular ejection fraction (LVEF) is a crucial indicator of cardiac function in the diagnosis and management of heart failure. Recent deep learning approaches have shown promising results for automated LVEF estimation from echocardiographic videos; most studies rely on spatiotemporal information, and few studies investigate the contribution of anatomical descriptors. In this study, we propose a comprehensive beat-level framework for EF estimation and heart failure classification using echocardiographic videos. Representative cardiac beat clips were extracted using tracing-derived annotations. Three architectures were investigated: (a) EchoEF-Net, a spatiotemporal Convolution Neural Network (CNN), which is a video-only model; (b) GeoFusionEF-Net, a multimodal deep learning feature-level fusion of geometry and video features for enhanced robustness and calibration; and (c) CrossAttnFusionEF-Net, which employs a staged cross-attention multimodal architecture to integrate the video and geometry interactions adaptively. Experiments were conducted on the publicly available EchoNet-Dynamic dataset (10,030 videos). On the official test set, the proposed GeoFusionEF-Net model effectively captures interactions between the video and geometry modalities, significantly reducing mean absolute error from 4.61 (EchoEF-Net) and 1.99 (CrossAttnFusionEF-Net) to 0.648, while improving root mean squared error and Pearson correlation to 1.413 and 0.99, respectively. The estimated EF predictions were further utilized to automate heart failure stratification into reduced (HFrEF), mid-range (HFmrEF), and preserved (HFpEF) EF categories. The meta-classifier of all three EF estimators achieved 0.97 accuracy and 0.99 Area Under the Curve (AUC). The proposed framework, combining beat-level motion analysis and geometric features with Shapley Additive Explanations- and Gradient-Weighted Class Activation Mapping-based explainability, enables quantification of ejection fraction and clinically reliable heart failure classification, thereby scaling echocardiographic decision-making.

Keywords: echocardiography, ejection fraction estimation, deep learning, cross-attention fusion, heart failure stratification

1. Introduction

Despite all efforts, cardiovascular disease remains a major cause of mortality worldwide, with heart failure (HF) in 64 million leading to 1.9 million deaths as per the World Heart Report 2025 [1]. The timely evaluation of cardiac performance and function is critical for both diagnosis and efficient treatment. Left ventricular ejection fraction (LVEF) is an essential metric for assessing cardiac performance and is widely used in the diagnosis and monitoring of heart dysfunction, disease progression, and HF risk stratification.

Several methods have been applied to the automatic evaluation of ejection fraction (EF) using extracted cardiac motion information from echocardiography videos, including conventional neural networks, spatiotemporal models, and new transformer-based architectures. Despite many studies in this area, previous research has focused on videos and overlooked other factors that could be used in the clinical evaluation of cardiac function.

To address these challenges, we have proposed a framework for analyzing cardiac beats to estimate EF and classify HF. Relevant cardiac beats in echocardiography videos are extracted. Three distinct features have been evaluated: EchoEF-Net, which relies on video information; GeoFusionEF-Net, which depends on tracing geometric features and video features as a feature-level fusion strategy; and CrossAttnFusionEF-Net, which employs the interaction between video and geometrical features by a cross-attention mechanism.

In addition to EF estimation, this work proposes a meta-learning HF classification pipeline that uses the outputs of EF estimation models to categorize patients into three classes: normal, HFmrEF, and HFrEF. In addition, Gradient-Weighted Class Activation Mapping (Grad-CAM) and Shapley Additive Explanations (SHAP) approaches are being applied to enhance the model's interpretation.

***Corresponding author:** Chaithra C. S., Department of Computer Applications, JSS Science and Technology University, India. Email: chaithracs@jssstuniv.in

The key contributions of this work are:

a. EchoEF-Net: motion-aware EF estimation using fixed-length beat-level motion-centered clip extraction.
b. GeoFusionEF-Net: A multimodal geometry-guided feature fusion architecture that integrates video motion embeddings and physiological geometrical descriptors of the left ventricle for efficient EF estimation.
c. CrossAttnEF-Net: A stage-level trained multimodal model using an adaptive cross-attention mechanism using motion features with geometrical priors for reliable EF estimation.
d. Heart Failure Classification: Hence, EF is the primary clinical indicator used to diagnose and stratify HF severity. We used a meta-stacked layer comprising multiple EF estimators, with and without geometric features, for HF classification.

2. Literature Review

2.1. Deep learning-based EF estimation

LVEF is an important index used in diagnosing and managing HF [2–4]. EchoNet-Dynamic is a large-scale video dataset released for EF estimation and has demonstrated the effectiveness of video-based deep learning for beat-to-beat cardiac function assessment [5]. Similarly, several studies have been explored for EF estimation, including real-time EF estimation networks, image quality assessment [6, 7], Bayesian video analysis frameworks, representation learning approaches [8], and convolutional video-based models [9]. The feature quality and data variability were further highlighted as important parameters for reliable EF prediction in clinical validation studies [10].

2.2. Geometry-aware and attention-based EF analysis

While video-based models are good at capturing cardiac motion, they generally ignore structural information routinely used in clinical evaluation. Geometry-aware methods (EchoGNN) utilized structural cardiac representations to improve EF estimation and interpretability with explainable AI. Studies on ventricular structure have highlighted the clinical relevance of structural information measurements [11]. More recently, transformer-based architectures such as EchoCoTr, ViVIEchoFormer, and hierarchical vision transformers have been proposed to address dependencies between spatial and temporal features in echo videos [12–15]. However, the systematic evaluation of geometry-aware and transformer-based models under a unified model is limited.

2.3. Heart failure stratification

HF is classified as HFrEF, HFmrEF, and HFpEF according to LVEF cut-offs [16]. Echocardiography is the image modality to assess ventricular heart function and guide clinical management [17–19]. Automatic identification of HF and EF-based classification methods have been explored in recent studies [20–23].

Although many recent studies address an automated echocardiographic analysis, some limitations are noticed. A very limited study considered video-only, geometry-aware, and attention-based EF estimation paradigms as a single unified model. The fusion strategies and geometric fusion studies remain unaddressed. There is a lack of explicit anatomical information, which is crucial for the model to learn the structural aspects of cardiac motion that contribute to EF estimation in complex cases. Most approaches are finalized as EF estimation and do not investigate HF classification. To address these gaps, this work investigates EchoEF-Net, GeoFusionEF-Net, and CrossAttnFusionEF-Net. Further introduces a meta-classifier for an HF classification framework, supported by explainable AI and statistical validation.

3. Methodology

3.1. Dataset

A standard benchmark dataset from Stanford Health Care, EchoNet-Dynamic [24], is used, consisting of 10030 unique de-identified echocardiographic videos, each comprising 4 apical chamber views of cardiac function. The dataset contains videos in.avi format with multiple cardiac cycles, used for LV assessment. The dataset has two CSV files. The first is metadata like frame height, width, number of frames per second, and total number of frames of each video with ground truth EF, video ID, and standard splits for training, validation, and testing. The second CSV file contains expert-annotated LV segmentation boundaries at end systole and end diastole, which can be used to compute volume and EF.

3.2. Frame preprocessing and beat-clip extraction

Similar preprocessing is used across all three models, and the data is handled consistently for fair comparison. Figure 1 shows the process of frame preprocessing, and the beat-clip extraction echocardiogram has a four-chamber epicardial view, which is in grayscale:

$$V = \{f_1, f_2, \ldots, f_T\}, \quad f_t \in R^{H_0 \times W_0} \tag{1}$$

where T is the number of frames in one cycle, which varies due to duration and heart rate. No artificial augmentation techniques were employed. Only deterministic preprocessing operations, including temporal clip extraction of 32 frames, spatial resizing to 112 × 112, and intensity normalization [0, 1], were performed to preserve relevant cardiac information. Each frame is resized to 112 × 112 pixels using bilinear interpolation and intensity-scaled to the range [0, 1]. $\mathcal{N}(x) = \frac{x}{255}$ is used to produce a standardized video by fixing the brightness level and image size while preserving natural cardiac motion.

Representative cardiac cycles were selected using ventricular area dynamics calculated from LV tracings provided by experts. For each annotated frame t, the area of the ventricular cavity contour was reconstructed from tracing coordinates and converted into a polygon. The ventricular area was calculated using the shoelace formula:

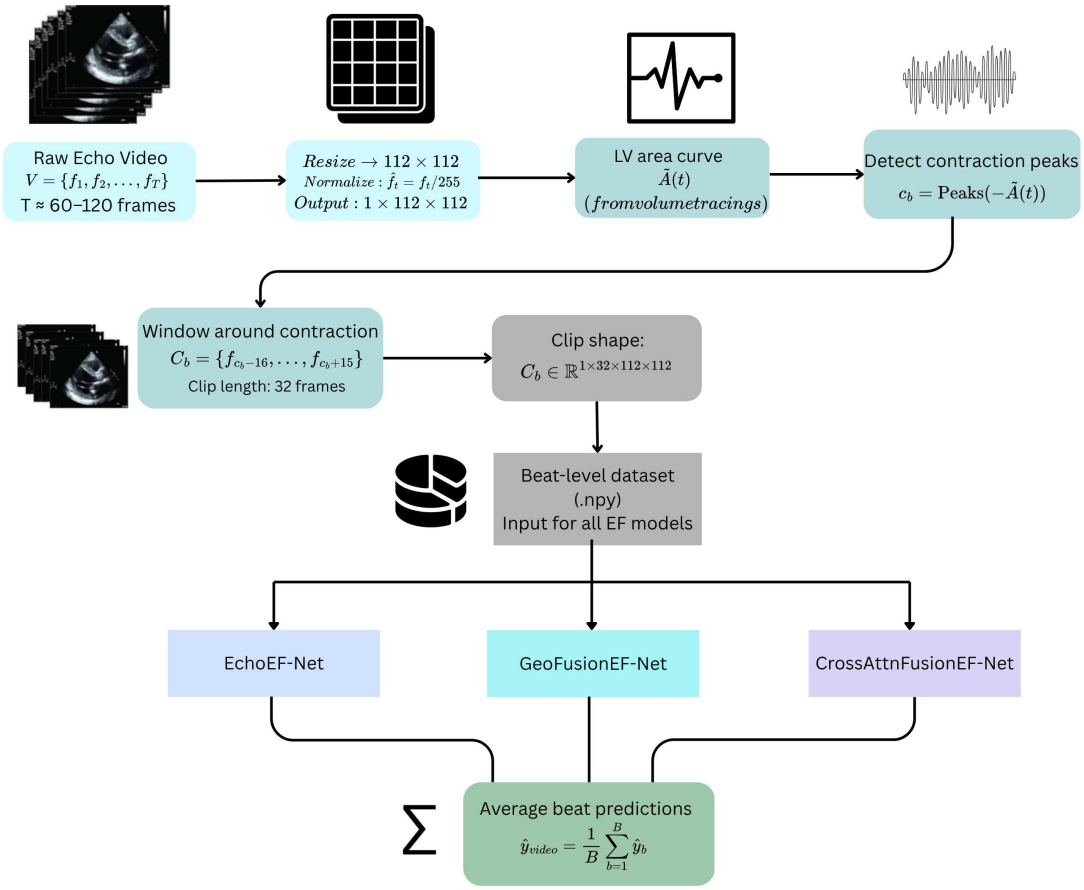
$$A(t_k) = \frac{1}{2}\left|\sum_{i=1}^{n}(x_i y_{i+1} - x_{i+1} y_i)\right| \tag{2}$$

where $A(t_k)$ represents the ventricular cavity area for the frame t_k and (x_i, y_i) denotes ordered contour coordinates defining the ventricular boundary. Since ventricular area measurements were available for only traced frames, the sparse area measurements were interpolated to generate a continuous LV area curve to make it smooth across each frame for the entire video sequence:

$$\tilde{A}(t) = \text{Interp}\left(A(t_k)\right), \quad t = 1, \ldots, T \tag{3}$$

T denotes the total number of video frames. This output provides the temporal variation in ventricular size over one

Figure 1
Frame preprocessing and beat-clip extraction process



cardiac cycle. Cardiac contraction points were identified by detecting peaks in the inverted area curve.

$$c_b = \text{Peaks}\left(-\tilde{A}\left(t\right)\right) \tag{4}$$

Here c_b is the center frame of the detected beat. Since the peak is detected using a negative area value, the identified points represent the minimum ventricular area during the most contracted phase of the heart (end systole). For each detected beat center, a fixed-length clip of $L = 32$ frames was extracted centered on end systole (contraction), rather than performing full-cycle segmentation.

$$Cb = \left\{\hat{f}_{c_b-\frac{L}{2}}, \ldots, \hat{f}_{c_b+\frac{L}{2}-1}\right\} \tag{5}$$

which yields $Cb = \left\{\hat{f}_{c_b-16}, \ldots, \hat{f}_{c_b+15}\right\}$. Here, f denotes resized video frame, 16 frames before contraction and 16 frames after contraction, for a total of 32 frames. Each extracted clip shows the heart's full contraction and relaxation, which is crucial to predict EF. This captures systole and nearby diastole with consistent temporal resolution across patients. To avoid redundancy and to preserve representative cardiac motion, a maximum of three beat-centered clips were selected from each video. Each clip was resized to 112×112 pixels and represented as a 4-D tensor $C_b \in R^{1\times32\times112\times112}$. All clips from each video are stacked into a single beat-level dataset $\mathscr{D} = \{C_b\}_{b=1}^{B_{\text{total}}}$, with $B_{\text{total}} = 30039$ clips from 10030 videos. This dataset is used as input to all three models—EchoEF-Net, GeoFusionEF-Net, and CrossAttnFusionEF-Net—for training and evaluation. The proposed tracing-guided beat extraction strategy enables models to focus on clinically meaningful contractions rather than processing redundant frames.

3.3. Geometrical features extraction

In addition to spatiotemporal video features, we computed four geometric features to incorporate anatomical information describing left ventricular (LV) morphology and contractile function, derived from expert-annotated ventricular contours in the dataset. We extracted tracing-derived geometrical features. These features are clinical descriptors of the left ventricle that characterize its size and shape. The features are ES_area, ED_area, Fractional Area Change (FAC), and Sphericity. For each annotated frame, a left ventricle boundary was shown as a closed polygon created by tracing coordinates. The area of the ventricular cavity was calculated using the shoelace formula:

$$A = \frac{1}{2}\left|\sum_{i=1}^{n}\left(x_i y_{i+1} - x_{i+1} y_i\right)\right| \tag{6}$$

where A denotes the ventricular cavity area, $(\boldsymbol{x_i}, \boldsymbol{y_i})$ and B is a set of boundary points of contour coordinates. For each video, the frames with the minimum and maximum ventricular areas are identified as end systole (ES) $\boldsymbol{ES_{Area}} = \min\left(\boldsymbol{A_t}\right)$ and end diastole (ED) $\boldsymbol{ED_{Area}} = \max\left(\boldsymbol{A_t}\right)$, respectively. Here $\boldsymbol{A_t}$ is the ventricular cavity area at frame t.

Using these measurements, four geometry descriptors were extracted as given in Table 1. The end-systolic area is the

Table 1
Geometrical features derived from the given LV tracings

Features	Description
$ESA = ESArea$	Smallest size of ventricular cavity during heart contraction
$EDA = ED_{Area}$	Largest size of ventricular cavity during filling
FAC	Systolic performance of the right ventricle
Sphericity S	LV spherical shape

smallest ventricular size calculated as $\boldsymbol{ESA = ES_{Area}}$, which means stronger contraction, and the largest size of ventricular area, while filling is represented as $\boldsymbol{EDA = ED_{Area}}$. To quantify ventricular contraction, the FAC was computed using

$$\boldsymbol{FAC = \frac{EDA - ESA}{EDA}} \tag{7}$$

This shows a relative decrease in ventricular cavity area between diastole and systole. To describe the changes in ventricular shape during the cardiac cycle, we computed Sphericity as

$$\boldsymbol{S = \frac{ESA}{EDA}} \tag{8}$$

where s denotes the normalized relationship between systolic and diastolic ventricular area. The final geometry feature vector for each beat clip is denoted as $\boldsymbol{G = [A_{ED}, A_{ES}, FAC}$, Sphericity $\boldsymbol{] \in R^4}$. These geometric features are derived from the dataset-provided tracings and are integrated with deep spatiotemporal video representations in the GeoFusionEF-Net and CrossAttnFusionEF-Net architectures to provide complementary anatomical information for EF estimation.

4. Model Architecture

a. EchoEF-Net: a partially reproduced video-only spatiotemporal model.
b. GeoFusionEF-Net: a fusion model that combines video and geometric features.
c. CrossAttnFusionEF-Net: a geometry-guided temporal model.

4.1. EchoEF-Net

EchoEF-Net is a baseline video-only model for EF estimation. Figure 2 illustrates the overall architecture of the EchoEF-Net model, which receives a beat-centered echo clip of 32 grayscale frames and learns ventricular dynamics directly from spatiotemporal clips. To efficiently estimate EF from video-only information, the model employs R(2 + 1)D, a 3D convolutional backbone that decomposes 3D convolutions into separate 2D spatial and temporal convolutions to capture ventricular contraction dynamics. The extraction is defined as

$$X' = \sigma(W_t \times \sigma(W_s \times X)) \tag{9}$$

where X' *is the* input feature map and W_s and W_t are spatial and temporal convolution kernels. $\sigma(\cdot)$ denotes the ReLU activation function. This enables the network to learn heart systolic contraction, diastolic relaxation, and valve motion dynamics over time. Heart function always depends on motion between frames, not on single static frames, which a 1D CNN can easily learn. R (2 + 1) D is much faster than 3D CNN, reducing parameters and enabling faster training. Thus, a single R(2 + 1)D block is computed X', which improves gradient flow, reduces parameters, and separates the structural information of the heart from motion dynamics. To facilitate stable optimization, each block is embedded with a residual connection:

$$Y = X + F(X) \tag{10}$$

where $F(X)$ convolution layers learn the residual transformation. Throughout the network, residual connections run, which help the gradient flow more effectively, enabling a deeper understanding of video patterns. The resulting spatiotemporal feature maps are aggregated through global average pooling to achieve a compact beat-level representation as

$$z_b = \frac{1}{CT'H'W'} \sum \mathrm{F_b} \tag{11}$$

where z_b is the latent feature vector of the b^{th}clip. A regression head turns this into a single predicted EF value for the clip.

$$\hat{y_b} = Wz_b + b \tag{12}$$

where z_b is the pooled feature vector, W is the learned weight matrix, b is the bias term, and $\hat{y_b}$ is the predicted EF for the clip. Multiple clips' estimated EF is averaged into a single video-level EF by

$$\hat{y}_{video} = \frac{1}{B} \sum_{b=1}^{B} \hat{y_b} \tag{13}$$

to reduce the variance, avoid dependency on a single temporal window. The network is optimized using the mean squared error (MSE) loss function

$$\mathscr{L}_E = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{14}$$

y_i is the ground truth EF, and $\hat{y}_i$ is the predicted EF. MSE penalizes large errors more than minor ones, which is standard in regression. The network is trained using the Adam optimizer with a learning rate of 1×10^{-4}. To improve generalization and reduce overfitting, weight decay regularization 1×10^{-5} was applied. Early stopping was adopted based on validation loss with a batch size of 16.

4.2. GeoFusionEF-Net

The proposed methodology is an extension of the EchoEF-Net model, which fuses video embeddings with derived geometric features. These features define the chamber geometry, and this model predicts the EF from a video 3D CNN backbone along with four physiological geometric features, such as Sphericity and FAC. Figure 3 presents the architecture of GeoFusionEF-Net. The model is fed with beat-level clips and four geometrical features extracted from LV tracings. For video feature extraction, an R(2 + 1)D spatiotemporal CNN is used $v = \phi_v(C_b) \in R^{16}$, which is a compact motion descriptor learned from a clip. The geometrical features are projected into a shared 16-dimensional embedding space$g = W_g x_g + b_g \in R^{16}$, which maps the geometry into a feature space compatible with video features. Both geometrical and video embeddings were concatenated $f \in R^{16+16} = R^{32}$, which occurs after the video backbone and before the regression head.

Figure 2
Architecture of the EchoEF-Net model

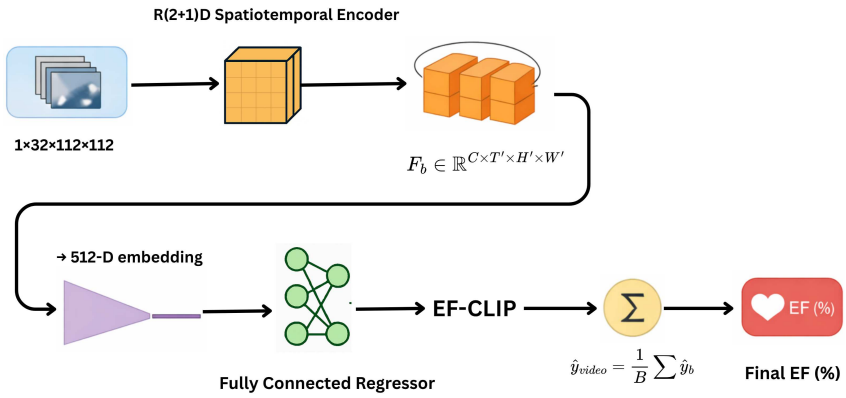
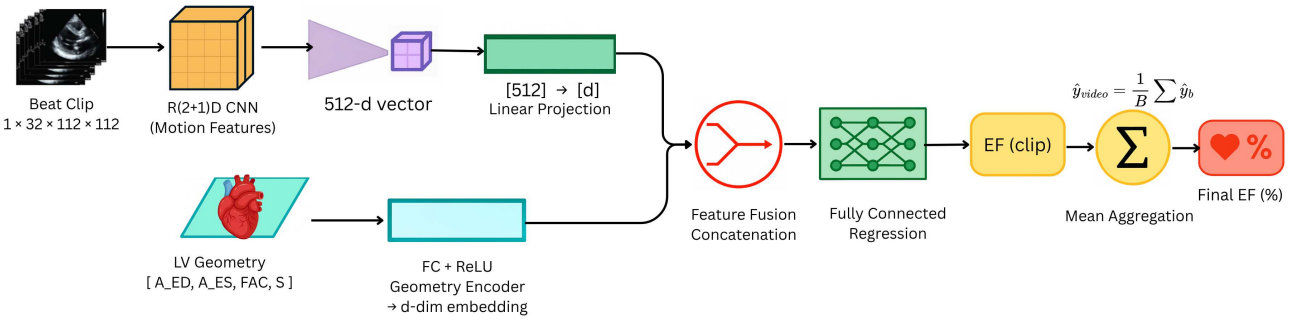


Figure 3
GeoFusionEF-Net architecture



This model concatenates at the feature level both 16-dimensional compact spatial and temporal embeddings from R(2 + 1)D CNN (i.e., $v \in R^{16}$) and geometry embedding, which is produced by a small multilayer perceptron $g \in R^{16}$, a feature vector of length 16 from the video encoder. The concatenated embedding $f = v\ g \in R32$, which represents the first 16 dimensions as motion representations and the last 16 dimensions as structural representations, is passed to the regression head of fully connected layers $h_b = \sigma(W_1 f + b_1)$, which produces beat-level EF values $\hat{y_b} = W_2 h_b + b_2$ for each video of three clips. Video-level EF estimation is aggregated as in the EchoEF-Net model $\widehat{y_{video}} = \frac{1}{B}\sum_{b=1}^{B} \hat{y_b}$. The model is trained using MSE, Adam optimization (which automates learning rate selection), and the ReLU activation function, which enhances sparse activations and mitigates gradient problems. It uses a batch size of 16 and trains for 30 epochs with a learning rate of $1e^{-4}$. We introduced a novel multimodal data fusion method that improves EF estimation compared to the EchoEF-Net model.

4.3. CrossAttnFusionEF-Net

Figure 4 illustrates the architecture of CrossAttnFusion-Net, which investigates whether explicit interactions between spatiotemporal video features and geometric descriptors could further improve EF estimation. A cross-attention mechanism is adapted to enable the model to dynamically determine which geometric features are most relevant to specific cardiac motion. A cross-attention mechanism is adopted. This cross-attention allows the geometric features to guide the network in learning clinically relevant motion patterns, rather than merely being added at the final regression step. An R2 + 1D CNN is employed to process beat-level clips; both video representations $v_b \in R^{d_v}$ from the video encoder and corresponding geometry feature vectors $g_b \in R^{d_g}$ from the geometry encoders are extracted from the beat clips. Before applying attention, the two embeddings are projected to a shared attention space, $z_v = W_v v_b, \quad z_g = W_g g_b$ with aligned projections as $z_v, z_g \in R^d$ before fusion. To model cross-modal interactions, query, key, and value representations are generated as

$$Q = W_Q z_v,\ K = W_K z_g,\ V = W_V z_g \tag{15}$$

Attention weights are computed using a scaled dot-product attention:

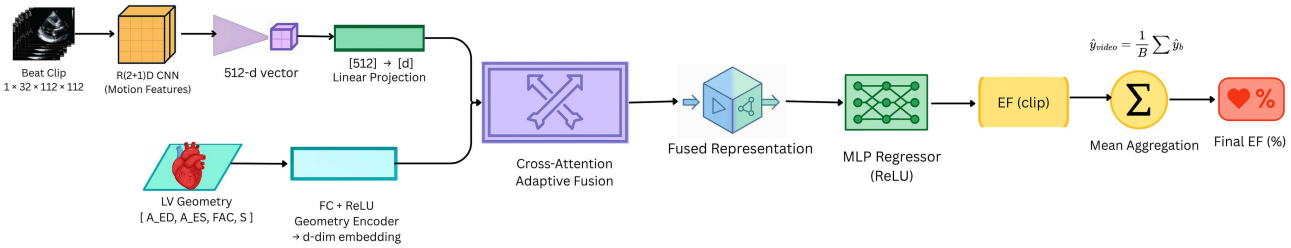
$$\alpha = \text{softmax!}\left(\frac{QK^\top}{\sqrt{d}}\right) \tag{16}$$

This determines how much geometry features influence video representation. This produces geometric information $g_{att} = \alpha V$ after filtering the video; attention is fused with the video representation via concatenation.

$$h_b = [\, z_v \,|\, g_{\text{att}} \,] \tag{17}$$

which creates a representation of cardiac motion and structural features. The attention learns which geometry matters for which motion. Hence, it is still raw concatenation, which is transformed and compressed before giving it to the regression head. The fusion vector is given to fully connected dense layers, the first dense layer $u_b = \sigma(W_1 h_b + b_1)$, which fuses video and geometry, learns non-linear interactions and enhances relevant patterns, and the second

Figure 4
CrossAttnFusionEF-Net architecture



dense layer $z_b = \sigma(W_2 u_b + b_2)$ compresses by removing duplicates and stabilizes the regression task. Finally, the regression head $\hat{y}_b = W_o z_b + b_o$ predicts the EF for beat clips b. At the end, the beat-level EF is aggregated $\widehat{y_{\text{video}}} = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_b$ using the same procedure as for the EchoEF-Net and GeoFusionEF-Net models.

4.3.1. Model training

The proposed model is trained at the stage level up to three stages. In the first stage, the cross-attention parameters were frozen,

$$\theta_g^* = argmin_{\theta_g} \frac{1}{N}\sum_{i=1}^{N} \left(f_{\theta_v^*, \theta_g}(C_i, g_i) - y_i \right)^2 \quad (18)$$

means not active, which enables the model to learn video-driven EF representations before geometry representations. In the second stage, the video encoder is frozen.

$$\theta_g^* = argmin_{\theta_g} \frac{1}{N}\sum_{i=1}^{N} \left(f_{\theta_v^*, \theta_g}(C_i, g_i) - y_i \right)^2 \quad (19)$$

The parameters were activated to learn geometric features using queries, keys, and values to avoid the dominance of geometric features. The third stage involves a cross-attention module, a regression head, and projection layers, which were fine-tuned at a low learning rate, while the video layers remained stable.

$$\theta^* = argmin_{\theta} \frac{1}{N}\sum_{i=1}^{N} \left(f_{\theta}(C_i, g_i) - y_i \right)^2 \quad (20)$$

The final model is given as:

$$\hat{y}_b = \psi\left(\mathcal{A}\left(\phi_v(C_b), \phi_g(g_b) \right) \right) \quad (21)$$

where $\phi_v(C_b)$ is the video encoder, $\phi_g(g_b)$ is the geometry encoder, $\mathcal{A}(\cdot)$ cross-attention fusion, and $\psi(\cdot)$is the regression head. Stage-wise training in a supervised setting, using MSE loss and an Adam optimizer, was adopted to ensure stable model behavior. ReLU activation functions were applied in the projection, attention, and regression layers to enable the learning of complex nonlinear relationships between video and geometric representations. The initial learning rate was 1×10^{-4}, which was later reduced during training. A batch size of 16 was adopted for EchoEF-Net and GeoFusionEF-Net, whereas CrossAttnFusionEF-Net uses a batch size of 8 due to the increased memory requirements of the cross-attention mechanism. Models were trained for approximately 25–30 epochs, with early stopping on validation loss to prevent overfitting, and staged optimization was applied to CrossAtnnFusionEF-Net. The proposed model is only effective for optimization, not for inference-time architecture. All models were implemented in PyTorch and executed on NVIDIA GPU A100 hardware.

5. Heart Failure Stratification

One of the major reasons to estimate EF is to understand which HF category a patient belongs to, so that clinicians can make treatment decisions. The standard cardiology guidelines [16] provide three categories of HF based on LVEF values, as Table 2 shows.

Table 2
Heart failure classification categories for various EF ranges

Heart failure category	EF range	Class
HFrEF: reduced EF	EF (<40%)	0
HFmrEF: moderate EF	EF (41–49%)	1
HFpEF: preserved EF (normal)	EF (>=50%)	2

Similar to clinicians' routine after estimating EF, they analyze HF risk stratification. This motivates us to extend the EF-estimated predictions for HF classification. HF categories are defined using standard EF thresholds. However, because predicted EF values may contain errors in the near-boundary region, a classifier is used to improve robustness and reduce misclassification. We utilized trained EF models to classify HF. Target variables are assigned to each sample based on standard cardiology guidelines $y_i \in$ {HFrEF, HFmrEF, Normal}. We evaluated five experimental configurations: EchoEF-Net-only, GeoFusionEF-Net-only, CrossAttnFusionEF-Net-only; a fourth is a meta-learning model that combines all three models' predictions; and the final configuration is meta-learning augmented with geometrical descriptors. A feature vector x_i is constructed for each patient i for a selected configuration. A multinomial logistic regression classifier is trained using 5-fold stratified cross-validation

$$p(y_i = k|x_i) = \exp\left(w_k^T x_i + b_k\right) / \sum_{j=1}^{3} \exp\left(w_j^T x_i + b_j\right) \quad (22)$$

for each experiment. Each configuration isolates motion only, geometry-guided video, attention-fused, and ensemble learning for HF risk stratification. The model yields robust, clinically meaningful results, given the small dimensions and moderate dataset (1276 test videos), which is well-suited to selecting a shallow classifier and helps avoid overfitting. The first configuration $x_i = \left[\widehat{EF_i^{\text{EchoEF}}}\right] \in R^1$, which uses EchoEF-Net predictions to classify HF using only cardiac motion information. The second configuration uses the predictions from the geometry-guided GeoFusionEF-Net model $x_i = \left[\widehat{EF_i^{\text{GeoFusion}}}\right] \in R^1$ to classify

HF. CrossAttnFusionEF-Net model's EF predictions are used to classify HF risk $x_i = \left[\widehat{EF_i^{\text{CrossAttn}}}\right] \in R^1$ to verify how well attention-guided fusion performs for HF stratification. Meta-learning that leverages the strengths of all three EF predictors is evaluated to assess how well the model learns a linear combination of predictions for HF classification.

$$x_i^{EF} = \left[\widehat{EF_i^{\text{EchoEF}}}, \widehat{EF_i^{\text{GeoFusionEF}}}, \widehat{EF_i^{\text{CrossAttn}}}\right] \in \mathbb{R}^3 \quad (23)$$

Along with these combination geometrical features, $x_i^{geom} = [A_{ED}, A_{ES}, FAC, S] \in \mathbb{R}^4$, the fifth configuration classifier is also added explicitly, given as:

$$x_i = \left[x_i^{EF} \mid x_i^{geom}\right] \in \mathbb{R}^7 \quad (24)$$

All configurations are trained on 80% of the training samples and tested on 20% of the test samples (1276), with 5-fold stratified cross-validation, and evaluated for accuracy, macro-F1, and weighted-F1.

6. Experimental Setup

6.1. Architectural specifications of proposed models

All three models employ the same R(2 + 1)D-18 backbone to extract a 512-dimensional spatiotemporal representation from echo videos. EchoEF-Net is a video-only baseline that directly regresses EF with a regression head of (512→128→64→1). GeoFusionEF-Net uses four geometric descriptors, which are processed through a geometry branch (4→128→128). The output is concatenated with video features, yielding a fused representation of 640 dimensions, which is regressed by a prediction head with 640→256→1 dimensions. Here, a feature-level fusion strategy is used. CrossAttnFusionEF-Net uses the same video and geometrical branches but adds an 8-head cross-attention module to model between-modality interactions before regression via a 640→256→1 output layer. This model incorporates attention-based fusion.

6.2. Implementation and reproducibility

All experiments were implemented using PyTorch and executed on an NVIDIA A100 GPU with CUDA acceleration. The echo videos were transformed into a 32-frame grayscale clip and downscaled to a constant size of 112 × 112 pixels, and pixel intensities were scaled to the [0, 1] range. All models have been applied to the standard split from the standard dataset and evaluated at the video and beat-level aggregations to improve clinical relevance and remain consistent with echo reporting. All experiments were performed with a fixed random seed of 42 to ensure reproducibility. The model was trained using the AdamW optimizer with a warm-up learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. Early stopping was applied based on validation mean absolute error (MAE). To compare the fusion strategies, EchoEF-Net, GeoFusionEF-Net, and CrossAttnFusionEF-Net used the same R(2 + 1)D-18 backbone. Table 3 summarizes the training configurations and reproducibility settings used across all models for fair comparison.

Table 3
Training and configuration settings

Parameter	Value
Dataset	EchoNet-Dynamic
Total videos	10,030
Total beat clips	30,074
Input resolution	112 × 112
Clip length	32 Frames
Backbone	R(2 + 1)D-18
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-5}
Random seed	42
Epoches	30
Batch size	16 (EchoEF, GeoFusion), 8 (CrossAttn)
Loss function	Mean squared error (MSE)
Hardware	NVIDIA A100 GPU
Framework	PyTorch
Device	CUDA

6.3. Evaluation metrics

1) Regression

Model performance was evaluated using MAE and root mean squared error (RMSE) to quantify prediction error. To assess linear agreement, the Pearson correlation coefficient (r), coefficient of determination (R^2), and concordance correlation coefficient (CCC) are evaluated to measure agreement with the ground truth EF. Clinical reliability is assessed using the percentage of predictions within ±5% and ±10 of the EF of the reference measurement. Stratified performance of the model is evaluated to assess robustness across clinical relevance ranges, such as reduced EF (<40%), mid-range EF (41–49%), and normal EF (>=50%).

2) Heart Failure Classification

The HF classification performance was evaluated using accuracy, which measures the overall correctness of classification; precision and recall, which measure the reliability and sensitivity of class predictions; F1-score, which is a balanced measure of precision and recall when there is class imbalance; and AUC, which measures the model's capability to discriminate between HF at different decision thresholds.

7. Experiment Results and Analysis

7.1. EF estimation performance

We evaluated the EchoEF-Net, GeoFusionEF-Net, and CrossAttnFusion-Net models for estimating EF from videos using beat-level aggregation on the official test cohort of n = 1276 echo videos. As shown in Table 4, GeoFusionEF-Net has the strongest performance among the three regression models. GeoFusionEF-Net has achieved MAE of 0.648%, RMSE of 1.413%, Pearson r of 0.993, R^2 of 0.987, and CCC of 0.993. EchoEF-Net achieved an MAE of 4.64% and an R^2 of 0.735, while CrossAttnFusionEF-Net achieved an MAE of 1.99% and an R^2 of 0.937, showing that adding geometric information boosts accuracy compared to

Table 4
Quantitative performance comparison of EF estimation models

Evaluation metric	EchoEF-Net	GeoFusionEF-Net	CrossAttnFusionEF-Net
MAE (%)	4.61	0.648	1.99
RMSE (%)	6.29	1.413	3.08
Pearson *r*	0.86	0.993	0.968
R^2	0.73	0.987	0.937
CCC	0.83	0.993	0.967
Within ±5 EF (%)	65.28	98.51	94.98
Within ±10 EF (%)	89.11	99.84	98.82

a video-only model. The calibration slope of 0.983 and intercept of 0.915 prove the Geo-FusionEF-Net model's reliability, closely matching the ideal calibration of slope=1 and intercept=0. The EchoEF-Net model has an intercept of 18.06 and a slope of 0.678, indicating bias and underfitting, suggesting it did not learn LV dynamics effectively. To prove clinical robustness, we measure clinically acceptable EF tolerance in which GeoFusionEF-Net achieved 98.5% accuracy within ±5% and 99.8% within ±10 % outperforming both EchoEF-Net (65.3 within ±10 %) and GeoFusionEF-Net (95.0% within ±5%), which results proved that geometry-guided fusion is reliable for clinical precision in EF estimation.

Calibration results in Table 5 and Figure 5 revealed that EchoEF-Net exhibited a strong positive bias, underestimating high EF and overestimating low EF. GeoFusionEF-Net demonstrates near calibration, confirming accurate EF predictions across clinical ranges. CrossAtnnFusionEF-Net achieved strong calibration with minor bias at extreme EF values, which is robust but yields a minimally conservative prediction in high EF ranges, as described in the calibration plots.

Table 5
Calibration characteristics of EF prediction models

Model	Slope	Intercept
EchoEF-Net	0.678	18.06
GeoFusionEF-Net	0.983	0.915
CrossAttnFusionEF-Net	0.924	3.99

Figure 6 shows parity plots indicating that GeoFusionEF-Net aligns with the ground truth, indicating near-perfect agreement. EchoNet-Fusion shows bias and significant scatter. CrossAttnFusionEF-Net showed minimal deviation, indicating better overall alignment.

The Bland–Altman analysis in Figure 7 shows that EchoEF-Net exhibited systematic bias and greater error.

GeoFusionEF-Net showed an error centered at zero and very narrow agreement limits, which is excellent from a clinical perspective. CrossAttnFusionEF-Net revealed a slight negative bias and moderate spread, but good agreement.

The error distribution histogram in Figure 8 shows EchoEF-Net with a wide spread of high variability and large errors with long tails. GeoFusionEF-Net shows a narrow, centered peak with highly precise predictions with minimal error. CrossAttnFusionEF-Net shows moderate spread with improved accuracy. Figure 9 shows the agreement, error distribution, and calibration comparison results across EF prediction models.

7.2. Stratified EF analysis

Table 6 and Figure 10 exhibit stratified EF estimation performance across clinically relevant EF categories. For EchoEF-Net, the prediction error increased as ventricular function decreased. The GeoFusionEF-Net achieved the lowest error across all EF categories, demonstrating the benefit of integrating tracing-derived geometric descriptors with video representations. CrossAttnFusionEF-Net also improved significantly over the video-only baseline but remained behind GeoFusionEF-Net. This indicates that geometry-aware multimodal learning improves EF estimation across clinical ranges, whereas simple feature-level fusion outperforms the more complex attention strategy in the fusion model.

Figure 5
Calibration plots of EF prediction models

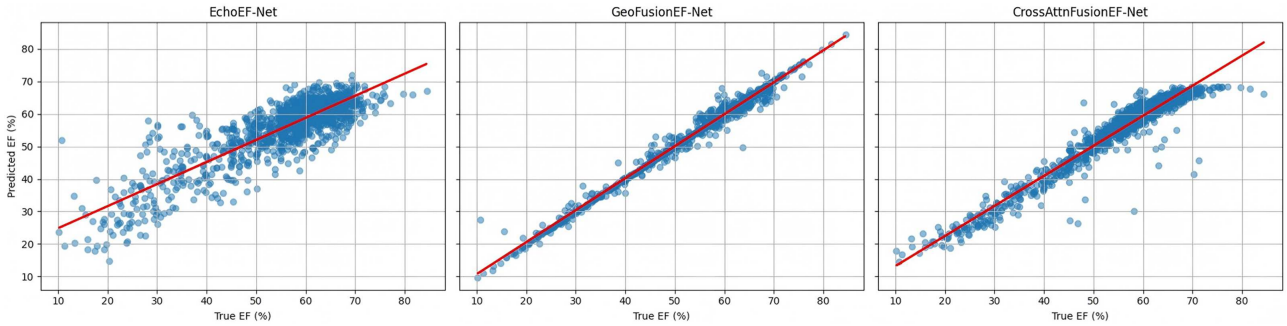


Figure 6
Parity plots comparing predicted and ground truth EF values

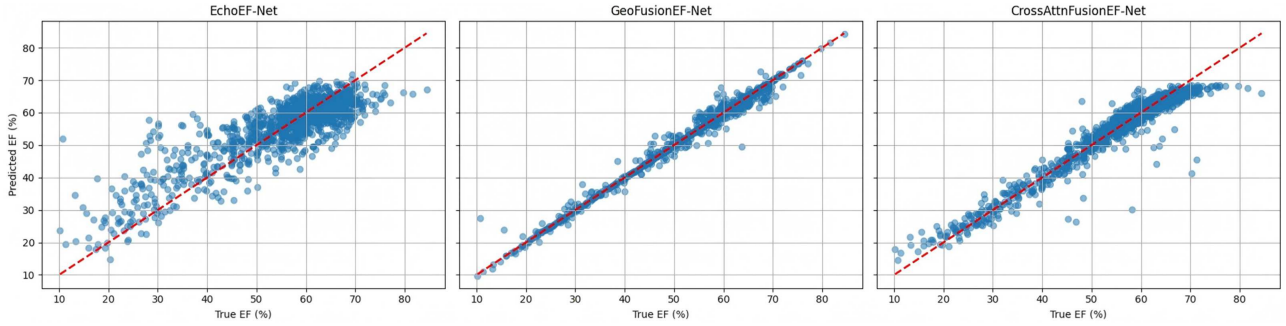


Figure 7
Bland–Altman analysis for agreement between predicted EF and ground truth EF

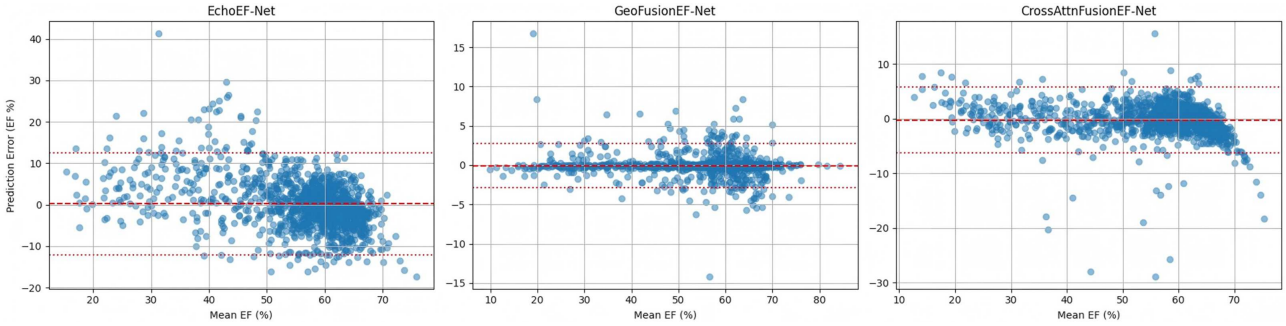


Figure 8
Error distribution plots across models

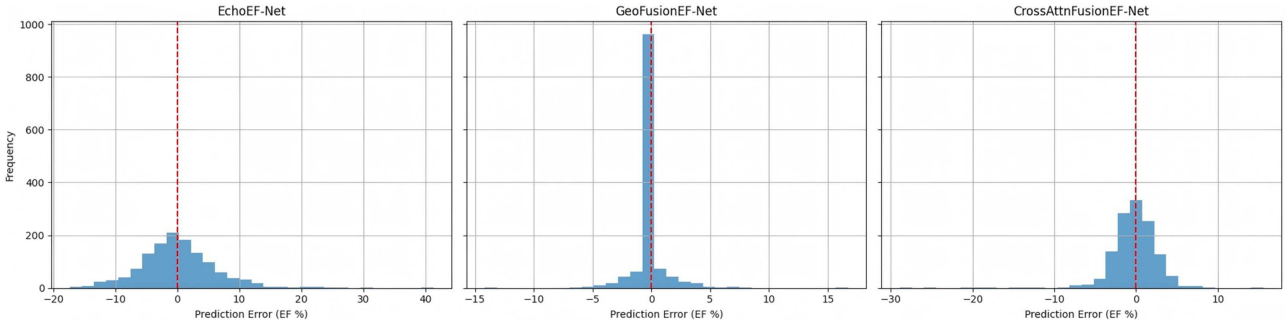


Figure 9
Agreement, calibration, and error distribution of EF prediction models

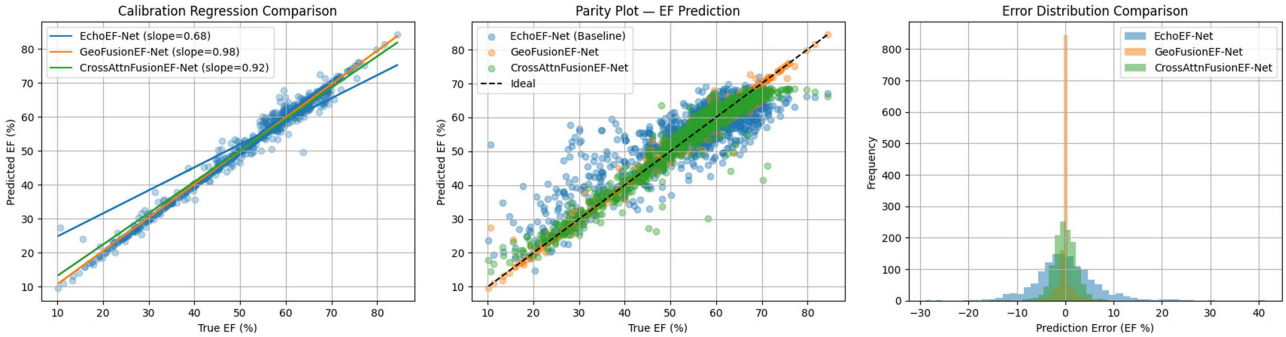
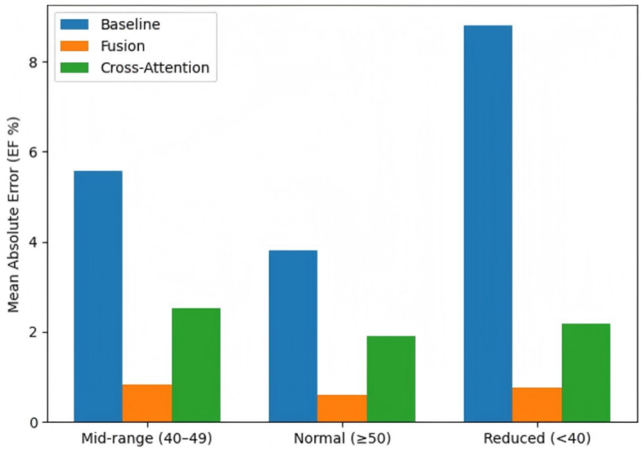


Table 6
Stratified EF estimation performance

EF category	EchoEF-Net	GeoFUsionEF-Net	CrossAttnFusionEF-Net
Reduced EF (<40%)	8.812	0.772	2.193
Mid-range EF (40–49%)	5.579	0.824	2.521
Normal EF (>=50%)	3.815	0.605	1.902

Figure 10
Stratified EF prediction error by clinical ranges



7.3. Ablation study

The ablation study results in Table 7 revealed that the video-only ablation model achieved an MAE of 4.614%, indicating that video-only is insufficient for accurate EF estimation. The geometry-only experiment model achieved an MAE of 2.023%, significantly outperforming the video-only EchoEF-Net.

Table 7
Fusion strategy ablation result

Fusion strategy	Model	MAE
Video Only	EchoEF-Net	4.614
Geometry Only	Geometry Branch	2.023
Late Fusion	LateFusionEF-Net	1.85
GatedFusion	GatedFusionEF-Net	3.067
Feature-Level Fusion	GeoFusionEF-Net	0.648
Cross-Attention Fusion	CrossAttnFusionEF-Net	1.99

GeoFusionEF-Net achieved the best overall performance of MAE 0.648%, outperforming more complex CrossAttnFusionEF-Net and GatedFusionEF-Net. The geometry-only model achieved an MAE of 2.03, while LateFusionEF-Net achieved a moderate improvement, and GatedFusionEF-Net showed degraded performance with an MAE of 3.067%. The cross-attention mechanism has performed well in large-scale multimodal learning problems; its effectiveness depends on the availability of sufficiently diverse complementary modalities. It also adds an additional layer of 1.05 million parameters, increasing optimization complexity and risk of overfitting. The gated fusion learned to suppress useful geometric information and did not effectively combine the two modalities. In the current study, geometrical descriptors provide more informative low-dimensional representations that are strongly related to LVEF. The results demonstrated that simple feature-level fusion is enough to integrate video and geometric information.

7.4. Statistical validation

To evaluate statistical robustness, 5000 bootstrap resamples were generated and used to estimate 95% confidence intervals (CIs) for MAE, RMSE, R^2, and Pearson correlation. Table 8 shows 95% bootstrap CIs of EF estimation models obtained from 5000 resampling iterations. The intervals quantify the uncertainty associated with each performance metric and provide an estimate of model stability across different samples. GeoFusionEF-Net achieved the narrowest CIs and the best predictive performance, indicating high robustness and low estimation uncertainty. CrossAttnFusionEF-Net also performed well compared to the EchoEF-Net baseline, while the baseline exhibits the largest uncertainty and prediction variability.

7.5. Statistical significance analysis

Table 9 presents both parametric and nonparametric paired statistical tests, which confirmed that the observed performance differences were statistically significant. The p-values obtained from both tests (all $p < 0.001$) indicate that observed performance differences between models are statistically significant. The comparison between EchoEF-Net and GeoFusionEF-Net yielded the largest effect size (Cohen's $d = 0.925$), demonstrating a substantial practical improvement through geometry-guided feature fusion. Observation between EchoEF-Net and CrossAttnFusionEF-Net, Cohen's $d = 0.587$, showed moderate improvement over video-only features. Cohen's $d = -0.560$ reflects the strong performance of GeoFusionEF-Net, showing that feature-level fusion provided more effective integration than the complex cross-attention mechanism.

7.6. Uncertainty analysis

Residual uncertainty analysis demonstrated substantially reduced prediction variability for GeoFusionEF-Net compared with the baseline model. Table 10 represents the residual uncertainty characteristics of the evaluated models. GeoFusinEF-Net demonstrated the lowest residual error, indicating the highest prediction stability and lowest uncertainty. All models exhibited minimal systematic prediction bias, demonstrating the robustness of the proposed feature-level fusion strategy.

7.7. Computational complexity

All proposed architectures exhibited similar computational complexity, as shown in Table 11, despite incorporating geometry-aware fusion modules compared with EchoEF-Net.

Table 8
Bootstrap confidence intervals (95%)

Model	MAE	RMSE	R^2	Pearson
EchoEF-Net	4.61(4.38–4.86)	6.29 (5.92–6.70)	0.735(0.700–0.766)	0.861 (0.840–0.879)
GeoFusionEF-Net	0.65 (0.58–0.72)	1.41 (1.21–1.64)	0.987(0.982–0.990)	0.993 (0.991–0.995)
CrossAttnFusionEF-Net	2.00 (1.87–2.13)	3.08(2.70–3.48)	0.937(0.918–0.952)	0.968 (0.959–0.976)

Table 9
Statistical significance analysis

Model	Paired *t*-test *p*-value	Wilcoxon *p*-value	Cohen's *d*
EchoEF-Net vs GeoFusionEF-Net	1.36×10^{-173}	5.98×10^{-182}	0.925
GeoFusionEF-Net vs CrossAttnFusionEF-Net	1.51×10^{-78}	1.22×10^{-131}	–0.56
EchoEF-Net vs CrossAttnFusionEF-Net	4.90×10^{-84}	4.13×10^{-102}	0.587

Table 10
Uncertainty analysis

Model	Residual mean	Residual SD	Residual variance	95^{th} percentile error
EchoEF-Net	–0.211	6.289	39.545	12.385
GeoFusionEF-Net	0.047	1.413	1.996	3.080
CrossAttnFusionEF-Net	0.250	3.071	9.431	4.981

Table 11
Computational complexity analysis

Model	Parameters(M)	GFLOPs	Inference time(ms/video)
EchoEF-Net	31.296	81.099	7.400
GeoFusionEF-Net	31.740	81.099	7.555
CrossAttnFusionEF-Net	32.528	81.099	7.475

The additional attention module increased the parameter count of CrossAttnFusionEF-Net by about 1.23M parameters, while inference latency remains similar.

7.8. Heart failure classification results

HF classification is performed using predicted EF values from EchoEF-Net, GeoFusionEF-Net, and CrossAttnFusionEF-Net under five different configurations. The first two configurations are EF predictions from individual models, the next configuration is the meta learner across all three models, and the last configuration is the fusion of geometric features with the meta learner. Table 12 shows the classification performance of the EchoEF-Net, yielding an accuracy of 0.87 and a macro-F1 of about 0.63, indicating a high error rate compared with the regression results. GeoFusionEF-Net classification results represent the benefits of geometry-guided EF estimation by achieving an accuracy of 0.97 and a macro-F1 of 0.95. The CrossAttnFusionEF-Net achieved strong results, with an accuracy of 0.95 and a macro-F1 of 0.89, demonstrating the benefits of multimodal feature integration. A meta-classifier combining the predicted EF values from three models improved classification performance, achieving an accuracy of 0.97 and a macro-F1 of 0.94. The fusion of geometric features with a meta-classifier achieved an accuracy of 0.97. This demonstrates that the GeoFusionEF-Net model alone effectively captures

Table 12
Heart failure stratification comparison results of various categories

Model configuration	Accuracy	Macro-F1	Weighted-F1	AUC (OvR)	Balanced Acc
HF-EchoEF-Net only	0.870	0.628	0.844	0.925	0.624
HF-GeoFusionEF-Net only	0.976	0.950	0.976	0.996	0.944
HF-CrossAttnFusionEF-Net only	0.952	0.890	0.951	0.985	0.881
HF-All3-MetaModel	0.974	0.945	0.974	0.996	0.939
HF-All3-Meta-Plus-Geom	0.970	0.934	0.970	0.995	0.927

geometric information, whereas EF estimation adds redundancy to HF classification. Both meta-classifiers benefit from geometric feature fusion, demonstrating model efficiency. The GeoFusionEF-Net has demonstrated strong clinical-relevance performance, achieving a sensitivity of 98.9% for HFrEF, 85.6% for HFmrEF, and 99.0% for HFpEF, with a consistent specificity of 94%, indicating fewer false positives.

Figures 11 and 12 show the confusion matrix and the Receiver Operating Characteristic (ROC) curve for all five configurations for heart failure classification. EchoEF-Net reveals misclassification for the HFmrEF class, which is derived from video-only EF results. GeoFusionEF-Net shows near-perfect discrimination in all classes, negligible off-diagonal errors, and unity of curves. CrossAttnFusionEF-net shows reliable results, showing slight confusion in mid-range EF, which is very hard to classify clinically. The meta classifier reduces errors and shows balanced performance across all heart failure categories, with high sensitivity and specificity. A meta classifier with geometric fusion shows additional improvement, indicating that EF prediction has already been learned within EF prediction models. Hence, errors concentrated in the HFmrEF category, because HFmrEF (40–49%) is a transitional state between reduced EF and preserved EF, overlapping ventricular function and remodelling patterns, leading to misclassification into other heart failure categories.

8. Comparative study with state-of-the-art methods

Recent deep learning approaches for automated EF estimation from Echo videos have shown promising results, as mentioned in Table 13. EchoNet-Dynamic established a state-of-the-art (SOTA) method for beat-level video analysis, achieving an MAE of 4.1%; a transformer-based model, EchoCoTr, achieved an MAE of 3.89%. A very recent study, ViViEchoFormer, reported an error rate of 2.83%, and CoReEcho recorded a 3.02% error rate in estimating EF through enhanced spatiotemporal modeling. In comparison, GeoFusionEF-Net drastically improves predictive accuracy and reduces the error rate to an MAE of 0.65%. By integrating video features with LV geometrical features, the multimodal representation of CrossAttnFusionEF-Net achieves an MAE of 1.99%, successfully incorporating geometrical guidance for motion-based analysis. These proposed model results indicate that integrating geometric or physiological features with motion-based representations strengthens data-driven prediction and physiological assessment of cardiac function for EF estimation. Most EF estimation methods rely solely on echocardiographic videos to learn cardiac motion patterns; in contrast, the present study derives geometric features from available LV tracings to incorporate physiological structure alongside cardiac motion information. This design improves robustness to image variability and reduces ambiguity by explicitly combining ventricular morphology and functional dynamics to estimate EF. This study incorporates explainability analysis to verify that predictions originate from clinically relevant LV structures, thereby supporting transparency and trust in clinical AI systems.

9. Explainable AI

9.1. Grad-CAM

To improve interpretability and clinical trust, we employed Grad-CAM on the final convolutional layers of each network to

Figure 11
Confusion matrix and Receiver Operating Characteristic (ROC) curve of EF prediction models for heart failure classification

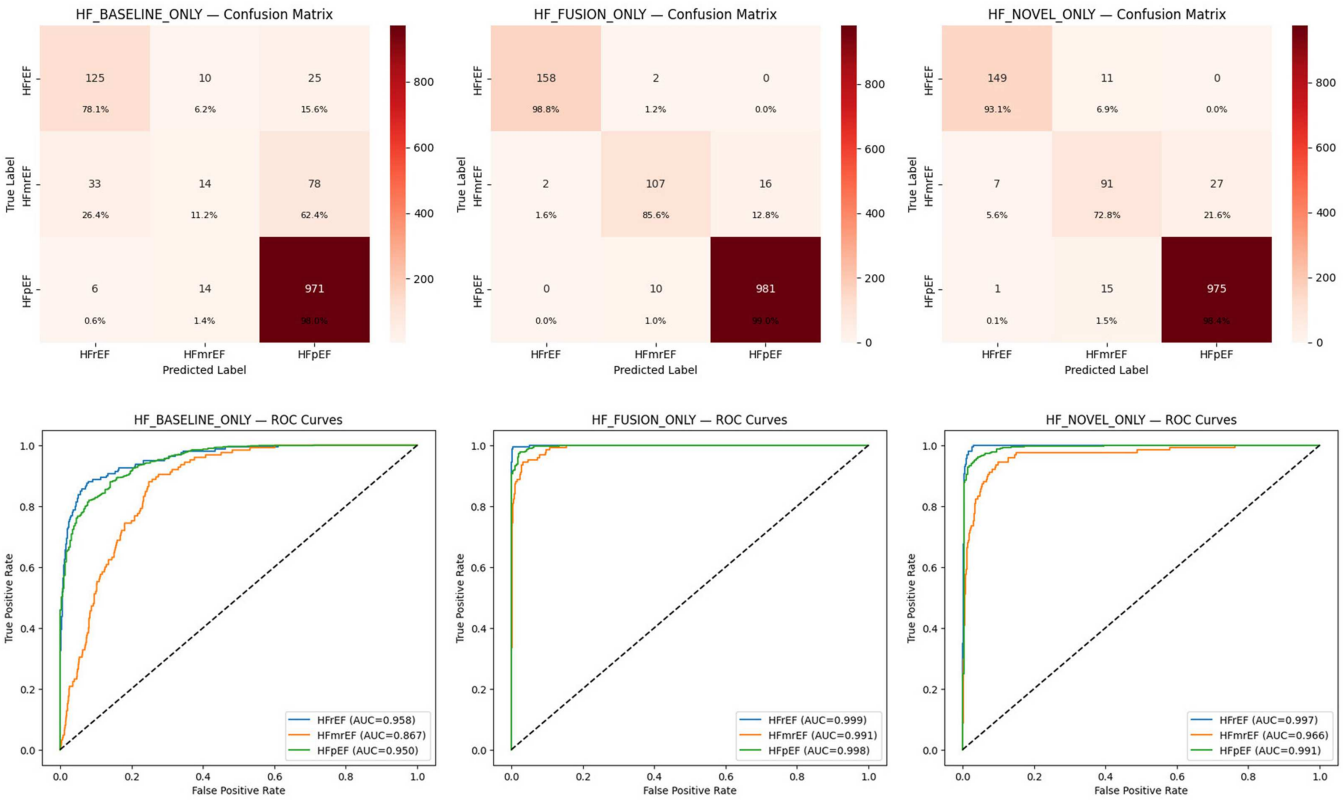


Figure 12
Confusion matrix and ROC curve of the meta-classifier model for heart failure classification

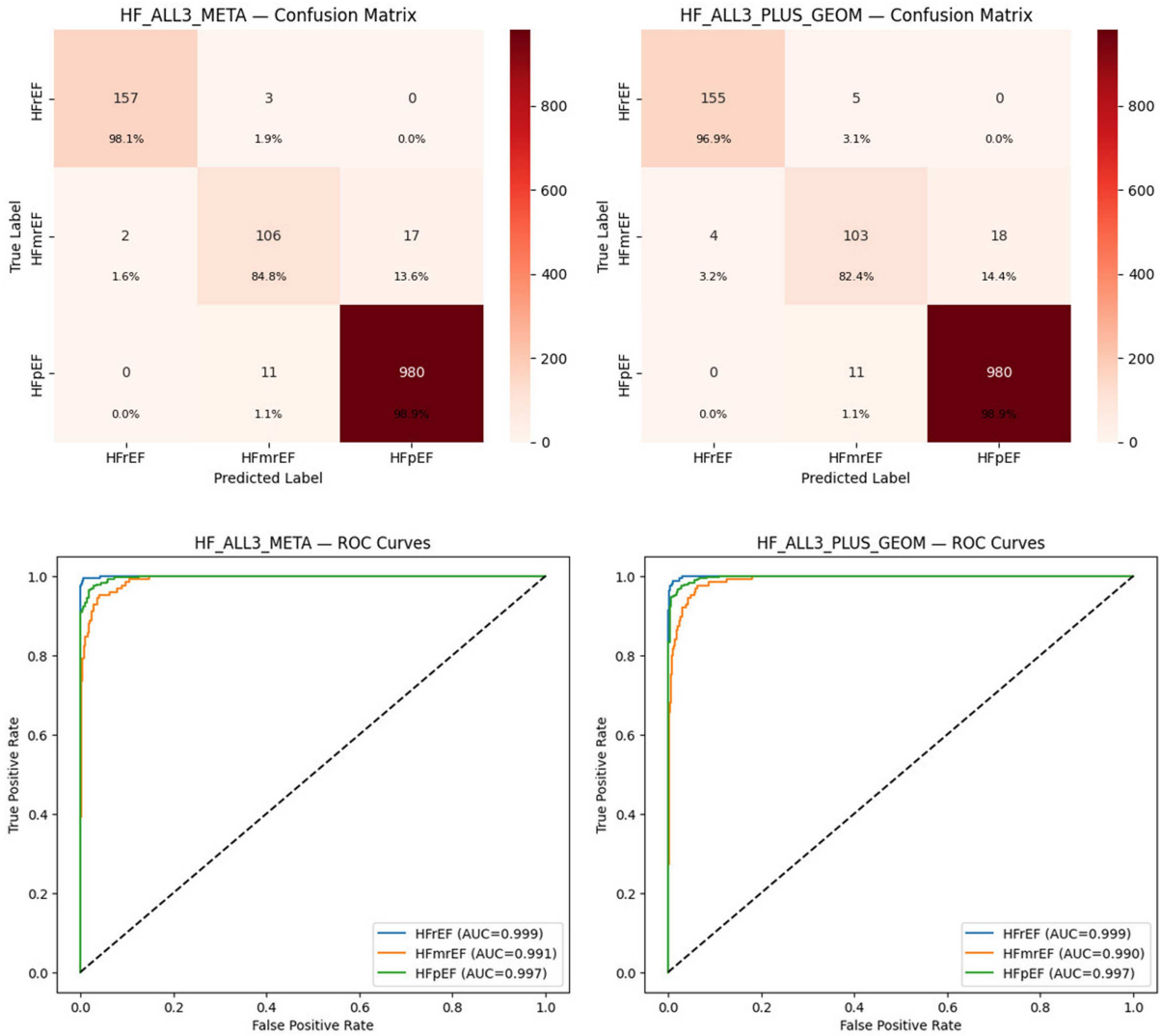


Table 13
Comparison of SOTA methods for EF estimation

Model	Dataset	Method	MAE (%)	R^2	Correlation
Ouyang et al. [5]	EchoNet-Dynamic Dataset	R(2 + 1) D CNN	4.1	0.74	0.93
EchoCoTr [13]		Transformer-based video encoder	3.89	0.79	–
ViViEchoFormer [12]		Transformer-based video regressor	2.83	–	0.95
CoReEcho [8]		Continuous representation learning	3.02	–	–
Frazry et al. [25]		Hierarchical vision transformer	3.5	–	–
Smistad et al. [7]	Clinical Echo Exams private dataset	CNN	6.3	–	–
Snapshot AI [26]	Multi-institutional echo images	Single frame	4.0	–	0.89

(*Continued*)

Table 13
(*Continued*)

Model	Dataset	Method	MAE (%)	R^2	Correlation
		Proposed model			
EchoEF-Net	EchoNet-Dynamic Video only	R(2 + 1)D CNN	4.61	0.73	0.86
GeoFusionEF-Net	EchoNet-Dynamic dataset Video + geometry	Multimodal deep learning fusion network	0.65	0.98	0.99
CrossAttnFusionEF-Net	EchoNet-Dynamic dataset Video + geometry	Cross-attention multimodal deep learning network	1.99	0.93	0.97

visualize critical regions that contribute to EF estimation across all three models. Figure 13 visualizes that EchoEF-Net diffuse attention activations across the cardiac chamber and surrounding tissues, suggesting meaningful anatomy is subordinated to broad intensity patterns. GeoFusionEF-Net focused on activations with endocardial borders and in the LV cavity, which are crucial areas that influence LV volume changes, demonstrating clinically meaningful spatial attention. CrossAttnFusionEF-Net shows excellent spatial selectivity with strong signals contained within the LV cavity while reducing irrelevant regions, suggesting improved spatial selectivity. Hence, geometry-guided fusion improves anatomical localization and raises the EF estimation accuracy.

9.2. SHAP

To ensure transparency, EF estimation is a clinically crucial measure in diagnosing HF. Black box predictions cannot be acceptable to clinicians. In this regard, we experimented with SHAP for the geometry branch, as shown in Figures 14 and 15, which illustrate how geometric features contribute to EF estimation. The results show that end-diastolic area (ED_area) and end-systolic area (ES_area) dominate the estimation of EF in GeoFusionEF-Net. ED_area and ES_area show a direct relationship with LV volume changes, whereas Sphericity and FAC contribute minimally, indicating that area-based metrics provide the most crucial anatomical information for EF prediction.

Figure 13
Grad-CAM visualizations highlighting the region influencing EF estimation for EF prediction models

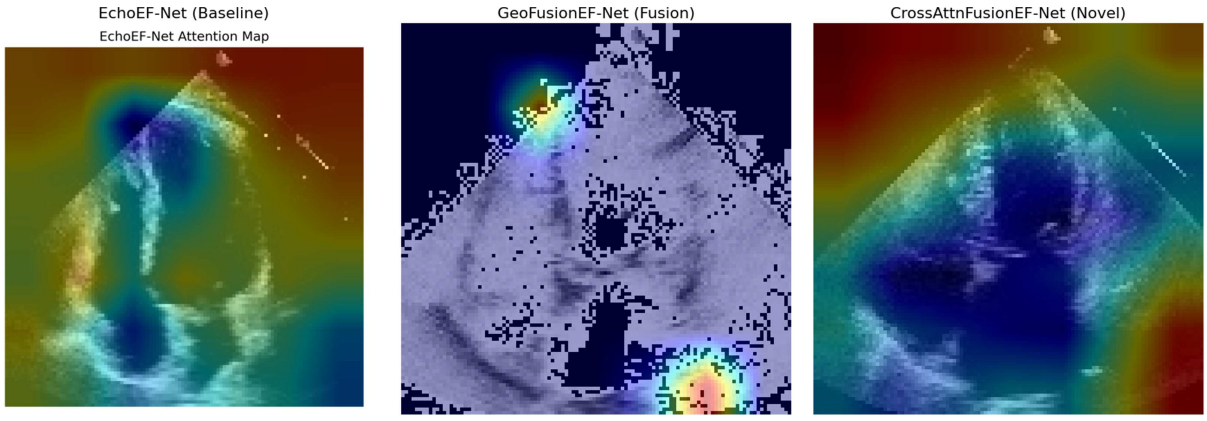


Figure 14
SHAP summary plots highlighting feature contribution for the GeoFusionEF-Net model

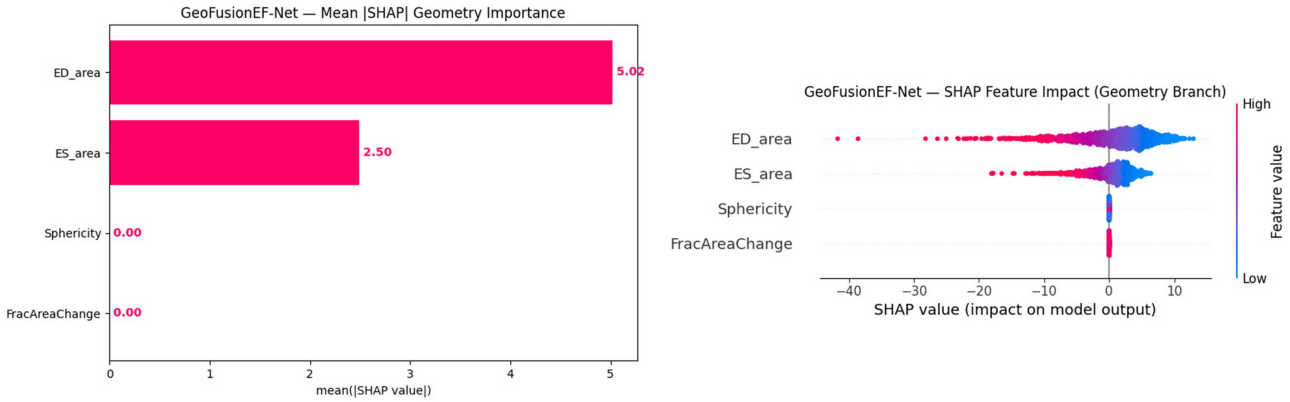


Figure 15
SHAP summary plots highlighting feature contribution for the CrossAttnFusionEF-Net model

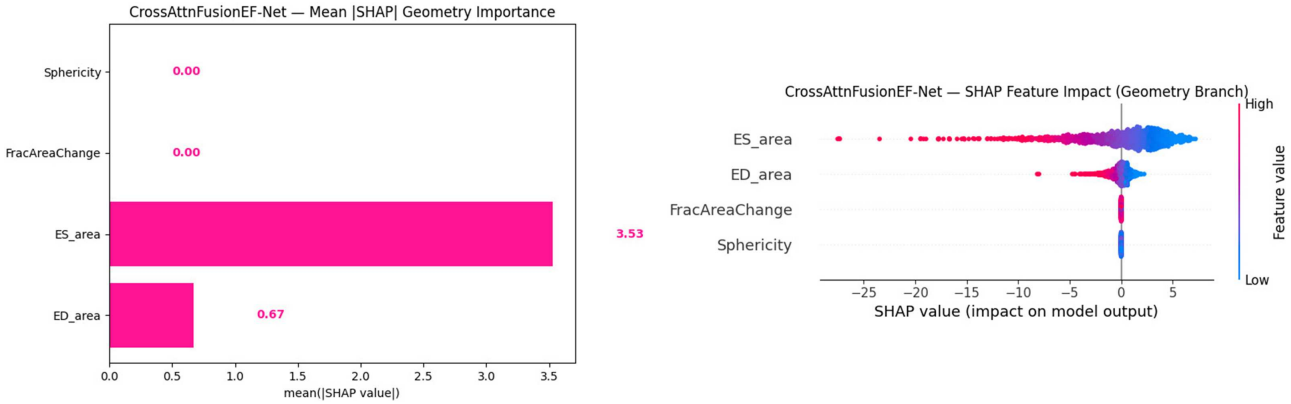
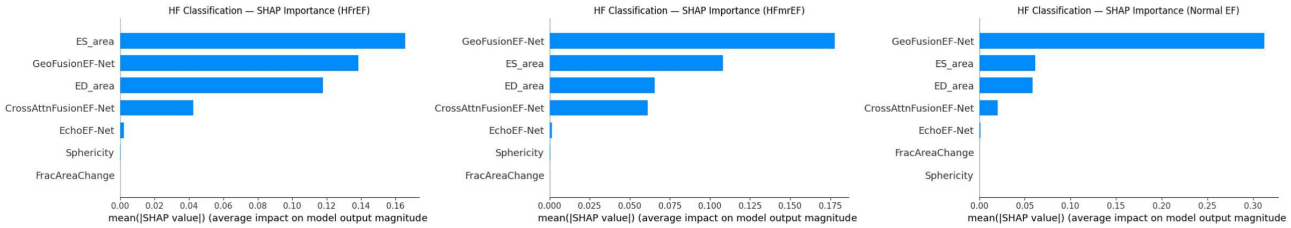


Figure 16
SHAP importance features for the HF_All3_ Meta_Plus_Geom HF classification model



The ES_area contributes more in CrossAttnFusionEF-Net, which shows that attention-guided fusion highlights contraction-related anatomical information.

To understand the model's decisions and clinical interoperability, and to identify the features that contributed to HF classification, explainable AI methods are used. The meta-classifier's decision for the HF category is explained using SHAP, as shown in Figure 16, for the final HF classification model. The results show that geometrical features provide additional structural information that improves classification and that EF estimation models serve as the primary discriminative signal for classification.

10. Discussion

Experimental results show that anatomical information greatly improves the EF estimation performance compared with video-only learning. EchoEF-Net achieved competitive performance by learning spatiotemporal cardiac motion patterns directly from echo videos, but prediction errors increased substantially in patients with impaired ventricular function. GeoFusionEF-Net, which integrated both geometric features and video representations, consistently outperformed in terms of MAE across all EF ranges. This indicates that anatomical measurements provide complementary information that improves robustness beyond what video information alone provides.

The geometrical features ES_{Area}, ED_{Area}, FAC, and $Sphericity$ provide the information of ventricular size, contractility, and morphological deformation during the cardiac cycle, which are physiologically related to ventricular function and therefore strongly related to EF. Since geometrical features are derived from expert-provided ventricular tracings, the relationship with EF has to be carefully considered. The superior results of GeoFusionEF-Net are attributed to the tracing-based geometric descriptors used in this study. However, results for the geometry-only modality showed that neither modality alone could match the performance of the multimodal framework. This indicates that the proposed model benefits from geometrical and video features rather than directly reconstructing EF from geometric measurements. Thus, GeoFusionEF-Net can be viewed as a geometry-assisted EF estimation method.

The ablation study on geometric features alone demonstrated the predictive value of these features, whereas fusion experiments showed that integrating geometry with video features yields significantly better performance than either modality alone. Feature -level fusion obtained the best overall performance among fusion methods evaluated, whereas CrossAttnFusionEF-Net was able to model the interactions between video and geometrical features; it did not outperform GeoFusionEF-Net. This suggests that the feature space used in this study may not require a complex attention mechanism for effective fusion. The additional flexibility introduced in cross-attention probably complicated the optimization and did not provide sufficient information for accurate predictions; however, it outperformed the video-only model.

The analysis shows that EchoEF-Net demonstrates spatiotemporal motion only through beat clips and is not reliable for more accurate EF estimation. GeoFusionEF-Net strongly reduces EF error by leveraging geometric features, demonstrating the strength of structural cardiac cycle information for estimating EF. While CrossAttnFusionEF-Net, which performs adaptive feature weighting on video, achieves marginal results compared to GeoFusion and superior results compared to the EchoEF-Net model, this suggests that the attention mechanism requires a larger dataset to fully learn the structural information. The overall results show that GeoFusionEF-Net provides reliable estimates of EF in ventricular function. The proposed fusion models leverage

geometric features, namely, expert-annotated ventricular contours. This reflects the usage of high-quality annotations available in the dataset to enable controlled evaluation of geometry-guided learning.

To enhance model interpretability, the proposed framework maps HF classification from EF estimates into three classes: HFrEF, HFmrEF, and HFpEF. Fusion-integrated EF models demonstrated strong performance in classifying HF categories, highlighting the clinical value of EF estimation. Despite strong results from the classification models, HRmfEF remained a challenging class, as expected given its transitional physiology, and the inclusion of geometric features has improved classification performance. Moreover, Grad-CAM and SHAP analyses enhance the models' transparency by highlighting clinically meaningful regions and features that affect the predictions.

10.1. Limitations

Despite the strong results of the models, several limitations should be noted. First, the experiments were primarily performed on a single EchoNet-Dynamic dataset; external validation on an independent dataset from a different center is required to further assess generalizability across imaging methods, machine types, and patient populations. The geometric feature extraction used in fusion models relies on dataset-provided LV tracings. Therefore, they are dependent on the availability and quality of tracings. The dataset is originally from a single source and relies on manually annotated LV tracings used in GeoFusionEF-Net. Inaccuracies in the annotations or discrepancies in manual tracing could introduce noise that affects predictions. CrossAttnFusionEF-Net introduces additional complexity and hyperparameter sensitivity, which limit its generalizability to automated clinical pipelines. Future work will focus on automated geometry extraction and real-time deployment in clinical environments.

11. Conclusion

This study introduces a motion-aware, anatomically guided framework for estimating LVEF and classifying HF from echocardiographic video. The EchoEF-Net model predicted EF using only spatiotemporal cardiac information from beat-level videos. The integration of geometrical descriptors derived from expert LV tracings in the dataset has achieved very good performance. The proposed GeoFusionEF-Net shows nearly perfect alignment with the ground truth, achieving an MAE of 0.648 and r = 0.993, thereby representing ventricular structural information and reducing the error of the video-only model. CrossAttnFusionEF-Net also achieved clinically interpretable results by improving motion–structure interactions and spatial selectivity. The estimated EF results are used to classify HF categories, with fusion-based models showing diagnostic performance. Explainability using Grad-CAM and SHAP shows that the model captured only clinically relevant LV regions, with the end-systolic and end-diastolic areas contributing more to EF estimation and HF classification, indicating physiologically meaningful behavior. The overall results indicate that integrating LV structural descriptors along with spatiotemporal cardiac motion provides a clinically reliable framework that enhances EF estimation and improves HF classification reliability.

Recommendations

The results showed that incorporating geometric features and multimodal fusion for EF estimation improves accuracy. Research studies need to focus on developing an automated pipeline and a dependable segmentation model for geometric feature extraction. The developed model can be used in real-time clinical settings and on portable ultrasound systems to quickly assess cardiac function.

Acknowledgement

We acknowledge the developers of the publicly available EchoNet-Dynamic dataset used in this study.

Ethical Statement

The authors declare that this study does not involve any direct human or animal experimentation and doesn't require formal ethical approval for secondary analysis, as it is based on a publicly available anonymized echocardiographic dataset (EchoNet-Dynamic). As the data were de-identified, it does not require Institutional Review Board approval or informed consent. The dataset complies with institutional and international ethical standards.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in the EchoNet-Dynamic repository at https://stanfordaimi.azurewebsites.net/datasets/834e1cd1-92f7-4268-9daa-d359198b310a.

Author Contribution Statement

Chaithra C. S.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Siddesha S.:** Conceptualization, Validation, Formal analysis, Investigation, Writing – review & editing, Supervision, Project administration. **Vinay Kumar N.:** Validation, Formal analysis, Writing – review & editing. **V. N. Manjunath Aradhya:** Validation, Formal analysis, Writing – review & editing.

References

[1] World Heart Federation. (2025). *World Heart Report 2025*. https://world-heart-federation.org/resource/world-heart-report-2025/

[2] Rosano, G. M., Teerlink, J. R., Kinugawa, K., Bayes-Genis, A., Chioncel, O., Fang, J., . . . , & Coats, A. J. (2025). The use of left ventricular ejection fraction in the diagnosis and management of heart failure. A clinical consensus statement of the Heart Failure Association (HFA) of the ESC, the Heart Failure Society of America (HFSA), and the Japanese Heart Failure Society (JHFS). *Journal of Cardiac Failure*, *27*(7), 1174–1187. https://doi.org/10.1016/j.cardfail.2025.03.014

[3] Lam, C. S., & Solomon, S. D. (2021). Classification of heart failure according to ejection fraction: JACC review topic of the

week. *Journal of the American College of Cardiology*, *77*(25), 3217–3225. https://doi.org/10.1016/j.jacc.2021.04.070

[4] Alhakak, A. S., Teerlink, J. R., Lindenfeld, J., Böhm, M., Rosano, G. M., & Biering-Sørensen, T. (2021). The significance of left ventricular ejection time in heart failure with reduced ejection fraction. *European Journal of Heart Failure*, *23*(4), 541–551. https://doi.org/10.1002/ejhf.2125

[5] Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C. P., . . . , & Zou, J. Y. (2020). Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, *580*(7802), 252–256. https://doi.org/10.1038/s41586-020-2145-8

[6] Szijártó, Á., Merkely, B., Kovács, A., & Tokodi, M. (2025). Deep learning-enabled echocardiographic assessment of biventricular ejection fractions: The dual-task QUEST-EF model. *European Heart Journal-Cardiovascular Imaging*, *26*(8), 1402–1405. https://doi.org/10.1093/ehjci/jeaf147

[7] Van De Vyver, G., Måsøy, S. E., Dalen, H., Grenne, B. L., Holte, E., Olaisen, S. H., . . . , & Smistad, E. (2025). Regional image quality scoring for 2-D echocardiography using deep learning. *Ultrasound in Medicine & Biology*, *51*(4), 638–649. https://doi.org/10.1016/j.ultrasmedbio.2024.12.008

[8] Maani, F. A., Saeed, N., Matsun, A., & Yaqub, M. (2024). Coreecho: Continuous representation learning for 2d+ time echocardiography analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 591–601. https://doi.org/10.1007/978-3-031-72083-3_55

[9] Rahman, S., Haque, R., Swapno, S. M. R., Islam, M. B., & Nobel, S. N. (2023). Deep learning-based left ventricular ejection fraction estimation from echocardiographic videos. In *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques*, 1–6. https://doi.org/10.1109/EASCT59475.2023.10392607

[10] Luong, C. L., Jafari, M. H., Behnami, D., Shah, Y. R., Straatman, L., Van Woudenberg, N., . . . , & Tsang, T. (2024). Validation of machine learning models for estimation of left ventricular ejection fraction on point-of-care ultrasound: Insights on features that impact performance. *Echo Research & Practice*, *11*(1), 9. https://doi.org/10.1186/s44156-024-00043-2

[11] Kusunose, K., Zheng, R., Yamada, H., & Sata, M. (2022). How to standardize the measurement of left ventricular ejection fraction. *Journal of Medical Ultrasonics*, *49*(1), 35–43. https://doi.org/10.1007/s10396-021-01116-z

[12] Akan, T., Alp, S., Bhuiyan, M. S., Helmy, T., Orr, A. W., Bhuiyan, M. M. R., . . . , & Bhuiyan, M. A. N. (2025). ViViEchoformer: Deep video regressor predicting ejection fraction. *Journal of Imaging Informatics in Medicine*, *38*(4), 2041–2052. https://doi.org/10.1007/s10278-024-01336-y

[13] Muhtaseb, R., & Yaqub, M. (2022). Echocotr: Estimation of the left ventricular ejection fraction from spatiotemporal echocardiography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 370–379. https://doi.org/10.1007/978-3-031-16440-8_36

[14] Thomas, S., Cao, Q., Novikova, A., Kulikova, D., & Ben-Yosef, G. (2024). EchoNarrator: Generating natural text explanations for ejection fraction predictions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 634–644. https://doi.org/10.1007/978-3-031-72083-3_59

[15] Mandal, S., Ghosh, S., Das, S., & Mukhopadhyay, J. (2026). Automated ejection fraction estimation from echocardiogram videos using foundation models. In *2026 9th International Conference on Electronics, Materials Engineering & Nano-Technology*, *9*, 1–6. https://doi.org/10.1109/IEMENTech202669403.2026.11434251

[16] McDonagh, T. A., Metra, M., Adamo, M., Gardner, R. S., Baumbach, A., Böhm, M., ..., & Group, E. S. D. (2023). 2023 focused update of the 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC. *European Heart Journal*, *44*(37), 3627–3639. https://doi.org/10.1093/eurheartj/ehad195

[17] Savarese, G., Gatti, P., Benson, L., Adamo, M., Chioncel, O., Crespo-Leiro, M. G., . . . , & Lund, L. H. (2024). Left ventricular ejection fraction digit bias and reclassification of heart failure with mildly reduced vs reduced ejection fraction based on the 2021 definition and classification of heart failure. *American Heart Journal*, *267*, 52–61. https://doi.org/10.1016/j.ahj.2023.11.008

[18] Weiss, A., Flowers, A., Eisenberg, S. D., Wu, L., Rodriguez, M., Khawaja, M., . . . , & Krittanawong, C. (2025). A contemporary review on heart failure with improved ejection fraction. *npj Cardiovascular Health*, *2*(1), 56. https://doi.org/10.1038/s44325-025-00093-3

[19] Dini, F. L., Cameli, M., Stefanini, A., Aboumarie, H. S., Lisi, M., Lindqvist, P., & Henein, M. Y. (2024). Echocardiography in the assessment of heart failure patients. *Diagnostics*, *14*(23), 2730. https://doi.org/10.3390/diagnostics14232730

[20] Nargesi, A. A., Adejumo, P., Dhingra, L. S., Rosand, B., Hengartner, A., Coppi, A., . . . , & Khera, R. (2025). Automated identification of heart failure with reduced ejection fraction using deep learning-based natural language processing. *Heart Failure*, *13*(1), 75–87. https://doi.org/10.1016/j.jchf.2024.08.012

[21] Decoodt, P., Sierra-Sosa, D., Anghel, L., Cuminetti, G., De Keyzer, E., & Morissens, M. (2024). Transfer learning video classification of preserved, mid-range, and reduced left ventricular ejection fraction in echocardiography. *Diagnostics*, *14*(13), 1439. https://doi.org/10.3390/diagnostics14131439

[22] Sidie, H., Wen, Z., Yidi, Z., Yun, L., & Hao, L. (2025). Risk stratification and survival prediction in heart failure: From grades to scores. *Frontiers in Cardiovascular Medicine*, *12*, 1–10. https://doi.org/10.3389/fcvm.2025.1676441

[23] Zhang, W., Zhao, H., Tian, Z., Fu, W., Li, Z., Geng, Z., . . . , & Zheng, M. (2026). Machine learning–based risk stratification identifies heart failure with preserved ejection fraction as an independent predictor of adverse outcomes in hypertrophic cardiomyopathy. *Scientific Reports*, *16*, 1–14. https://doi.org/10.1038/s41598-026-46573-z

[24] Ouyang, D., He, B., Ghorbani, A., Lungren, M. P., Ashley, E. A., Liang, D. H., & Zou, J. Y. (2019). Echonet-dynamic: A large new cardiac motion video data resource for medical machine learning. *In NeurIPS ML4H Workshop*, *5*, 2.

[25] Fazry, L., Haryono, A., Nissa, N. K., Hirzi, N. M., Rachmadi, M. F., & Jatmiko, W. (2022). Hierarchical vision transformers for cardiac ejection fraction estimation. In *2022 7th International Workshop on Big Data and Information Security*, *9*, 1–6. https://doi.org/10.1109/IWBIS56557.2022.9924664

[26] Malins, J. G., Anisuzzaman, D. M., Jackson, J. I., Lee, E., Naser, J. A., Rostami, B., . . . , & Attia, Z. I. (2025).

Snapshot artificial intelligence—Determination of ejection fraction from a single frame still image: A multi-institutional, retrospective model development and validation study. *The Lancet Digital Health*, *7*(4), 255–263. https://doi.org/10.1016/j.landig.2025.02.003

How to Cite: S., C. C., S., S., N., V. K., & Aradhya, V. N. M. (2026). Beat-Level Echocardiographic Clip Analysis with Geometry-Video Fusion and Cross-Attention for Accurate Ejection Fraction Estimation and Heart Failure Stratification. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62029883