# A Review of CardiacField for Left Ventricular Ejection Fraction Estimation from Echocardiography

Eda Özkaynar
Department of Electrical and Electronics Engineering
Middle East Technical University, Ankara, Türkiye
eda.ozkaynar@metu.edu.tr

***Abstract*—Left ventricular ejection fraction (LVEF) is a critical metric for assessing cardiac function. One of the most commonly used imaging methods for LVEF assessment is echocardiography. However, determining the ejection fraction through echocardiography can be affected by variability between different observers or within the same observer over time. The accuracy of this assessment is often dependent on the operator's expertise. To overcome these challenges, previous studies have developed various deep learning techniques. Most of the techniques use 2D echocardiography images to estimate LVEF. However, CardiacField approaches the task as an inverse problem by reconstructing a patient-specific 3D heart model from multiple 2D views.This reconstruction allows for direct volume calculation, resulting in more precise and physiologically consistent EF estimation.**

## I. Clinical Motivation and Background

Heart failure is a serious health problem affecting more than 26 million people worldwide, and its prevalence is expected to increase as the global population ages. In the United States, heart failure with reduced ejection fraction accounts for nearly half of all heart failure cases [3].

Heart failure occurs when the heart is no longer able to pump blood effectively enough to meet the body's needs. As a result, tissues may not receive adequate oxygen and nutrients [6]. Common symptoms include fatigue, shortness of breath, and other related issues [3].

Clinically, heart failure is typically classified into two main types: with reduced ejection fraction (HFrEF) and with preserved ejection fraction (HFpEF). Ejection fraction (EF) represents the percentage of blood volume pumped out of the left ventricle with each heartbeat. For example, if 70% of the blood in the ventricle is ejected, the EF is 70%. Normal EF values range between 52–72% for men and 54–74% for women. An EF of 40% or lower is generally considered sufficient to diagnose HFrEF. EF is often measured from the left ventricle because right-sided heart failure is frequently a consequence of left-sided dysfunction [3].

LVEF can be measured using various imaging methods, but echocardiography is the most common non-invasive option. During an echocardiographic evaluation, end-diastolic volume (EDV) and end-systolic volume (ESV) are obtained, and EF is calculated using the following formula:

$$EF = \frac{\text{EDV - ESV}}{\text{EDV}} \times 100 \tag{1}$$

To estimate LVEF correctly, it is essential to accurately detect the endocardial border of the left ventricle. This can be done manually, semi-automatically, or fully automatically. Manual measurements can vary significantly depending on who performs them, as different operators may trace the borders differently or use different techniques. In automatic or semi-automatic methods, variations usually result from algorithmic differences. Furthermore, identifying the correct end-diastolic and end-systolic frames is critical. Accurate EF estimation depends on having sufficient temporal resolution to capture the maximum and minimum ventricular volumes [5].

To deal with the challenges of accurate and reliable ejection fraction estimation, several studies have proposed deep learning–based approaches using two-dimensional echocardiography (2DE), typically limited to a single apical four-chamber (A4C) view. For example, Ouyang et al. used a DeepLabV3 network to segment the left ventricle and then predicted EF using a 3D convolutional R2+1D model trained on video clips [8]. The final EF value was obtained by combining the outputs of these two models. Similarly, Batool et al. extracted clinically relevant features from segmented A4C images and used them as inputs to both deep learning and classical machine learning models for LVEF prediction [9].

Both of these methods were developed and evaluated using the EchoNet-Dynamic dataset [7], a large public dataset provided by the Stanford AIMI Center. This dataset contains 10,030 echocardiography videos, all captured using a single A4C view. Using a fixed view makes it easier to collect data and train models. However, it limits our understanding of space and makes it harder to apply our findings to different situations or understand the complex 3D shape of the heart. These issues show that we need better methods for estimating ejection fraction (EF) that consider spatial awareness more effectively.

### A. Clinical Background

Various methods have been developed for estimating the left ventricular ejection fraction (LVEF), depending on the echocardiographic imaging technique used. The estimation of left ventricular volume requires different mathematical models and formulas according to the imaging modality [5]. The most commonly used technique in clinical practice is the biplane method of disks (also known as the modified Simpson's method), which is recommended by the American Society of Echocardiography [10].

In this approach, the endocardial boundaries of the left ventricle are carefully traced in both apical four-chamber (A4C) and apical two-chamber (A2C) views, at both end-diastolic and end-systolic phases. Then, the ventricle is virtually divided

along its long axis into approximately 20 thin disks. Using the diameter and thickness of each disk, the end-diastolic volume (EDV) and end-systolic volume (ESV) are computed. Finally, the EF is calculated using Equation 1. As a result, the modified Simpson's method provides an approximation of the three-dimensional ventricular volume using a limited number of two-dimensional cross-sectional images.

## II. Related Work

Early deep learning approaches for EF prediction from echocardiography focused on segmenting the left ventricle and calculating volumes using rule-based or regression models. For instance, EchoNet-Dynamic [7] employs a DeepLabV3 model with ResNet and ASPP to segment end-systolic and end-diastolic frames. A second network with spatiotemporal convolutions then estimates EF from 32-frame clips using test-time augmentation. Batool et al. [9] extract structural features from segmentation masks and use Simpson's method to compute EDV/ESV, followed by regression models such as SVR, RF, or LSTM.

Beyond segmentation-based models, some recent works eliminate the segmentation step and directly regress EF from echocardiographic video data. Reynaud et al. [11] propose UltraSound Video Transformers (UVT), where video frames are encoded using a ResNet autoencoder and passed to a Transformer-based model to predict EF. Fazry et al. [4] introduce UltraSwin, a 3D patch-based Transformer model that leverages spatiotemporal video tokens instead of individual frames to achieve robust EF prediction.

Segmentation-based methods are easy to understand but can make mistakes and only work from one perspective. Transformer-based methods perform well but do not clearly show the structure of the heart. Because of these issues, researchers developed methods like CardiacField. This method uses several 2D views and physically informed modeling to create a 3D model of the heart, leading to accurate and reliable estimates of ejection fraction (EF).

## III. Our Implementation and Experiments

In our project, we started by experimenting with the original EchoNet-Dynamic framework [7], using the publicly available implementation for segmentation and ejection fraction (EF) prediction from A4C echocardiography videos. To build upon this, we reimplemented the method described by Batool et al. [9], which involved segmenting the left ventricle with a DeepLabV3 model and extracting features based on the modified Simpson's method. These features, such as Simpson's discs and left ventricle (LV) axis length, were then used as inputs for several regression models, including RNN and LSTM, which we trained from scratch.

In addition, we explored a Transformer-based approach by adapting the ViViT (Video Vision Transformer) architecture [1] to directly predict EF from raw videos without requiring segmentation. Although the ViViT model showed potential, our results were heavily impacted by data imbalance in the EchoNet-Dynamic dataset, with most samples clustered around the normal EF range (55–70%). This made it difficult for the model to generalize across lower or higher EF values, especially in clinical edge cases.

Overall, our implementations helped us understand the practical challenges associated with segmentation-based and end-to-end learning approaches, especially when working with limited or imbalandced datasets. Based on our observations, we decided to look into CardiacField. This is a self-supervised method that uses information from different views and a physics-based model to estimate ejection fraction (EF). It does this without needing segmentation or a lot of labeled data.

## IV. Review of CardiacField

CardiacField is a novel computational echocardiography framework that aims to estimate left and right ventricular ejection fractions (LVEF and RVEF) using only two-dimensional (2D) ultrasound probes.This method solves important problems in current ejection fraction estimation methods, which often rely on precise segmentation, single-view data, and lack of generalizability due to data-driven training. Unlike many past studies that use supervised learning or fixed shape models, CardiacField uses a self-supervised implicit neural representation (INR) network. This network builds a customized 3D model of the heart from multiple 2D echocardiogram views, guided by physically-informed constraints.

## V. Method Overview

CardiacField consists of three main components: the physical imaging model for 2D echocardiography (2DE) acquisition, the representation of positional parameters, and the implicit neural representation network, which is guided by a physics-informed loss function.

First, they develop the physical imaging model for 2DE acquisition. Next, they specify the representation of positional parameters and the image selection strategy that will be used within this model. Finally, they construct the implicit neural representation of a 3D heart and design a physics-informed loss function based on the imaging model to effectively guide the training of CardiacField.

### A. Physical Imaging Model of 2DE Acquisition

Before constructing the neural representation for the 3D cardiac volume, CardiacField introduces a forward imaging model that describes how two-dimensional echocardiographic (2DE) images are generated from the 3D heart. This model relies on two coordinate systems: a 3D world coordinate system to describe the heart's spatial structure, and a 2D image coordinate system to describe each ultrasound frame.

The heart is represented as a continuous intensity function $O(\vec{X})$, where $\vec{X} = [x, y, z]^T$ denotes a point in 3D space and $O(\vec{X})$ returns the intensity at that location.

Each 2D ultrasound image can be interpreted as a cross-sectional slice through this 3D volume. This process is described by two operators:

- **Slicing operator** ($\mathcal{S}_M^N$): This operator reduces an $N$-dimensional function to $M$ dimensions by zeroing out the final $N - M$ coordinates. For example, $\mathcal{S}_2^3 \circ f(x_1, x_2, x_3) = f(x_1, x_2, 0)$.
- **Transforming operator** ($\mathcal{B}$): This applies a geometric transformation to the function's input, representing

the orientation and position of the probe. It is implemented as a $3 \times 4$ matrix combining a rotation matrix and a translation vector.

With these definitions, the imaging process of a 2DE frame is expressed as:

$$P(\vec{u}) = \mathcal{S}_2^3 \circ \mathcal{B} \circ O(\vec{X}),$$

where $\vec{u} = [u, v]^T$ denotes the pixel coordinates in a 2D image, $P(\vec{u})$ is the observed image intensity, and $\mathcal{B}$ maps the 3D structure to the 2D imaging plane. This physical model forms the basis for supervising the training of CardiacField without requiring ground truth annotations.

### B. Positional Parameters and Image Selection

In CardiacField, accurately reconstructing the 3D heart model requires knowing where each 2D ultrasound image was taken from. To achieve this, the system estimates the position and orientation of the ultrasound probe at the moment each image was captured. This information is referred to as the *positional parameters*.

Initially, approximate positions are estimated using a method called PlaneInVol, which is trained to predict where in 3D space a 2D image comes from. These estimates include how the probe was rotated and where it was located during acquisition.

After these initial predictions, CardiacField further refines the position of each image during training. This is done by optimizing the reconstruction so that the generated 3D model best explains all 2D input images.

To ensure high-quality reconstruction, not all images are used. The system collects a large number of 2D views by rotating the probe around the apex of the heart. Then, 120 images are selected based on how well they match predefined viewing directions. Images that are too noisy or captured from poor angles are automatically excluded.

This selection and refinement process helps CardiacField create a consistent and accurate 3D representation of the heart using only high-quality, well-positioned 2D images.

### C. Implicit Neural Representation of 3D Heart

An *Implicit Neural Representation* (INR) is a method of representing complex signals or structures, such as images or 3D shapes, using a continuous function parameterized by a neural network. Instead of storing explicit data like voxel grids or point clouds, an INR maps a spatial coordinate $\vec{X} = [x, y, z]^T$ directly to a value such as intensity or occupancy via a neural network $F_\theta(\vec{X})$. In the context of CardiacField, INR is used to represent the 3D structure of the heart by learning a function that predicts the image intensity at any given point in space.

In CardiacField, the 3D structure of the heart is modeled as a continuous function that maps a spatial coordinate $\vec{X} = [x, y, z]^T$ to an intensity value. This function is represented by a small neural network called a multilayer perceptron (MLP), denoted as $F_\theta(\vec{X})$, where $\theta$ includes the trainable weights of the network.

To improve the model's ability to represent fine details and sharp anatomical boundaries, CardiacField uses a multiresolution hash table in combination with the MLP. The 3D space is divided into small grid cells, and each cell is assigned a learnable feature vector. For any input point $\vec{X}$, the model finds the eight closest grid points and applies trilinear interpolation to combine their vectors into a single input. This interpolated vector is then passed through the MLP to predict the intensity at that location.

### D. Physics-Informed Loss for Unsupervised Learning

CardiacField does not rely on annotated 3D ground truth data for training. Instead, it uses a physics-informed loss function that leverages the known geometry of 2D echocardiography imaging to supervise the model in a self-supervised manner.

Given a set of 2D echocardiographic images $\{P_i\}_{i=1}^N$, the system aims to reconstruct a 3D heart volume via an implicit neural representation $F_\theta(\vec{X})$, where $\vec{X} \in \mathbb{R}^3$ is a spatial coordinate and $\theta$ are the network parameters. The key idea is to simulate how each 2D ultrasound frame would have been formed from the 3D structure, based on physical imaging principles.

Each 2D frame is assumed to be a planar slice of the 3D heart, obtained by applying a spatial transformation $\mathcal{B}_i$ (which includes rotation and translation) followed by a slicing operation $\mathcal{S}_2^3$. This physically-motivated image formation process is expressed as:

$$P_i^{\text{pred}} = \mathcal{S}_2^3 \circ \mathcal{B}_i \circ F_\theta(\vec{X}).$$

The predicted image $P_i^{\text{pred}}$ is then compared with the actual ultrasound image $P_i$ using the S3IM loss function, which measures structural similarity between the two:

$$\mathcal{L} = S3IM\left(\mathcal{S}_2^3 \circ \mathcal{B}_i \circ F_\theta(\vec{X}) - P_i\right).$$

Although S3IM itself is not physics-based, it is applied to outputs generated through a physics-informed process. Thus, the entire learning procedure is governed by imaging physics, making it a true physics-informed optimization. Both the network parameters $\theta$ and the positional parameters $\mathcal{B}_i$ are optimized jointly:

$$\theta^*, \mathcal{B}^* = \arg\min_{\theta, \mathcal{B}} \mathcal{L}.$$

## VI. Results and Discussion

Table I presents the performance of various LVEF prediction methods reported in the literature, based on evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination ($R^2$). Among the methods compared, EchoNet [7] shows a relatively strong baseline with an MAE of 4.22 and $R^2$ of 0.79. Transformer-based models such as UVT and UltraSwin offer slightly improved performance over traditional 3D convolutional models like R3D and MC3, but still face limitations in

TABLO I: *Comparison of LVEF Prediction Metrics Reported in the Literature*

| Method | MAE ↓ | RMSE ↓ | $R^2$ ↑ |
|---|---|---|---|
| EchoNet [8] | 4.22 | 5.56 | 0.79 |
| R3D [8] | 7.63 | 9.75 | 0.3 |
| MC3 [8] | 6.59 | 9.39 | 0.42 |
| UVT [?] | 5.32 | 7.23 | 0.64 |
| UltraSwin-base [4] | 5.59 | 7.59 | 0.59 |
| CardiacField (LVEF) | 2.48 | 3.05 | Not reported |

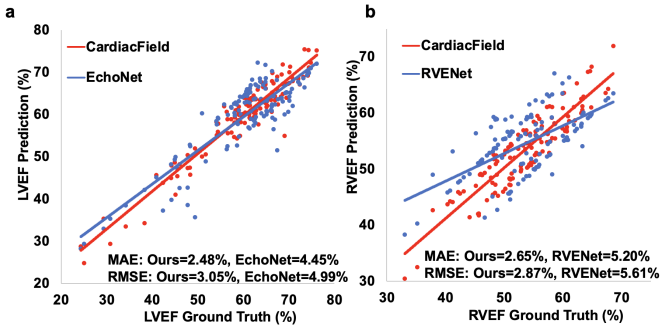


Figure 1: Comparison of Cardiac Field left ventricular ejection fraction results with those of EchoNet

modeling complex spatial relationships due to their reliance on individual frames or 3D patches.

CardiacField shows a clear improvement over earlier methods, achieving the lowest Mean Absolute Error (MAE) of 2.48 and Root Mean Square Error (RMSE) of 3.05 among all the techniques assessed. Although the original study does not provide the $R^2$ value, the low error rates indicate a good match between the predicted and actual ejection fraction (EF) values, as also illustrated in Figure 1. Importantly, CardiacField achieves this strong performance without using segmentation masks or ground truth volumes. It relies on a self-supervised setup guided by a physics-informed image formation model. .

These results show the benefits of including knowledge about imaging physics and multi-view geometry in deep learning systems. By treating ejection fraction (EF) estimation as an inverse problem, where we reconstruct a 3D heart volume from 2D echocardiograms, CardiacField can directly calculate changes in volume. This leads to more accurate and realistic EF estimates. This approach not only improves the quality of the estimates but also follows well-known principles in computational imaging. Therefore, CardiacField is especially promising for clinical use, where consistency, reliability, and minimal effort for annotations are important.

## REFERENCES


[1] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A Video Vision Transformer," arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2103.15691

[2] Shen C, Zhu H, Zhou Y, Liu Y, Yi S, Dong L, Zhao W, Brady DJ, Cao X, Ma Z, Lin Y. CardiacField: computational echocardiography for automated heart function estimation using two-dimensional echocardiography probes. Eur Heart J Digit Health. 2024 Sep 24;6(1):137-146. doi: 10.1093/ehjdh/ztae072. PMID: 39846074; PMCID: PMC11750196.

[3] M. Bloom, B. Greenberg, T. Jaarsma, et al., "Heart failure with reduced ejection fraction," Nat Rev Dis Primers, vol. 3, no. 17058, 2017. [Online]. Available: https://doi.org/10.1038/nrdp.2017.58

[4] L. Fazry, A. Haryono, N. K. Nissa, Sunarno, N. M. Hirzi, M. F. Rachmadi, and W. Jatmiko, "Hierarchical Vision Transformers for Cardiac Ejection Fraction Estimation," in *IWBIS 2022 - 7th International Workshop on Big Data and Information Security, Proceedings*, pp. 39-44, IEEE, 2022. [Online]. Available: https://doi.org/10.1109/IWBIS56557.2022.9924664

[5] T. A. Foley, S. V. Mankad, N. S. Anavekar, C. R. Bonnichsen, M. F. Morris, T. D. Miller, and P. A. Araoz, "Measuring Left Ventricular Ejection Fraction - Techniques and Potential Pitfalls," *European Cardiology Review*, vol. 8, pp. 108-114, 2012.

[6] D. L. Mann, *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*, R. O. Bonow, D. L. Mann, D. P. Zipes, and P. Libby, Eds., pp. 487–504, Elsevier Saunders, 2012.

[7] D. Ouyang, B. He, A. Ghorbani, M. P. Lungren, E. A. Ashley, D. H. Liang, and J. Y. Zou, "Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning," in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[8] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, "Video-based AI for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, pp. 252–256, 2020. [Online]. Available: http://dx.doi.org/10.1038/s41586-020-2145-8

[9] Batool S, Taj IA, Ghafoor M. Ejection Fraction Estimation from Echocardiograms Using Optimal Left Ventricle Feature Extraction Based on Clinical Methods. Diagnostics (Basel). 2023 Jun 24;13(13):2155. doi: 10.3390/diagnostics13132155. PMID: 37443550; PMCID: PMC10340260.

[10] Roberto M. Lang, Luigi P. Badano, Victor Mor-Avi, Jonathan Afilalo, Anderson Armstrong, Laura Ernande, Frank A. Flachskampf, Elyse Foster, Steven A. Goldstein, Tatiana Kuznetsova, Patrizio Lancellotti, Denisa Muraru, Michael H. Picard, Ernst R. Rietzschel, Lawrence Rudski, Kirk T. Spencer, Wendy Tsang, Jens-Uwe Voigt, Recommendations for Cardiac Chamber Quantification by Echocardiography in Adults: An Update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging, Journal of the American Society of Echocardiography, Volume 28, Issue 1, 2015, Pages 1-39.e14, ISSN 0894-7317, https://doi.org/10.1016/j.echo.2014.10.003.

[11] Reynaud, H., Vlontzos, A., Hou, B., Beqiri, A., Leeson, P., and Kainz, B., "Ultrasound video transformers for cardiac ejection fraction estimation," in Medical Image Computing and Computer Assisted Intervention –MICCAI 2021, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds. Cham: Springer International Publishing, 2021, pp. 495–505.