Iranian Journal of Electrical and Electronic Engineering

Journal Homepage: ijeee.iust.ac.ir

Estimating Ejection Fraction from Echocardiogram Data: A Narrative Review on Deep-Learning-Based Methods

Roshanak Ghorbani* and Hamid Behnam*(C.A.)

Abstract: In recent years, many studies have used deep learning (DL) methods to estimate ejection fraction (EF) from two-dimensional echocardiogram data. DL methods offer several advantages. They automate EF estimation, overcome the limitations of manual or semi-automatic methods, and improve estimation accuracy and computational speed. Despite significant advances in DL-based EF estimation methods, the number of review papers examining their pipelines and performances is limited. In the current study, we categorized the DL-based EF estimation methods. We also reviewed their pipelines, architectures, performances, advantages, and limitations. Moreover, we discussed the current status, research trends, challenges, and future direction of this field. Finally, we addressed several open challenges for future studies. The results revealed that the field is mature yet evolving. DL-based methods have shown great potential for estimating EF. Additionally, many recent studies have focused on the clinical integration and evaluation of these methods.

Keywords: Artificial Intelligence, Deep Learning, Echocardiography, Ejection Fraction, Machine Learning, Review.

In Press

1 Introduction

Ejection fraction (EF) is one of the most important parameters for evaluating cardiac function and the severity of heart diseases. This parameter is used in diagnosing and prognosing heart failure (HF) [1]. Despite the improvement in mortality rates in recent years, cardiac diseases have remained the major causes of death [2]. Therefore, their early treatment and diagnosis are crucial. In this regard, EF is an important tool for patient management and trial design. The reason is its simple calculation and the fact that its improvement over time is related to better clinical outcomes [3]. These reasons make it useful for tracking treatment response in clinical applications. EF is defined as the fraction of blood ejected from a heart chamber (ventricle or atrium) in every heartbeat, relative to its end-diastolic volume (EDV). EF can be calculated as ($EF = \frac{EDV-ESV}{EDV}$), where the ESV corresponds to its end-systole volume. The normal range of EF is often between 50% to 70% [4]. To conclude, EF is an important clinical parameter for assessing heart function and cardiac diseases.

Considering the clinical importance of EF, various methods have been proposed to calculate it. Among them, cardiac magnetic resonance imaging (CMRI) is the gold standard for assessing the heart morphology and function [5]. However, many studies have used two-dimensional (2D) echocardiogram data to estimate EF. It is because the echocardiography is more accessible and suitable for repeated use in periodic patient monitoring [6]. However, estimating EF as a volumetric parameter from 2D echocardiography videos has some challenges. These challenges include the heart's out-of-plane motion, the captured videos' low quality [7], and the captured signals' low signal-to-noise ratio (SNR).

Previously proposed methods for estimating EF from 2D echocardiogram data can be broadly classified into three categories: manual, semi-automatic, and fully automatic methods. Manual EF calculation methods are often tedious, time-consuming, experience-dependent,

Iranian Journal of Electrical & Electronic Engineering, 2027.
Paper first received 5 April 2026 and accepted 27 October 2026.
* The authors are with the Department of Biomedical Engineering, School of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran.
E-mail: behnam@iust.ac.ir.
Corresponding Author: Hamid Behnam.

prone to human error, and poorly reproducible. To address these challenges, various studies attempted to automate EF estimation. Among them, deep learning (DL) methods have attracted great attention recently. The development of Artificial Intelligence (AI) presented a promising opportunity to automate the EF estimation process by overcoming the limitations of traditional methods and enhancing estimation accuracy and speed. Over time, AI-based EF estimation methods evolved from handcrafted machine learning (ML) approaches to more advanced DL-based methods and end-to-end frameworks.

Given the growing number of DL-based EF estimation approaches, analyzing and comparing them is important and helps improve them in future studies. However, previous review papers on AI applications in EF estimation (such as [8-11]), mainly summarized the literature and highlighted areas where AI can be applied in echocardiography. In this study, we provide a comprehensive analysis of several existing DL-based methods for estimating EF from 2D echocardiogram data. We present a thematic categorization of these methods. We also analyze their pipelines, architectures, performances, advantages, and limitations. However, the overall conclusion is challenging due to the variety of existing studies in terms of datasets, model architectures, and evaluation protocols. To the best of our knowledge, there is no previous review paper that specifically analyzes the DL-based methods proposed for estimating EF on 2D echocardiography videos to this extent. By synthesizing existing studies, this paper helps many researchers in this field. It helps AI researchers understand the current status of the field and identify research gaps to develop new methods. It also helps physicians and cardiac imaging professionals to gain deeper insights into the methods developed in the field.

The structure of the current paper is as follows. Section 2 presents the methodology applied to find relevant papers. Section 3 summarizes previous EF calculation methods. Section 4 categorizes previously proposed methods and discusses their objectives, results, and limitations. Finally, Section 5 discusses the current status, research trends, challenges, gaps, and future direction of the field, followed by the conclusion in Section 6.

2 Materials and Methods

The current study presents a narrative review of the DL-based methods proposed for estimating EF from 2D echocardiogram data. To extract relevant papers, we searched the Web of Science Core Collection (WOSCC) and gathered 49 articles by January 2026. We applied a search strategy in the titles as follows: ((Echocardiograph*) AND ("Deep*" OR "Artificial Intelligence" OR "AI" OR "Machine Learning" OR "neural network*" OR "Convolutional*" OR "CNN" OR "Recurrent*" OR "RNN" OR "LSTM" OR "Transformer*") AND ("Ejection Fraction" OR "EF" OR "LVEF" OR "RVEF")). Then, we scanned these papers and excluded irrelevant publications with exclusion criteria as follows: 1) non-original publications, including review papers, protocols, and editorials, 2) publications that did not use echocardiogram data, and 3) non-English papers. Finally, we included 23 original articles that were published in English. We reviewed the selected articles and narratively synthesized their key themes and findings to summarize the current status of the field. In this study, we focused exclusively on publications that used DL methods to estimate EF and did not consider studies using DL for cardiac chamber volume estimation (without addressing EF).

For further analysis of DL-based EF estimation methods and to support our narrative review, we also conducted a light bibliometric analysis of extracted papers to assess the current status and research hotspots in the field more deeply. To achieve this, we used the RStudio software (version 2025) and the "bibliometrix" library for descriptive statistical analyses.

3 Background: Traditional Methods

In clinical applications, experienced experts may be able to approximate EF visually from 2D echocardiogram videos. However, this method is experience-dependent and prone to human errors. The quantitative calculation of EF has several classic approaches based on manual methods. Among them, methods such as the modified bi-plane Simpson (MBPS) or bi-plane area-length (BPAL) are the most widely used.

The MSBP method models the target cardiac chamber (ventricle or atrium) as a stack of disk-shaped elements. It calculates the diameter of these disks from both the apical four-chamber (A4C) and apical two-chamber (A2C) views. For each view, the end-systole (ES) frame is identified by an echocardiographer. The endocardium borders are manually traced. Next, the chamber's long axis is determined and divided into twenty (the most recommended number) equal segments. The diameter of each disk (i-th disk) is measured in the A4C view (a_i) and in the A2C view (b_i). These paired diameters are then applied to Equation (1) [12] to calculate the target chamber volume at the ES phase.

$$Volume = \frac{\pi}{4}\sum_{i=1}^{20} a_i\, b_i \frac{L}{20} \quad (1)$$

where L represent the length of the chamber's long axis. The same procedure is then repeated for the end-diastole (ED) frame. Finally, the EF is computed using Equation (2) [12].

$$EF = \frac{EDV-ESV}{EDV} \times 100 \quad (2)$$

In Press

Where ESV and EDV indicate the end-systolic and end-diastolic volumes, respectively.

On the other hand, the BPAL method considers the target chamber's cross-sectional area as circular. In this method, the ES frame is first identified by an echocardiologist. Next, the chamber's cross-sectional area from the short-axis view (A) and the chamber's long axis length from the A4C view (L) are approximated and applied to Equation (3) [12].

$$Chamber's\ volume = \frac{8(A)^2}{3\pi L} \quad (3)$$

Then, these steps are repeated for the ED frame. Finally, the EF is achieved using Equation (2).

In clinical applications, the MBPS and BPAL methods are often used to calculate chamber volumes from two views obtained from separate heart cycles. Using two very different cardiac cycles with irregular rhythms may introduce errors in volume calculation. In these cases, the single-plane Simpson method or the single-plane area-length method can be used. These single-plane-based methods use only one view (typically A4C). They are helpful when video from one view is unavailable or difficult to capture. Acquiring only one view (instead of two) also reduces imaging time. However, these single-plane methods assume a circular geometry. This assumption may not be fully established, leading to approximation errors, especially in irregular-chamber-shaped cases. Moreover, since only a single view is recorded, poor video quality can introduce errors in the calculation.

To conclude, calculating EF manually from 2D echocardiogram videos is challenging. A major challenge is the need for manual myocardium tracing, which is time-consuming. Other challenges include high potential for human error and low repeatability. These challenges led to the development of methods to automate EF estimation. In automating EF estimation, DL methods have offered several promising advantages. One is reducing operator dependency. Others include higher accuracy, rapid computation, reduced inter- and intra-observer variability, and more repeatability. In recent years, many studies have focused on developing DL-based methods to estimate EF from 2D echocardiogram videos.

4 Thematic Categories of Existing DL-Based and ML-Based Methods

In this section, we categorize the DL-based EF estimation methods proposed in previous papers. We categorize them according to task (regression-based and classification-based), feature extraction (handcrafted-based and DL-based), approach (segmentation-based and end-to-end), and input (image-based and video-based), as illustrated in Fig. 1. We will discuss these categories below.

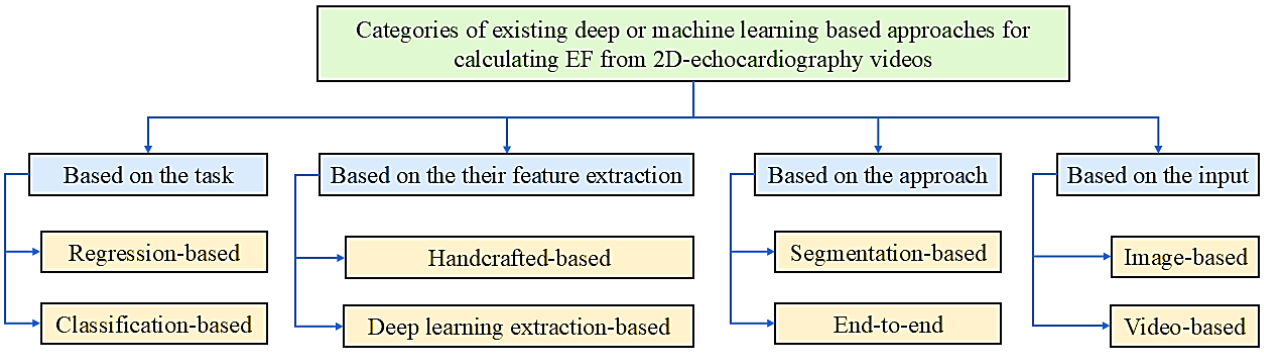


Fig 1. Schematic of thematic categorization of existing DL-based EF estimation methods using 2D echocardiogram data.

4.1 Classification-Based vs. Regression-Based Methods

EF is a continuous parameter and can be estimated using regression. Many existing software estimate EF directly using a regression model. Previous regression-based EF estimation approaches can be broadly categorized into two scenarios: 1) models trained to estimate end-systolic and end-diastolic chamber volumes and then calculate EF using Equation (2), and 2) models trained to estimate EF as a direct regression task. The results of Ref. [13] showed that the DL model trained on the first scenario performed better than the one trained on the second scenario.

Whereas, several previous studies treat EF as a target for categorization of heart failure severity. Heart failure can be classified based on the left ventricular ejection fraction (LVEF) value into three main classes: HF with reduced EF ($EF \leq 40\%$), heart failure with mildly reduced EF ($40\% < EF < 50\%$), and HF with preserved EF ($EF \geq 50\%$) [14]. Despite this three-class categorization, previous studies have categorized the EF value range into different numbers of classes, including two-class and four-class. A similar categorization has been done for other cardiac chambers' EF values. As an example, Ref. [15] proposed an ML-based pipeline to classify pulmonary regurgitation patients'

echocardiography images into two classes based on the right ventricle (RV) EF: ($RV\ EF < 47\%$) and ($RV\ EF \geq 47\%$), and to detect RV dysfunction ($RV\ EF < 47\%$). As another example, Ref. [16] proposed a DL-based model to classify the M-mode images in PLAX view into two classes: ($EF > 50\%$) and ($EF \leq 50\%$). The detailed pipeline of these two publications will be discussed in the next subsection.

Moreover, several publications have used DL-based methods to detect heart diseases from echocardiogram data. For example, they detected heart failure with preserved ejection fraction (HFpEF) as a two-class classification task. Using DL-based models for detecting HF enables the extraction of novel features that are not commonly used in clinical examinations.

In general, treating EF estimation as a classification task often requires less data than regression. This is important, especially when training large DL models that require large datasets. However, discretizing EF as a continuous variable for classification results in quantitative information loss. For example, a person's EF reduction may go unnoticed if its categorization does not change. In addition, when categorizing EF, selecting appropriate thresholds for each class is important. For example, selecting an inappropriate threshold may result in patients with very close EF values being assigned to different categories and treated differently in clinics. In contrast, regression approaches result in more detailed information. However, regression requires the model to learn the relationship between videos and continuous targets rather than limited classes in classification. Therefore, choosing regression or classification involves a trade-off between clinical information, data requirements, and model complexity. Moreover, another challenge of EF classification is that datasets are often unbalanced (the preserved ejection fraction class often has more samples). This should be considered when reporting the models' results.

4.2 Handcrafted Features vs. Deep-Based Extracted Features Methods

In estimating EF from echocardiogram data, feature extraction is a key step. Extracting appropriate features improves estimation accuracy, increases the model's generalizability, and reduces estimation sensitivity to noise and environmental changes in data. It also reduces data dimensionality, which is important for high-dimensional echocardiogram data.

Previous studies have used various features to estimate EF, from handcrafted features (including statistical parameters, shape, motion, frequency, texture, and gradient-based) to features extracted by DL models. Very early EF estimation methods relied on handcrafted features, and their pipeline had two stages: feature extraction, followed by classification or regression. These pipelines are referred to as ML-based EF estimation methods. Among them, early pipelines used simple handcrafted features (such as target chamber area or volume, texture-based features, or speckle-based features), followed by an ML model (such as linear regression, support vector machine (SVM), or k-nearest neighbor model) for regression or classification.

As an example of ML-based EF estimation methods, Ref. [15] proposed a pipeline based on the handcrafted features. It used the gradient boosted trees (XGBoost) algorithm for RV dysfunction prediction ($EF < 47\%$) in pulmonary regurgitation patients. The authors investigated two scenarios: 1) using guideline-directed RV echocardiogram measures (such as fractional area change, RV EDV, and RV ESV), and 2) using 44 features, including demographic variables, clinical, and echocardiogram features (such as age, sex, and height). They also used the EF values obtained using CMRI imaging as a reference. The results showed that the first scenario performed better than the second one. However, it was not sufficiently accurate compared to the selected reference and clinicians' visual diagnosis. Therefore, further improvement is needed, which may be achieved by using more appropriate features. Overall, their pipeline has some advantages and disadvantages. An advantage is its ability to estimate RV volumes from 2D echocardiogram data. Since estimating RV EF is challenging and often correlates poorly with MRI due to the RV's anatomical features. Another advantage is its very light and interpretable structure (compared to other automatic EF estimation methods). However, a limitation is its dependence on the extracted features (as demonstrated by the results) and requirement for strong feature engineering. In conclusion, the study demonstrates the potential of automated models to estimate RV EF from echocardiographic data. Additionally, it highlights the need to develop more accurate models for RV EF estimation in patients with pulmonary regurgitation.

As another example of ML-based EF estimation methods, Ref. [17] used inter-beat dynamics (length, area, width, and volume) of LV and left atrium (LA) to detect HFpEF. The authors proposed a pipeline that included a segmentation step (LV and LA), a value extraction step (length, area, width, and volume), followed by a classification step (HFpEF or no HFpEF). In the pipeline, they employed the U-Net model (with four down-convolution and four up-convolution layers) to segment LV. Also, they used a convex-hull algorithm to extract target values. Finally, they employed four 1D CNNs for classification. The classification model takes extracted values (from 60 frames) as input signals and predicts HFpEF. The results showed an acceptable correlation with the expert's manual calculations. Overall, their pipeline presents both advantages and

In Press

disadvantages. An advantage is the use of temporal information through inter-beat dynamics. It also used 1D CNNs, which are simple. In addition, it is highly interpretable. However, the pipeline's performance is dependent on the accuracy of segmentation and value extraction. An error in any of these steps propagates to subsequent stages, affecting the overall accuracy. Additionally, it requires 60 frames to ensure one cardiac cycle. This can limit its clinical applications for patients with bradycardia or for high-frame-rate imaging. Therefore, the pipeline requires an extra step to ensure it receives a full cardiac cycle. This step can be a frame-rate reduction step. However, the frame-rate reduction may lead to temporal information loss. Moreover, a limitation of the study is the use of a narrow set of heart diseases for training the pipeline. Using a more diverse dataset might increase the model's generalizability. To conclude, the study demonstrates the potential of AI-based HFpEF detection methods using only a 2D echocardiography video from the A4C view. This is important since detecting HFpEF is clinically challenging and time-consuming. Adding data from other views may improve the detection of HFpEF.

In conclusion, the performance of ML-based EF estimation methods is dependent on handcrafted features. Handcrafted features are highly sensitive to noise and data quality. To address the limitations of ML-based EF estimation methods, subsequent studies have used DL models. DL models extract features directly from the input videos. They can extract novel features that are rarely considered in handcrafted features.

As an example of using DL-based extracted features, Ref. [18] used an ML-based algorithm (AutoEF, BayLabs inc, San Francisco, CA) to extract features directly from echocardiogram images. The algorithm estimated EF without first estimating the left ventricular end-diastolic volume (LVEDV) and left ventricular end-systolic volume (LVESV). The authors modeled LV volume using longitudinal and radial contraction coefficients (C_L and C_r, respectively). This allowed them to calculate EF based on the minimum values of these two coefficients, without first estimating LVESV and LVEDV (as shown in Equation (4)). They trained a DL-based pipeline to estimate the coefficients directly from A2C and A4C views.

$$EF = 1 - (C_{R(\min)} \times C_{L(min)}) \quad (4)$$

They also investigated the algorithm's performance in detecting ($EF \leq 35\%$). They used the average of EF values calculated by experts using the BPS method as a reference. Their results correlated closely with the selected reference. Their pipeline entails both benefits and limitations. A benefit is that it does not rely on any segmentation. Additionally, the authors used a relatively large dataset (compared to other studies) for training. However, their test dataset was very small. Therefore, their pipeline requires further evaluation, especially on data from other centers. Another limitation is that the authors did not analyze the pipeline's details, and its data is not publicly available. To improve the model's accuracy, they also proposed a detrending method to correct the algorithm's error relative to the reference EF. However, the effectiveness of this method requires further investigation on data from other centers. To conclude, the study indicates the potential of DL-based models to extract features from echocardiogram data. It also shows their potential to estimate EF directly, without first estimating volumes.

Among the DL-based models employed to extract features for estimating EF, convolutional neural networks (CNNs) are the most frequently used. As an instance, Ref. [16] proposed a model based on ResNet50 to extract features directly from M-mode images in PLAX view and predict ($EF > 50\%$) as a two-class classification task. Its pipeline included a region of interest (ROI) cropping step, followed by the ResNet50 model. The EF values estimated by the BPS method were used as a reference. Overall, a benefit of ResNet50 is its skip connections, which mitigate the vanishing of gradients during DL model training. The pipeline also included batch normalization layers to slow training by an internal covariance shift and ReLU (rectified linear unit) as the activation function. Also, dropout and regularization were used to prevent overfitting. The results showed good accuracy in classifying EF and reduced inter- and intra-observer variability in estimations. However, a limitation is that the authors used a limited number of test samples, which may lack sufficient variety. Therefore, the results may not fully reflect the model's generalizability. To conclude, the study demonstrates the potential of CNNs in estimating EF.

Generally, CNNs have several advantages for estimating EF. The advantages included hierarchically extracting features from images using local filters and preserving the image's spatial structure. These advantages make CNNs suitable for tasks such as estimating the area in an image. Additionally, feature extraction in CNNs is translation-invariant. This is important for estimating EF because imaging conditions (such as probe placement and patient characteristics) can cause displacement, rotation, or scaling in captured videos. Finally, CNNs are often simple and have a small number of parameters (compared to other DL models, such as transformers and hybrid models). This feature makes them suitable for medical tasks where a limited amount of labeled training data is available. Despite these advantages, using CNNs for EF estimation also has some disadvantages. A major disadvantage of early 2D CNNs is their inability to capture temporal dependencies of video frames. It is worth noting that in EF estimation,

In Press

the final aim is to estimate EF accurately, not to extract temporal features. However, temporal information in echocardiogram data captures the heart's motion, improving estimation of EF as a volumetric parameter.

To address the 2D CNNs' limitation in extracting temporal dependencies in echocardiogram data, 3D CNNs have been employed. 3D CNNs extract both spatial and temporal information. As an example, Ref. [19] employed a 3D CNN model to diagnose HFpEF from the A4C view data and classify the data into three classes: nondiagnostic, no HFpEF, and HFpEF. Its model consists of three 3D convolution layers, each with two convolutions and a 3D filter, batch normalization, ReLU, and a max-pooling layer. The model took 30 frames with a stride of 10 as input. In the last layer, it used softmax as the activation function to classify the data into the HFpEF and non-HFpEF categories. The authors also added a nondiagnostic output (based on the entropy and prediction uncertainty). The results showed an acceptable accuracy. An advantage of the model is its ability to diagnose HFpEF (which is clinically challenging) by using only a single video from the A4C view. The A4C view is easy to acquire and is often captured during the clinical routines. Moreover, the authors used data gathered from multiple centers in the United Kingdom and the United States. Multi-center data can increase the data diversity and the model's generalizability. However, the model required each input to contain at least 30 frames. It may limit its application for videos with fewer frames. Another limitation is the model's poor explainability. The extracted features for the decision-making process are not fully clear. This limitation may reduce its clinical reliability. To conclude, the study demonstrates the potential of fully automatic DL-based models in detecting HFpEF from A4C view data.

Generally, a disadvantage of 3D CNNs is their inability to model long-temporal dependencies (such as several cardiac cycles). Moreover, the CNN's ability to capture global context is limited and requires multiple layers to model. However, it would not be important for estimating EF from echocardiogram images. Unless in situations where the imaging field of view is such that the target chamber occupies most of the image. This situation may cause a challenge due to the CNNs' limited receptive field. In this situation, transformer models may be better, since they can consider both global and temporal dependencies.

To address the 3D CNN's limitation in capturing long-term dependencies, several studies have employed recurrent neural networks (RNNs), especially in combination with CNNs. RNNs often have fewer parameters than transformers and therefore require less training data. However, their ability to capture long temporal dependencies is limited compared to that of transformers.

In recent years, transformers have attracted significant attention. Various transformer models have been proposed to estimate EF. A strength of transformers is their ability to capture both spatial and long-term temporal dependencies. It is useful for investigating the movement of the target chamber muscles in echocardiogram data. However, they often have high computational costs and require large datasets for training. This is challenging in medical tasks, where the labeled data is limited.

In estimating EF recently, many studies have employed hybrid models (such as CNN and RNN, or CNN and transformer). Hybrid models help to combine the strengths of different architectures. Among them, Ref. [20] combined CNN and transformer layers and proposed a DL model (QHAutoEF) to segment LV for EF estimation. It used a hybrid model to overcome CNNs' limitation in extracting global features and transformers' limitation in handling local features. Its pipeline had several steps: detection of ES and ED frames, segmentation of the LV, followed by estimation of EF. The segmentation model consists of an encoder to extract multi-level features, followed by two head networks (a fine branch and a coarse branch) for LV segmentation. The fine-aligned branch used a channel-spatial attention-based module to capture global dependencies and extract detailed alignment features. The features were then fed to the coarse branch to enhance the coarse mask. The refined coarse mask was then adjusted using a feature-positioning module to fuse features and suppress noise. Finally, a better LV mask was obtained by applying a convolution layer on the fused features. In other words, the pipeline used the CNN as the backbone and the transformer in one of its branches. The authors used the ES and ED frames, borders, and EF values estimated manually by an expert as a reference. The segmentation model achieved better performance than U-Net, ACNN, attention U-Net, and U-Net^{++} segmentation models on the CAMUS dataset. An advantage of the pipeline is its lightness and having a low computational cost (compared to common CNN-transformer hybrids). Also, an advantage of the study is considering implementation requirements in designing the model and integrating it into an echocardiography machine. However, a limitation is that the authors excluded images with foreshortening and low quality, patients under 18 years old, and those with thoracic deformity. This limitation may reduce the generalizability of the results. The authors also did not mention the ES and ED detection method of their automatic pipeline. Another limitation of this pipeline is the need for manual adjustment to select more appropriate frames for the ES and ED phases. To conclude, the study indicates the potential of CNN-transformer hybrids for estimating EF and software-hardware implementation in clinical applications.

In Press

Generally, combining multiple models increases the architectural complexity in hybrid models. This complexity may lead to a large number of parameters to be tuned and a computationally expensive training process.

4.3 Segmentation-Based vs. End-to-End Methods

Among the previously proposed EF estimation approaches, many studies have developed models taking the target chamber's segmented masks as input. They often had a two-stage pipeline: a segmentation model followed by regression or classification. Among them, early pipelines used non-DL segmentation methods (such as threshold-based, region-based, or active-contour-based), followed by a multi-stage EF estimation algorithm similar to manual methods (such as MBPS or BPAL). These approaches are highly interpretable. However, using multi-stage crafted pipelines may lead to errors introduced in any stage propagating to subsequent stages and reducing overall performance. Later studies employed a non-DL segmentation method, followed by a DL-based EF classification or regression model. Although these methods were less interpretable, they had fewer stages and less potential for error propagation.

To address the limitations of conventional segmentation methods, recent ML-based EF estimation studies have developed several DL models to segment borders and improve the segmentation accuracy. For example, Ref. [21] employed a modified version of U-Net (DPS-Net) for left ventricle segmentation. Its EF estimation pipeline had two steps: the DPS-Net for LV segmentation, followed by the BPS method for EF estimation. Its segmentation model was trained to segment data from multiple views (A2C, A3C, and A4C). Its pipeline used a feature pyramid module (including pyramidal dilated convolutions) to extract and fuse multi-scale features (for refining the segmentation borders) and multi-level features (for extracting LV's global features). It also used dilated convolutions with different dilation rates to increase the receptive field. Moreover, it used a pyramid pooling module (including parallel pooling and convolutions) to capture contextual information at multiple scales. Finally, it used deep supervision with composite losses (Dice, weighted binary cross-entropy, and modified mean absolute error) at multiple levels to prevent gradient vanishing during training. The authors used the EF values calculated using the BPS method as a reference. The results show that the DPS-Net performed acceptably in segmenting the LV. Overall, an advantage of the study is that the authors evaluated the pipeline's performance on different heart diseases (including hypertrophic cardiomyopathy, atrial fibrillation, and dilated cardiomyopathy) and data from different centers. However, the authors developed the pipeline for 2D echocardiography data, without considering other modalities (such as three-dimensional or contrast echocardiography). This pipeline calculated EF using both A2C and A4C views that were acquired from different heart cycles. This may lead to errors in patients with irregular heart cycles. Overall, the pipeline's EF estimation is highly affected by errors in segmentation and calculation steps. To conclude, the study indicates the potential of DL-based models in segmenting cardiac chambers.

As another example of using DL models for segmentation, Ref. [22] proposed a framework (EchoEFNet) to segment LV and extract the echocardiographic landmarks' (cardiac apex and mitral annulus) positions. The authors then used the extracted data by the BPS method to calculate LV volume and EF. Generally, they used the EchoEFNet workflow to obtain the segmentation mask and echocardiographic landmarks in ES and ED frames. The workflow included two parts: a ResNet50 model with dilated convolution as the backbone to extract features, followed by two network branches: one for segmenting the LV and the other for extracting landmarks' locations from the backbone model's output. Each of these two branches had a decoder structure, including an atrous spatial pyramid pooling (ASPP) to capture features from different receptive fields, followed by a multi-scale feature fusion (for fusing high- and low-level features) up-sampling. The pipeline performed acceptably in segmenting LV and in calculating EF. Overall, this pipeline presents both advantages and disadvantages. An advantage of the multi-task workflow is that it produces both segmentation masks and the landmarks' positions. This reduces errors in positioning landmarks on the segmentation mask. Another advantage is the dilated convolution layer in two residual layers of the backbone. It prevents loss of spatial feature and maintains spatial resolution during image down-sampling. Moreover, to reduce human labeling, the authors proposed an automated labeling method for training their landmark detection branch. Their labelling method was based on searching three anatomically designed patches in the segmentation mask to find the echocardiographic landmarks. Overall, their developed EF estimation pipeline is relatively interpretable. However, it depends on the accuracy of the segmentation and landmark detection steps. The accuracy of these steps may be affected by low image quality, abnormal LV shapes, and blurry borders. A limitation of this workflow for clinical applications is the lack of an automatic ES and ED frames detection step, which was not included in this paper. Overall, the study indicates the potential of DL-based models in segmentation and landmark extraction.

Furthermore, Ref. [23] proposed a segmentation-based pipeline to estimate EF. Its pipeline was based on segmenting the left ventricle and using the SPAL method and EF calculation formula. Its pipeline had

In Press

several steps: segmenting the LV in all video frames (using an atrous CNN model), then estimating the LV area (by counting the number of pixels in the segmented mask). Followed by identifying the ES and ED frames (using a peak-detection method on the LV area curve) and estimating the LV length (by using a regression ensemble learning model). The authors then employed the extracted data and the SPAL method to calculate LVEDV, LVESV, and then EF using Equation (2). The segmentation atrous CNN model used in the study contained ResNet50 and atrous convolutions, followed by ASPP. Additionally, the ensemble learning model was based on regression models (including Lasso, Extra Trees, and Adaboosting). The authors reported the average of LVEF across multiple cycles according to the recommendations of the American Society of Echocardiography. They further employed the estimated LVEF for categorization of heart failure with reduced EF (LV$EF < 40\%$). The results showed an acceptable performance for both LVEF estimation and reduced EF categorization. Overall, the pipeline has several pros and cons. The authors used LVEF averaged over multiple heart cycles rather than a single cycle. Their results showed that it is more accurate for the EF reduction detection step. They also trained and tested their pipeline on the EchoNet-Dynamic dataset. Using publicly available datasets makes comparison of models easier. However, the pipeline had multiple steps that are prone to error. An error in each of these steps propagates to subsequent ones and affects the estimated EF. To conclude, this paper indicates the potential of combining DL-based LV segmentation models with the manual EF estimation methods.

In general, segmentation-based EF estimation methods have strengths and weaknesses. The advantages of these methods over traditional methods are automating the segmentation process, reducing EF estimation time, and relieving echocardiographers of the time-consuming segmentation process. However, a limitation is that the accuracy of the segmentation step affects the overall process. Segmentation accuracy may be affected by image quality, which depends on the imaging device and the operator's experience. Using low-quality images may lead to errors in segmentation and EF estimation.

To prevent segmentation errors in low-quality images, several studies have added a quality-assessment pre-processing step to remove data with inadequate image quality. For example, Ref. [24] employed a five-point DL-based quality-scoring model to improve the accuracy and repeatability of LVEF estimation. The scoring method was based on LV clarity in the image, the recognizability of ventricles and atria, the LV-to-LA area ratio, and the segmentability of LV endocardium and epicardium. These parameters are important when segmenting the LV using an automatic method. The authors proposed a pipeline consisting of an encoder (based on the YOLOX model's backbone) to extract features, followed by three decoders. The first decoder (based on deconvolution layers) was used to segment the LV endocardium and epicardium, followed by two CNN networks (one to assess LV endocardium clarity and the other to identify the integrity of the segmentation mask). The clarity classification network was then used to divide the LV endocardium segmentation results into five parts. It then used the clarity of these parts to determine whether the image demonstrated a clear view of the LV. The second decoder (based on fully connected neural networks) was used to classify views (A2C and A4C). The third decoder (based on YOLOX's detection head) was used to detect the cardiac anatomical structures, including the LV, LA, and RV. The segmentation masks of images with adequate quality were then used to identify frames with the minimum and maximum LV area as ES and ED frames, respectively. These masks were also used to calculate EF using the BPS method and the EF calculation formula. In a video with multiple cardiac cycles, the method reported the largest LVEF. The authors employed the ES and ED frames extracted from the echocardiogram (ECG) and the LVEF calculated using the BPS method by experts as a reference. Their quality scoring and LVEF estimation results showed acceptable agreement with the reference. Overall, an advantage of the study is the use of data from multiple centers. However, the pipeline for clinical application requires evaluation on a larger and more diverse dataset. Additionally, this pipeline can be extended to other echocardiogram views. To conclude, the results of the study indicate the potential of DL-based quality-scoring models for improving LVEF estimation.

To further decrease the error in segmentation-based EF estimation models, several studies have added a quality-improvement step. For example, Ref. [25] added a denoising (based on wavelet transform) step to improve training of the CNN segmentation model, and a feature-extraction step (based on an autoencoder) to reduce data dimensionality. Its pipeline took an image (ES or ED frame) as input, and applied wavelet-based denoising with soft thresholding. It used the encoder part of an autoencoder for feature extraction, and then applied a CNN (based on the U-Net model) to segment the LV. The segmentation masks were then used to calculate LVESV and LVEDV by the BPS method. EF was then calculated using Equation (2). The model was trained to segment the A2C and A4C views separately. The authors used an expert's manual segmentation mask as a reference. Their EF estimation results showed an acceptable correlation with the reference. However, a limitation of the study is that the authors trained and tested their pipeline on a limited dataset. Therefore, its generalizability for different heart diseases requires

In Press

further examination. Additionally, eliminating human intervention in the pipeline requires adding a step to automate the identification of ES and ED frames. To conclude, the study indicates the potential of preprocessing methods in improving LV segmentation and EF estimation.

In general, segmentation-based EF estimation methods have several limitations. One limitation is that various factors may affect the segmentation of echocardiogram data. These factors include speckle noise, low image quality, low clarity of cardiac borders, and structural abnormalities in heart chambers. Another limitation is that DL-based segmentation models often require large labeled datasets for training. However, labelling a lot of data can be challenging for echocardiographers. Since segmenting cardiac borders is tedious and time-consuming. To address this challenge, several studies have developed automatic labeling methods.

To address the challenges of segmentation-based EF estimation methods, recent studies have developed several end-to-end models. These models estimate EF directly from echocardiogram videos, without segmentation. As an example, Ref. [26] proposed a fully automatic pipeline combining CNN and transformer models. It proposed a multi-stream CNN-transformer model to estimate EF from four views (A2C, apical three-chamber (A3C), A4C, and PLAX). The pipeline was an eight-stream CNN model with transformer layers. It employed a two-stream I3D model (an action recognition model) for data from each view. It captured both spatial and temporal information in the videos. Each two-stream model contained two 3D CNNs: one to process the original image, and the other to process the image's optical flow. The outputs (for all eight streams) were combined in a linear layer. Overall, this pipeline has both advantages and disadvantages. An advantage is capturing temporal information in videos by using a transformer attention mechanism and adding a bidirectional encoder representation from the transformer (BERT) layer at the end of each I3D model. Another advantage is the use of a classification model (an Inception2D model as a preprocessing step) to extract the target views from clinical examination data. Additionally, the authors used shared weights in the model because images acquired from different views contain common information. They did that to enhance the model's robustness to misclassified views and reduce the computational complexity. They also used EF values estimated by eyeballing, the BSP method, or the Teichholz method as a reference. The results showed an acceptable correlation between the EF values estimated by the pipeline and the reference. Moreover, an advantage of the study is the evaluation of the pipeline on two publicly available datasets (EchoNet-Dynamic and CAMUS ("Cardiac Acquisition for Multi-Structure Ultrasound Segmentation"), despite testing on in-house data. Testing the pipeline on public data makes it easy to compare the pipeline with other developed methods evaluated on these datasets. However, a limitation of the pipeline is that the authors reduced the frame rate of the input videos by much less than the common clinical values to optimize the computational complexity. This can lead to the loss of dependencies between frames and loss of temporal information. Therefore, the pipeline requires adding a preprocessing step to ensure it captures the whole cardiac cycle. It is especially important in patients with bradycardia. To conclude, the study indicates the potential of the combined CNN-transformer models in capturing both spatial and temporal information in echocardiogram data.

Generally, end-to-end EF estimation methods have both benefits and drawbacks. One benefit is that they do not require preprocessing. It reduces the model's sensitivity to preprocessing methods. However, this may make end-to-end methods require a large training dataset for proper generalizability.

4.4 Image-Based vs. Video-Based Methods

Many early DL-based EF estimation methods were image-based. They took ES and EF frames as input. These models required an extra preprocessing step to identify ES and ED frames. Therefore, they had a two-stage pipeline: an ES and ED frames identification step, followed by a DL-based regression or classification model to estimate EF. Among the ES and ED frames identification methods, several early studies used non-DL methods. They included methods such as the correlation between the heart cycle's frames [27], a combination of the Isomap method with manifold learning [28], and the non-negative matrix factorization (NMF) method [29]. In clinical applications, ES and ED frames are often detected using the simultaneously acquired ECG signals. This method is currently employed in various clinical software programs. However, echocardiography and acquiring ECG simultaneously increases the imaging equipment. This is challenging, especially in point-of-care ultrasonography (POCUS) systems and handheld devices.

As an example of ES and ED frames detection methods, Ref. [23] used a peak-detection method on the LV area curve of each video to detect ES and ED frames. In other words, it used the frame with maximum LV area as ED and the frame with minimum LV area as ES. Generally, its LVEF estimation pipeline included the LV segmentation step (using an atrous CNN model) and the LV area estimation step (counting the segmentation mask's pixels). Additionally, the LV length quantification step (using an ensemble learning), followed by the SPAL method and the EF formula (Its detailed pipeline was discussed in Section 4.3). More specifically, its ES and ED detection method was based

In Press

on first segmenting all frames in a video. Then, calculating the LV area in each frame and obtaining the LV area curve. Finally, using the peaks of the LV area curve to detect ED and ES frames. The authors further improved their ES and ED detection method by proposing an approach to remove frames with segmentation error. Overall, their detection method had acceptable accuracy and high interpretability. However, segmenting and quantifying LV in all video frames is time-consuming and highly sensitive to segmentation error. To conclude, the study indicates the potential of automatic interpretable ES and ED detection methods.

In recent years, many DL-based ES and ED detection methods have been developed. The DL methods reduce detection time and increases repeatability. However, they often require large datasets for training to achieve appropriate generalizability.

To address the challenges of image-based approaches, researchers have developed several video-based methods. These models directly take videos as input and do not require a separate ES and ED identification step. They use temporal dependencies in frames. As an example, Ref. [19] employed a 3D CNN model that took a video with at least 30 frames from the A4C view to detect the HFpEF. As another example, Ref. [26] used an eight-stream CNN and transformer model that took 18-frame videos from all four A2C, A3C, A4C, and PLAX views as input, and analyzed the images and their optical flow to estimate EF. It used transformers and attention mechanisms to capture temporal information in videos. The pipelines of these two studies were discussed in Section 4.3. To conclude, video-based methods often reduce the input video's frame rate significantly to optimize computational time and complexity. It may lead to loss of temporal information and frames' dependencies.

In general, both image-based and video-based EF estimation methods have their own pros and cons. The video-based DL models are often larger and more complex than image-based models. This may lead to longer processing times. However, image-based models require an extra preprocessing step to identify ES and ED frames. Additionally, image-based models capture only the end-diastolic and end-systolic phases of the target chamber. In contrast, video-based models capture temporal information and the target chamber's motion pattern throughout the whole cardiac cycle.

Table 1 summarizes the papers discussed in this section. It lists the extracted papers that have proposed a new pipeline for EF estimation. (The papers that evaluated the performance of commercial software programs are discussed in the next section.)

Table 1. Summary of the extracted papers proposed a new deep- or machine-learning-based pipeline to estimate EF from 2D echocardiogram videos.

Ref	Year	Architecture (Model Name)	Dataset	Input (View)	Task	Performance	Highlights
[18]	2019	An ML-based algorithm (AutoEF, BayLab inc, San Francisco, CA)	In-house data (Tr: more than 50000 studies, Te: 99 patients)	Videos (A2C and A4C)	Regression	For regression: MAD: 2.9±2% Compared to the reference: r: 0.95 ICC: 0.92 bias: 1% LOA: 12.1 For detecting ($EF \leq 35\%$): Acc: 0.92 Se: 0.9 Sp: 0.92 κ: 0.806	Extracted features directly from data. Estimated EF directly without first estimating LVESD and LVEDV. Used the average of EF values calculated by experts using the BPS method as a reference. Evaluated for detecting ($EF \leq 35\%$). Time of EF calculation per sample: 1 to 5 seconds on a personal computer
[16]	2020	A CNN model based on the ResNet50	In-house data (Tr: 400 M-mode images, including 200 samples for ($EF > 50$) and 200 samples for ($EF \leq 50$), Te: 60 samples)	M-mode images (PLAX)	Two-class classification ($EF > 50$) (Includes cropping the ROI)	Acc: 98% F1 for ($EF > 50$): 77% F1 for ($EF \leq 50$): 77%	Aimed to reduce the time of EF estimation. Used the BPS method as a reference.
[21]	2021	A modified	In-house data	ES and	Regression	For	Used dilated dense blocks.

In Press

		version of the U-Net model (DPS-Net) for segmentation + The BPS method	(36890 frames of 340 patients from three different centers (Tr: data from 100 individuals, Te: data from 240 studies)) and public dataset (CAMUS and EchoNet-Dynamic)	ED images (A2C and A4C)	(Includes segmentation)	segmentation: Dice for In-house data: (0.931 for A2C view and 0.933 for A4C view) Dice for CAMUS: (0.932 for ED and 0.928 for ES data) Dice for EchoNet-Dynamic: 0.925 For regression: ICC: 0.998 MAE: 1.2% RMSE: 1.3% Pearson r: 096 Pearson r for averaging EF for multiple cycles: (0.97 for AF, 0.97 for HCM, and 0.98 for DCM patients)	Evaluated on data captured by three different devices and from three heart disease phenotypes. Average segmentation time per person: 15.2±8.1 seconds Average EF estimation time per person: 8±3.3 seconds Achieved higher segmentation accuracy than the EchoNet-dynamic algorithm.
[17]	2021	U-Net for segmentation + four 1D CNNs	In-house data (1041 patients with HFpEF and 1263 asymptomatic individuals, Further test: 315 patients with HFpEF and 150 patients with SCOPD)	Images (A4C)	Classification (Includes segmentation and value extraction)	For the LA area: Acc: 0.91 Se: 0.96 Sp: 0.85 F1: 0.91 κ: 0.81 AUROC: 0.95 On further test dataset for LV volume: Acc: 0.85 Se: 0.79 Sp: 0.89 F1: 0.78 κ: 0.67 Highest AUROC: 0.95	Designed to detect HFpEF. Used inter-beat dynamics (length, area, width, and volume of both the left ventricle and the left atrium. Evaluated on patients from another hospital as test data. Average classification time per input: less than three minutes.
[25]	2022	A CNN model based on the U-Net For LV segmentation + The BPS method	In-house data (Tr: 70 patients, Te: 12 patients)	Images (A2C and A4C)	Regression (Includes segmentation, denoising, and unsupervised feature extraction)	For segmentation: Dice: 0.937 Jaccard index: 0.89 ConC: 0.852 HSD: 3.738 MASD:3.892 For regression: Pearson r for ES volume: 0.97±2% Pearson r for ED volume: 0.986±2% Pearson r for	Used denoising and feature extraction steps before the segmentation model to enhance its accuracy. Used a wavelet-based denoising method. Used the encoder of an autoencoder to extract features and reduce data dimensionality.

In Press

						EF: 0.88±3%	
[22]	2023	A DL-based framework (EchoEFNet) for LV segmentation and EF estimation	Te: public dataset (CAMUS), and private dataset (CMUEcho (contains 191 patients))	ES and ED images (A2C and A4C)	Regression (Includes segmentation)	On the CAMUS: Pearson's r: 0.854 bias: -0.374 LOA: 13.223 PELOA: 25.391 ICC: 0.854 On the CMUEcho: Pearson's r: 0.916 bias: -1.491 LOA: 6.758 PELOA: 12.992 ICC: 0.902	Used to segment LV and to extract echocardiographic landmarks (cardiac apex and mitral annulus). Used the BPS method and EF formula to calculate EF. Used ResNet50 with dilated convolution layers as the backbone. Automatically labelled the landmarks' locations for the model's training.
[19]	2023	A 3D CNN model (AI HFpEF)	In-house data (Tr and Val: 6756 patients, Te: 1284 patients)	Videos (A4C)	Three-class classification for (HFpEF, no HFpEF, and nondiagnostic) detection	Tr AUROC: 0.97 Val AUROC: 0.95 Te AUROC: 0.91 Te Se: 87.8% Te Sp: 81.9%	Classified patients into those with HFpEF and those without HFpEF. Training time by employing a single Tesla V100 GPU: 8 hours Identified patients with significantly higher mortality rates.
[26]	2024	An eight-stream CNN-transformer model	In-house data Tr: 8853 examinations, Val: 1898 examinations, Te: 1898 Examinations and also Public data (EchoNet-Dynamic and CAMUS datasets)	Videos (four views: A2C, A3C, A4C, and PLAX)	Regression (Includes view classification)	On all test data: R^2: 0.84 MAE: 4% bias: 0.1±5.2 On CAMUS (a four-stream version): R^2: 0.94 MAE: 1.9 bias: -0.1±2.4 On EchoNet-Dynamic (a two-stream version): R^2: 0.81 MAE: 4.1 bias: -0.3±5.4	Took videos that included 20 frames from all four views. Processed the original images and their optical flow. Employed the attention mechanism by adding a BERT layer at the end of their 3D CNN models to capture temporal information. Robust to misclassified views. Training time on a Nvidia DGX-2: about 5.6 days
[23]	2024	An Atrous CNN for segmentation + The single-plane area-length method	Public data (EchoNet-Dynamic and CAMUS)	ES and EF images (A4C)	Regression (Includes segmentation)	For segmentation: Dice on EchoNet-Dynamic: 0.922 For regression: Pearson r on EchoNet-Dynamic: 0.84 For categorization ($LVEF < 40\%$): AUROC on EchoNet-Dynamic: 0.98	Evaluated for categorizing reduced EF (EF<40%). Detected ES and ED frames using peak detection on the LV area curve. Used only the A4C view. Averaged LVEF across multiple cycles according to the American Society of Echocardiography.

In Press

						AUROC on CAMUS:0.9	
[24]	2025	A DL-based model for quality scoring and segmentation + The BPS method	In-house data (Tr: 4736 echocardiograms from 2461 participants, Te: 175 participants)	(A2C and A4C)	Regression (Includes view classification, quality scoring, and segmentation)	IoU for LV segmentation: more than 0.8 MA prediction for anatomical structure detection: 0.91 Acc for view classification: 0.96 ± 0.05 WA precision for quality assessment: (0.5 to 0.6) WA recall score for quality assessment: (0.5 to 0.6) ICC for regression: 0.821 Pearson r for regression on high-quality data: 0.90 ± 0.04	Proposed a five-point image quality scoring DL-based model. Used participants' data from 10 centers to train and from two centers to test. Achieved close quality scores to those of experts. Used the ES and ED frames extracted from the electrocardiogram signal as reference.
[20]	2025	A hybrid CNN-transformer model (QHAutoEF)	In-house data (2613 videos) and Public data (CAMUS dataset for further testing)	ES and ED images (A4C)	Regression (Includes segmentation)	For segmentation: Dice: (0.91 for ES and 0.93 for ED) MAE: 3.67% IoU: (0.84 for ES and 0.87 for ED) MSE: 20.31% RMSE: 4.51% MAPE: 6.72% R^2_{adj}: 0.82 RAE: 0.62 RSE: 0.21 ICC for regression: 0.9 For detecting (EF≥60%): AUROC: 0.92 Se: 0.88 Sp: 0.8	Combined CNN and transformer models to enhance the image analysis. Integrated the model as software into the echocardiography hardware. Model size: 10.2 MB Computational cost: (93.86 G FLOPS). Calculation time: 22ms Achieved better results than the U-Net, U-Net^{++}, ACNN, and attention U-Net on the CAMUS dataset.
[15]	2025	An ML algorithm (XGBoost algorithm)	In-house data (232 echocardiograms paired with CMRI)	Handcrafted features (First scenario:	Two-class classification for (EF<47) detection	AUROC for RV dysfunction prediction (EF<47) by the first scenario:	Estimated the right ventricle's EF. Used the CMRI data as a reference. Used data from pulmonary

In Press

	guideline-directed RV echocardiographic measures, Second Scenario: 44 features, including demographic variables, clinical, and echocardiographic features)	0.71±0.09 AUROC for RV dysfunction prediction (EF<47) by the second scenario: 0.66±0.09 ICC between the model and the CMRI: 0.83	regurgitation patients. Employed for diagnosing RV dysfunction (EF<47) prediction. The second scenario did not outperform the first scenario in RV dysfunction detection.

Ref: Reference, ML: Machine learning, Tr: Train, Te: Test, ES: End-systole, ED: End-diastole, EF: Ejection fraction, A2C: Apical two-chamber view, A4C: Apical four-chamber view, MAD: Mean absolute deviation, r: Correlation coefficient, ICC: Intra-class correlation coefficient, LOA: Limits of agreement, Acc: Accuracy, Se: Sensitivity, Sp: Specificity, κ: Cohence's kappa coefficient, LVESV: Left ventricle end-systolic volume, LVEDV: Left ventricle end-diastolic volume, BPS: Biplane Simpson, CNN: Convolutional network, PLAX: Parasternal long axis view, ROI: Region of interest, MAE: Mean absolute error, RMSE: Root mean square error, AUROC: Area under the receiver operating characteristic curve, AF: Atrial fibrillation, HTCM: Hypertrophic cardiomyopathy, DCM: Dilated cardiomyopathy, 1D: one-dimensional, HFpEF: Heart failure with preserved ejection fraction, SCOPD: Stable chronic obstructive pulmonary disease, LV: Left ventricle, ConC: conformity coefficient, DL: Deep learning, PELOA: Percentage of LOA, 3D: Three-dimensional, AI: Artificial intelligence, Val: Validation, GPU: Graphic processing unit, R^2: Coefficient of determination, A3C: Apical three-chamber view, IoU: Intersection over union, MA: Mean absolute, WA: Weighted average, FLOPs: floating point operations, MSE: Mean squared error, MAPE: Mean absolute percentage error, R^2_{adj}: Adjusted coefficient of determination, RAE: Relative absolute error, RSE: Relative squared error, XGBoost: Gradient Boosted Trees, CMRI: Cardiac magnetic resonance imaging, RV: Right ventricle, RVEDV: Right ventricle end-diastolic volume.

Table 1 shows that AI advances have significantly improved EF estimation methods in terms of accuracy and speed. Many AI-based models have been proposed recently. However, a fair comparison of these models based on their reported information is challenging. It is because they were often very diverse in terms of training and testing datasets and performance metrics. This table shows that regression-based methods often require more training data than classification-based models. It is because regression-based methods predict continuous values. Additionally, they are more sensitive to noisy data. The table also indicates that the end-to-end models often require more training data. It is because they often have more parameters to learn both the features and the prediction mapping than models based on handcrafted features. Also, end-to-end models often have more complex architectures than segmentation-based models. Furthermore, the table shows that various studies have combined segmentation methods with the BPS, the BPAL methods, or DL-based regression models to estimate EF. To conclude, these results are helpful for future development of EF estimation methods.

Table 1 also shows that the A4C is the most commonly used imaging view for automatically estimating EF from echocardiogram data. It indicates that recent studies have focused more on using data from multiple centers. Although several publicly available datasets (such as EchoNet-Dynamic and CAMUS) exist, their diversity, data labels, and sample size per disease are limited. In general, dataset bias is important for assessing DL-based models' generalizability, and current publicly available datasets often have biases that limit the generalizability of models trained on them. These datasets are often collected under specific conditions in terms of imaging equipment, patient samples, and data quality. They are also often annotated by a limited number of experts, which may cause variation in estimated values. For example, the EchoNet-Dynamic dataset contains 2D echocardiography data from the A4C view collected during a limited time period at a single center (Stanford University Hospital). These features may limit the EchoNet-Dynamic dataset's representation of other imaging modalities, imaging views, clinical practice, and patient population from other hospitals. In addition, the EchoNet-Dynamic dataset provides limited information on image-quality labels, patients' other cardiovascular diseases, and patient demographic information, which limits subgroup analysis across age, sex, cardiovascular disease, and image-quality groups. Similarly, the CAMUS dataset was collected from a single center, with annotations mostly provided by a single expert, and half

IUST PRESS

of the patients have reduced ejection fraction, which may not truly reflect an unselected population in common clinical conditions. For these reasons, current available datasets may not fully reflect the diversity of real-world data. As a result, dataset characteristics, including imaging systems, data quality, patient features, and disease distribution, may affect the representation learned by a DL-based model and cause the model to learn patterns that are specific to the underlying conditions. Therefore, good performance of DL-based models on a benchmark dataset does not guarantee clinical generalizability. DL-based models require extensive clinical evaluation for clinical integration; this is particularly important for devices imported from countries with different patient populations and disease patterns. The requirements for clinical deployment of DL-based algorithms are further discussed later in this paper. Therefore, developing more diverse datasets is essential. The table indicates that the correlation coefficient, area under the receiver operating characteristic curve (AUROC), and Dice similarity coefficient are the most frequently used parameters in reporting the results of regression, classification, and segmentation models, respectively. Studies often used the results of the BPS method or CMRI as the reference for evaluating their model's performance for estimating EF from echocardiogram data. Finally, the table shows that the quality of the acquired videos affects DL models' performance; therefore, several studies added a quality assessment step to address this problem.

To further analyze the models proposed in papers reviewed in Table 1, we summarized the loss functions of these models in Table 2. In general, an appropriate selection of a loss function is important in EF estimation because EF ranges are often distributed unevenly in datasets.

Table 2. Summary of loss functions used in papers presented in Table 1.

Reference	Task	Loss function	Loss type	Rationale for loss selection	EF imbalance handling
Asch et al. (2019) [18]	Regression	Not reported	Not reported	Not reported	Not reported
Sarkar et al. (2020) [16]	Two-class classification	Not reported	Not reported	Not reported	Used balanced dataset
Liu et al. (2021) [21]	Segmentation + Regression	Not reported	Not reported	Not reported	Not reported
Chiou et al. (2021) [17]	Classification + Segmentation	Not reported	Not reported	Not reported	Not applicable
Kulkarni et al. (2022) [25]	Segmentation + Unsupervised feature extraction + Regression	$L_{CNN} = L_{CE} + L_{Dice}$ (For Unsupervised feature extraction: $L_{AutoEncode} = MSE$)	Hybrid loss	Dice for segmentation: To give more weight to segmentation predictions that have a high overlap with the ground-truth labels.	Not reported
Li et al. (2023) [22]	Segmentation + Regression	$L_{segmentation} = L_{BCE}$ ($L_{landmark\ detection} = MSE$) ($L_{total} = \frac{1}{2}\sum_{i=1}^{2}(L_{seg} + L_{det})$)	Hybrid loss	Not reported	Not reported
Akerman et al. (2023) [19]	Three-class Classification	Cross-entropy	Binary classification loss	Not reported	Not applicable
Alvén et al. (2024) [26]	View classification + Regression	For view classification: Balanced Cross-entropy For LVEF prediction: MSE	For view classification: Binary classification loss	For view classification: Cross-entropy loss adjusted based on the number of samples in each class	Handled through loss-based adaptive sampling

In Press

Zhang et al. (2024) [23]	Segmentation + Regression	Not reported	Not reported	Not reported	Not reported
Li et al. (2025) [24]	View classification + Quality scoring + Segmentation + Regression	Not reported	Not reported	Not reported	Not reported
Lin et al. (2025) [20]	Segmentation + Regression	Not reported	Not reported	Not reported	Not reported
Duong et al. (2025) [15]	Two-class classification	Binary logistic	Binary: logistic	Not reported	Not applicable

EF: Ejection fraction; CE: Cross-entropy; MSE: Mean squared error; BCE: Binary cross-entropy; seg: Segmentation; det: landmark detection; LVEF: Left ventricle Ejection fraction.

Table 2 shows that despite the importance of the loss function for optimization and reproducibility of DL-based models, most of the papers did not report it. In addition, reviewed studies rarely report the effect of the loss function and EF imbalance handling method on their models' performance. Moreover, this table shows that for EF regression, the reviewed papers often used mean-squared error (MSE) as the loss function, which gives more weight to large errors. Additionally, they often used Dice Similarity (DS) Loss and cross-entropy loss, respectively, for segmentation and classification. Table 2 also shows that the reviewed papers used different loss functions, and they rarely reported the results of comparing different loss functions for their model. Additionally, the reviewed articles differ in terms of parameters, including data, data preprocessing, model, and evaluation methods. These parameters can affect final results, making it challenging to select a loss function that is optimal for all models. Overall, reporting the complete objective function of DL-based models, including the loss function and learning rate, is important for their reproducibility. Therefore, studies should provide more information about their selected loss function and the rationale for their selection.

In addition to the loss function, another important aspect of evaluating DL-based models and integrating them into clinical settings is the models' computational complexity. Therefore, we summarized the computational complexity and deployment characteristics of the reviewed models in Table 1 and present this summary in Table 3.

Table 3. Summary of deployment characteristics and computational complexity of the models proposed in papers presented in Table 1.

Reference	Model	Task	Frame/Image dimensions	Parameters	FLOPs	Hardware	Inference time
Asch et al. (2019) [18]	An ML-based algorithm (AutoEF, BayLab inc, San Francisco, CA)	Regression	Not reported	Not reported	Not reported	Not reported	1 to 5 seconds on a standard personal computer
Sarkar et al. (2020) [16]	A CNN model based on ResNet50	Two-class classification	Not reported	Not reported	Not reported	Not reported	Not reported
Liu et al. (2021) [21]	A modified version of the U-Net model (DPS-Net) for segmentation + The BPS method	Segmentation + Regression	256×256	Not reported	Not reported	Xeon 2.1 GHz CPU and one NVIDIA Titan XP GPU under Ubuntu 18.04	Average for segmentation: 15.2 ± 8.1 seconds Average for EF estimation: 8 ± 3.3 seconds
Chiou et al. (2021) [17]	U-Net for segmentation + Four 1D CNNs	Segmentation + Classification	128×128	Not reported	Not reported	Not reported	less than three minutes

In Press

Kulkarni et al. (2022) [25]	A CNN model based on the U-Net for segmentation + The BPS method	Segmentation + Denoising + Unsupervised feature extraction + Regression	658×658	5,420,360 trainable	Not reported	A single 12GB Tesla K80 GPU	Not reported
Li et al. (2023) [22]	A DL-based framework (EchoEFNet) For LV segmentation and EF estimation	Segmentation + Regression	224×224	Not reported	Not reported	Two Intel Xeon Gold 5215 2.5 GHz CPUs and one NVIDIA Tesla T4 GPU	Not reported
Akerman et al. (2023) [19]	A 3D CNN model (AI HFpEF)	Classification (Three-class)	256×256	Not reported	Not reported	Three NVIDIA Tesla V100 GPUs, each with 32 GB RAM; training time: 8 hours	Not reported
Alvén et al. (2024) [26]	An eight-stream CNN-transformer model	View classification + Regression	223×169	Not reported	Not reported	An NVIDIA DGX-2 with 22 CPUs and 4 GPUS, 218 GB RAM and 19 GB per GPU; training time: 136 hours	Not reported
Zhang et al. (2024) [23]	An Atrous CNN for segmentation + The single-plane area-length method	Segmentation + Regression	112×112	Not reported	Not reported	Not reported	Not reported
Li et al. (2025) [24]	A DL-based model for quality scoring and segmentation + The BPS method	View classification + Quality scoring + Segmentation + Regression	Not reported	Not reported	Not reported	The NNVIDIA Jetson Orin NX edge computing module	3 milliseconds per frame
Lin et al. (2025) [20]	A hybrid CNN-transformer model (QHAutoEF)	Segmentation + Regression	Not reported	Model size: 10.2 MB	93.86 G	Intel® Core ™ i7-10700F CPU @ 2.9 GHz and NVIDIA GeForce RTX 3090 GPU (24 G)	22 milliseconds
Duong et al. (2025) [15]	An ML algorithm (XGBoost algorithm)	Two-class classification	Not applicable	Not reported	Not reported	Not reported	Not reported

FLOPs: Floating-point operations per second; CNN: Convolutional neural network; CPU: Central processing unit; GPU: Graphics processing unit; BPS: Biplane Simpson's; GB: Gigabytes; LV: Left ventricle; EF: Ejection fraction; RAM: Random access memory; FPH: Frames per heartbeat; MB: Megabytes.

Table 3 shows that the reviewed models have different computational complexities, and only a few studies reported all computational requirements for their models. Table 3 also highlights that the reviewed papers rarely reported the inference time of their model. While inference time shows how long it takes for the model to generate the output, the number of parameters indicates the model's size and memory requirements, and Floating-point operations per second (FLOPs) indicates the amount of computation for inference. These

In Press

parameters are important for integrating DL-based models in clinical settings. In addition, Table 3 shows that models with higher performance often require more computational resources. The high computational requirements of DL-based models can limit their application in smaller devices, such as handheld ultrasound devices and POCUS systems. However, since the evaluation of the reviewed models was not the same, the information in Table 3 is not sufficient for a comparison between model complexity and accuracy. Additionally, the results indicate that the reviewed articles mostly focused on reporting performance parameters, such as MSE, Dice similarity coefficient, and accuracy; however, computational complexity and deployment requirements are important for models that are designed to integrate into clinical settings. Overall, a fair clinical applicability evaluation of DL-based models requires studies to report all computational parameters and implementation requirements for their models.

We categorized these extracted papers using the categories proposed in Fig. 1. The results are presented in Table 4. In this table, we considered the proposed pipeline in Ref. [16] as an end-to-end method. It is because, although it requires an ROI-cropping step, it is a preprocessing stage. It may be considered a conflict with the end-to-end meaning in general. However, unlike segmentation-based EF estimation methods, it does not require segmentation. Therefore, considering it as an end-to-end method is valid.

Table 4. Categorization of the papers presented in Table 1 using the thematic categorization proposed in Fig. 1.

Based on the task		Based on their feature extraction		Based on the approach		Based on the input	
Regression-based	Classification-based	Handcrafted-based	DL-extracted based	Segmentation-based	End-to-end	Image-based	Video-based
Asch et al. (2019) [18]	_	_	Asch et al. (2019) [18]	_	Asch et al. (2019) [18]	_	Asch et al. (2019) [18]
_	Sarkar et al. (2020) [16]	_	Sarkar et al. (2020) [16]	_	Sarkar et al. (2020) [16] (requires an ROI cropping step)	_	_
Liu et al. (2021) [21]	_	Liu et al. (2021) [21]	_	Liu et al. (2021) [21]	_	Liu et al. (2021) [21]	_
_	Chiou et al. (2021) [17]	Chiou et al. (2021) [17]	_	Chiou et al. (2021) [17]	_	_	Chiou et al. (2021) [17]
Kulkarni et al. (2022) [25]	_	Kulkarni et al. (2022) [25]	_	Kulkarni et al. (2022) [25]	_	Kulkarni et al. (2022) [25]	_
Li et al. (2023) [22]	_	Li et al. (2023) [22]	_	Li et al. (2023) [22]	_	Li et al. (2023) [22]	_
_	Akerman et al. (2023) [19]	_	Akerman et al. (2023) [19]	_	Akerman et al. (2023) [19]	_	Akerman et al. (2023) [19]
Alvén et al. (2024) [26]	_	_	Alvén et al. (2024) [26]	_	Alvén et al. (2024) [26]	_	Alvén et al. (2024) [26]
Zhang et al. (2024) [23]	_	Zhang et al. (2024) [23]	_	Zhang et al. (2024) [23]	_	Zhang et al. (2024) [23]	_
Li et al. (2025) [24]	_	Li et al. (2025) [24]	_	Li et al. (2025) [24]	_	Li et al. (2025) [24]	_
Lin et al. (2025) [20]	_	Lin et al. (2025) [20]	(Not mentioned)	Lin et al. (2025) [20]	_	Lin et al. (2025) [20]	_
_	Duong et al. (2025) [15]	Duong et al. (2025) [15]	_	_	_	_	_

DL: Deep learning, ROI: Region of interest.

5 Clinical Evaluation of Existing Software Programs

Despite studies proposing new EF estimation methods, several publications have focused on evaluating commercially available software programs in clinical applications and for real-world patients. Table 5 presents some of these publications. These publications can be broadly categorized into three groups: evaluating the software's overall performance, evaluating its performance for special patient populations (such as very ill or very young patients), and evaluating its

In Press

performance for POCUS systems and handheld devices data. These groups are being discussed in the following subsections.

5.1 Evaluating Overall Performance

In EF estimation software, accuracy and computational speed are two critical parameters. The computational speed should significantly reduce the expert's estimation time. In recent years, several publications have evaluated these programs against other EF estimation methods, including the BPS method, experts' diagnosis, or CMRI.

CMRI is the gold-standard imaging for estimating EF. It is important to examine whether software using 2D echocardiogram data can produce results close to this reference. As an example, Ref. [30] evaluated the performance of an AI-based EF estimation software program (Ligence Heart, version 3.42.0, Vilnuis Lithvania). The authors considered two scenarios: 1) using single-loop analysis guided by the expert (choosing the best loop manually, including the loop with the lowest foreshortening and the best image quality), and 2) using multi-loop analysis fully automated. They achieved a closer correlation with the CMRI results in the first scenario. The result contrasted with the theoretical benefits of using multi-loops. However, a limitation of the algorithm is its indirect estimation of EF. In other words, the program first estimated the LVEDV and LVESV, then employed them to calculate EF. Additionally, the results showed a range-dependent bias in the program's estimations (it overestimates or underestimates EF compared to the CMRI). Another limitation is that the authors analyzed only a limited amount of data with little patient diversity and image quality. To conclude, the study indicates the potential and feasibility of AI-based EF estimation programs.

As another example, Ref. [31] compared a DL-based software's (LVivo™ module, DiA Imaging Analysis, Beer-Sheva, Israel) results to EF values estimated by experts using the MBPS method and EF values estimated from CMRI data. The results showed comparable performance to the BPS method and the CMRI data. The software's results were closer (compared to the MBPS method) to those of CMRI in patients with low image quality, reduced EF, and abnormal wall motion. Overall, the software has several advantages and disadvantages. The authors mentioned that using software took longer than calculating EF manually using the MBPS method. Another limitation is its feasibility rate, as it was unable to classify views or estimate EF for several samples. Additionally, the software was unable to estimate EF from contrast agent-based echocardiogram data. Therefore, the authors did not use echocardiogram images acquired using contrast agents. The study also has both strengths and weaknesses. The authors did not analyze the software's pipeline or its training data. Also, generalizing the study's results to more specific patient populations (such as children or less common heart diseases) needs further examination. To conclude, the study indicates the potential of DL-based models to achieve results comparable to the manual method and CMRI.

Additionally, it is important to examine whether the EF estimation software's results are close to the experts' calculations. For example, Ref. [32] compared an AI-based clinical software (Ultromics EchoGo Core, Oxford, UK) with the manual BPS method in trastuzumab-treated patients. The results showed an acceptable correlation with the BPS method for A2C and A4C views. However, they showed low correlation with the BPS method for A3C data. A limitation of the study is that the authors used the data from only a single center. This limited the variety of heart diseases in the dataset. To conclude, the study shows the potential of AI-based programs in monitoring trastuzumab-treated patients.

It is also important for ES estimation software to perform adequately when used on videos acquired by inexperienced users. For example, Ref. [33] evaluated the performance of an AI-based software (model from Ref. [18]) on images captured by first-year medical students. The students used an ML-based algorithm (Caption Health Inc., Brisbane, CA, USA) to help them acquire videos by using quality assessment (guiding them to tilt, move, or rotate the transducer to acquire images). Three experts also captured videos from the patients and calculated EF by the BPS method. For each sample, the three experts' calculations were averaged and employed as a reference. The software results on the medical students' acquired data showed a strong correlation to the reference. An advantage of the software was its ability to estimate EF using only the PLAX view. However, the study was a primary study. It considered only a limited number of patients and heart diseases. For example, only one female participated in the study. This may introduce bias in results. Therefore, more variational data is needed to confirm the feasibility of the pipeline. To conclude, the study indicates the potential of fully automatic AI-based programs to estimate EF from data captured by novices.

As another example, Ref. [34] evaluated the performance of an AI-based software (Vscan app and Caption AI AutoEF software) that guides data acquisition and EF estimation. It examined whether the software can help a general practitioner achieve acceptable results with a handheld ultrasound device (Vscan AIR SL handheld device, GE HealthCare, Waukesha, WI). It used the EF values estimated from standard echocardiography by an expert as a reference. The results showed an acceptable correlation with the

In Press

reference. A limitation of the study is the small sample size. The authors selected patients by convenience. Also, they used only one general practitioner. These can limit the generalizability of the results. To conclude, the study indicates the potential of AI-based programs to estimate EF from data captured by low-experienced users using POCUS systems.

In general, some software may perform acceptably, but take longer to estimate EF than manual methods. For example, Ref. [35] evaluated the performance of an automated AI application (LVivo Seamless, DIA imaging Analysis, Beer-Sheva, Israel) for view selection and EF estimation. The authors used the EF values calculated by an expert using the MBPS method as a reference. The results had an acceptable correlation with the selected reference. However, the authors mentioned that using the application took longer than an expert's manual calculation using the BPS method. Generally, software's longer processing time may lead experts to prefer calculating EF manually rather than using the automatic software. Therefore, accuracy and speed are two important parameters in automatic software programs. Overall, the application has both pros and cons. The authors test-retested several randomly selected patients to evaluate the reliability and repeatability of the AI application. The results showed lower test-retest variability than the MBPS method. However, a limitation of the application is the feasibility rate. The application was unable to classify views and estimate EF for several patients. It limits its generalizability and reduces its reliability in clinical application. To conclude, the study highlights the potential of AI-based applications in estimating EF in realistic clinical applications.

5.2 Evaluating for Special Patient Populations

Table 5 indicates that the results of EF estimation software programs often correlate acceptably with the BPS method and experts' calculations. However, some programs perform poorly for less common heart diseases. It may be because they trained on low-diversity datasets, their generalizability is limited to less common heart diseases. Generally, training software programs on diverse datasets from different centers may improve their generalizability to different heart dysfunctions. Therefore, clinical software programs require further investigation and improvement for special patient populations, including diverse heart diseases, abnormal anatomies, and heterogeneous patient features. In this regard, the development of large public datasets may be helpful.

For example, Ref. [36] evaluated the performance of an ML-based algorithm (Real-Time EF, GE Healthcare, Chicago, USA) installed on an ultrasound device (Venue, GE Healthcare) for critically ill patients. The software first estimated the LVEDV and LVESV using the single-plane Simpson disk method (from A4C view), then calculated the LVEF. The results showed an acceptable correlation with EF values calculated manually by experts. Overall, the study has some strengths and weaknesses. The authors estimated the LVEF using the software by both experts and novices. The results of detecting the ($EF < 50\%$) had acceptable specificity, but low sensitivity for both experts and novices. It indicates the need to improve sensitivity. Moreover, the study examined the software only on the hemodynamically stable patients. Further studies are needed for hemodynamically unstable patients. To conclude, the study indicates the potential of AI-based software programs for estimating EF with better reproducibility than manual calculation.

Examining previous publications indicates that the datasets are often gathered from adult patients. Therefore, it is important to evaluate the performance of the EF estimation software for children. For example, Ref. [37] evaluated the performance of an ML-based software program (offline AutoLV, TomTec Imaging Systems, Unterschleissheim, Germany) on young adults and children. The results showed an acceptable correlation with the EF values calculated by an expert using the BPS method. The software significantly reduced EF estimation time compared to manual calculation. However, the authors used a limited sample size. They included a small number of young adults with normal hearts from a single center. Further evaluation is needed on a more diverse group of patients. To conclude, the study indicates the potential of ML-based EF estimation software for children.

5.3 Evaluating for POCUS Systems and Handheld Devices

In recent years, many software programs have been developed to estimate EF from data acquired by POCUS systems and handheld devices. The POCUS systems and handheld devices often produce low-quality images. Therefore, EF estimation programs need to perform acceptably on low-quality images. Additionally, POCUS systems and handheld devices are often used by inexperienced users. Therefore, EF estimation programs need to be easy to use. For inexperienced users, adding extra steps (such as view recognition and quality assessment) to the EF estimation pipeline can help them acquire better data.

As an instance, Ref [38] evaluated the performance of an AI-based clinical software program (LVivo EF, DiA Imaging Analysis Ltd) on images acquired by medical students using a handheld ultrasound device (Vscan Extend with Dual Probe, General Electric). The authors used EF values estimated by an echocardiographer in a formal echocardiography using the Simpson method and eyeballing as a reference. They considered three scenarios to estimate the EF: 1) visually

In Press

by cardiologists, 2) visually by the medical students, and 3) using the AI-based program on medical students' captured data. The results of the third scenario showed an acceptable correlation with the first scenario and the reference. The study has some pros and cons. Generally, estimating EF is easier for men than women. In the study, most participants were men, which may have led to biased results. Additionally, the program was unable to estimate the EF in several conditions, including low image quality or low LVEF. Moreover, the authors used patients from a single center. They also only examined videos from the A4C view. To conclude, the study indicates the potential of AI-based software to estimate EF from data acquired by low-experienced users using handheld ultrasound devices.

As another example, Ref. [39] assessed the performance of a DL-based EF estimation model (Caption Health, Inc, San Francisco, CA) for classifying left ventricular dysfunction from data captured by a point-of-care ultrasound device (Terason uSmart 3200t Plus). The authors considered two protocols: 1) using data from the clinical examinations at three centers captured by commonly used ultrasound devices, and 2) using data captured by POCUS systems using image acquisition guidance. They used EF values calculated by an echocardiographer using the biplane method of disks as a reference. The authors classified the estimated values from the first protocol into four classes: ($EF < 30\%$), ($30\% \leq EF \leq 52\%$), ($53\% \leq EF \leq 73\%$), and ($EF > 73\%$). They then evaluated the model's accuracy in detecting ($EF < 53\%$). They also compared the results with the experts' visual estimations and achieved acceptable accuracy. The results of both protocols showed acceptable agreement with the reference. Overall, the software has both pros and cons. The program was primarily developed for POCUS systems. Point-of-care systems are often used by low-experienced users who may have challenges in capturing the A2C view. Unlike some EF estimation programs that require images from both A2C and A4C views, this software can estimate EF from a single A2C, A4C, or PLAX view or any combination of them. This helps users select the view with the best quality. However, the authors did not explain or analyze the software's pipeline or training data. To conclude, the study indicates the potential of DL-based models in estimating EF and detecting ventricular dysfunctions from POCUS data.

Furthermore, Ref. [40] evaluated the performance of an AI-based software (Auto EF Quantification, Philips North America, Cambridge, MA) for estimating EF from echocardiogram videos with adequate and inadequate quality acquired by a handheld ultrasound device. The authors used the EF values estimated by using the Simpson's method as a reference. The results showed an acceptable correlation with the reference for videos with adequate image quality. However, the software performed poorly on videos with inadequate image quality (even worse than experts' visual estimation). In the study, the authors included a limited number of samples with low EF values. Moreover, all the videos were captured by a single expert. It may reduce the generalizability of the results. To conclude, the study highlights that the performance of EF estimation software may be affected by image quality. In these software programs, low image quality can lead to estimation errors. Therefore, it is necessary to develop robust methods. Also, to enhance the image quality in POCUS systems and handheld devices.

Table 5. Summary of the extracted papers evaluated the performance of deep- or machine-learning-based software programs for estimating EF from 2D echocardiogram data.

Ref	Year	Architecture (Model Name)	Dataset	Imaging View	Task	Performance	Highlights
[33]	2021	An AI-based software (the model from Ref. [18])	In-house data (Test: 14 consecutive patients, and three healthy volunteers)	A2C, A4C, PLAX	Regression	For A2C view: Pearson r: 0.84 bias: 3.94% (±7.8) For A4C view: Pearson r: 0.87 bias: 3.05% (±8.1) For PLAX view: Pearson r: 0.79 bias: 1.35% (±5.63)	Used by 19 first-year medical students (at least three novices scanned each patient). Compared the results with those of three experts as ground truth. Able to estimate LVEF using only the PLAX view.
[39]	2021	A DL-based algorithm (Caption Health, Inc, San Francisco, CA)	In-house data (Test: 166 clinical examinations, and 67 patients	A2C, A4C, PLAX, or any combination of them	Regression	ICC of their protocol one: (0.86-0.95) biases<2% ICC of their protocol two: 0.84	Developed for POCUS data. Considered two protocols: 1) used data acquired by commonly used ultrasound devices.

In Press

			whose data were acquired using POCUS systems)			biases: (2.5±6.4%)	2) used data acquired by POCUS systems. Used EF values estimated by echocardiographers using the biplane method of disks as a reference. Evaluated for detecting ($EF < 53\%$). The LVEF estimation time per data for each view: 1 to 5 seconds
[37]	2022	An ML-based software program Including AI and speckle tracking (AutoLV, TomTec,Imaging Systems, Unterschleissheim, Germany)	In-house data (Test: 50 normal volunteers)	A2C, A4C, Long axis	Regression (Includes segmentation)	Between the software estimation and reference: r: 0.8373 Software's analysis time: 111.9±35.7 seconds The BPS method's analysis time: 155.2±37.6 seconds	Investigated the software's reliability for children and young adults. Used by a sonographer's manual calculation using the BPS method as a reference. Achieved a good correlation between the program's estimations and the expert's calculations.
[36]	2022	An ML-based algorithm (Real-Time EF, GE Healthcare, Chicago, USA), which has been installed on a cart-based ultrasound system (Venue, GE Healthcare)	In-house data (Test: 95 ICU stay patients, 15 patients excluded because of low image quality)	A4C	Regression (Includes view classification)	Between the results of novices using the software and the reference: r: 0.81 bias: 0±7% Between the results of experts' using the software and the reference: r: 0.86 bias: 0±6% For detecting $LVEF < 50\%$: Experts' Se: 70% Experts' Sp: 98% Novices' Se:73% Novices' Sp: 98%	Evaluated for the critically ill patients. Compared experts' and novices' results using the software. Used the EF values calculated manually by experts as a reference. Achieved acceptable specificity in detecting ($LVEF < 50\%$), but its specificity needed improvement.
[35]	2023	An automated AI application (LVivo Seamless, DIA imaging Analysis, Beer-Sheva, Israel)	In-house data (Test: 1083 consecutive patients, but a total of 889 echoes were analyzed)	A2C, A4C	Regression (Includes view selection)	For LVEF: Pearson r: 0.87 bias: +4.5% For LVESV: Pearson r: 0.93 bias: -14 mL For LVEDV: Pearson r: 0.89 bias: -12 mL	Used the experts' results using the modified BPS method as a reference. Achieved a longer median processing time (4.5 minutes) than the expert using the BPS method (2.85 minutes).
[32]	2023	An AI-based clinical	In-house data	A2C, A3C,	Regression	Pearson r: 0.895	Examined whether this software is

In Press

		software (Ultromics EchoGo Core, Oxford, UK)	(Test: 142 trastuzumab-treated patients underwent a total of 648 echocardiographic studies)	A4C		for A3C view: Pearson r: 0.423 feasibility rate: 51%	feasible for EF assessment. Showed an acceptable correlation with the manual method. Compared the results with the BPS method. Achieved the lowest correlation with manual calculation for the A3C view. Achieved a significantly lower feasibility rate than manual calculation.
[38]	2023	An AI-based clinical software program (LVivo EF, DiA Imaging Analysis Ltd) on a handheld ultrasound device (Vscan Extend with Dual Probe, General Electric)	In-house data (Test: 88 patients, but only successful in 82 cases)	A4C	Regression	Pearson r between the students' results using AI software and the experts' estimations: 0.83 Pearson r between the students' visual estimations and the experts' estimations: 0.51	Evaluated on data acquired by medical students using POCUS systems. Used the EF values calculated by echocardiographers using the Simpson method, and eyeballing as a reference. Compared three EF estimation scenarios: 1) visually by the students 2) by students using the software 3) visually by a cardiologist. Achieved better performance compared to medical students' visual estimation.
[31]	2024	A DL-based software (LVivoEF™, DiA Imaging Analysis, Beer-Sheva, Israel)	In-house data (229 patients)	A2C and A4C	Regression (Includes view selection)	Pearson r between the results of the software and the modified BPS method: 0.93 Pearson r between the results of the software and CMRI: 0.89 Pearson r between the results of the BPS method and CMRI: 0.89 For the software: EF estimation: Bias compared to CMRI: 2.2% LOA: (-13.5 to 17.9%) ICEE: 1 TTR variability:	Used the EF values estimated using CMRI as a reference. Achieved results closer (than the BPS method) to CMRI for patients with low image quality, reduced EF, and abnormal wall motions. Achieved better ICEE results than the BSP method. Achieved lower TTR variability than the BPS method. Feasibility rate for the tested data: 87% Processing time: three to six minutes. Median processing time of the BPS method: 2.6 min Software's processing

In Press

						2.5% For the modified BPS method: EF estimation: Bias compared to CMRI: -2.2% LOA: (18.6 to 14.1%) ICEE < 0.91 TTR variability: 7.9%	time: 4 to 7 minutes
[40]	2024	An AI-based model (Auto EF Quantification, Philips North America, Cambridge, MA)	In-house data (54 clips from 28 patients)	A4C	Regression	Pearson r between the software's results and the reference: 0.56 Pearson r between the expert's results and the reference: 0.8 Pearson r between the software's results and the reference for 27 adequate quality clips: 0.86 Pearson r between the software's results and the reference for 27 inadequate quality clips: 0.46	Used the EF values calculated using the BPS method as a reference. Included data of adult patients who were scheduled to undergo lung or heart transplant, aortic, coronary bypass graft, or valve surgery. Achieved lower performance for cases with inadequate image quality.
[30]	2025	An AI-based software program (Lingence Heart, version 3.42.0, Vilnuis Lithvania)	In-house data (Test: 118 patients)	A2C, A4C	Regression	For multi-loop: Pearson r: 0.87 κ: 0.68 For one-loop: Pearson r: 0.89 κ: 0.75	Compared the software's results with the cardiac magnetic resonance as a reference. Acceptable correlation with CMRI in both single-loop and multi-loop scenarios. Achieved better results in a single loop compared to multiple loops.
[34]	2025	An AI-based clinically available software (Caption AI AutoEF software)	In-house data (Test: 25 patients)	A2C, A4C, PLAX	Regression (Includes AI-guided image acquisition assistance)	For regression: Pearson r: 0.95 For classification of the estimated values into three classes: Se: 90% Sp: 100%	Examined whether a general practitioner can accurately estimate EF using the software. Used the EF values from a standard echocardiography view as a reference. Achieved acceptable quality for 19 cases. Estimated EF values were classified into:

In Press

	($EF < 40\%$), ($40\% \leq EF \leq 49\%$), and ($EF \geq 50\%$).

Ref: Reference, AI: Artificial intelligence, A2C: Apical two-chamber view, A4C: Apical four-chamber view, PLAX: Parasternal long axis view, r: Correlation coefficient, LVEF: Left ventricle ejection fraction, DL: Deep learning, ICC: Intra-observer correlation coefficient, POCUS: Point-of-care ultrasonography, ML: Machine learning, Se: Sensitivity, Sp: Specificity, LVESV: Left ventricle end systolic volume, LVEDV: Left ventricle end diastolic volume, A3C: Apical three-chamber view, BPS: Biplane Simpson method, EF: Ejection fraction, ICEE: Inter-observer correlation coefficient, CMRI: Cardiac magnetic resonance imaging, TTR: True test-retest, LOA: Limits of agreement, κ: Cohen's kappa coefficient.

Overall, Table 5 indicates that commercially available EF estimation software programs often have limitations. These limitations may reduce their reliability, leading cardiologists to prefer calculating EF manually. Therefore, widespread clinical application of these programs is associated with improvements in their accuracy and speed. These programs also have strong potential to help novices accurately estimate EF, especially when using POCUS systems.

6 Results and Discussion

In this study, we conducted a narrative review of DL-based EF estimation methods that use 2D echocardiogram data. We did it by analyzing 23 original articles from WOSCC by the end of January 2026. We also categorized these methods, highlighted their features, compared their performance, and analyzed their advantages and limitations. The results reveal that in previous studies, a broad range of DL architectures (from basic CNNs and transformers to hybrid and multi-stream models) have been employed to estimate the EF. In the context of DL development, recent publications have mostly focused on developing hybrid DL-based models to improve the speed and accuracy of EF estimation. Additionally, in the context of clinical validation, several studies have concentrated on the clinical examination of these models for various image qualities, diseases, and abnormal heart anatomies. We also conducted a light bibliometric analysis, revealing that the synthesized data in this review reflect current research trends and existing gaps in the field. The subsections below present the findings.

6.1 Current Research Status and Hotspots

We conducted a light bibliometric analysis of the extracted papers in the Method section. The results are presented in Fig. 2. They indicate that in these studies, the core concept is to automate the quantification of LVEF using AI in echocardiography. Many previous studies have focused on estimating LVEF. It is because of the importance of LVEF in circulating blood through the whole body. However, several recent studies (such as [41]) have focused on estimating the EF of other cardiac chambers. The reason is that in blood circulation, the performance of other parts of the cardiovascular system is also important.

Analyzing the journals shows that these papers have been published in both clinical cardiology journals and interdisciplinary AI journals. This indicates that the field has attracted the attention of both medical professionals and computer scientists. Additionally, the fact that these papers are often published open-access suggests that the field is rapidly growing and often supported by substantial funding.

The research theme analysis shows a shift in publications from conceptual studies to clinical studies on the use and adaptation of the methods in clinical devices. This indicates that the field is mature and still evolving. Additionally, we performed a keyword co-occurrence network analysis and clustered the papers. We also interpreted and labeled the clusters based on the papers' content into five categories. The results are presented in Table 6. They indicate that the United States and China are dominant in the number of publications. Moreover, European countries mostly focused on clinical validation studies. These results are helpful for future collaboration and funding in the field.

To conclude, the results of the bibliometric analysis indicate that the synthesis of publications in our narrative review highly reflects the current research status, trends, and future direction of the field.

Table 6. Thematic clusters of extracted publications on deep- and machine-learning methods for estimating EF from 2D echocardiogram data were identified through bibliometric analysis on the authors' keywords.

Category number	Cluster theme	Highlights
1	The segmentation-based DL models for EF estimation	The models proposed in these papers often have an acceptable accuracy. However, their performance is sensitive to the segmentation accuracy.
2	The end-to-end DL models for EF estimation	The proposed models in these papers often have low explainability and require large datasets.
3	The transformers and hybrid models	The models introduced in these publications are new and often offer great potential for future advancements.
4	Clinical diagnosis	These papers often use the ejection fraction not as an output, but as a feature

In Press

		for further diagnosis.
5	Validation methods against the gold standard	These publications have evaluated the performance of the previous EF estimation methods for clinical applications.

DL: Deep learning.

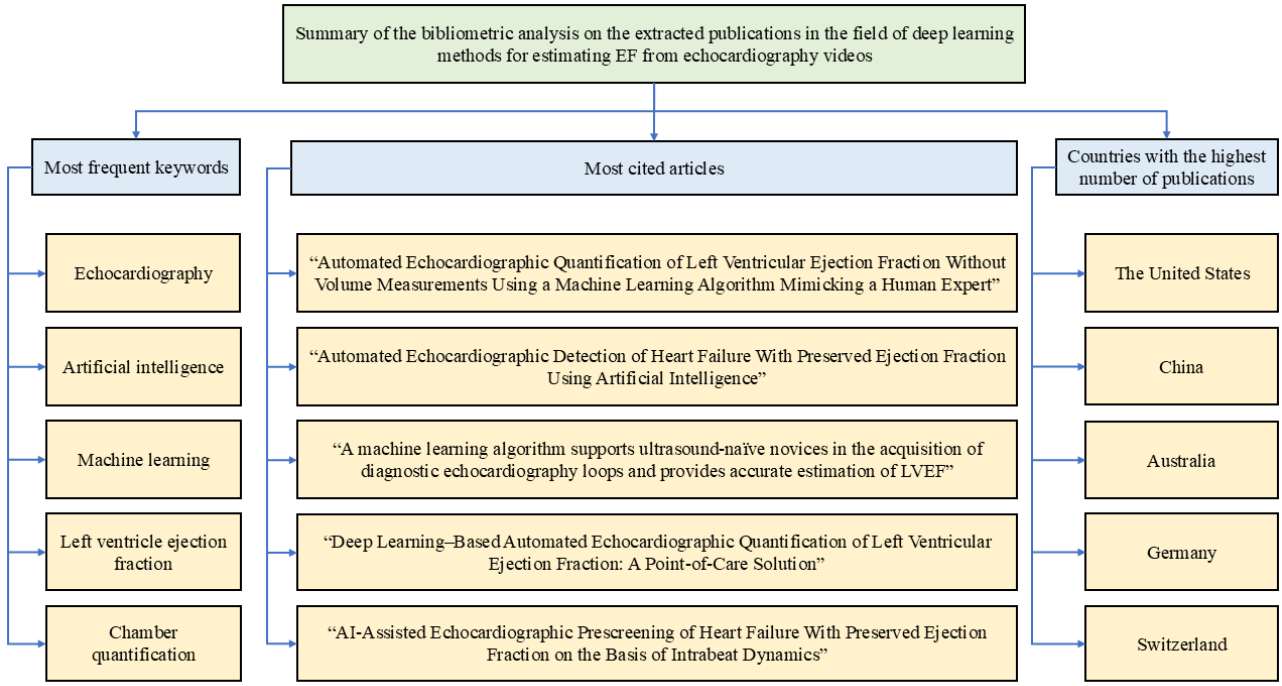


Fig 2. Summary of the bibliometric analysis conducted on the extracted publications about deep- and machine-learning EF estimation methods that use 2D echocardiogram data.

6.2 Challenges and Limitations

In recent years, numerous studies have developed DL-based EF estimation methods that show good results on benchmark datasets. However, their good performance on benchmark datasets alone is not sufficient for integrating these algorithms into clinical settings. Their clinical integration requires extensive regulatory evaluation and technical validation across different data acquisition conditions, patient populations, and imaging systems. Based on how the DL-based algorithms are used in clinical settings, there are several requirements, including clinical performance, cybersecurity, and risk management, that should be addressed. Overall, regulatory considerations are important for integrating DL-based methods in clinical settings, and AI-based algorithms must receive the required regulatory approval before they can be used in clinical settings. The required approval depends on the country where the algorithm will be used and may include AI-specific requirements and device regulation [42]. In the United States (U.S.), the main pathway for AI-enabled medical equipment by the Food and Drug Administration (FDA) is the FDA 510(k) pathway. This pathway is based on showing that the new device is similar to a device that already exists in the market. AI-based cardiovascular devices' premarket evaluation and requirements before clinical use can include clinical validation and algorithm performance [43]. For AI-based algorithms in echocardiography such as EF estimation methods, randomized controlled trials (RCTs) are the gold standard for their clinical evaluation because clinical decisions may be affected by these algorithms. In the United Kingdom (U.K.) and Europe, similar regulatory pathways are used through, respectively, the Medicines and Healthcare products Regulatory Agency (MHRA) and the Medical Device Regulation (MDR) [42]. Some of the regulatory requirements are needed even after AI-based medical devices have entered markets, including quality requirements and post-market reporting. The FDA is developing guidance about the changes that apply to improve the performance of AI-based algorithms after they enter the market [43]. Overall, AI applications in echocardiography and DL-based approaches are a rapidly evolving field; as a result, regulatory requirements for these algorithms are still developing.

Additionally, several DL-based EF estimation software programs (such as Ultromics EchoGo Core [44], Philips HeartModel [45], and TomTec AutoLV [37]) have been developed in recent years. However, the

IEEE

widespread applicability of these programs in clinical diagnoses is limited. It is because these programs have some challenges. One challenge is that the programs often rely on endocardium segmentation. However, their segmentation is often inaccurate and requires the user to adjust the borders. Additionally, the programs typically require the operator to select the ES and ED frames manually or to synchronize the video with the acquired ECG. Moreover, fully automatic programs often show suboptimal performance in special patient populations. To conclude, the widespread use of these programs requires addressing these challenges.

Despite significant advances in EF estimation programs, their use in clinical applications faces several challenges. One main challenge is limited generalizability. In previous studies, DL models were often trained and tested on datasets with limited data variation. When applied to real-world data with different imaging settings, operator experience, patient features (such as body mass index [46], obesity, and certain pulmonary diseases [47]), and anatomic abnormalities, they may show lower performance. The low generalizability of these programs limits their clinical application. Another challenge is low interpretability and operation as a black box. These programs often lack explainability in feature extraction and decision-making. This limits their reliability in clinical applications. Furthermore, the low quality of the videos often reduces the programs' performance. The low quality of data captured by POCUS systems or handheld devices necessitated the development of programs robust to video quality. To conclude, further use of DL models in clinical applications relies on addressing these challenges.

6.3 Research Gaps and Future Direction

The field of applying AI to estimate EF is rapidly evolving. The findings of this review paper suggest several promising research areas for future studies. One area is developing DL-based pipelines that optimize hardware cost, power consumption, memory footprint, and inference time. This area is widely useful for developing more efficient handheld ultrasound devices. Another area is the development of explainable DL-based models specifically tailored for echocardiogram data. It is important because previously developed models are often not specifically designed for echocardiogram data. Explainable models can improve the programs' clinical reliability.

Furthermore, another area is developing a fully automatic pipeline to estimate EF from raw RF data. Recently, many studies have focused on fully automating EF estimation by adding steps, such as guided probe placement [48], view classification [49, 50], and image quality assessment [51]. However, these studies often focused on processing images or videos. Future studies can focus on leveraging beamforming-aware DL-based models that can directly estimate EF from raw RF signals with an attention mechanism. These models will reduce the sensitivity of the EF estimation to image reconstruction and post-processing methods. Moreover, another area is developing models for captioning echocardiogram videos using estimated EF data and specialized terminology familiar to experts in this domain. Finally, another area is the development of DL-based methods for data encryption. This helps to ensure the security of data transmission and patients' willingness to share their data for research purposes in this field.

7 Conclusion

To conclude, the results of this study indicate that the field of applying DL-based methods to estimate EF from 2D echocardiogram data is mature yet rapidly evolving. Several previous studies have focused on developing new pipelines to improve the estimation accuracy. Other studies have evaluated commercially available software programs in clinical settings for special patient populations, including children and those with less common heart diseases. It is anticipated that future studies will focus more on designing pipelines with optimal implementation (improving processing speed and software performance for large-scale data). Additionally, they will further focus on evaluating EF estimation programs for less common patient populations, especially those who require repetitive echocardiogram follow-up.

The importance of reviewing the DL-based EF estimation methods lies in their scientific, clinical, and policy-related impacts. From a scientific perspective, the review reveals performance, limitations, and research gaps. The results guide researchers to develop more robust and reliable software programs. From a clinical perspective, the review reveals the performance of these DL-based models on different data quality and cardiac pathologies. This helps to evaluate their reliability and prevent diagnostic errors. Finally, from a policy-related perspective, the review reveals the potential of AI advances in medical decision-making. Bringing more attention to investigations in this field, including handheld devices and POCUS systems. To conclude, the results of this study are valuable for researchers to understand research gaps, for clinicians to gain insight into advances in this field, and for investors to inform relevant investment decisions.

Overall, the current study has several limitations worth noting. One limitation is that we only extracted the relevant publications from the WOSCC. Although WOSCC is one of the most commonly used resources, excluding other major databases (such as Scopus and PubMed) may result in missing some relevant

In Press

publications. Moreover, we limited our results to January 2026. This may cause the absence of the most recent papers. Additionally, we conducted a light bibliometric analysis using a limited number of publications. Therefore, the results may not comprehensively represent the entire contributions in this field. Moreover, another limitation is that the fair comparison of the previously proposed ML- and DL-based EF estimation pipelines' performance and reproducibility is challenging. It is because they are often trained and tested on private (in-house) datasets. To address this limitation, using a common publicly available dataset to evaluate and report the results of these pipelines is helpful. Also, the lack of standardized evaluation metrics across studies limits the direct comparison of models' performance. It highlights the requirement for unified evaluation criteria.

In conclusion, despite significant progress in this field, several challenges and research gaps remain and should be addressed in future studies. It is also worth noting that when experts interpret echocardiogram data, they often consider additional patient information, such as medical history, symptoms, and medications. DL-based methods may overlook this additional information in their decisions because they often take limited patient data as input.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

R. Ghorbani: Idea and Conceptualization, Research and Investigation, Analysis, Original Draft Preparation. **H. Behnam:** Idea and Conceptualization, Supervision, Revise and Editing, Verification.

Funding

No funding was received for this work.

Informed Consent Statement

Not applicable.

Declaration of generative AI and AI-assisted technologies

The authors declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

Acknowledgement

This work is based upon research funded by Iran National Science Foundation (INSF) under project No 40501921.

References

[1] C. R. Passos, A. A. Moreira, R. F. Reis, R. W. dos Santos, M. Lobosco, and B. M. Rocha, "A Coupled Model of the Cardiovascular and Immune Systems to Analyze the Effects of COVID-19 Infection," *BioTech*, vol. 14, 2025.

[2] S. C. W. Tan, B.-B. Zheng, M.-L. Tang, H. Chu, Y.-T. Zhao, and C. Weng, "Global Burden of Cardiovascular Diseases and its Risk Factors, 1990-2021: A Systematic Analysis for the Global Burden of Disease Study 2021," *QJM : monthly journal of the Association of Physicians*, 2025.

[3] I. Gotsman, D. R. Zwas, A. Keren, O. Amir, and D. Leibowitz, "Beyond the Initial Insult: Clinical Characteristics and Prognosis of Heart Failure with Improved Ejection Fraction," *Cardiology*, vol. 150, pp. 147 – 154, 2024.

[4] X. Chen *et al.*, "Optimization of Clinical Trial Design and Decision-Making for Heart Failure with Preserved Ejection Fraction (HFpEF): A Meta-Analysis Based on a Placebo Response Model," *Cardiovascular Therapeutics*, vol. 2025, 2025.

[5] A. Barison, A. Aimo, G. Todiere, C. Grigoratos, G. D. Aquaro, and M. Emdin, "Cardiovascular magnetic resonance for the diagnosis and management of heart failure with preserved ejection fraction," *Heart failure reviews*, vol. 27, no. 1, pp. 191–205, 2022.

[6] N. Mumoli and S. Marengo, "Clinical utility of echocardiography in internal medicine: a narrative review," *Italian Journal of Medicine*, 2024.

[7] Y. Nagata *et al.*, "Impact of image quality on reliability of the measurements of left ventricular systolic function and global longitudinal strain in 2D echocardiography," *Echo Research & Practice*, vol. 5, no. 1, pp. 28–39, 2018.

[8] T. Barry *et al.*, "The role of artificial intelligence in echocardiography," *Journal of imaging*, vol. 9, no. 2, p. 50, 2023.

[9] J. M. Farina *et al.*, "Artificial intelligence-based prediction of cardiovascular diseases from chest radiography," *Journal of Imaging*, vol. 9, no. 11, p. 236, 2023.

[10] C. Krittanawong *et al.*, "Deep learning for echocardiography: introduction for clinicians and future vision: state-of-the-art review," *Life*, vol. 13, no. 4, p. 1029, 2023.

[11] G. Zamzmi, L.-Y. Hsu, W. Li, V. Sachdev, and S. Antani, "Harnessing machine intelligence in automatic echocardiogram analysis: Current status, limitations, and future directions," *IEEE reviews in biomedical engineering*, vol. 14, pp. 181–203, 2020.

[12] R. M. Lang *et al.*, "Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the

In Press

American Society of Echocardiography and the European Association of Cardiovascular Imaging," *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.

[13] A. Ghorbani *et al.*, "Deep learning interpretation of echocardiograms," *NPJ digital medicine*, vol. 3, no. 1, p. 10, 2020.

[14] G. Savarese, D. Stolfo, G. Sinagra, and L. H. Lund, "Heart failure with mid-range or mildly reduced ejection fraction," *Nature Reviews. Cardiology*, vol. 19, pp. 100 – 116, 2021.

[15] S. Q. Duong *et al.*, "Machine learning prediction of right ventricular volume and ejection fraction from two-dimensional echocardiography in patients with pulmonary regurgitation," *The International Journal of Cardiovascular Imaging*, pp. 1–14, 2025.

[16] P. Sarkar and V. Chandra, "A novel approach for detecting abnormality in ejection fraction using transthoracic echocardiography with deep learning," 2020.

[17] Y.-A. Chiou, C.-L. Hung, and S.-F. Lin, "AI-assisted echocardiographic prescreening of heart failure with preserved ejection fraction on the basis of intrabeat dynamics," *Cardiovascular Imaging*, vol. 14, no. 11, pp. 2091–2104, 2021.

[18] F. M. Asch *et al.*, "Automated echocardiographic quantification of left ventricular ejection fraction without volume measurements using a machine learning algorithm mimicking a human expert," *Circulation: Cardiovascular Imaging*, vol. 12, no. 9, p. e009303, 2019.

[19] A. P. Akerman *et al.*, "Automated echocardiographic detection of heart failure with preserved ejection fraction using artificial intelligence," *JACC: Advances*, vol. 2, no. 6, p. 100452, 2023.

[20] M. Lin *et al.*, "A combined system with convolutional neural networks and transformers for automated quantification of left ventricular ejection fraction from 2D echocardiographic images," *Intelligent Medicine*, vol. 5, no. 1, pp. 46–53, 2025.

[21] X. Liu *et al.*, "Deep learning-based automated left ventricular ejection fraction assessment using 2-D echocardiography," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 321, no. 2, pp. H390–H399, 2021.

[22] H. Li, Y. Wang, M. Qu, P. Cao, C. Feng, and J. Yang, "EchoEFNet: Multi-task deep learning network for automatic calculation of left ventricular ejection fraction in 2D echocardiography," *Computers in Biology and Medicine*, vol. 156, p. 106705, 2023.

[23] Y. Zhang *et al.*, "Development of automated neural network prediction for echocardiographic left ventricular ejection fraction," *Frontiers in Medicine*, vol. 11, p. 1354070, 2024.

[24] X. Li *et al.*, "An automatic and real-time echocardiography quality scoring system based on deep learning to improve reproducible assessment of left ventricular ejection fraction," *Quantitative Imaging in Medicine and Surgery*, vol. 15, no. 1, pp. 770–785, 2025.

[25] P. Kulkarni and D. Madathil, "Fully automatic segmentation of LV from echocardiography images and calculation of ejection fraction using deep learning," *International Journal of Biomedical Engineering and Technology*, vol. 40, no. 3, pp. 241–261, 2022.

[26] J. Alvén, E. Hagberg, D. Hagerman, R. Petersen, and O. Hjelmgren, "A deep multi-stream model for robust prediction of left ventricular ejection fraction in 2D echocardiography," *Scientific Reports*, vol. 14, no. 1, p. 2104, 2024.

[27] N. Kachenoura, A. Delouche, A. Herment, F. Frouin, and B. Diebold, "Automatic detection of end systole within a sequence of left ventricular echocardiographic images using autocorrelation and mitral valve motion detection," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007: IEEE, pp. 4504–4507.

[28] A. Shalbaf, H. Behnam, P. Gifani, and Z. Alizadeh-Sani, "Automatic detection of end systole and end diastole within a sequence of 2-D echocardiographic images using modified Isomap algorithm," in *2011 1st Middle East Conference on Biomedical Engineering*, 2011: IEEE, pp. 217–220.

[29] H. Dröge, B. Yuan, R. Llerena, J. T. Yen, M. Moeller, and A. L. Bertozzi, "Mitral valve segmentation using robust nonnegative matrix factorization," *Journal of imaging*, vol. 7, no. 10, p. 213, 2021.

[30] P. Mołek-Dziadosz *et al.*, "Left ventricular ejection fraction assessment: artificial intelligence compared to echocardiography expert and cardiac magnetic resonance measurements," *POLISH ARCHIVES OF INTERNAL MEDICINE-POLSKIE ARCHIWUM MEDYCYNY WEWNETRZNEJ*, vol. 135, no. 9, 2025.

[31] K. M. Sveric *et al.*, "Improved assessment of left ventricular ejection fraction using artificial

In Press

intelligence in echocardiography: A comparative analysis with cardiac magnetic resonance imaging," *International Journal of Cardiology*, vol. 394, p. 131383, 2024.

[32] J. Jiang, B. Liu, Y. Li, and S. Hothi, "Clinical service evaluation of the feasibility and reproducibility of novel artificial intelligence based-echocardiographic quantification of global longitudinal strain and left ventricular ejection fraction in trastuzumab-treated patients," *Frontiers in cardiovascular medicine*, vol. 10, p. 1250311, 2023.

[33] M. Schneider *et al.*, "A machine learning algorithm supports ultrasound-naïve novices in the acquisition of diagnostic echocardiography loops and provides accurate estimation of LVEF," *The International Journal of Cardiovascular Imaging*, vol. 37, no. 2, pp. 577–586, 2021.

[34] R. E. Harskamp, "Left Ventricular Ejection Fraction Assessed by a General Practitioner Using AI-Guided Point-of-Care Ultrasonography: Feasibility and Diagnostic Agreement With Conventional Echocardiography," *Journal of Diagnostic Medical Sonography*, p. 87564793251390476, 2025.

[35] K. M. Sveric *et al.*, "Single-site experience with an automated artificial intelligence application for left ventricular ejection fraction measurement in echocardiography," *Diagnostics*, vol. 13, no. 7, p. 1298, 2023.

[36] R. Varudo *et al.*, "Machine learning for the real-time assessment of left ventricular ejection fraction in critically ill patients: a bedside evaluation by novices and experts in echocardiography," *Critical Care*, vol. 26, no. 1, p. 386, 2022.

[37] L. Li *et al.*, "Machine learning-enabled fully automated assessment of left ventricular volume, ejection fraction and strain: experience in pediatric and young adult echocardiography," *Pediatric cardiology*, vol. 45, no. 6, pp. 1183–1191, 2024.

[38] Z. Dadon *et al.*, "Empowering Medical Students: Harnessing Artificial Intelligence for Precision Point-of-Care Echocardiography Assessment of Left Ventricular Ejection Fraction," *International Journal of Clinical Practice*, vol. 2023, no. 1, p. 5225872, 2023.

[39] F. M. Asch *et al.*, "Deep learning–based automated echocardiographic quantification of left ventricular ejection fraction: a point-of-care solution," *Circulation: Cardiovascular Imaging*, vol. 14, no. 6, p. e012293, 2021.

[40] E. J. MacKay *et al.*, "Pragmatic evaluation of a deep-learning algorithm to automate ejection fraction on hand-held, point-of-care echocardiography in a cardiac surgical operating room," *Journal of Cardiothoracic and Vascular Anesthesia*, vol. 38, no. 4, pp. 895–904, 2024.

[41] M. Murayama *et al.*, "Deep learning to assess right ventricular ejection fraction from two-dimensional echocardiograms in precapillary pulmonary hypertension," *Echocardiography*, vol. 41, no. 4, p. e15812, 2024.

[42] S. Bennett *et al.*, "Artificial intelligence in echocardiography: a position statement from the British Society of Echocardiography," *Echo Research & Practice*, vol. 13, no. 1, p. 26, 2026.

[43] M. Saini, G. Kc, A. J. Williams, P. M. Coplan, and L. E. Gressler, "Regulatory challenges and opportunities: A review of US food and drug Administration-Approved artificial intelligence and machine Learning-Enabled cardiovascular devices," *Therapeutic Innovation & Regulatory Science*, vol. 60, no. 2, pp. 393–422, 2026.

[44] J. I. Cotella *et al.*, "Artificial intelligence based left ventricular ejection fraction and global longitudinal strain in cardiac amyloidosis," *Echocardiography*, vol. 40, no. 3, pp. 188–195, 2023.

[45] J. Morrissey *et al.*, "Assessment of HeartModel® Automated Left Ventricular Ejection Fraction for Patients with Hypertrophic Cardiomyopathy," *medRxiv*, p. 2025.10. 30.25339136, 2025.

[46] M. Yaqub, B. Kelly, J. A. Noble, and A. T. Papageorghiou, "The effect of maternal body mass index on fetal ultrasound image quality," *American Journal of Obstetrics & Gynecology*, vol. 225, no. 2, pp. 200–202, 2021.

[47] S. G, U. Gopalakrishnan, R. K. Parthinarupothi, and T. Madathil, "Deep learning supported echocardiogram analysis: A comprehensive review," *Artificial Intelligence in Medicine*, vol. 151, p. 102866, 2024/05/01/ 2024, doi: https://doi.org/10.1016/j.artmed.2024.102866.

[48] S. Ferraz, M. Coimbra, and J. Pedrosa, "Assisted probe guidance in cardiac ultrasound: A review," *Frontiers in Cardiovascular Medicine*, vol. 10, p. 1056055, 2023.

[49] L. Wu *et al.*, "Standard echocardiographic view recognition in diagnosis of congenital heart defects in children using deep learning based on knowledge distillation," *Frontiers in Pediatrics*, vol. 9, p. 770182, 2022.

[50] J. A. Naser *et al.*, "Artificial intelligence-based classification of echocardiographic views,"

In Press

European Heart Journal-Digital Health, vol. 5, no. 3, pp. 260–269, 2024.

[51] A. H. Abdi *et al.*, "Automatic quality assessment of echocardiograms using convolutional neural networks: feasibility on the apical four-chamber view," *IEEE transactions on medical imaging*, vol. 36, no. 6, pp. 1221–1230, 2017.

Biographies:



R. Ghorbani was born in Arak, Iran, in 1998. She received the B.Sc. and M.Sc. degrees in biomedical engineering from Amirkabir University of Technology (AUT), Tehran, Iran in 2020 and 2023, respectively. Currently, she is pursuing the Ph.D. degree in biomedical engineering at Iran University of Science and Technology (IUST), Tehran, Iran. Her research interests include medical image processing and deep learning.



H. Behnam received the B.Sc. degree in electrical engineering from Iran University of Science and Technology (IUST), Tehran, Iran, in 1988. Additionally, he received the M.Sc. degree in medical engineering from Sharif University of Technology, Tehran, Iran, in 1992, and the Ph.D. degree in applied electronics from Tokyo Institute of Technology, Tokyo, Japan, in 1998. Since then, he has been a Researcher with Iran Research Organization for Science and Technology, Tehran, Iran. In addition, he has been a Faculty Member with IUST, Tehran, Iran from 2004. Currently, he is an Associate Professor of Biomedical Engineering with IUST. His research interests include medical image processing, ultrasound in medicine, and medical signal processing.

In Press