CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Quantification of Left Ventricle Structure and Function from Echocardiogram Videos

by

Samana Batool

A dissertation submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Faculty of Engineering
Department of Electrical Engineering

2024

Quantification of Left Ventricle Structure and Function from Echocardiogram Videos

By

Samana Batool

(DEE181001)

Dr. Andrew Ware, Professor

University of South Wales, UK

(Foreign Evaluator 1)

Dr. José Valente de Oliveira, Senior Researcher

University of Lisboa, Portugal

(Foreign Evaluator 2)

Dr. Imtiaz Ahmad Taj

(Research Supervisor)

Dr. Noor Muhammad Khan

(Head, Department of Electrical Engineering)

Dr. Imtiaz Ahmad Taj

(Dean, Faculty of Engineering)

DEPARTMENT OF ELECTRICAL ENGINEERING

CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY

ISLAMABAD

2024

I dedicate my dissertation work to my parents, whose lessons in perseverance, compassion, and humility have helped me navigate the challenges along this path. To my husband, whose support and encouragement have been the pillars of my strength. To my beloved children for being my constant source of joy and happiness amidst the challenges of this journey.

CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD
Expressway, Kahuta Road, Zone-V, Islamabad
Phone:+92-51-111-555-666 Fax: +92-51-4486705
Email: info@cust.edu.pk Website: https://www.cust.edu.pk

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the dissertation, entitled "**Quantification of Left Ventricle Structure and Function from Echocardiogram Videos**" was conducted under the supervision of **Dr. Imtiaz Ahmed Taj**. No part of this dissertation has been submitted anywhere else for any other degree. This dissertation is submitted to the **Department of Electrical Engineering, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Electrical Engineering.** The open defence of the dissertation was conducted on **August 23, 2024**.

Student Name : Samana Batool (DEE181001)

The Examination Committee unanimously agrees to award PhD degree in the mentioned field.

Examination Committee :

(a) External Examiner 1: Dr. Adil Masood Siddiqui
Professor
MCS, NUST, Islamabad

(b) External Examiner 2: Dr. Usman Akram
Professor
CEME, NUST, Islamabad

(c) Internal Examiner : Dr. Noor Muhammad Khan
Professor
CUST, Islamabad

Supervisor Name : Dr. Imtiaz Ahmad Taj
Professor
CUST, Islamabad

Name of HoD : Dr. Noor Muhammad Khan
Professor
CUST, Islamabad

Name of Dean : Dr. Imtiaz Ahmad Taj
Professor
CUST, Islamabad

AUTHOR'S DECLARATION

I, **Samana Batool** (Registration No. **DEE181001**), hereby state that my dissertation titled, "**Quantification of Left Ventricle Structure and Function from Echocardiogram Videos**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my PhD Degree.

(**Samana Batool**)

Dated: August 23, 2024

Registration No : DEE181001

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the dissertation titled "**Quantification of Left Ventricle Structure and Function from Echocardiogram Videos**" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete dissertation has been written by me.

I understand the zero-tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled dissertation declare that no portion of my dissertation has been plagiarized and any material used as reference is properly referred/ cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled dissertation even after award of PhD Degree, the University reserves the right to withdraw/ revoke my PhD degree and that HEC and the University have the right to publish my name on the HEC/ University Website on which names of students are placed who submitted plagiarized dissertation.

(**Samana Batool**)

Dated: August 23, 2024

Registration No : DEE181001

List of Publications

It is certified that following publication(s) have been made out of the research work that has been carried out for this dissertation:-

1. **Batool, S.**, Taj, I.A., and Ghafoor, M.(2023). "Ejection Fraction Estimation from Echocardiograms Using Optimal Left Ventricle Feature Extraction Based on Clinical Methods," *Diagnostics*, 13(13), 2155. DOI: *https://doi.org/10.3390/diagnostics13132155*

2. **Batool, S.**, Taj, I.A., and Ghafoor, M.(2024). "EFNet: A Multitask Deep Learning Network for Simultaneous Quantification of Left Ventricle Structure and Function," *Physica Medica*, vol. 125, p. 104505. DOI: *https://doi.org/10.1016/j.ejmp.2024.104505*

(Samana Batool)
Registration No: DEE181001

Acknowledgement

I extend my deepest gratitude to Dr. Imtiaz Ahmad Taj, my PhD supervisor, for his unwavering guidance, encouragement, and support throughout my doctoral journey. His patience and expertise have been instrumental in overcoming obstacles and achieving milestones, shaping me into the researcher I am today, for which I am forever grateful.

I also like to thank Dr. Mubeen Ghafoor for his invaluable assistance and collaboration in my research endeavors. His meticulous proofreading and validation of my work have greatly enhanced its quality and credibility.

My appreciation further extends to Dr. Shoab Ahmed, at Sir Syed CASE Institute of Technology, and Dr. Habib Ur Rehman, Consultant Cardiologist at Shifa International Hospital, Islamabad for graciously sharing their invaluable knowledge and experience that significantly aided my research.

I am also grateful to the members and colleagues of the Vision and Pattern Recognition (VisPRS) research group for their camaraderie, support, and guidance, which have enriched my academic experience.

To my beloved parents, husband, and children, I am indebted for their continuous prayers, boundless support, infinite patience, and constant encouragement throughout my PhD journey. Their love and faith in me have been my source of strength. I sincerely thank each and every individual who has played a role, big or small, in shaping my academic and personal growth during this transformative journey.

(Samana Batool)

Abstract

Echocardiography is one of the most commonly used imaging systems for assessing heart anatomy and function. It provides valuable information about the function and structure of the left ventricle (LV), the accurate quantification of which is important for the diagnosis and management of various cardiovascular conditions. In clinical practice, these tasks are performed manually which is time-consuming and prone to inter-observer and intra-observer variability due to human involvement. Additionally, echocardiogram studies often involve multiple videos, creating a large volume of complex data that is challenging to analyze. This highlights the need for automated machine learning methods to process such extensive datasets and identify intricate patterns in the quantification of heart structure and function that skilled observers might overlook, paving the way for computer-assisted diagnostics in this field.

In the first part of this study, ejection fraction (EF) is estimated from end-systolic and end-diastolic frames by extracting multiple features from the segmented images. These features are analyzed using both neural networks and machine learning algorithms. Results show that machine learning techniques not only automate the process but also deliver consistent and more accurate results, compared to clinical methods. The evaluations are performed on the publicly available EchoNet-Dynamic echocardiogram dataset.

To streamline and automate the process, a fully automated multitask network, the EchoFused Network (EFNet), is introduced, which simultaneously performs LV segmentation and EF estimation through cross-module fusion. The proposed model employs semi-supervised learning to estimate EF from the entire cardiac cycle, providing more reliable estimations and eliminating the need to identify end-systolic and end-diastolic frames. To optimize LV segmentation and EF estimation jointly, losses from task-specific modules are combined using a normalization technique, ensuring commensurability on a comparable scale. The proposed model is evaluated on two distinct datasets, EchoNet-Dynamic and CAMUS, demonstrating

its effectiveness in achieving superior outcomes, surpassing current state-of-the-art methods.

Lastly, expanding on the previous work, methods were explored to enhance LV segmentation based on insights from previous joint EF estimation and LV segmentation. To improve the quality and accuracy of LV delineation, it is proposed to include edge information through a multitask network that employs a common encoder for shared feature extraction from echocardiogram data. Separate decoder modules for semantic segmentation and edge prediction are utilized, each with its own cost function, which are combined to perform joint optimization within the network. The proposed method exhibits enhanced accuracy across multiple metrics, demonstrating the effectiveness of the proposed approach in overcoming the challenges associated with LV delineation.

Contents

List of Figures

List of Tables

Abbreviations

2DE	Two-dimensional echocardiography
A2C	Apical two-chamber
A4C	Apical four-chamber
A5C	Apical five-chamber
ACNN	Anatomically constrained neural networks
AI	Artificial Intelligence
ASE	American Society of Echocardiography
AUC	Area under the curve
BCE	Binary cross entropy
BEASM	B-spline explicit active surface model
BSA	Body surface area
CAD	Coronary Artery Disease
CAMUS	Cardiac Acquisitions for Multi-Structure Ultrasound Segmentation
CMR	Cardiovascular Magnetic Resonance
CNN	Convolutional neural network
CT	Computerized Tomography
CVD	Cardiovascular Diseases
DCM	Dilated Cardiomyopathy
DL	Deep Learning
DSC	Dice similarity coefficient
DT	Decision trees
EACVI	European Association of Cardiovascular Imaging
ED	End-diastolic

EDV	End-diastolic volume
EF	Ejection fraction
EFNet	EchoFused Network
ES	End-systolic
ESV	End-systolic volume
FCN	Fully connected network
FN	False negative
FP	False positve
FPN	Feature pyramid network
FPR	False positive rate
FS	Fractional Shortening
GLS	Global longitudinal strain
HCM	Hypertrophic Cardiomyopathy
HF	Heart failure
HFimEF	Heart failure with improved ejection fraction
HFmrEF	Heart failure with mid-range ejection fraction
HFpEF	Heart failure with preserved ejection fraction
HFrEF	Heart failure with reduced ejection fraction
IAS	Interatrial septum
IVC	Inferior Vena Cava
IAS	Interatrial septum
IC	Ischemic Cardiomyopathy
IoU	Intersection over union
IVS	Interventricular septum
IVUS	Intravascular ultrasound
LA	Left atrium
LR	Linear regression
LSTM	Long short-term memory
LV	Left ventricle
LVID	Left ventricle internal diameter
LVIDd	Left ventricle internal diameter at diastole

LVIDs	Left ventricle internal diameter at systole
MAE	Mean absolute error
MC3	Mixed convolutional network
MI	Myocardial Infarction
ML	Machine learning
MLD	Myocardial length at diastole
MLS	Myocardial length at systole
MSE	Mean squared error
MV	Mitral valve
NN	Neural network
NRMSE	Normalized root mean squared error
PLAX	Parasternal long axis
POCUS	Point of care ultrasound
PSAX	Parasternal short axis
R3D	3D-ResNet
RA	Right Atrium
RBF	Radial basis function
RF	Random forest
RMSE	Root mean squared error
RNN	Recurrent neural network
ROC	Receiver operating characteristic curve
ROI	Region of interest
RV	Right ventricle
SC	Subcoastal
SDCR	Systolic-diastolic cross ratio
SDCR$_{simp}$	Simpson's systolic-diastolic cross ratio
SGD	Stochastic gradient descent
SHG	Stacked hourglass
SIHD	Stable ischemic heart disease
SRF	Structured random forest
SV	Stroke volume

SVM	Support vector machine
SVR	Support vector regression
TEE	Transesophageal echocardiography
TN	True negative
TP	True positive
TPR	True positive rate
TTE	Transthoracic Echocardiogram
TV	Tricuspid valve
UVT	Ultrasound vision transformer
WMA	Wall motion abnormality

Symbols

A	LV area
A_d	LV area in diastole
a_i	Disk diameters in A4C
A_s	LV area in systole
b_i	Disk diameters in A2C
b_s	Bias for spatial convolution
b_t	Bias for temporal convolution
C_L	Contraction coefficient in longitudinal direction
$concat$	Concatenation layer
C_R	Contraction coefficient in radial direction
D_d	LV diameter during diastole
D_S	LV diameter during systole
f	Number of input frames to EFNet
f_{ReLU}	Non-linear activation function
K_D	Ratio between semi-axes in diastole
K_S	Ratio between semi-axes in systole
l	Height of a disk
L	LV major axis length
$\mathcal{L}$	Loss function
l_d	LV major axis length in diastole
l_s	Lv major axis length in systole
m	Downsampled layer
n	Dense block layer

N	Sample size
$Q1$	Lower quartile
$Q3$	Upper quartile
U	Set of ground truth LV contour points
$upconv$	transposed convolutional layer
V	Set of estimated LV contour points
W_s	Weights for spatial convolution
W_t	Weights for temporal convolution
$x^{m,n}$	Output from node $X_{m,n}$
y	Ground truth LV mask
$\hat{y}$	Estimated LV mask
z	Ground truth EF
$\bar{z}$	Ground truth EF mean
$\hat{z}$	Estimated EF
$\bar{\hat{z}}$	Estimated EF mean
z_r	Output from residual block
z_s	Spatial tensor
z_t	Temporal tensor
σ	Standard deviation

Chapter 1

Introduction

1.1 Background

Cardiovascular disease (CVD) stands as the primary global cause of mortality, contributing to over 17.9 million fatalities in 2019, with projections indicating an anticipated increase to more than 23.6 million by the year 2030 [1]. In a recent investigation targeting individuals aged 35 to 70, it was revealed that 40% of worldwide fatalities resulted from CVD, primarily affecting low- and middle-income nations [2]. The majority of these fatalities are linked to strokes and heart attacks. Heart disease encompasses a variety of disorders impacting both the function and structure of the heart. There is a growing interest in the detection of heart disease at an early stage as obesity, hypertension and metabolic disorders like diabetes are on the rise. A delay in diagnosis might result in a poor prognosis, which is frequently linked to permanent pathophysiologic alterations that develop over time.

Imaging modalities that have found use in cardiology include Cardiovascular Magnetic Resonance Imaging (CMR), Fundus Photography, Computerized Tomography (CT), Echocardiography, Intravascular Ultrasound (IVUS), and others. Despite the high imaging quality of CMR and CT, one of the most commonly utilized imaging systems for assessing heart anatomy and function is echocardiography,

mainly due to its mobility, availability, and cheaper cost as compared to other techniques. Echocardiography makes it possible to evaluate the structure and function of the heart without exposing patients to radiation or invasive treatments.

1.2 Artificial Intelligence in Echocardiography

Echocardiography presents challenges that are not straightforward to achieve for multiple reasons. Rather than comprising a single still image, an echocardiogram study can consist of several videos collected from multiple views. This vast amount of multidimensional data generated in each study is difficult to comprehend and, hence, not fully utilized. Additionally, measurements can differ from one video to another due to beat-to-beat inconsistency and variability resulting from the estimation of a three-dimensional object using two-dimensional images. Operator-dependent data acquirement, device inconsistency, and low image quality further restrict echocardiography. Given the presence of these constraints, it appears that echocardiography could benefit from the implementation of automated learning methods to aid human interpretation.

Progress in echocardiogram interpretation, standardization, and workflow, facilitated by automated monitoring and analytic techniques, holds considerable promise in minimizing variability in outcomes and providing high-quality, cost-effective healthcare, especially to individuals in resource-limited settings. Automated quantification and the identification of pathological features such as valve disease, regional wall motion abnormalities, and cardiomyopathies, coupled with rapid application at the point of care, will become feasible through these advancements [3]. These advancements not only enhance diagnostic accuracy and efficiency but also ensure that critical cardiac care becomes accessible to a broader population, ultimately improving overall health outcomes. Artificial intelligence (AI) techniques, which include machine learning (ML) and deep learning (DL), can play an effective role in achieving these objectives by providing precise, automated analysis and improving diagnostic accuracy.

1.3 Automation Tools in Echocardiography

Several automation tools and technologies have emerged to streamline and enhance the diagnostic process. These tools include software applications and platforms designed to automate tasks such as image acquisition, analysis, and interpretation. Some platforms offer workflow optimization solutions specifically designed for echocardiography departments. These tools streamline the entire echocardiography process, from appointment scheduling to image acquisition, interpretation, and reporting, optimizing resource utilization and improving patient throughput.

Telemedicine platforms with integrated echocardiography capabilities enable remote consultations and monitoring of patients with cardiovascular conditions. These platforms allow clinicians to perform echocardiographic exams remotely, review images in real-time, and provide timely interventions and recommendations, expanding access to cardiac care.

Point-of-care ultrasound (POCUS) devices equipped with automated features for image acquisition and analysis are becoming increasingly popular in clinical settings. These handheld devices offer portability, ease of use, and rapid imaging capabilities, making them suitable for point-of-care assessments in various healthcare settings. For example, a laptop-based echocardiography system has practically all 2D echocardiographic applications, although a pocket-sized ultrasound frequently lacks spectral Doppler and color flow features. A laptop-based and a pocket-size device are shown in Fig. 1.1.

POCUS has the potential to revolutionize bedside medicine and make physical examinations obsolete. Numerous investigations have conclusively demonstrated that POCUS is equally efficient and effective as traditional equipment. POCUS can also be utilized in more scenarios and by a broader spectrum of users than normal echocardiograms due to its availability. The handheld imaging devices, however, have restricted function with several challenges for their clinical application. The most important of them is the need for widespread standardization of training for extensive use of POCUS. AI can be used to address such limitations

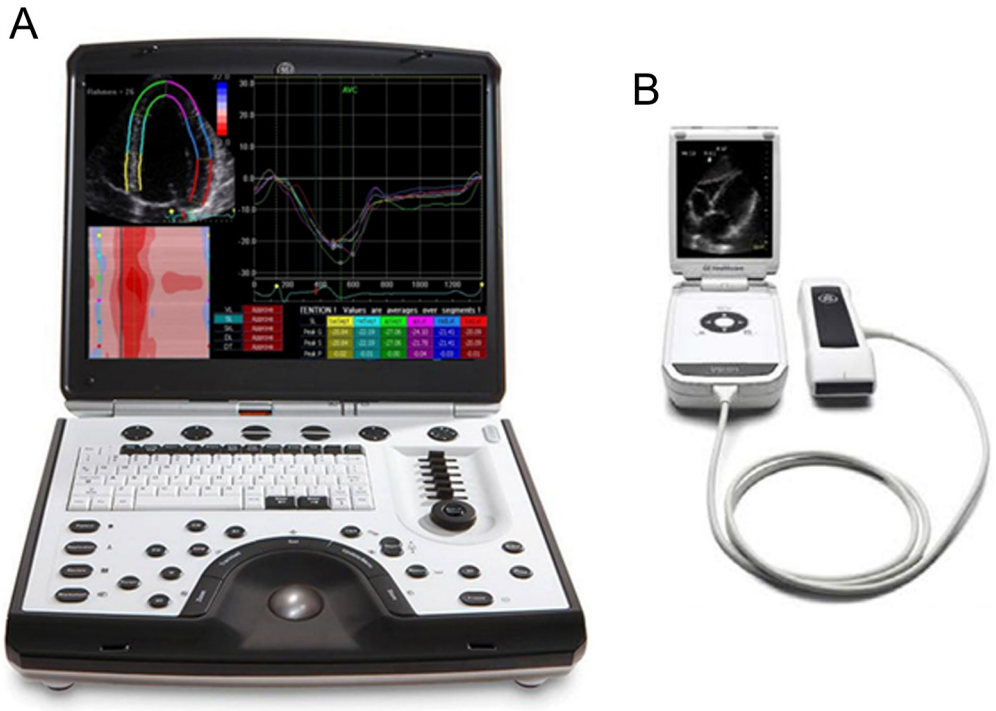


FIGURE 1.1: Handheld ultrasound machines; (A): a laptop-based equipment, (B): a pocket-size ultrasound [4].

by providing automated guidance and diagnostic support, thereby enhancing the accuracy and reliability of these portable tools in clinical settings. For instance, an AI-based automated left ventricle ejection fraction analysis program for PUS images has been created ("LVivo by DiA Imaging Analysis Ltd.", and "Vscan by GE Healthcare"). Recently, various devices the size of a smartphone have been made available. These devices employ AI based methods in order to assess cardiac function. Some of these gadgets are said to be inexpensive and useful to physicians in practice (e.g. "Vscan GE Healthcare", "Butterfly IQ", "Philips Lumify").

Zhang et al. [5] proposed that those who are not even experts in this field could use portable devices at point-of-care settings and then upload images to a cloud computing system which could perform a comparison with earlier studies. In this way, AI could prove to be useful in detecting early indicators of cardiac disorder in a cost-effective manner leading to a reduction of CVD-related morbidity and mortality. Furthermore, integrating these technologies into routine clinical practice could pave the way for more personalized and preventive cardiac care, significantly transforming the landscape of cardiovascular health management.

1.4 Echocardiography Principles and Methods: An Overview

Echocardiography is a non-invasive test in which high-frequency sound waves (ultrasounds) are used to produce an image of the heart. An echocardiogram is the name for such an image. Echocardiography can provide information to doctors regarding blood clots in the heart, difficulties with the aorta; the heart's major artery, problems with the function of heart valves, problems with the heart's contracting or relaxing function, and pressures in the heart. Fig. 1.2 depicts certain measurements that are generally obtained from echocardiograms during the occurrence of heart failure (HF).

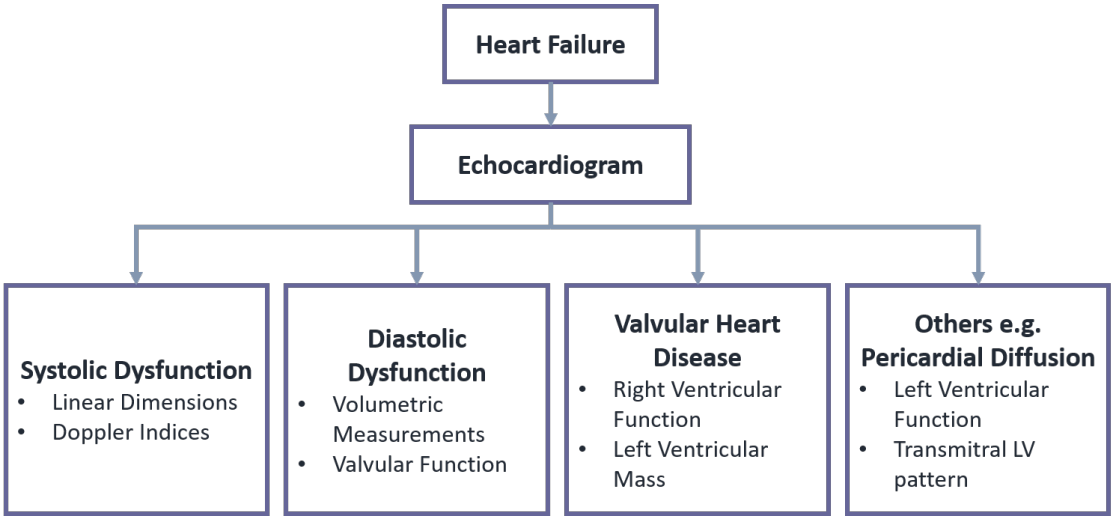


FIGURE 1.2: Role of echocardiography in heart failure [6].

1.4.1 Transthoracic Echocardiography (TTE)

There are several different kinds of echocardiograms which include Transthoracic Echocardiography (TTE), Transesophageal Echocardiography (TEE), Doppler, Stress, and the three-dimensional echocardiography depending on the method of obtaining them. The most frequent type of echocardiogram is a TTE, which involves placing a transducer on the chest wall, which emits sound waves (ultrasounds) that bounce off the heart structures and create still or moving images

of the internal regions of the heart. This allows healthcare providers to visualize the heart in real-time and evaluate its pumping ability, valve function, chamber size, and overall cardiac health. The TTE plays a crucial role in diagnosing a wide range of cardiac conditions, including coronary artery disease, heart valve disorders, heart failure, congenital heart defects, and pericardial diseases. It helps healthcare providers make informed decisions about treatment plans, monitor disease progression, and assess the effectiveness of interventions.

1.4.2 The Echocardiographic Views

Since an echocardiogram view only produces a two-dimensional image of an organ that is three-dimensional, various viewing angles are required to adequately visualize all cardiac components. Fig. 1.3 shows different scanning planes of the heart.

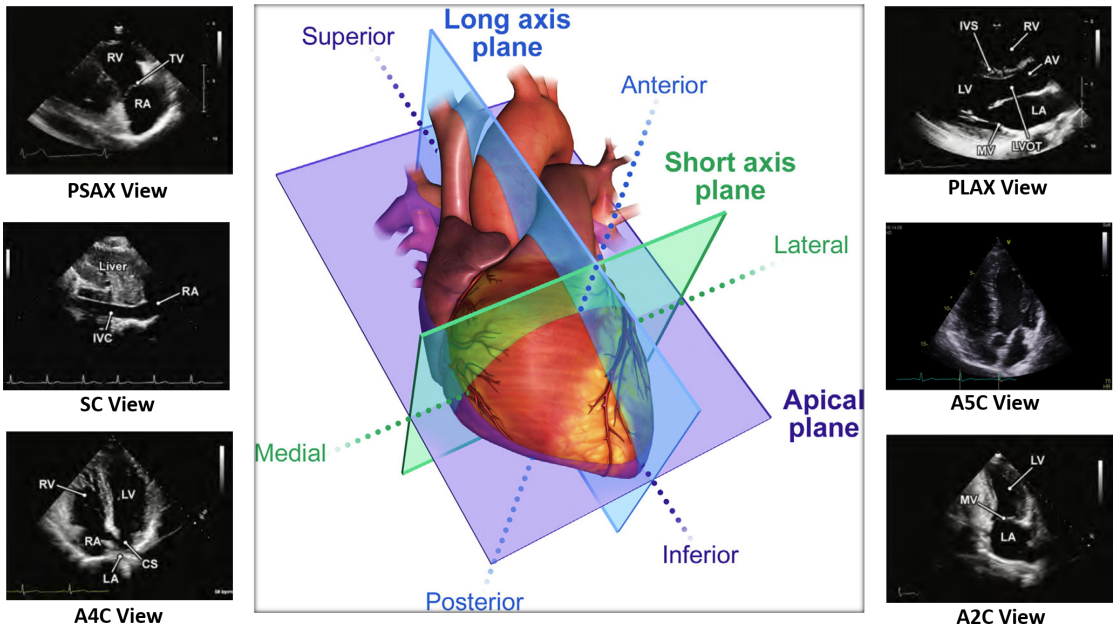
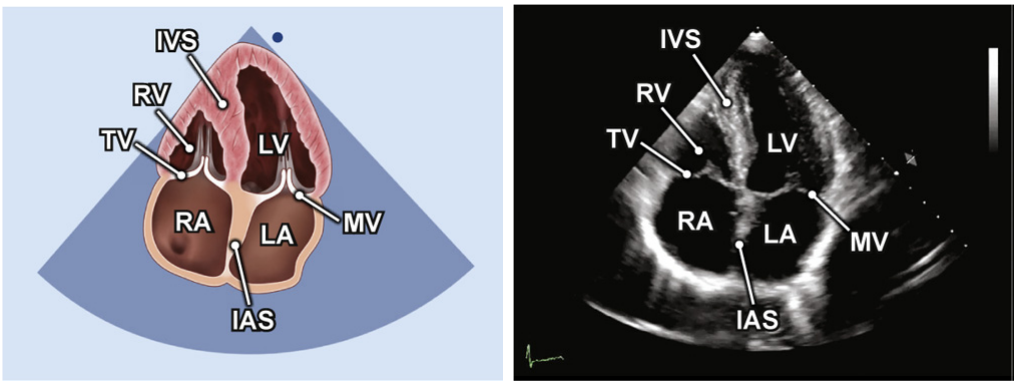


FIGURE 1.3: Different planes to scan the heart during an echocardiogram test. PLAX - Parasternal long axis, PSAX - Parasternal short axis, A2C - Apical two-chamber, A4C - Apical four-chamber, A5C - Apical five-chamber, SC - Subcostal[7]

Images obtained in the Parasternal long axis (PLAX) perspectives are shown in the long-axis plane. The images taken in the Parasternal short axis (PSAX) views are represented in the short-axis plane and the apical plane corresponds to images



(**a**) Anatomic Image (**b**) Echocardiographic Image

FIGURE 1.4: Apical four-chamber view [7]

acquired from the apical window. The major positions of the transducer include the apical, parasternal, suprasternal, and subcostal. Various tomographic images can be obtained by rotating and tilting the probe. Each will provide orientation information depending upon the scanning planes.

The data studied in this research comprises an apical four-chamber view (A4C), a critical perspective in echocardiography that provides comprehensive visualization of the heart's chambers and valves. For the acquisition of the A4C view, the apical window is placed beneath the left breast tissue and is where the apical impulse can be felt. The heart is shown upside down on the screen, as seen through the apically positioned transducer. The apex is at the top of the screen, while at the bottom are the atria. The interventricular septum (IVS) and interatrial septum (IAS) separate the left ventricle (LV) and left atrium (LA) on the right and the right ventricle (RV) and right atrium (RA) on the left. The tricuspid valve (TV) can also be seen located between the RA and RV, whereas the mitral valve (MV) is located between the LA and LV. This comprehensive view allows for detailed examination of the heart's structure and function, providing crucial insights into both the right and left sides of the heart. The A4C view, shown in Fig. 1.4, is one of the most appropriate for assessing ventricular function, making it invaluable in diagnosing and monitoring various cardiac conditions. The Fig. 1.4 (a) shows the anatomic view whereas Fig. 1.4 (b) shows its echocardiographic view, illustrating the practical application of this imaging technique in a clinical setting.

1.5 Quantification of Cardiac Chamber through Echocardiography

One of the most significant outcomes obtained from an echocardiogram is the quantification of the cardiac chamber which is fundamental in cardiac imaging. Consistency in chamber quantification methodology is upheld by the establishment and widespread dissemination of official guidelines. Adhering to these recommendations ensures consistency among practitioners and enhances effective communication. These standardized practices are crucial for comparing patient data across different healthcare settings and for longitudinal patient monitoring. The most recent recommendations for echocardiographic chamber quantification were jointly issued by the American Society of Echocardiography (ASE) and the European Association of Cardiovascular Imaging (EACVI) in 2015 [8]. These guidelines provide recommendations for the performance, interpretation, and clinical application of echocardiographic chamber quantification, ensuring uniformity in clinical practice.

1.5.1 The Left Ventricle

An important aspect of left ventricular structure quantification is the measurement of LV size. LV dimensions, which comprise linear internal dimensions, volumes, and wall thickness, are among the quantitative data generated from echocardiography that can influence patient therapy and serve as powerful predictors of outcomes. End-diastole and end-systole measurements are typically reported, and these are subsequently utilized to calculate global LV function parameters. There is a strong association between heart size and outcomes in people with stable ischemic heart disease (SIHD). LV size and ejection fraction (EF) are important indicators of survival not only in patients with HF but are also a salient feature for those who do not have any history of heart attack, according to data from the Framingham Heart Study. Moreover, accurate assessment of LV size can help guide clinical decisions, such as the timing of interventions and the adjustment of medical therapies, thus improving patient prognosis.

1.5.1.1 LV Size Assessment

LV size measurements include LV linear measurements and LV volume measurements. LV linear measurements include LV internal diameter at diastole (LVIDd) and LV internal diameter at systole (LVIDs). LV internal diameter (LVID) is the length taken from inner edge to inner edge, orthogonal to the major LV axis, at or proximately under the margin of tips of the MV leaflet. LVIDd is measured at end-diastole (specified as the frame with the greatest LV dimensions or volume or the first frame after the closure of the mitral valve.). LVIDs is measured at end-systole (specified as either the frame with the smallest LV dimensions or volume or the one after the closure of the aortic valve) [8].

For LV volume measurements, 2D echocardiography estimations are established on geometric techniques that use quantification of LV dimensions to compute volume. A preferred technique is the biplane method of disks, which is the modified Simpson's method [8]. There is an alternative area-length technique as well, which is seldom used. The endocardial-blood pool interface is traced on images with clear endocardial boundary delineation at end-diastole and end-systole using apical two- and four-chamber views (ideally an LV-focused view). When reaching the MV plane, the contour is completed by a straight line that joins the two opposing parts of the MV ring. The bisector between this line and the apex of the LV is the LV length. The volume measurements are calculated using these volume tracings and lengths. A brief description of these techniques is given below.

Area–Length Method The area-length method is a straightforward method to calculate LV volume. The area (A) of the LV in a four-chamber view and the length (L) of the ventricle (between MV and the apex) are used to determine volume, given by Eq. (1.1);

$$V = \frac{0.85 \times A^2}{L}. \tag{1.1}$$

This method assumes LV to be a bullet-like shape. However, it may encounter limitations since this assumption is not always applicable.

Biplane Method of Disks The ASE and the EACVI both recommend the use of the biplane method of disks (modified Simpson's method) [8]. This method manually delineates endocardial borders at the end-systole and end-diastole in two orthogonal views; apical four-chamber (A4C) and apical two-chamber (A2C). The LV is divided into a row of elliptical disks aligned perpendicular to the ventricle's major axis, and their respective volumes are added. A cross-section of one such disk is illustrated in Fig. 1.5. The end-diastolic volume (EDV) is determined from the end-diastolic (ED) frame. The end-systolic volume (ESV) is determined from the end-systolic (ES) frame.

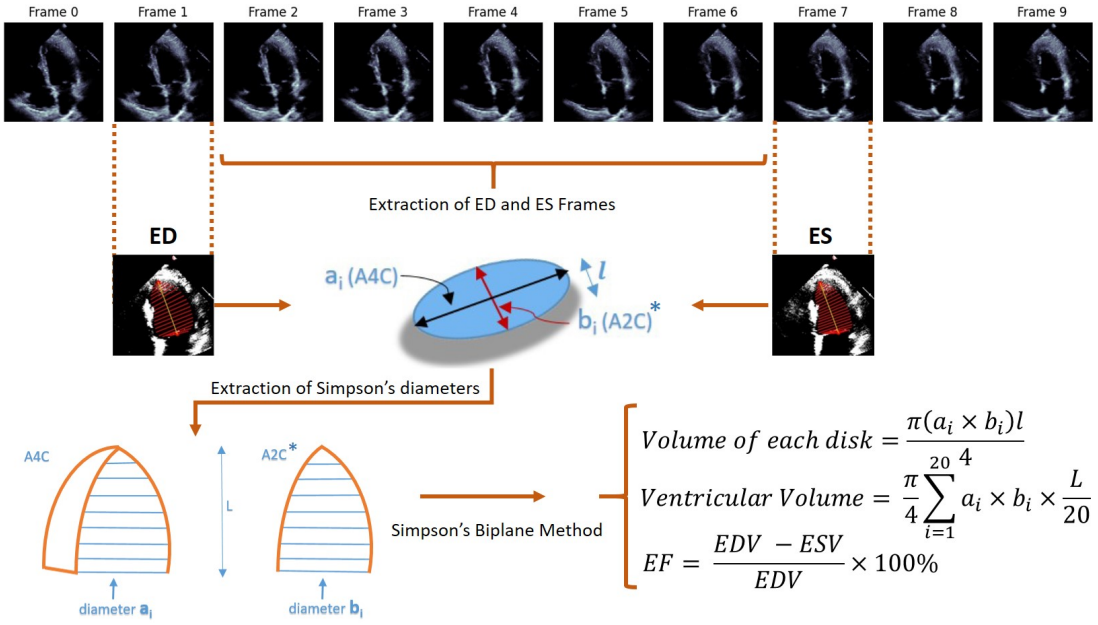


FIGURE 1.5: Simpson's Biplane Method (a_i—disk diameters in A4C view, b_i—disk diameters in A2C view, L—length of major axis, l—height of a single disk).

The *Volume of each disk* shown in Fig. 1.5 is given by Eq. (1.2);

$$Volume\,of\,each\,disk = \frac{\pi(a_i \times b_i)l}{4}. \tag{1.2}$$

where a_i and b_i are the semi-axis lengths of a disk, obtained from an A4C and A2C, respectively, and l is the height of the disk. Adding the volumes of these disks gives the total volume of the left ventricle. These volumes are then used to find an estimate of LV function.

The stroke volume (SV) can be calculated as given by Eq. (1.3);

$$SV = EDV - ESV(mL). \tag{1.3}$$

Because images must be optimized and endocardial borders must be precisely detected and traced according to standards, a lot of sonographer experience is necessary. This approach can be hampered by poor endocardial border characterization, foreshortened images, and inappropriate techniques.

1.5.1.2 LV Function Assessment

To measure global LV function for any 1-dimensional, 2D, or 3D parameter, the difference between the respective values at the end of diastole and systole is normalized by the end-diastolic value.

LV Ejection Fraction EF measures the percentage of blood given out of the LV during each cardiac cycle. EF is calculated from estimates of LV volumes obtained at the end of systole and diastole, using the formula in Eq. (1.4). LV volume estimates are derived either from Simpson's method of disks or the area-length method, the former being the preferable one.

The EF is given by;

$$EF = \frac{EDV - ESV}{EDV} \times 100\%. \tag{1.4}$$

Table 1.1 shows biplane LV EF generated from two-dimensional echocardiography (2DE), showing different ranges of values for normal, mild, and severely abnormal EF (%) per gender.

Fractional Shortening The heart's muscular contractility is measured by Fractional Shortening (FS). By the end of systole, the dimensions of the end-diastolic diameter have decreased substantially. The effectiveness of the heart in ejecting blood is reduced if the diameter does not shorten by at least 28 percent. In a

TABLE 1.1: Ranges of values for 2DE-derived LV EF and LA volume [8].

	Male		**Female**	
	EF (%)	**Max LA vol* BSA*** (mL/m^2)	**EF (%)**	**Max LA vol* BSA*** (mL/m^2)
Normal range	52 - 72	16 - 34	54 -74	16 - 34
Mildly abnormal	41 - 51	35 - 41	41 - 53	35 - 41
Moderately abnormal	30 - 40	42 - 48	30 - 40	42 - 48
Severely abnormal	< 30	> 48	< 30	> 48

*Max: Maximum, vol: Volume, BSA: Body Surface Area

symmetrically contracting ventricle, it is defined as the percentage change in the LV minor axis and can be obtained using the formula given in Eq. (1.5);

$$FS = \frac{LVIDd - LVIDs}{LVIDd} \times 100\%. \tag{1.5}$$

The normal range of FS is from 25% to 45%.

Global Longitudinal Strain Global longitudinal strain (GLS) is a measure used in echocardiography to assess the longitudinal deformation of the LV of the heart. Its use is increasingly recommended in clinical practice to improve the early diagnosis and management of various cardiac conditions. It is calculated by tracking the movement of specific points within the LV myocardium throughout the cardiac cycle. GLS represents the percentage change in the length of the LV myocardium from end-diastole to end-systole and provides information about the overall contractile function of the heart, given by Eq. (1.6);

$$GLS(\%) = \frac{MLs - MLd}{MLd}. \tag{1.6}$$

where ML is the myocardial length at end-systole (MLs) and end-diastole (MLd). Because MLs is smaller than MLd, peak GLS is a negative number.

Other measurements made as part of cardiac chamber quantification include RV size and function, RA and LA area and volume measurements, and Inferior Vena Cava (IVC) diameter measurements. These aspects are out of the scope of this research and will not be discussed in detail here.

1.6 Heart Failure and Cardiomyopathy

Heart failure can occur in two main forms: systolic failure, characterized by a weakened heart muscle, and diastolic failure, marked by stiffness impairing normal relaxation. Among the various causes contributing to HF, cardiomyopathy is one of the main causes. It refers to a group of diseases that affect the heart muscle, leading to abnormalities in its structure and function. These conditions can weaken the heart and impair its ability to pump blood effectively to the rest of the body. Cardiomyopathy can be caused by various factors, including genetic predisposition, infections, certain medications, toxins, and systemic diseases. It can result in symptoms such as shortness of breath, fatigue, swelling of the legs, and an irregular heartbeat. Depending on the type and severity, cardiomyopathy can lead to complications such as HF, arrhythmias, and even sudden cardiac death. Treatment typically focuses on managing symptoms, slowing disease progression, and reducing the risk of complications.

1.6.1 Role of Echocardiography in Detecting Cardiomyopathy

Echocardiography plays a crucial role in detecting cardiomyopathy by providing detailed images of the heart's structure and function in real-time. Through echocardiography, healthcare professionals can visualize the heart's chambers, walls, valves, and blood flow, providing insights into the condition of the heart muscle. This imaging technique allows for the identification of various features indicative of cardiomyopathy, such as chamber enlargement, wall thickness alterations, and impaired contractility. Additionally, echocardiography aids in assessing cardiac function, including ejection fraction, diastolic function, and myocardial strain, which are crucial parameters in diagnosing and monitoring cardiomyopathy. By capturing detailed images of the heart and providing quantitative data on its function, echocardiography assists clinicians in making accurate diagnoses and guiding treatment decisions for individuals with cardiomyopathy.

1.6.2 EF Based Classification of HF

The classification of HF based on left ventricular EF holds significance due to variations in prognosis, treatment response, and its role in patient selection for clinical trials, as most trials rely on EF criteria. HF with reduced EF (HFrEF), also known as systolic HF, occurs when the heart muscle fails to contract effectively, resulting in inadequate pumping of oxygen-rich blood to the body. An EF $\leq$35% or $\leq$40% is commonly referred to as HFrEF [9].

HF with preserved EF (HFpEF) accounts for at least half of the HF cases, and its incidence is on the rise [10]. HFpEF is classified based on different thresholds for EF, including >40%, >45%, or $\geq$50%. Since some patients in this category exhibit an EF that isn't entirely normal but doesn't show a significant reduction in systolic function, they are described as having preserved EF [11].

Patients whose EF falls between the ranges of HFrEF and HFpEF have been labeled as "HF with mid-range EF" or "HF with mildly reduced EF" [12]. They are classified as HF with mid range EF (HFmrEF) due to their lower-than-normal EF. Patients with HFmrEF typically exhibit a dynamic trajectory, either improving from HFrEF or deteriorating to HFrEF. Therefore, a single EF measurement may not suffice for these patients, and assessing the trajectory of EF over time and its underlying cause is crucial. Diagnosing HFmrEF and HFpEF can pose challenges. While classic signs and symptoms of HF, coupled with EF values of 41% to 49% or $\geq$50%, are essential for diagnosis, additional objective measures of cardiac dysfunction can enhance diagnostic specificity. Therefore, due to the complexities involved in diagnosing HFmrEF and HFpEF, in this study, we considered the case of HFrEF only for the classification of absence or presence of cardiomyopathy. HFrEF, which leads to morbidity and mortality, poses a significant public health concern. In recent years, there have been important scientific advancements in the management of HFrEF, resulting in better outcomes for patients. Some of the recent developments include SGLT2 inhibitors, vericiguat, and transcatheter mitral valve repair. These treatments have been shown to improve the prognosis of patients beyond the use of standard neurohormonal therapies. However, despite

these advancements, the morbidity and mortality rates associated with the disease remain high. After being hospitalized for HFrEF, the 5-year survival rate remains only 25% [13].

Fig. 1.6 illustrates EF based cardiomyopathy classification according to the guidelines provided by [11]. It also outlines a general diagnostic process, detailing subsequent clinical actions specifically for cases of HFrEF.

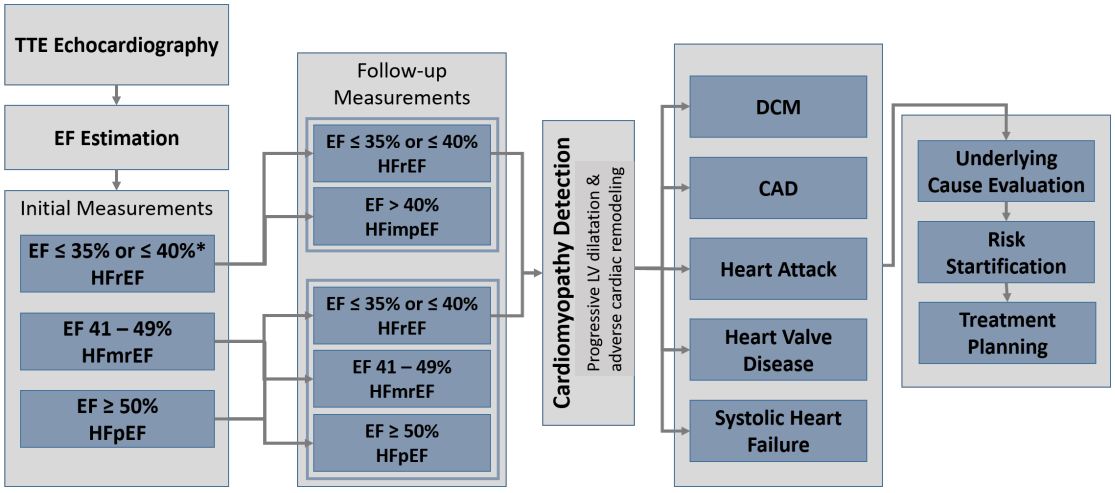


FIGURE 1.6: EF based disease classification [11]. HFrEF represents HF with reduced EF; HFmrEF, HF with mid range EF; HFpEF, HF with preserved EF; HFimpEF, HF with improved EF; DCM, Dilated Cardiomyopathy; CAD, Coronary Artery Disease. *Depends on the threshold suggested by the cardiologist

1.7 Echocardiography Datasets

1.7.1 EchoNet-Dynamic Dataset

In this study, the EchoNet-Dynamic dataset [14] is used, which contains over 10,000 A4C (2D B-mode) echocardiography recordings from individuals undergoing imaging between 2016 and 2018 during routine clinical care at Stanford University Hospital. The average age of the patients is 68 $\pm$ 21, and 49 percent of them are female. The training, validation, and testing sets have 7460, 1288, and 1277 patients, respectively. Each video includes a boundary tracing, also known as volume tracing, of the LV border at the end of systolic and diastolic frames.

Additionally, cardiac parameters such as EF, EDV, and ESV are provided for each video. These human expert annotations are derived from assessments conducted by a skilled cardiac sonographer, further reviewed by an imaging cardiologist, and considered as the ground truth values. This comprehensive dataset supports robust training and evaluation of machine learning models. The high-quality annotations ensure the reliability of the derived models for clinical application. The label variables included with the dataset are listed in Table 1.2 below.

TABLE 1.2: Dataset label variables

Variable	**Description**
FileName	Hashed file name used to link videos, labels and annotations
EF	Ejection fraction calculated from ESV and EDV
ESV	End systolic volume calculated by method of discs
EDV	End diastolic volume calculated by method of discs
Height	Video height
Width	Video width
FPS	Frames per second
NumFrames	Number of frames in whole video
Split	Classification of train/validation/test sets used for benchmarking

1.7.1.1 Volume Tracings

The tracings of the LV are obtained at the endocardial boundary at end-systole and end-diastole for each clip. An estimate of ventricular volume is obtained by using the tracings to integrate the ventricular area along the length of its main axis. A set of paired coordinates is used to represent expert tracings as shown in Fig. 1.7. The first pair of coordinates associated with each clip gives LV length (major axis length), while the rest of the coordinate pair set represents the minor axis lengths between the apex and the mitral valve. The file name and frame number from which the volume tracings are obtained are given with the volume tracings.

1.7.1.2 Dataset Statistics

The summary of statistics of the EchoNet-Dynamic dataset is given in Table 1.3.

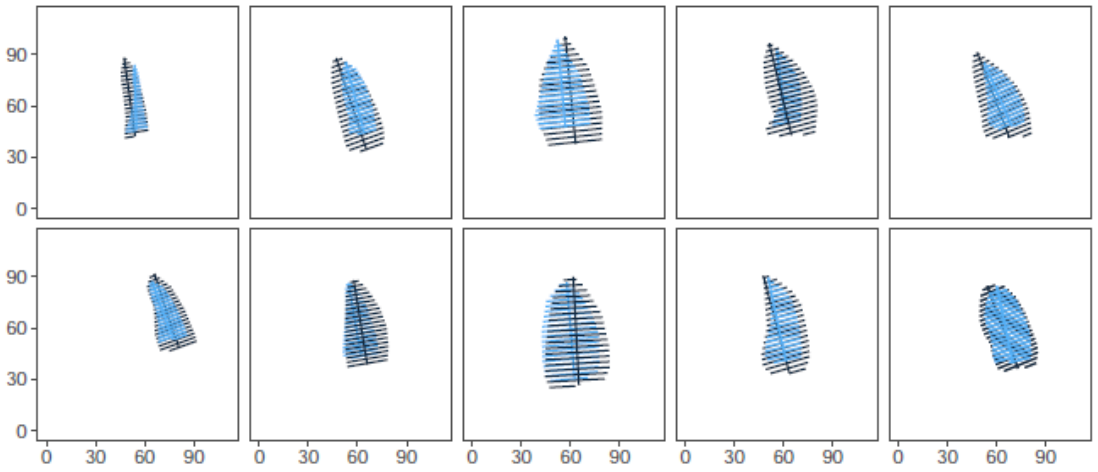


FIGURE 1.7: Human expert tracings [14]

TABLE 1.3: Dataset statistics [14]

Metric	**Total**	**Training**	**Validation**	**Test**
Number of Videos	10,036	7,465	1,289	1,282
Female (%)	4885 (48%)	3662 (49%)	611 (44%)	612 (44%)
Age (Years)	68 (21)	70 (22)	62 (18)	62 (17)
Frames Per Second	50.9 (6.8)	50.8 (6.7)	51.0 (6.5)	51.3 (7.3)
Number of Frames	175 (57)	175 (57)	176 (52)	176 (60)
Ejection Fraction (%)	55.7 (12.5)	55.7 (12.5)	55.8 (12.3)	55.3 (12.4)
End Systolic Volume (mL)	43.3 (34.5)	43.2 (36.1)	43.3 (34.5)	43.9 (36.0)
End Diastolic Volume (mL)	91.0 (45.7)	91.0 (46.0)	91.0 (43.8)	91.4 (46.0)

1.7.2 CAMUS Dataset

The second dataset employed in this study is the CAMUS (Cardiac Acquisitions for Multi-Structure Ultrasound Segmentation) dataset. This dataset comprises two and four-chamber view data from 500 patients acquired at the University Hospital of St Etienne (France). The training dataset is composed of 450 patients whereas the testing dataset comprises 50 patients. The dataset is provided with manual annotations from cardiologists for the LV endocardium and epicardium contour, and the LA [15].

The CAMUS dataset allows for robust training and evaluation of segmentation algorithms by providing well-annotated echocardiographic images. These annotations enable precise localization and measurement of cardiac structures, which are crucial for developing and validating methods for accurate cardiac assessment.

Fig. 1.8 illustrates some example frames from the datasets, showcasing the variety of data and annotations provided.

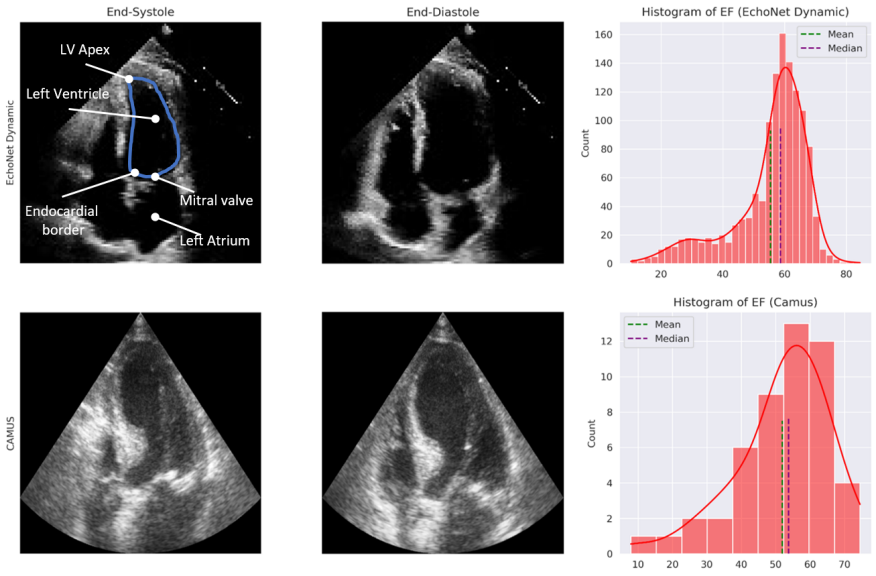


FIGURE 1.8: Left: End-diastolic and end-systolic A4C sample frames from EchoNet-Dynamic and CAMUS datasets. Right: EF distribution corresponding to the datasets

1.8 Research Objectives

The main challenge in employing ML methods to quantify the structure and function of the left ventricle lies in the exploitation of temporal information from echocardiogram data for precise LV EF estimation. Additionally, establishing consistency with recommended clinical methods to ensure transparency and reliability presents a significant challenge. Moreover, integrating interconnected information poses another problem. This entails combining various tasks with differing objective functions across different scales and ensuring convergence during joint training.

The main objectives of this research are:

- Development of a model for the quantification of LV structure and function that adheres to clinically recommended methods, ensuring transparency and reliability of results.

- Development of a model capable of simultaneously performing LV segmentation and EF regression from echocardiogram videos using multitask optimization. This involves integrating the objective functions from two distinct tasks with varying scales in a way that ensures convergence during joint training.

- Addressing the challenge of scarcity of medical data in training DL models by exploring techniques such as data augmentation and semi-supervised learning.

- Development of a robust segmentation method by integrating boundary information. This includes decoupling the mask and edge processing and fusing them in a way that complements the performance of each other to enhance the accuracy compared to traditional segmentation methods.

1.9 Research Contributions

The main contributions of this work are:

- Development of an ML-based method for estimating left ventricular EF from ED and ES frames. It performs LV segmentation followed by feature extraction using Simpson's method. To effectively leverage the temporal information present in the frames of an echocardiogram, the proposed method utilizes Simple Recurrent Neural Network (RNN) and Long Short-Term Memory Network (LSTM). Our contributions have been published in [16].

- A DL-based multitask model; EFNet is developed which enables concurrent LV segmentation and EF regression from echocardiogram videos, employing joint optimization of the objective function and leveraging their interconnectedness to enhance overall performance. Effective strategies are devised to seamlessly combine these objective functions, ensuring coherence and consistency in the model's training process. The proposed model undergoes training and evaluation using a larger dataset, enabling robust learning from

a diverse range of samples. Furthermore, the model is fine-tuned on a smaller dataset, investigating the potential benefits of leveraging DL techniques to train effectively on limited data resources. To enhance the dataset's size and diversity, data augmentation methods are utilized. Our contributions from this work have been published in [17].

- A multitask DL model featuring a common encoder for shared feature extraction from input data is developed to improve LV segmentation. This model integrates two distinct modules—the Mask Generation Decoder for mask segmentation and the Edge Predictor for boundary prediction. The incorporation of edge supervision from the Edge Predictor significantly improved the network's capability to preserve spatial boundary details, resulting in an enhancement in semantic segmentation performance. Additionally, the multitask model optimizes through joint training by combining losses from both the Mask Generation Decoder and Edge Predictor. Furthermore, extensive research on the structure of the Edge Predictor leads to the proposal of the optimal architecture, which demonstrates superior performance in the regression of edge coordinates.

1.10 Organization of the Dissertation

The first chapter of this dissertation provides an overview of the role of artificial intelligence in Echocardiography, followed by its principles and methods in clinical settings. It then outlines the aims and objectives of this research and a summary of the research contributions.

Succeeding this introductory chapter, a relevant literature review and established methodologies are presented in Chapter 2. Various aspects of automating echocardiography are discussed, with a focus on LV segmentation and EF estimation. A section on current studies involving multitask learning is also included. The detailed discussion on established work led to a comprehensive analysis of the gaps

within existing techniques, and subsequently the formulation of the problem statement for the dissertation.

Chapter 3 presents the estimation of left ventricular EF from ED and ES frames. This is accomplished by exploring different techniques such as polynomial regression and LSTM networks on features derived from segmented LV, adhering to clinical methods. The chapter explores both neural networks and traditional ML techniques investigated within the study.

Chapter 4 presents the simultaneous quantification of LV segmentation and EF estimation from the echocardiographic videos. The presented model utilizes multitask optimization based on DL techniques. The chapter begins with a brief overview of deep networks for segmentation and regression, followed by a detailed explanation of the proposed multitask model. Various experiments on normalization techniques used to combine the objective functions are analyzed in detail.

Since the performance of the multitask network in Chapter 4 relies heavily on accurate LV segmentation, Chapter 5 introduces a method to enhance segmentation by decoupling edge and mask information. It details the Decoupled Edge Guided Module's architecture, presents qualitative and quantitative results, and includes an analysis of ablation experiments with various encoders, loss functions, and edge module layers. Chapter 6 marks the final chapter of the dissertation, providing conclusions drawn from the research and discussing possible future extensions of the work done in this study.

1.11 Summary

This chapter provides an introduction to echocardiography, emphasizing its clinical significance and widespread adoption due to its non-invasive nature, cost-effectiveness, and ease of accessibility. It explores the potential of artificial intelligence to streamline various tasks within echocardiographic tests. It also provides

an overview of the TTE test, one of the most commonly utilized echocardiographic procedures, which will be employed in this research.

A comprehensive overview of the recommended clinical methods for quantification of cardiac chambers is presented with a main focus on left ventricular structure and function. A description of the datasets utilized in this research is also provided. Challenges encountered in automating echocardiogram analysis are discussed, leading to the formulation of research objectives. Finally, the chapter outlines the main contributions stemming from this research.

Chapter 2

Literature Review

AI plays a significant role in automating various aspects of echocardiography, enhancing efficiency, accuracy, and clinical decision-making in the evaluation and management of CVDs. It can provide image interpretation by analyzing echocardiographic images to identify anatomical structures, quantify cardiac function parameters such as EF, and detect abnormalities or pathology. Measurements traditionally performed manually by clinicians, such as chamber dimensions, wall thickness, and blood flow velocities, can be automated using AI. This reduces the time required for analysis and minimizes the risk of human error. Further, the quality of echocardiographic images can be assessed in real-time, flagging images with suboptimal quality for re-acquisition or further review. Moreover, AI-powered decision support systems can assist clinicians in interpreting echocardiographic findings, providing recommendations for diagnosis, risk stratification, and treatment planning based on established guidelines and evidence-based practices. AI can also aid in personalizing treatment plans by integrating echocardiographic data with patient-specific information, enhancing individualized patient care. Among these tasks, the primary focus of this research is the segmentation of the LV and the measurement of cardiac structural and functional parameters, leading to the early detection and diagnosis of cardiovascular diseases. The role of AI in the automation of different tasks performed by human experts in clinical investigation is illustrated in the flow diagram in Fig. 2.1.

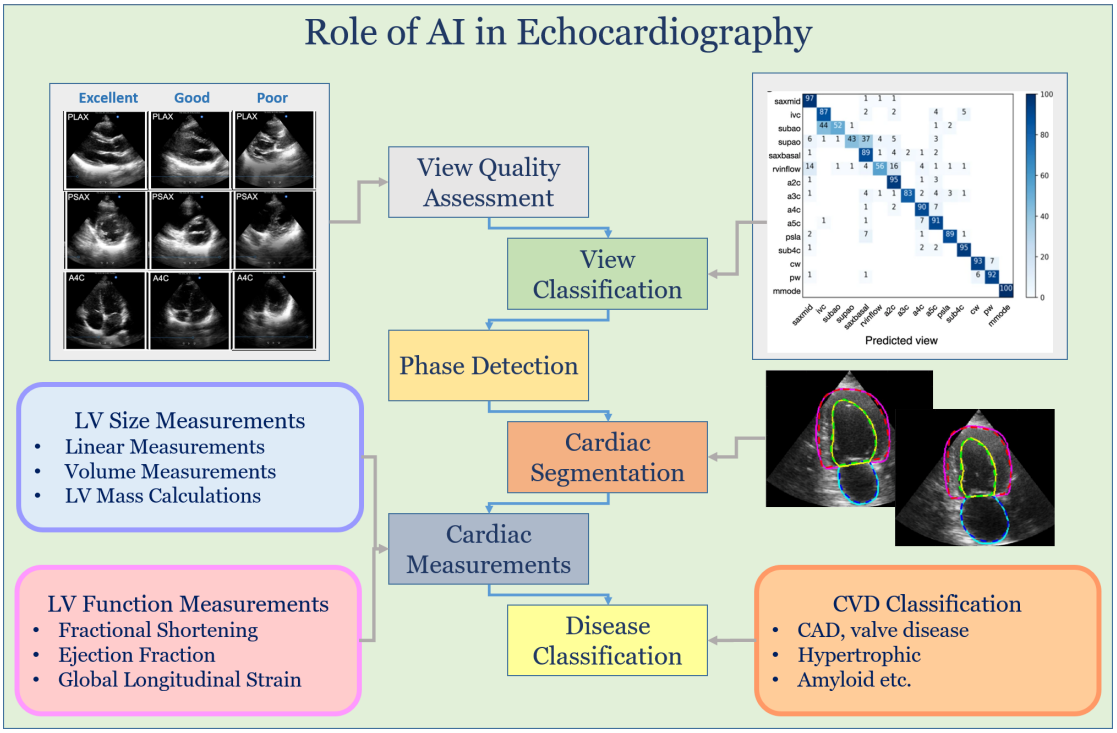


FIGURE 2.1: Role of AI in automation of echocardiography [14, 15, 18]

2.1 Quality Assessment and View Classification

In the automation of echocardiography, a lot of initial work has been done to assess the quality of the echocardiographic frame, which makes future jobs easier to analyze. According to studies, providing cardiologists with real-time feedback can help them improve image quality[19].

In [20], convolutional neural networks are used on the A4C view to do the echo quality assessment. The evaluation only used ES frames and did not take advantage of the information provided in sequential echo images. The authors further built on their previous work in [21] by assessing the quality of echo cine loops in five common imaging planes that improved the accuracy and gave real-time feedback to the user.

In [22], the authors proposed a "Quality Metric for Cardiac Ultrasound Videos (CUQI metric)" as a metric for assessing the quality. The CUQI metric assessed the quality of the cardiac video and measured motion information distortion and edge distortions between the reference and distorted footage.

View classification is another essential step in making an automated pipeline for interpreting an echocardiogram. An echocardiogram is obtained from a variety of angles, each of which reveals different aspects of the heart structure. In the past few years, a lot of work has been done on automatic classification of echocardiography views. For the categorization of echocardiographic videos of eight viewpoint classes, Gao [23] proposed a CNN architecture that included the fusion of selective as well as automatic DL networks. Notably, this two-strand CNN design gave a considerably good performance. To make the application of a DL approach in POCUS systems easier, view classification models must be able to be executed in real-time on mobile computing platforms (e.g., Android phones) with minimal computing capability. The work by Vaseli et al. [24] described a lightweight classification model six-fold faster than deep networks for echo view classification over twelve common views to fulfill the demand for speedy mobile apps for real-time POCUS diagnostics. In this work, knowledge distillation is used to train a series of lightweight models utilizing the learned knowledge of various DL architectures (teacher networks) based on the three most prevalent designs: VGG-16, DenseNet, and Resnet (student networks).

Madani et al. in [25] proposed a CNN model that classified 12 video views using cluster analysis. The dataset, which comprised still images as well as cines, was obtained from patients with different diseases. They trained their network to concurrently classify 15 standard views (12 B-mode clips, 3 stills). The previous

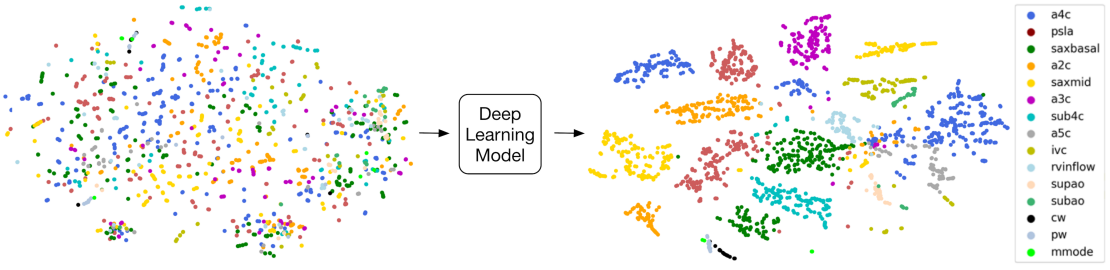


FIGURE 2.2: View classification performed on 15 standard echocardiographic views [25]

work [25] is expanded to automate cardiovascular disease prediction in [18] by employing both semi-supervised and supervised learning. They categorized different views as well as segmented out key characteristics in echocardiography images.

2.2 Segmentation of Cardiac Chambers

The current standard for border segmentation in an echocardiogram requires specialists to manually delineate before generating structural and functional indices using the traced boundaries. This procedure is time-consuming, prone to errors, and subject to significant intra- and inter-reader variance. Automating this procedure saves time by providing rapid, precise, and objective segmentation throughout the cardiac cycle. Following that, characteristics or cardiac indicators can be extracted from the segmented region, followed by CVD categorization.

To perform segmentation, certain conventional methods for image processing have been proposed, including a watershed algorithm for LV border segmentation [26], and K-means clustering [27]. These techniques are computationally efficient but have a high signal-to-noise ratio and fail to produce acceptable results when there are unclear borders and non-uniform regional intensities.

2.2.1 Spatial Segmentation

In ML techniques for segmentation, each pixel is labeled as a background or part of a chamber. By labeling each pixel with its respective class, semantic CNNs split the image into distinct regions. In [5], the structure of the heart chambers in four different perspectives are segmented using four different semantic UNet models [28]. The trained models performed with IoU values ranging from 73 to 92. The geometric dimensions, longitudinal strain, mass, volumes, and LV EF of each image are computed using the segmented heart chambers. These indicators are then utilized to evaluate the anatomy and function of the heart. As described in the study, the suggested automated system outperformed manual measurements across the board for all cardiac indicators. This improvement in accuracy underscores the potential of AI in providing reliable cardiac assessments. FCN and UNet were also employed for chamber segmentation in studies by Yue et al. [29] and Silva et al. [30]. In 2019, Leclerc et al. [15] made a significant contribution to echocardiographic segmentation by introducing the CAMUS dataset. This dataset

consists of 500 patients, with 450 patients allocated for training purposes, featuring publicly available expert annotations. Additionally, there are 50 testing patients. Each patient in the training dataset includes images and expert annotations of the LV, myocardium, and LA for both ED and ES frames of the cardiac cycle, captured in both A4C and A2C views. Alongside releasing the CAMUS dataset to the public domain, the authors also presented competitive results with various versions of a UNet [28] fine-tuned specifically for the task.

More recently, Leclerc et al. [31] introduced LU-Net, a segmentation network composed of two steps, influenced by the principles of Mask R-CNN. In this approach, the first network predicts a region of interest (ROI) around the heart, while the second network focuses on predicting accurate segmentations within the ROI. LU-Net demonstrated enhanced results compared to the authors' previous UNet architecture, even surpassing intra-observer accuracy for epicardial segmentation. Other studies have proposed modifications to the base UNet architecture to enhance its performance in 2D echocardiographic segmentation. For example, Moradi et al. [32], drawing inspiration from feature pyramid networks (FPN), developed MFP-Unet, which incorporates dilated convolutions to expand the receptive field and upscales feature maps to maximize the information available in the final layers.

Ouyang et al. released their dataset named EchoNet-Dynamic [33]. This dataset is designed specifically for assessing the EF and segmenting the LV in A4C sequences. In their next paper, the authors introduced a DL system [14] that first performed the segmentation of the LV using weak supervised learning. The labels provided by the clinical experts were taken as the ground truth. As a next step, the segmentation results were combined with predictions produced at frame level to provide an estimation of EF over every cardiac cycle. In a recent study by Jafar et al.[34], the authors used a combination of two CNNs, the YOLOv7 algorithm, and a UNet for performing segmentation of the LV endocardium, LV epicardium, and LA. Another work by Liao et al. [35] proposed two models using a Swin Transformer combined with K-Net and Segformer, for LV segmentation in echocardiography, aiming to address the limitations of CNNs' receptive fields.

For chamber segmentation, ML based techniques (both traditional and contemporary) beat human expertise. Building powerful ML based approaches, on the other hand, needs a big and well-annotated dataset.

2.2.2 Spatio-Temporal Segmentation

In addition to the above mentioned studies that have performed 2D segmentation, some studies have exploited the use of temporal information inherent in echocardiograms.

PV-LVNet, introduced by Ge et al. [36], instead of providing intermediate segmentation, focuses on predicting multiple indices associated with cardiac function. By leveraging temporal information through a recurrent network, PV-LVNet aims to accurately localize and isolate the LV across entire echocardiographic sequences. Subsequently, another recurrent network is employed to predict more precise indices, similar to LU-Net's approach to segmentation. Additionally, PV-LVNet incorporates multi-view information by providing cropped sequences from both A2C and A4C views to a single network for LV volume prediction. This approach highlights the integration of temporal dynamics and multi-view data to enhance the accuracy of LV volume estimation and cardiac functional assessment in echocardiography.

In another study by Li et al. [37], temporal information is utilized for segmentation from multi-view images in MV-RAN. They extracted multiscale features by utilizing dilated convolutions in the encoder. Two branches of the network, a 3D CNN and an LSTM are used. The 3D CNN classifies the cine's view and LSTM is used to perform segmentation based on spatial along with temporal information. While the 2D segmentation performance was reported for the ES and ED frames of the CAMUS dataset, the results regarding temporal consistency were only provided on a private dataset.

Recently, Wei et al. [38] presented CLAS, a 3D segmentation network designed to ensure temporal coherence using only ES and ED frames. CLAS achieves this by

predicting deformation fields and utilizing them for annotations during training. Their study demonstrated enhanced consistency between ES and ED predictions, aligning closely with intra-observer variability. However, for the temporal consistency across the cines, they provided only qualitative evaluations on a small patient cohort. Future studies could benefit from larger-scale quantitative assessments to validate the robustness of CLAS in clinical settings.

2.2.3 Edge Enhanced Segmentation

To delineate LV boundary, most studies have been primarily centered on using mask-based semantic segmentation techniques. There are a few studies in domains other than echocardiography that have also explored the concept of multi-task learning by integrating edge detection along with mask detection into neural networks. Such methodologies leverage image edge information to enhance the quality of segmentation in various domains. In a study by Lin G et al. [39], RefineNet is proposed, which aims to refine semantic segmentation results by exploiting multi-scale features at different levels. It utilizes boundaries as intermediate representations to refine segmentation results.

The work by [40] introduces a segmentation network for remote sensing. It integrates numerous weighted edge supervisions to preserve spatial boundary details. Their network accomplishes both semantic segmentation and edge detection simultaneously. By leveraging encoder edge loss to benefit from deep supervision in shallow layers and decoder edge loss to assist high-level semantic parsing, they demonstrate the significance of edge supervision in semantic segmentation. Another model, Gated-SCNN by Takikawa et al. [41], incorporates shape information into feature maps by incorporating the shape stream along with the regular stream. The gate mechanisms were used to define the information flow between the regular semantic stream and the shape stream, allowing for the extraction of targets and refinement of boundary predictions. In [42], the authors emphasized the decoupling of edge and mask prediction in the Mask R-CNN architecture. This highlighted the relevance of decoupling in the context of instance boundaries.

Their approach involved employing conventional edge detection filters on both predicted and ground truth masks to align them, ensuring rapid convergence during training. Similarly, in a study by Chang et al. [43], mask localization accuracy is improved by using object boundary information in the prediction process within the network. Zuo et al. in [44] used single-layer edge supervision to enhance the backbone network's perception of edge information, leading to improved segmentation accuracy, and further emphasizing the importance of edge information in segmentation tasks. Sui et al. [45] also integrated edge features into the network to optimize the loss function and enhance the segmentation results, aligning with the focus on leveraging edge information for guiding semantic segmentation tasks.

To the best of our knowledge, edge guided multitasking has not been explored for enhancing LV segmentation from echocardiographic data. Previous studies involving echocardiographic data predominantly concentrated on acquiring knowledge for semantic segmentation solely from ground truth masks. Employing multitask learning strategies to address semantic segmentation and boundary prediction simultaneously can enable models to leverage shared information, leading to improved performance and more accurate results.

2.3 EF Estimation

Zhang et al. in [5] developed a completely automated and extensible echocardiography interpreting process. As a part of their approach, they performed preprocessing complete echo studies, view classification, image segmentation, and detection of the cardiac cycle using CNNs. Using the segmentation output, LV length, area, volume, and mass were estimated, leading to the calculation of EF. The longitudinal strain was also calculated using particle tracking, followed by disease detection. It has been observed that automated algorithms overstate structural estimates, such as LV and LA volumes, whereas measures of function, such as EF showed better accuracy. While the results demonstrated convincing median absolute differences, large deviations could be seen in the results because of outliers.

The main reason behind extreme failure cases can be attributed to underlying complex segmentation tasks, where subtle variations and ambiguous boundaries challenge algorithmic precision, particularly in challenging clinical scenarios or unusual anatomical presentations.

The deep learning model based on echocardiography videos by Ouyang et al. in [14], automates the processes of segmentation of the LV, calculating EF, and evaluating cardiomyopathy. They used labels from human experts for weak supervised learning to conduct semantic segmentation of the LV for each frame of the video. Then, from native echocardiography recordings, a three-dimensional CNN is trained to estimate the EF for each frame. Finally, the segmentation findings are coupled with frame-level estimates of EF to generate a final evaluation for each cardiac cycle. Furthermore, heart failure with a low EF is detected. The dataset used in this research consisted of around ten thousand echocardiography recordings which are made publicly accessible by the authors.

However, EchoNet-Dynamic did not agree with the original expert labeled ground truth in a few videos. These films had incorrect human labeling, low visual quality, or arrhythmias and heart rate fluctuations. To prevent bias in measurements, several preprocessing procedures could be done to such videos to either increase their quality or automatically eliminate them from the training set. As a result, more research in a variety of clinical settings is required.

In [46] the authors extended previous analyses of [14] and used the same deep learning model EchoNet-Dynamic developed using echocardiography images in [14]. Their model was shown to be capable of classifying local cardiac structures as well as estimating volumetric measures and heart performance metrics. Left ventricular hypertrophy, an aberrant size of the left atrial, and the existence of devices like defibrillators and pacemakers were also identified. In addition, their model also gave predictions of certain demographic information from echocardiography images. The dataset used contained more than 2.6 million A4C view images from 2850 patients. When the model was trained to estimate EF directly, the findings were more accurate than when it was estimated using predicted volumes. In

this work, the prediction of EDV and ESV doesn't give promising results with low R^2 resulting in a higher bias in EF, which is calculated from ESV and EDV as is done in the clinical convention. To properly evaluate cardiac mobility and correlation in cardiac structures, future studies will require improved use of temporal information across frames.

In [47], an ML algorithm was developed to estimate EF without measuring LV volumes. The approach assumed that the ventricle contracts throughout the systole simultaneously along its long axis and in the radial direction, allowing the calculation of EF from the estimated contraction coefficients in the longitudinal and radial directions without measuring the volumes using Eq. (2.1).

$$EF = 1 - [C_{L-min}] \times [C_{R-min}]. \tag{2.1}$$

where;

$$C_L = \frac{l_d}{l_s} \;\; and \;\; C_R = \frac{D_d}{D_s}.$$

Here, l_s and l_d represent lengths of the LV, while D_s and D_d represent diameters of the LV during systole and diastole, respectively. For example, if during systole, the ventricle shortens by 14%, C_L (contraction coefficient in the longitudinal direction) would reach a minimum value of 0.86, and if at the same time, its radial dimension shortens by 30%, corresponding to the minimum C_R (contraction coefficient in the radial direction) value of 0.70, this would result in EF of 40%. The proposed ML algorithm is designed to train the computer to estimate the minimum values of the above-mentioned two contraction coefficients, $C_L - min$, and $C_R - min$, at the end of a contraction. To produce automated estimations of LV EF, the system is trained on a dataset of over 50,000 echocardiographic recordings, comprising numerous apical two and four-chamber views.

In [48], Liu et al. proposed a segmentation approach in which they improved the characteristics of regions with low contrast based on adjacent contexts while decreasing the detrimental influence of noise. Also, instead of independently predicting the class of each pixel, the results of neighboring pixels are also taken into

account explicitly. To achieve that, they presented the "deep pyramid local attention neural network (PLANet)" as a deep learning model in which they used supervised learning to make the FCN explicitly learn the pairwise interdependencies of labels. The learned label correlation is used as a weight to update segmentation by the adjacent prediction. They used the dataset of CAMUS and a subset of EchoNet-Dynamic to test their model. In a more recent work by Tokodi et al. [49] DL based network for estimation of RV EF from 2D echocardiographic videos without performing segmentation was proposed. In another study by Zeng et al., [50], LV segmentation was performed followed by EF estimation which required the identification of ES and ED frames.

2.4 CVD Classification

The identification or prediction of a particular heart illness based on visual characteristics or computed cardiac indices is known as CVD classification. The development of a CVD classification system that results in a completely automated diagnostic facility based on ML can pave the way for low-cost, high-grade medical care for patients in settings with limited resources. Such advancements can significantly improve early detection and treatment of cardiovascular diseases, thereby enhancing patient outcomes globally.

2.4.1 WMA and Cardiomyopathy Detection by Echocardiography

Two-dimensional echocardiography is widely used to identify and analyze wall motion abnormalities (WMA). Four terms are usually used in echocardiography to describe different types of WMA; Hypokinetic (reduced movement), akinetic (lack of movement), dyskinetic (abnormal movement), and aneurysm (abnormal wideness). Several heart disorders, such as cardiomyopathy and coronary artery disease (CAD), show these anomalies [7].

Cardiomyopathy is a heart muscle disorder that causes abnormal dilatation, stiffening, or loss of function of the heart's principal sectors. The three major cardiomyopathy disorders are dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), and ischemic cardiomyopathy (IC). DCM is a heart muscle condition that produces aberrant global motion by enlarging the LV wall. HCM is a muscular condition in which the heart muscle (myocardium) thickens, resulting in LV stiffness along with global and localized motion abnormalities. Ischemic cardiomyopathy (IC) results in heart muscle weakening [7]. CAD develops when the coronary arteries constrict or get clogged. Myocardial infarction (MI) is a dangerous heart condition caused by a severely constricted or blocked coronary artery. Automated CVD classification systems leveraging ML can enhance early detection and tailor treatment strategies for these complex cardiac conditions, thereby improving patient outcomes and optimizing healthcare resource allocation.

2.4.2 WMA and Cardiomyopathy Detection by ML Techniques

Several ML-based techniques have been reported in the literature that use automatically derived B-mode parameters (e.g., LV volume) or characteristics related to disease derived straight from the data to detect WMA, CAD, and cardiomyopathy disorders. Leung and Bosch [51], for example, presented an automated technique for assessing WMA. The proposed method is developed and evaluated on A2C and A4C echocardiographic images. The annotated contours are used to build an LV shape model, which is then analyzed with PCA to derive statistical parameters for anomaly categorization. The classifier is trained using a variety of PCA shape modes and parameters. In all circumstances, having fewer shape parameters results in a better classification rate. The trained binary classifier obtained an average accuracy of 91.1%. Similarly, Qazi et al. [52] employed a shape-based technique to determine the LV border in each frame automatically. The defined LV shape is then used to extract various cardiac structural and functional parameters, including circumferential and radial strains, as well as local,

segmental, and global Simpson volumes. The retrieved features are then reduced to the optimal features for training the classifier (Kolmogorov-Smirnov test). The trained classifier, which was evaluated on 220 instances, has a sensitivity of between 80% and 90% in identifying cases as normal or abnormal (hypokinetic, akinetic, dyskinetic, and aneurysm).

Shalbaf et al. in [53] established a quantitative regional index for WMA identification and CAD prediction. The ground truth labels comprised the LV area, landmarks, and levels of abnormalities. To define LV and produce a novel index for WMA classification, the proposed technique combines affine transformation and B-spline snake. The suggested index is calculated using the control points of the B-spline snake model. Two threshold values are used for classification, which are computed using the quantitative regional indices of all images in the training set. The established thresholds are used to classify the 125 instances in the testing set as normal or abnormal (hypokinetic, akinetic, dyskinetic, aneurysm). The proposed index's abnormality score and the ground truth had an absolute agreement of 83% and a relative agreement of 99%.

For CAD risk assessment, Araki et al. [54] introduced a method for classifying patients as high or low risk. The method begins by extracting 56 different grayscale features from the image that represent the coronary texture. Gray-level co-occurrence matrix, grey-level run length matrix, intensity histogram, grey-level difference statistics, neighborhood grey-tone difference matrix, invariant moment, and statistical feature matrix are some examples of these features. After that, six feature combinations are created, and the best one is selected based on classification accuracy. For CAD risk assessment, the best set is used to train a Support Vector Machine (SVM). A total of 2865 B-mode frames were gathered from 15 patients to test the method. In classifying patients as low-risk or high-risk, the suggested technique had an average accuracy of 94.95% and an AUC of 0.95. Other ML methods for CAD detection and risk assessment can be found in [55] (first-order statistical features, ANOVA for reduction, and NN classifier), [56] (trace transform and fuzzy texture), [57] (discrete wavelet transform and marginal fisher analysis), and [58] (GLCM and SVM).

In [59] and [60] automated techniques for identifying and diagnosing dilated cardiomyopathy (DCM) and hypertrophic cardiomyopathy (HCM) are proposed. The automated approach in [59] uses Fuzzy c-means (FCM) to perform frame-level segmentation which is then used to extract cardiac characteristics such as volume and EF. Principal component analysis (PCA) and discrete cosine transform (DCT) techniques are employed to extract features that are used with NN, SVM, and combined K-NN for DCM and HCM diagnosis. The PCA features with the NN classifier had the best performance in identifying normal and afflicted hearts, according to the experimental data. It also revealed that PCA characteristics were superior to DCT and cardiac indices (e.g., EF) for diagnosing DCM and HCM.

To discriminate hypertrophic cardiomyopathy (HCM) from physiological hypertrophy in athletes (ATH), Narula et al. [60] utilized three ML classifiers. Random forests (RF), SVM, and neural networks were employed as part of a classification ensemble. Using commercial software, many geometric and mechanical indices were retrieved from the defined chamber. The information gain method was then used to reduce this information further. It was found that mid-left ventricular segmental, volume, and longitudinal strain were the important characteristics or predictors for distinguishing between HCM and ATH, according to the results of the IG algorithm. The authors claimed that ML algorithms can reliably distinguish between healthy and pathological hypertrophic remodeling patterns. However, the dataset used in this study for testing purposes was relatively small. In addition, only systolic frames were utilized, thus restricting the exploitation of complete information that could have been done by including diastolic frames as well.

Table 2.1 summarizes the objectives, techniques employed, datasets, and results of the current work using ML and DL in echocardiography.

2.5 Multitask Learning

In computer vision, multitask learning has been widely employed to learn multiple related tasks simultaneously. Recently, a lot of work has been done in multitask

Table 2.1: ML and DL in echocardiography

Author	Objective	Technique Used	Dataset	Results
Abdi et al. [20]	QA on A4C	DCNN using Particle Swarm Optimization	A4C images, Train 6916, Test 1386	MAE 0.71 ± 0.58
Abdi et al. [21]	QA on 5 Views	DNN including LSTM	Train 4675, Test 1144	Accuracy 85%
View Classification				
Gao et al. [23]	8 View Classification	DNN incorporating Spatial and Temporal info	Train 280, Test 152	Accuracy 92.1%
Madani et al. [25]	Classification of 15 views and LV Hypertrophy	Ensemble of 3 CNN Models	A4C 267 and 455 cases	Accuracy 80% Accuracy 92.3%
Vaseli et al. [24]	12 View Classification	Lightweight models from distillation of VGG-16, DenseNet and ResNet	Train 9967 Validation 3322 Test 3322	Accuracy 88.1%
Segmentation & EF Prediction				
Melo et al. [26]	Segmentation	Watershed Algorithm	4C long axis 900 frames	RMSD 2.14 Corr 0.985
Zhang et al. [5]	View Classification	VGG Network	7168 videos	Accuracy 84%

Table 2.1 continued from previous page

	LV Segmentation	UNet	791 images	IoU 72% - 90%
	LV Size and Function		8666 cases	MAD 15% - 17%
	EF Prediction		6407 cases	MAD 9.7%
Asch et al. [47]	EF Prediction	Algorithm devised using CNNs	A2C, A4C 50,000 cases, Test 297 cases	Accuracy 0.92 MAD 2.9%
Ouyang et al. [14]	LV Segmentation	EchoNet-Dynamic	A4C videos	DSC 0.92
	EF Prediction		Train 7465	MAE 4.1%
	HF Classification		Test 1288	AUC 0.97
Leclerc et al. [31]	LV Segmentation	UNet based multitask network	CAMUS	MAE 1.5mm HD 5.1 mm
Li et al. [37]	Multiview Segmentation	Multiview recurrent aggregate network	CAMUS Curated data	DSC 0.92
Ghorbani et al. [46]	Pacemaker	DCNNs	A4C	AUC 0.89
	Enlarged LA		Data 2.6 Million	AUC 0.86
	LVH		Train 2546 cases	AUC 0.75
	EDV,ESV		Test 337 cases	R^2 0.74, R^2 0.70
	EF			R^2 0.50
	EF from volumes			R^2 0.33

Table 2.1 continued from previous page

Liu et al. [48]	LV Segmentation EF Estimation	Pyramid local attention	CAMUS sub EchoNet-Dynamic	DSC 0.956 Corr 0.882 Corr 0.869
CVD Classification				
Leung et al. [51]	WMA Classification	PCA	Train 65, Test 64	Accuracy 91.10%
Qazi et al. [52]	WMA Classification	Bayesian Network	220 cases	Accuracy 80%-90%
Shalbaf et al. [53]	WMA Classification CAD Prediction	Affine transformation and B-spline snake	125 cases	Accuracy 83%
Araki et al. [54]	CAD risk assessment	SVM	2865 B-mode frames 15 cases	Accuracy 94.95% AUC 0.95
Balaji et al. [59]	DCM, HCM Classification	PCA with BPNN	Normal 20 videos DCM 30 videos HCM 10 videos	Accuracy 92.04%
Narula et al. [60]	DCM, ATH Classification	Ensemble of SVM, RF & ANN	HCM 62 cases ATH 77 cases	AUC 0.80
Smitha et al. [58]	Plaques Classification	SVM	Carotid B-mode	Accuracy 97.92%

learning in different domains, such as image classification, pose estimation, and action recognition. By jointly learning related tasks, multitask learning enhances the model's performance, enables knowledge transfer, and promotes a deeper understanding of visual data. A detailed review of multitask learning in deep neural networks is given in [61]. It discusses different learning strategies, which include task-specific layers, hard parameter sharing, and soft parameter sharing. Hard parameter sharing involves sharing the network's lower layers among all tasks, but each task has its own output layer. Soft parameter sharing allows tasks to share parts of the layers while still maintaining their individual sets of parameters. Another work by Zhang et al. [62] also discusses both traditional and DL-based approaches for multitask learning, including problem formulation, optimization methods, and evaluation metrics.

In work by Amyar et al., [63] a method is proposed for the detection of COVID-19 pneumonia from chest CT scan. The proposed method uses a multitask DL architecture that integrates segmentation, classification, and reconstruction tasks. Other works which have utilized multitask in medical imaging include [64], [65] and [66] for Alzeihmer's disease prediction and progression,[67] in ECG, [68], [69],[70] and [71] in cardiac imaging.

In multitask optimization, task-specific models with distinct weights are employed, and a combined cost function is utilized. This allows the models to jointly optimize a single objective function while ensuring similarity in their parameters. Multitask optimization provides various benefits, including efficient data utilization, accelerated model convergence, and mitigation of model overfitting through shared representations.

2.6 Gap Analysis

Recent research on echocardiograms has emphasized the utilization of independent DL models for various tasks, including view classification, LV segmentation, and EF estimation. The accuracy attained by these methods depends on several

factors, including the clarity and quality of the data, the volume of available data, and the accuracy of the clinical ground truth. In automated pipelines designed to accomplish different tasks, the accuracy of each task is significantly influenced by the accuracy of the preceding one. For instance, in much of the existing literature, EF estimation relies heavily on the precise delineation of the LV boundary and the accurate identification of ES and ED frames. Moreover, the scarcity of publicly available medical data limits the achievable accuracy with data-intensive DL techniques.

The main limitations of earlier work are summarized as follows:

- Among the techniques commonly employed in clinical practice for assessing LV volumes and left ventricular EF, Simpson's method stands out as the preferred choice for evaluating LV EF [8]. However, most studies that have utilized clinical methods to determine LV volumes and EF from segmented LV have relied on the area-length method due to its simplicity. This method assumes the shape of the LV to be bullet-shaped, which may not always hold true [8]. Limited research has been conducted on exploration of Simpson's method for extracting LV features from segmented chambers and leveraging these features in conjunction with ML techniques to estimate LV volumes and EF. Therefore, scant attention has been given to the development of automated approaches based on the underlying principles of Simpson's method.

- Some methods may lack interpretability or transparency in their decision-making process, making it challenging to understand and trust the generated segmentations and predictions. Furthermore, there is a lack of emphasis on aligning with clinical workflow in existing methods. Adapting methodologies or techniques to match clinical practices not only ensures clinical relevance but also improves visibility, interpretability, and the ability to trace errors back more easily.

- Existing LV segmentation algorithms utilizing DL predominantly concentrate on pixel classification within the object's body, prioritizing low-level

features such as color, shape, and texture. However, they often overlook high-level details like edges and boundaries, leading to less accurate detection of LV borders. The commonly utilized encoder-decoder based architectures, known for their efficiency in generating semantic segmentations, suffer from the loss of important spatial information necessary for segmentation. The down-sampling and up-sampling operations in these architectures may result in the loss of fine-grained details, making it difficult to precisely outline object boundaries. Consequently, segmentation outputs from these methods may exhibit fuzzy or imprecise object boundaries. In delineating the LV, it is important to establish clear demarcations between adjacent structures, minimizing pixel ambiguity near the LV border. This includes accurately identifying specific boundary points such as the LV apex and the mitral valve annulus [8]. Furthermore, echocardiographic data present further challenges such as low signal-to-noise ratio, indistinct borders, low contrast, and organ variation. Ensuring accurate delineation of the LV boundary is essential for cardiologists to obtain precise clinical insights.

- Estimating EF necessitates measurements from systolic and diastolic frames within the cardiac cycle. Existing literature predominantly requires the identification of these specific frames for evaluation, which involves additional processing steps. This approach can be time-consuming and prone to inaccuracy, especially when conducted from a single view without the complementary orthogonal view. The exploration of utilizing the entire echocardiographic cine (video) for both LV segmentation and EF estimation, aiming to eliminate the need for frame identification and encompass the variations throughout the entire cardiac cycle, remains largely unexplored in existing research.

- Lastly, in the majority of existing studies on echocardiograms, LV segmentation, and EF regression have typically been addressed as separate tasks. The concept of simultaneous feature learning from segmentation and regression models is relatively novel and has not been extensively investigated. Training

these tasks concurrently allow models to leverage shared information, potentially leading to enhanced performance and more accurate results. Despite their distinct objectives and outcome requirements, LV segmentation and EF estimation tasks are closely interconnected and mutually reliant. Therefore, exploiting their interdependencies holds significant potential. However, cross-module learning via multitask optimization to exploit this shared information remains unexplored in this specific context.

Overall, the gap analysis highlights the need for more robust, standardized, and clinically validated automated methods for EF estimation from echocardiographic videos, addressing challenges related to performance variability, cardiac dynamics modeling, and clinical translation.

2.7 Problem Statement

Despite recent advancements in echocardiogram analysis, there remains a critical need for automated methods capable of precisely estimating left ventricular EF from echocardiographic videos. Existing approaches often lack consistency with clinical workflow and may overlook crucial temporal information inherent in echocardiographic frames. Furthermore, most studies treat LV segmentation and EF estimation as independent tasks, neglecting the potential benefits of simultaneous feature learning. Simultaneous training of segmentation and regression models could leverage shared information and lead to enhanced performance and accuracy. However, cross-module learning via multitask optimization remains unexplored in this context.

The deficiency in the existing literature highlights the imperative for a comprehensive and standardized approach that integrates LV segmentation, feature extraction, and ML methods to accurately estimate LV EF from echocardiographic videos. The incorporation of clinical methods, such as the preferred Simpson's

method, into the underlying principles of automated methodologies, would enhance transparency, reliability, and traceability, which are essential for seamless integration into clinical settings.

Moreover, the inherent relationship between LV segmentation and EF estimation emphasizes the need to explore the potential of jointly exploiting their interconnectedness using DL models to enrich the content and quality of information. In this regard, exploring segmentation and regression techniques beyond the traditional approaches may further contribute to enhancing accuracy.

2.8 Research Methodology

The research methodology used in this study, as depicted in Fig. 2.3, offers a systematic framework to develop an efficient and accurate model for automating tasks integral to an echocardiographic test. By following this methodology, the aim is to streamline and enhance the process of echocardiographic analysis, particularly in tasks related to LV segmentation and EF estimation.

Through the utilization of basic neural networks and traditional ML techniques, this research explores the feasibility of automating clinical methods using AI. This initial phase involves extracting pertinent features from segmented LV images based on Simpson's method, fundamental for quantifying cardiac function.

Moreover, the temporal information inherent in echocardiogram is exploited through the employment of LSTMs for EF estimation. Central to this investigation is the evaluation of the ML model's accuracy based on the recommended Simpson's method, contrasted against the direct application of this method as traditionally practiced by cardiologists in clinical settings. Through iterative refinement and validation against clinical standards, this study aims to establish the reliability and scalability of AI-driven cardiac assessments.

Building upon the insights gained from this foundational research, the subsequent objective is to advance toward the development of a more robust and automatic

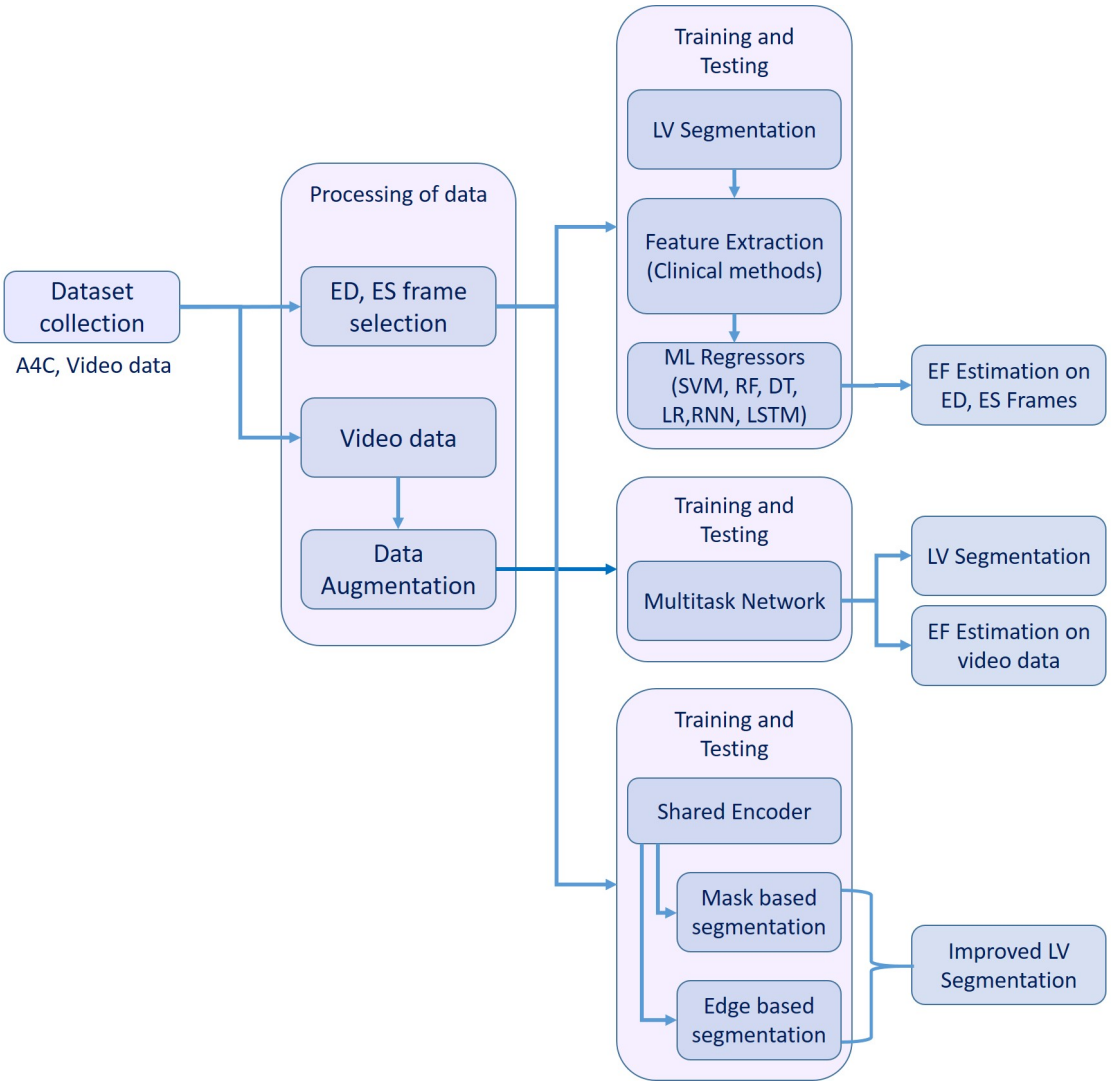


FIGURE 2.3: Research methodology

model capable of operating seamlessly on entire video sequences, thus circumventing the need for laborious detection of ES and ED frames. To address this, a multitask model named EchoFused Network (EFNet) is developed, which simultaneously performs LV segmentation and EF estimation. It used a semi-supervised DL approach, followed by cardiomyopathy detection. Integral to this task is the implementation of various normalization techniques aimed at scaling the objective functions from distinct tasks before training, ensuring smooth convergence of disparate components.

Additionally, a refinement in LV segmentation is pursued by decoupling edge and mask information. This involves employing an encoder-decoder based model with

separate decoders dedicated to edge-based and mask-based segmentation. By combining outcomes from both decoders, a final improved output is achieved. Rigorous evaluation through ablation experiments validates the reliability and effectiveness of the proposed methodology in automating echocardiographic tasks, promising significant strides in clinical efficiency and diagnostic precision.

In conclusion, the methodology used in this research provides a comprehensive and systematic framework for the development of an efficient and accurate automatic model for echocardiogram analysis. By automating key tasks involved in echocardiographic testing, the clinical workflows are enhanced, improving diagnostic accuracy, and ultimately, optimizing patient care.

2.9 Summary

This chapter provides a detailed literature review on the use of AI in automating different tasks performed as a part of an echocardiographic test. The existing studies such as view classification, quality assessment, cardiac segmentation, EF estimation, and CVD detection are investigated in detail. Furthermore, the chapter examines the utilization of multitask learning in medical imaging, providing insights into its relevance and applications. A detailed analysis of gaps in the current methodologies is presented followed by the problem statement. Finally, the chapter outlines the research methodology to address the identified gaps and challenges.

Chapter 3

Quantification of LV Function from Segmented Frames

This chapter describes the first contribution of this work; the quantification of LV function (EF estimation) based on features extracted from segmented LV frames using ML techniques. EF is an important clinical variable assessed from echocardiography via the measurement of LV parameters. Significant inter-observer and intra-observer variability is seen when EF is quantified by cardiologists using huge echocardiography data. ML algorithms can analyze vast datasets and detect complex patterns in the structure and function of the heart. This ability is valuable as it allows for computer-assisted diagnostics, supplementing the expertise of skilled human observers.

In this section of the research, LV segmentation is performed on ES and ED frames, followed by feature extraction from the LV based on clinical methods. Various approaches were explored to estimate the EF from these extracted features. Initially, we investigated different methods of combining diverse sets of features to create unified features. These consolidated features were then subjected to polynomial regression to estimate EF, along with end-systolic and end-diastolic volumes. The proposed feature functions included the Systolic-Diastolic Cross Ratio ($SDCR$) and Simpson's Systolic-Diastolic Cross Ratio ($SDCR_{simp}$).

In the subsequent part of the research, the features extracted from segmented LV underwent analysis using both NNs and traditional ML algorithms to estimate the EF. The findings from this approach suggest that employing ML techniques on the extracted features from the LV yields higher accuracy compared to utilizing Simpson's method for estimating the EF. The evaluations are performed on a publicly available echocardiogram dataset, EchoNet-Dynamic.

3.1 LV Volume and EF Estimation using Area-Length and Simpson's Method

As a first step, the two conventional methods most widely used by cardiologists to find the volume of the chambers; namely the area-length and Simpson's method of disks were employed. In order to accurately quantify volumes using these methods, the foremost and essential step is to find the length and diameter of the chambers correctly. The length of the LV is taken as the distance between the mitral annulus to the apex in the A4C view, whereas the diameter is the width at the mid-cavity level in the A4C view. Finding accurate LV length is found to play a very vital role in finding the end-systolic and end-diastolic LV volumes and hence the EF accurately. Therefore, several experiments were performed to find optimum accuracy in the estimation of LV length. The most accurate results were obtained by using a rectangular bounding box around the LV as shown in Fig. 3.1. The length and width of the bounding box served as the length and diameter of the LV, respectively.

LV volumes (EDV, ESV) are then found using the area and the length. The first approach employed is the area-length method, which is based on Eq. (1.1). The second approach used Simpson's method, which is the most widely used method in routine clinical practice to find LV volume from the volume tracings [8]. This method, also known as the biplane method of disks, divides LV into a number of elliptical disks and sums up their volumes to find the total volume of LV as described in section 1.5.1.1.

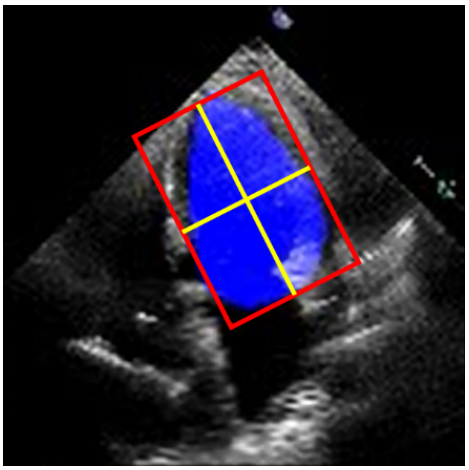


FIGURE 3.1: Minimum area rectangular bounding box

Finally, the volumes computed using both methods are employed to determine the LV EF using Eq. (1.4). The flow diagram depicted in Fig. 3.2 illustrates the procedural steps of this approach. It begins with the input of ES and ED frames, which are processed to derive an estimated EF as the output. This systematic framework underscores the integration of volumetric data analysis into EF assessment.

3.2 LV Volume and EF Estimation using Polynomial Regression

Polynomial regression is employed to explore the relationship between different features extracted from LV and LV measurements such as EF, ESV, and EDV using the EchoNet-Dynamic dataset. These features include the area, diameter, and length of the LV. Three models are trained by applying polynomial regression on a combination of the above mentioned features to;

(i) Predict EF from ESV and EDV,

(ii) Predict EF directly

Different sets of features were used in polynomial regression which include

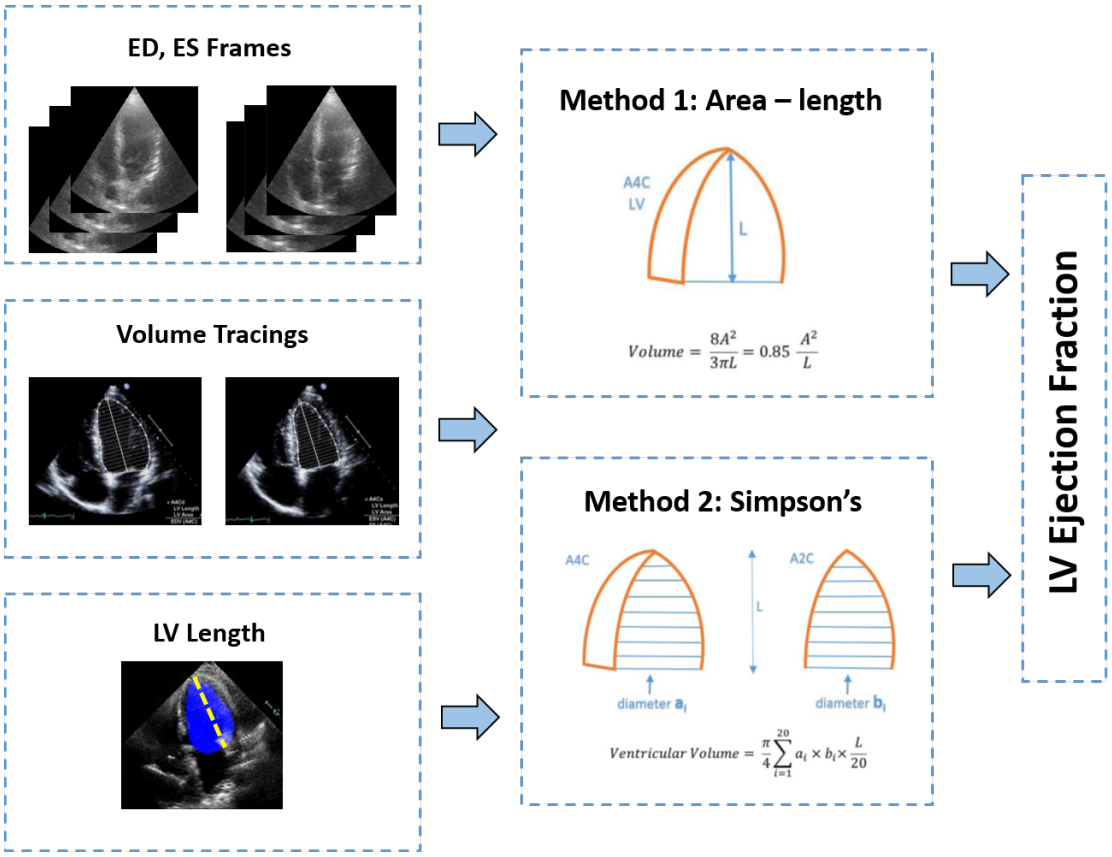


FIGURE 3.2: EF estimation using area-length and Simpson's method

(i) Multiple features,

(ii) $SDCR$ and

(iii) $SDCR_{simp}$

where the multiple features include respective areas, lengths, and diameters of end-systolic and end-diastolic frames to be used as features in polynomial regression. $SDCR$ stands for the Systolic – Diastolic Cross Ratio; the proposed feature function obtained from the area-length method and $SDCR_{simp}$ stands for the Simpson's Systolic – Diastolic Cross Ratio; the second proposed feature function derived from Simpson's method. The trained models are also tested against the CAMUS dataset.

3.2.1 Polynomial Regression on $SDCR$

$SDCR$, a new feature function that we have proposed, has been derived using the basic area-length formula and by studying the cross-correlation between different features and the ground truth EF and is defined by Eq. (3.1);

$$SDCR \equiv \frac{A_s^2 \times l_d}{A_d^2 \times l_s}. \tag{3.1}$$

Here, A_s and A_d indicate the areas obtained at the end-systole and end-diastole respectively, and l_s and l_d denote the LV lengths. This ratio simplified the regression process and its interpretation when used as a feature in polynomial regression in comparison to when multiple features were used.

3.2.2 Polynomial Regression on $SDCR_{simp}$

$SDCR_{simp}$ has been derived using Simpson's biplane method of disks. According to this method, the LV is divided into a row of n elliptical disks aligned perpendicular to the ventricle's major axis. Adding the respective volumes of these disks gives the total volume of the LV. The volume of a single disk in Simpson's biplane method is given by Eq. (1.2). We define k as the ratio between the semi-axes of the disks. For this work, we assume that k is some constant for a particular patient. Therefore;

$$Let \frac{b_i}{a_i} \equiv k.$$

Using this assumption in Eq. (1.2) and integrating over n-disks, we find the LV volume as given by Eq. (3.2);

$$v_i = \frac{\pi(a_i^2 \times k)l}{4},$$

$$V = \frac{\pi k L}{4n} \sum_{i=1}^{n} a_i^2. \tag{3.2}$$

Here, we also make another assumption that the constant k_D; defined as the ratio between semi-axes in end-systolic, remains the same as k_S; the ratio between semi-axes in end-diastolic i.e.

$$k_S = k_D.$$

Using this assumption in (1.4) yields;

$$EF = \frac{\left(\frac{\pi k_D L_D}{4n} \sum_{i=1}^{n} a_{Di}^2\right) - \left(\frac{\pi k_S L_S}{4n} \sum_{i=1}^{n} a_{Si}^2\right)}{\left(\frac{\pi k_D L_D}{4n} \sum_{i=1}^{n} a_{Di}^2\right)} \times 100\%,$$

$$EF = (1 - \frac{L_S \sum_{i=1}^{n} a_{Si}^2}{L_D \sum_{i=1}^{n} a_{Di}^2}) \times 100\%.$$

where we define $SDCR_{simp}$ as;

$$SDCR_{simp} \equiv \frac{L_S \sum_{i=1}^{n} a_{Si}^2}{L_D \sum_{i=1}^{n} a_{Di}^2}. \tag{3.3}$$

where L_S and L_D are the lengths of LV in the end-systolic and end-diastolic frames respectively. Similarly, a_{Si} and a_{Di} are the semi-axes lengths of an i^{th} disk shown in Fig. 1.5 in end-systolic and end-diastolic frames, respectively. Polynomial regression is then employed on these feature functions to find estimates of EF. The results obtained are presented in section 3.4.

3.3 EF Estimation using ML and NN Techniques

The subsequent part of the chapter describes the use of Simpson's method for the extraction of structural features of the LV. ML techniques are then employed on these features to estimate EF.

The proposed method first performs LV segmentation on an A4C view of an echocardiogram, then extracts pertinent features based on Simpson's method from

segmented images. This approach not only provides detailed insights into the intermediate steps involved but also remains consistent with the workflow of the methods used in clinical settings. By explicitly incorporating the segmentation and feature extraction steps, we can better understand and interpret the underlying processes, enhancing the transparency and interpretability of the EF estimation process. To estimate the EF, ML methods are applied to features obtained from the LV, and their efficacy, in this case, is investigated. Simple Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) are employed to leverage the temporal information present in the frames of an echocardiogram effectively.

3.3.1 Traditional Machine Learning Techniques

Traditional ML techniques for regression that are used in this study to estimate EF include Support Vector Regression (SVR), random forest (RF), decision tree (DT), and linear regression (LR). These traditional ML methods each have their strengths and weaknesses and are suitable for different types of regression problems.

3.3.1.1 Linear Regression

LR is a basic and widely used regression method that models the relationship between the independent variables and the dependent variable as a linear function. It assumes a linear relationship between the input features and the target variable, with the goal of minimizing the difference between the predicted values and the actual values. This simplicity and interpretability make LR particularly valuable for initial modeling and understanding the directional impact of predictors on the target variable.

3.3.1.2 Support Vector Regression

SVR is a supervised learning algorithm that aims to find the best hyperplane that maximizes the margin between the predicted values and the target variable. It finds a subset of training examples, called support vectors, that are used to define the hyperplane. SVR can handle linear and non-linear relationships using different kernel types, such as linear, polynomial, or radial basis function (RBF) kernels. This flexibility allows SVR to effectively model complex datasets and improve prediction accuracy across diverse applications.

3.3.1.3 Decision Trees

DTs represent a flowchart-like structure where each internal node represents a feature or attribute and each leaf node represents a predicted value. The DT splits the data based on the values of the features, aiming to minimize the variance within each resulting subset. DTs are easy to interpret and can handle both numerical and categorical data. However, DTs can be susceptible to overfitting, meaning they may not generalize well to unseen data and may instead memorize the training data. To mitigate this issue, the maximum depth of the DT can be set. If the maximum depth is set too high, the tree can become overly complex and capture noise or irrelevant patterns in the training data, leading to overfitting. In this study, the maximum depth of the DT was carefully selected to achieve a balance between capturing important relationships in the data and avoiding overfitting.

3.3.1.4 Random Forest

RF is an ensemble learning method that combines multiple DTs to make predictions. It works by creating a set of DTs using random subsets of the training data and random subsets of features. Each tree in the RF independently predicts the target variable and the final prediction is obtained by averaging or voting

on the predictions of individual trees. RFs are known for their ability to handle high-dimensional data along with non-linear relationships and provide feature importance measures.

3.3.2 Neural Network-Based Techniques

Neural networks (NN) are computational models inspired by the human brain. They consist of interconnected nodes called neurons organized into layers. The network receives input signals, processes them using activation functions, and produces output signals. By adjusting the connections between neurons, known as weights, NNs can learn patterns and make predictions. They are commonly used in ML and have been successful in tasks such as classification and regression.

RNNs and LSTMs are popular neural network architectures used in ML for regression tasks involving sequential data. These networks are effective in capturing temporal patterns and dependencies in sequential data, making them valuable tools for regression tasks involving time series or other sequential data formats.

3.3.2.1 RNN

RNNs are designed to handle sequential data by maintaining a hidden state that captures information from previous time steps. This allows them to capture temporal dependencies and model the dynamics of the sequence. In the context of regression, RNNs can learn to predict the next value in a sequence based on the previous values.

3.3.2.2 LSTM

LSTM is a variant of RNN that addresses the issue of vanishing gradients, which can occur with RNNs. LSTM introduces a memory cell that can retain information over long periods of time, allowing the network to capture dependencies over longer

sequences. It achieves this by using gates that control the flow of information into and out of the memory cell.

For regression tasks, RNNs and LSTM networks are trained to predict continuous values based on the input sequence. The input sequence is represented as a time series, where each element corresponds to a specific time step. The network processes the sequence iteratively, updating its hidden state and producing an output at each time step. The final output can be used as the regression prediction.

3.3.3 Proposed EF Estimation from LV Features

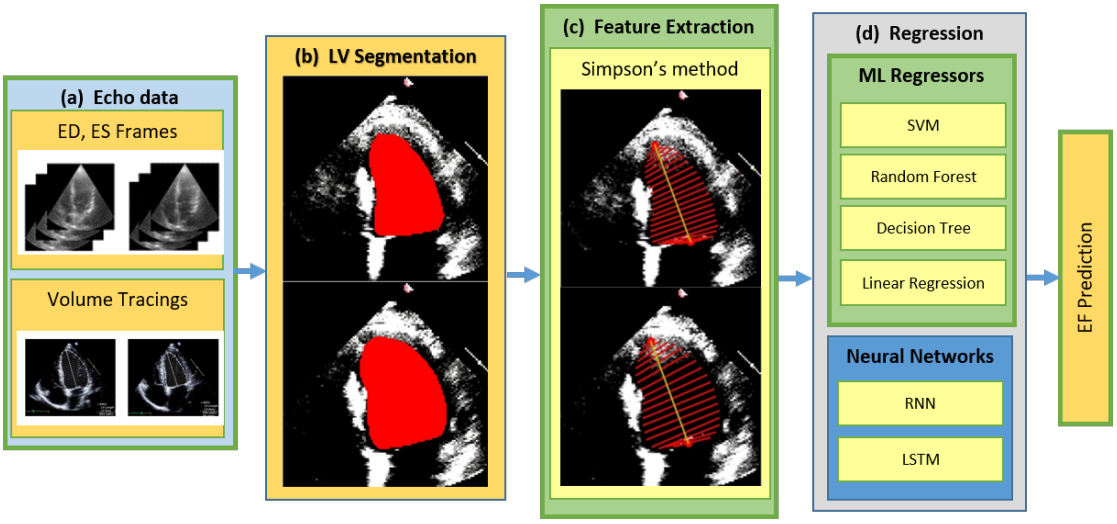


FIGURE 3.3: Proposed Method; (**a**) ES, ED input frames; (**b**) LV segmentation performed with DeepLab; (**c**) Simpson's diameters extracted from LV; (**d**) Regression performed using ML and NN algorithms.

The proposed method for estimating EF from the LV involves two main steps. Firstly, LV segmentation is performed on the A4C views of the dataset using DL techniques. The resulting segmented LV masks are then utilized for feature extraction based on monoplane Simpson's method. Regression techniques are subsequently applied to these extracted features in order to estimate EF. These techniques encompass both traditional ML approaches, chosen for their simplicity and ease of implementation, as described in section 3.3.1, as well as NNs outlined in section 3.3.2. The overall framework of the proposed method is illustrated in

Fig. 3.3. The different blocks of the framework are described in detail in sections 3.3.3.1–3.3.3.3.

Traditional ML algorithms present certain advantages. They offer interpretability, require fewer computational resources, and exhibit reliable performance even on smaller datasets with simpler relationships. These algorithms can provide insights and understanding into the underlying patterns within the data. NNs, on the other hand, demonstrate their strength in capturing intricate patterns and modeling non-linear relationships. They offer the advantage of requiring less explicit feature engineering and exhibit strong performance on large and complex datasets. NNs have showcased exceptional performance across various domains and exhibit good generalization capabilities when appropriately trained, albeit with the caveat of potential overfitting if the model capacity is not effectively controlled. By employing both NNs and traditional ML algorithms, the aim is to leverage the respective strengths of each approach and explore their effectiveness in addressing the research objectives.

3.3.3.1 LV Segmentation

The input to the LV segmentation module, as shown in Fig 3.3, consists of the ES and ED frames. These frames are processed by the LV segmentation module to extract the left ventricle region of interest. In this work, DeepLab is employed for the semantic segmentation of the LV chamber from an echocardiographic image. The use of atrous convolutions, commonly referred to as dilated convolution, is the primary component of the DeepLab model [72]. They enable the model to effectively capture features at different scales. By using the DeepLab model with the ResNet architecture as its backbone, the approach employs atrous convolution in a cascading or parallel manner. This allows for the capturing of multi-scale context by utilizing different atrous rates [72]. Additionally, the inclusion of the Atrous Spatial Pyramid Pooling module, along with image-level features, further enhances the performance of the model.

For the segmentation training, the PyTorch framework with pre-trained weights for DeepLabv3 was utilized. Resnet50 was used as the backbone. The model underwent training for a total of 45 epochs, employing a batch size of 16. The SGD optimizer was utilized with an initial learning rate set to 10^{-3}. To facilitate learning rate decay, a 'Reduce Learning Rate On Plateau' strategy was employed, decreasing the learning rate when no improvement was observed for a consecutive number of epochs known as 'patience'. In our case, the 'patience' value was set to 3.

During the segmentation process, the input image size was standardized to 112 $\times$ 112 pixels. The models were trained for A4C views from the EchoNet-Dynamic dataset by using the training, validation, and testing sets as mentioned in section 1.7. The model implementation was carried out using Python version 3.10.10 and PyTorch version 2.0.0, respectively. The experiments were conducted on a server equipped with an NVIDIA Tesla P100 GPU.

3.3.3.2 Feature Extraction

After LV segmentation, feature extraction is performed on segmented LV based on the monoplane Simpson's method. The EDV and ESV found using this method are used to calculate EF using Eq. (1.4), given in Chapter 1.

In our case, only A4C measurements are available; hence, the monoplane Simpson's method is utilized, which uses measurements obtained from A4C only. To extract features from the segmented images, the contour points (volume tracings) are derived from the segmented mask representing the LV at the ES and ED frames, respectively. The localization process identifies the mitral valve and apex points, which enable the determination of the major axis of the LV. The major axis is the length between these two points. Subsequently, the disks are positioned orthogonal to the major axis, spaced at the positions obtained by dividing the major axis into 20 equal parts. At each disk position, a region of interest (ROI) is defined as a disk shape centered at that position, and the pixels within this ROI are extracted. To calculate the diameter of each disk, the maximum distance between any two

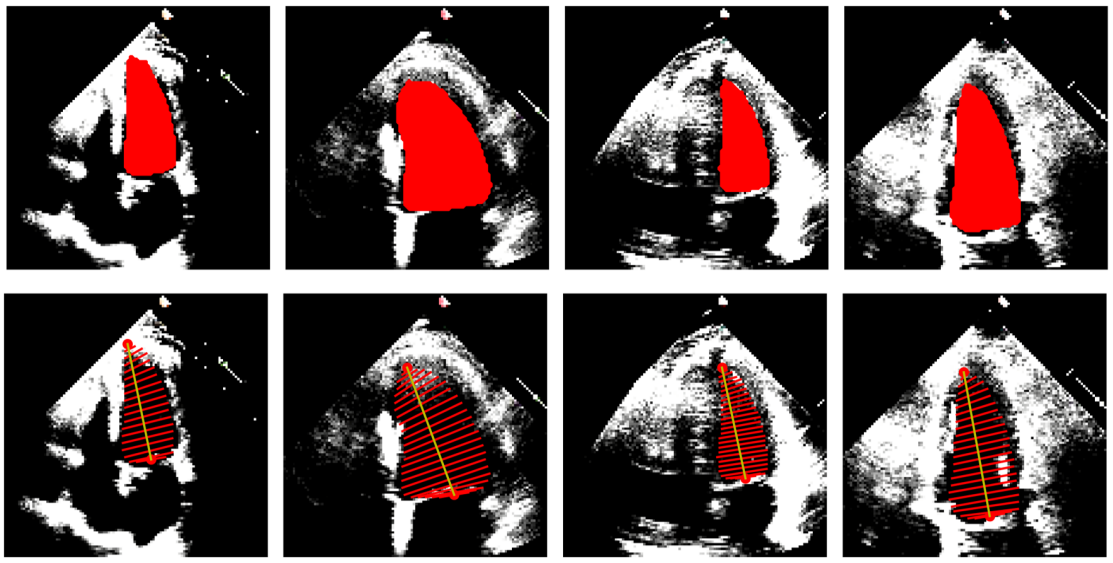


FIGURE 3.4: LV segmentation and feature extraction on systolic frames. Top: LV segmentation masks. Bottom: Diameter tracings.

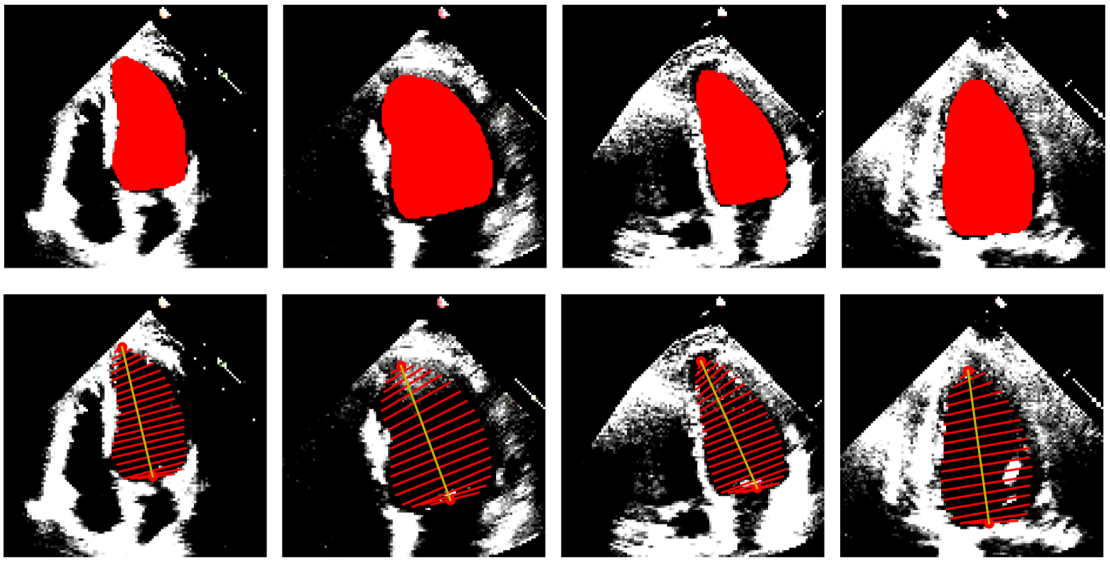


FIGURE 3.5: LV segmentation and feature extraction on diastolic frames. Top: LV segmentation masks. Bottom: Diameter tracings.

points within the ROI is computed. Figs. 3.4 and 3.5 show the segmented images and their corresponding feature extraction at the end-systole and end-diastole, respectively.

3.3.3.3 EF Estimation

The features extracted from the dataset are utilized as the training set for both ML algorithms, discussed in section 3.3.1 and NNs, described in section 3.3.2. To evaluate the performance of the model and optimize hyperparameters, a k-fold cross-validation is conducted on a validation set. The trained model is then assessed on a separate test set, employing various metrics, including the correlation coefficient (Corr), mean absolute error (MAE), and root mean squared error (RMSE).

Recognizing the significance of temporal information encompassing the ES and ED frames in echocardiogram outcomes, RNNs and LSTMs are employed to capture such temporal dependencies and investigate whether this method could yield improved results compared to the traditional ML techniques.

The RNN is a variant of neural networks that is created to effectively handle sequential data. RNNs exhibit remarkable suitability for tasks involving the analysis and generation of sequences, owing to their unique ability to retain internal memory or context. A distinguishing characteristic of RNNs lies in their recurrent connections, which facilitate the transmission of information from one step in the sequence to the subsequent step. This mechanism empowers the network to capture dependencies and discern patterns that unfold over time or sequence. During each time step, the RNN receives an input, which is processed alongside the internal memory derived from the preceding step. The network subsequently produces an output, along with an updated internal memory. This iterative process continues throughout each step in the sequence, establishing a temporal relationship that links the inputs and outputs. The internal memory of an RNN assumes the form of a hidden state, evolving as the network traverses each element of the sequence. This hidden state acts as a repository, preserving pertinent information related to prior inputs and their influence on the current step.

The LSTM network is a common version of RNNs. When propagating information over lengthy sequences, ordinary RNNs may have the issue of vanishing gradients.

To manage information flow and solve the vanishing gradient issue, LSTMs employ specialized memory cells and gating systems. This makes LSTM a preferred choice when exploiting both spatial and temporal information from the ES and ED frames.

3.3.3.4 Hyperparameter Tuning

In this study, the best hyperparameters and the corresponding model were obtained using a grid search with five-fold cross-validation.

TABLE 3.1: Hyperparameters of neural networks.

Hyperparameter	**Simple RNN**	**LSTM**
Model Layers	3	3
Dense Layer	1	1
Nodes in Layers 1, 2 and 3	128, 32, 16 units	64, 32, 16 units
Loss Function	MSE	MSE
Optimizer	Adam	SGD
Batch size	32	32
Epochs	70	100
Learning Rate	0.001	0.001
Activation Function	tanh	tanh

Table 3.1 gives the hyperparameter values for the neural networks. For both simple RNNs and LSTMs, sequential models were constructed using the Keras library, comprising three layers of each followed by a dense layer. To determine the optimal number of hidden units, different configurations were tested. The models' hyperparameters were fine-tuned to optimize their performance. The models were then compiled using the loss function and optimizer as mentioned for each NN in Table 3.1. This hyperparameter configuration was chosen based on the goal of minimizing the mean squared error (MSE) between the predicted and the ground truth values.

The hyperparameter values for the ML models along with their descriptions are provided in Table 3.2. For each model, various hyperparameters are listed, including kernel type, degree, gamma, C (regularization parameter), and epsilon

(tolerance around the ground truth) for SVR; number of estimators, maximum depth, and features to consider for best split for RF; and maximum depth of the tree, minimum samples required at a leaf node, and minimum samples to split at an internal node for DT. The table also specifies the values for grid search and the values ultimately selected for each hyperparameter. This information aids in understanding the parameter configurations utilized for training the respective machine learning models.

3.4 Results

3.4.1 Evaluation Metrics

The dice similarity coefficient (DSC) is used to evaluate the performance of the segmentation tasks by finding the similarity or overlap between the segmented image and the ground truth mask [28, 73, 74]. It is particularly useful for evaluating the accuracy of binary segmentation masks, as is the case in this study. The formula for calculating the DSC is given in Eq. (3.4).

$$DSC = \frac{2|\hat{y} \cap y|}{|\hat{y}| + |y|}. \tag{3.4}$$

where $\hat{y}$ and y represent the estimate of the segmented mask and the ground truth, respectively. $|\hat{y}|$ and $|y|$ represent the sizes of these masks, and $\hat{y} \cap y$ represents the intersection of these masks. The DSC ranges from 0 to 1, where 0 indicates no overlap or dissimilarity between the sets, and 1 indicates a perfect match or complete overlap between the sets.

To evaluate how accurately features have been extracted from LV masks based on monoplane Simpson's method, the Hausdorff Distance (HD) and MAE between contour points are used. The HD between sets U and V can be calculated as follows:

$$hausdorff(U, V) = \max\left(\max_{u \in U} \min_{v \in V} \|u - v\|, \max_{v \in V} \min_{u \in U} \|u - v\|\right). \tag{3.5}$$

Table 3.2: Hyperparameters of ML algorithms.

Model	Hyperparameter	Definition	Values for Grid Search	Values Selected
SVR	Kernel type	The kernel function	[Linear, RBF]	RBF
	Degree	Degree of the kernel function	[2–6]	-
	Gamma	Kernel Coefficient	[0.001, 0.01, 0.1, 1]	0.01
	C	Regularization Parameter	[0.1, 1, 10]	10
	Epsilon	Tolerance around the ground truth	[0.1, 0.01, 0.001]	0.1
RF	n_estimators	Number of trees in the Forest	[100–500]	400
	max_depth	Maximum depth of the tree	[None, 5, 10]	None
	max_features	Features to consider for best split	[auto, sqrt, log2]	Auto (40)
DT	max_depth	Max depth of the tree	[5–10]	9
	min_samples_leaf	Minimum samples required at a leaf node	[5–20]	17
	min_samples_split	Minimum samples to split at internal node	[5–20]	15

where U represents the set of contour points obtained from the ground truth LV mask and V represents the set of contour points obtained from the predicted LV mask. The HD and MAE in this case are given in millimeters (mm).

The accuracy of the regression estimates is measured using MAE and RMSE, which are the percentage errors in the case of EF. MAE serves as a metric to compute the aggregate of errors between estimated and actual values. It is calculated by taking the average of the absolute differences between predicted and actual values. MAE provides a measure of the average prediction error without considering the direction of the errors. RMSE gives comparatively large weight to big errors because the errors are squared before being averaged. In our case, undesirably larger errors are present mostly because of a few erroneous files in the data. Hence, MAE in this case is a more desirable metric from an interpretation point of view. The MAE and RMSE are given in Eqs. (3.6) and (3.7).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{z}_i - z_i|, \tag{3.6}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{z}_i - z_i)^2}. \tag{3.7}$$

where $\hat{z}$ and z represent estimated and ground truth EF respectively and N is the total number of data points.

To access the agreement between the estimated EF and the ground truth value of EF, the Pearson correlation coefficient is used, which is given in Eq. (3.8).

$$Corr = \frac{\sum_{i=1}^{N}(z_i - \bar{z})(\hat{z}_i - \bar{\hat{z}})}{\sqrt{\sum_{i=1}^{N}(z_i - \bar{z})^2 \sum_{i=1}^{N}(\hat{z}_i - \bar{\hat{z}})^2}}. \tag{3.8}$$

where z_i and $\hat{z}_i$ are the individual data points and $\bar{z}$ and $\bar{\hat{z}}$ are the averages of the estimated and ground truth EFs, respectively. A perfect correlation has a value of 1, whereas 0 indicates no correlation between z and $\hat{z}$.

3.4.2 Results of Area-Length and Simpson's Method

As a first step; area-length and Simpson's method were employed on Echonet-Dynamic dataset to find the end-systolic and end-diastolic volumes. EF is then calculated from these volumes using Eq. (1.4). Table 3.3 shows the results for RMSE and MAE obtained for EF on both Echonet-Dynamic and CAMUS dataset. As can be seen from Table 3.3, minimum MAE was achieved using Simpson's method, indicating that it models the geometry of LV better than the area-length method.

TABLE 3.3: Area-length and Simpson's method for EF prediction.

Dataset	**Area-length**			**Simpson's Method**		
	MAE	RMSE	Corr	MAE	RMSE	Corr
EchoNet-Dynamic	11.304	14.209	0.548	9.492	13.300	0.638
CAMUS	13.271	15.687	0.432	10.707	14.989	0.487

RMSE gives comparatively large weight to big errors because the errors are squared before being averaged. In our case, undesirably larger errors are present mostly because of a few erroneous files in the data, which the introduction of some pre-processing step could easily remove. Hence, in this case, larger RMSE is neither particularly undesirable nor is it affecting the result much. Also, MAE in our case is a more desirable metric from an interpretation point of view.

3.4.3 Results of Polynomial Regression

Table 3.4 shows the results of applying polynomial regression to estimate EF using various combinations of LV features as well as proposed feature functions. The results are obtained on the EchoNet-Dynamic dataset. As can be seen from Table 3.4; the best MAE is obtained by applying second-order polynomial regression to the proposed feature function; $SDCR_{simp}$.

Here, multiple features include: A_s – systolic area, A_d – diastolic area, D_s – systolic diameter, D_d – diastolic diameter, L_s – systolic length and L_d – diastolic length.

TABLE 3.4: Polynomial regression for EF prediction.

Features	**Degree**	**MAE**	**RMSE**	**Corr**
Multiple Features	1	12.691	14.139	0.552
	2	11.112	12.911	0.635
	3	9.189	11.577	0.648
$SDCR$	1	15.392	16.371	0.332
	2	10.342	12.309	0.525
	3	9.436	12.149	0.527
$SDCR_{simp}$	1	13.065	15.950	0.406
	2	8.715	10.446	0.655
	3	9.042	11.042	0.651

For the combination of multiple features, the MAE decreases from 12.691 to 9.189 as the polynomial degree increases from 1 to 3. Using $SDCR$ as a single feature, the MAE decreases from 15.392 to 9.436 as the polynomial degree increases from 1 to 3. This suggests that higher-order polynomial regression provides better accuracy in EF prediction using multiple features as well as $SDCR$. However, for the case of $SDCR_{simp}$, the MAE decreases from 13.065 to 8.715 as the polynomial degree increases from 1 to 2. However, there is a slight increase in MAE to 9.042 when the polynomial degree is further increased to 3. This suggests that the optimal polynomial degree for $SDCR_{simp}$ is 2, providing the lowest MAE compared to degrees 1 and 3. This phenomenon could be attributed to overfitting when using a polynomial degree of 3, leading to a degradation in accuracy.

For second order polynomial, using $SDCR$ as a feature function, the best fit is given by Eq. (3.9);

$$EF_{est} = 2.22(SDCR)^2 - 83.7(SDCR) + 92. \tag{3.9}$$

For second order polynomial, the best fit using $SDCR_{simp}$ is given by Eq. 3.10;

$$EF_{est} = 37.91(SDCR_{simp})^2 - 128.13(SDCR_{simp}) + 104.13. \tag{3.10}$$

The models given in Eqs. (3.9) and (3.10) using proposed features $SDCR$ and

$SDCR_{simp}$ respectively, are also evaluated against the CAMUS dataset. The results are shown in Table 3.5.

TABLE 3.5: EF prediction using proposed models on CAMUS.

Features	**MAE**	**RMSE**	**Corr**
$SDCR$	11.669	14.243	0.479
$SDCR_{simp}$	9.744	12.828	0.603

Here too, as shown in Table 3.5 the best results are obtained by employing second-order polynomial using the proposed feature; $SDCR_{simp}$.

To provide a comprehensive comparison of the performance of different methods, Fig. 3.6 presents a plot that visually contrasts the MAE and RMSE values of various estimation techniques using distinct markers.

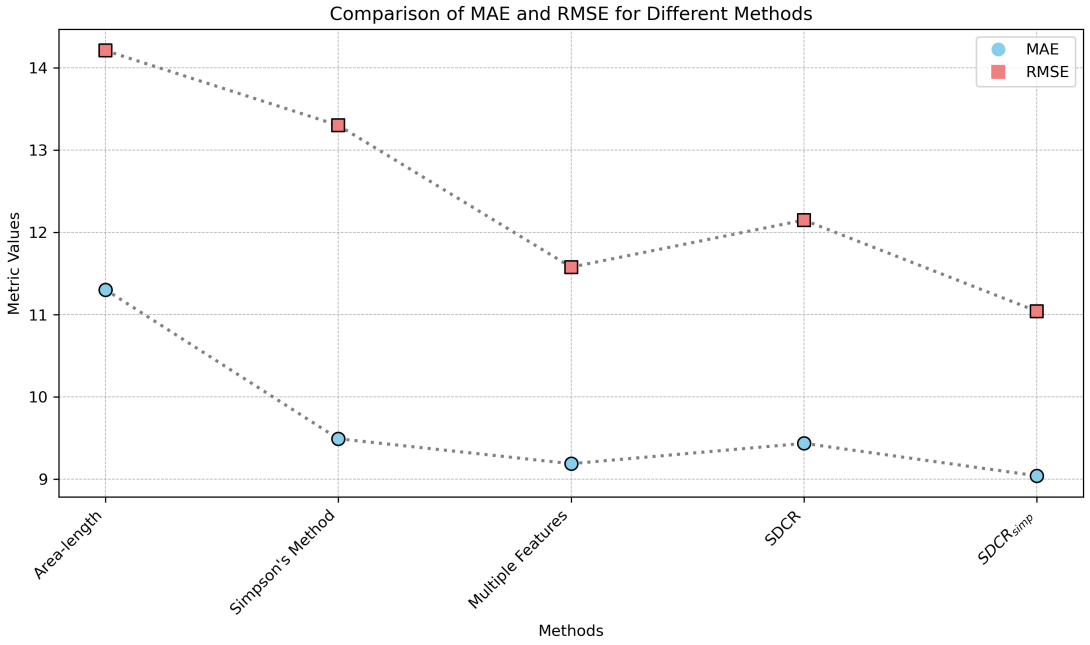


FIGURE 3.6: A comparison of MAE and RMSE of clinical methods with polynomial regression on different sets of features.

Another set of tests was run using two sets of features and varying polynomial degrees to estimate ESV and EDV, respectively. The results are given in Table 3.6. In this case, it can be seen from the results that the RMSE and MAE obtained are high in value showing that polynomial regression failed to capture the underlying distribution for the volumes. One of the main reasons is that it is not sufficient to

predict the volume of a 3D shape (LV) from 2D data that is available to us from the A4C view. Availability of an orthogonal view i.e. A2C in addition to A4C is a necessary requirement to provide a better prediction of the volumes. However, we were able to estimate EF with quite a reasonable accuracy as shown in Table 3.4 since it depends on the ratio of volumes rather than their absolute value.

TABLE 3.6: Polynomial regression for volume prediction.

Features	**Degree**	**RMSE**			**MAE**
		ESV	EDV	EF	EF
$A_s, A_d, D_s, D_d, L_s, L_d$	1	42.7	25.3	21.6	20.3
	2	38.6	24.3	20.7	17.9
	3	40.3	22.6	16.9	15.8
$A_s, A_d, D_s, D_d, \frac{1}{L_s}, \frac{1}{L_d}$	1	43.6	26.1	21.1	19.3
	2	41.5	25.7	19.0	17.6
	3	43.5	24.2	16.4	15.3

3.4.4 Results of ML and NN Techniques

The overall DSC obtained for LV segmentation is 0.92, which indicates a considerably reasonable similarity index. The DSCs for segmentation of systolic and diastolic frames are given in Table 3.7. The table also includes the HD and MAE for extracted features (volume tracings) in systolic and diastolic frames.

TABLE 3.7: LV segmentation and feature extraction.

	DSC	**HD (mm)**	**MAE (mm)**
Segmentation—Systolic	0.930	-	-
Segmentation—Diastolic	0.911	-	-
Volume Tracings—Systolic	-	6.324	6.704
Volume Tracings—Diastolic	-	7.280	5.716

To estimate EF, the set of features derived from monoplane Simpson's method comprises the diameters of the disks, along with the length of LV. A combination of regression techniques was applied to this set of features. LSTM and RNN were also used to estimate EF. Table 3.8 shows the results obtained.

TABLE 3.8: EF estimation from the extracted features.

Regressors	**MAE**	**RMSE**	**Corr**
LSTM	5.736	7.726	0.777
Simple RNN	6.489	9.180	0.746
SVR	6.727	8.908	0.689
RF	6.799	9.022	0.677
DT	6.865	9.084	0.671
LR	6.736	8.954	0.683
Simpson's Method	9.492	13.300	0.638

LSTM demonstrates the lowest MAE and RMSE among all models, indicating superior accuracy in EF estimation. The correlation coefficient of 0.777 suggests a strong linear relationship between the predicted and ground truth EF values. Simple RNN exhibits slightly higher MAE and RMSE compared to LSTM, indicating slightly inferior performance. However, it still maintains a relatively strong correlation with the ground truth EF values (Corr = 0.746). SVR, RF, DT, and LR show similar performance in EF estimation, with moderate MAE and RMSE values and correlation coefficients ranging from 0.671 to 0.689. While they may not achieve the same level of accuracy as LSTM and Simple RNN, they still demonstrate reasonable predictive capability. In contrast to the machine learning models, Simpson's method yields higher MAE and RMSE values, indicating lower accuracy in EF estimation. This is further illustrated in Fig. 3.7, which visually compares the MAE and RMSE obtained from both clinical and machine learning methods.

Overall, the results highlight the effectiveness of LSTM and other machine learning approaches in accurately estimating EF from extracted features, offering promising prospects for clinical applications in cardiac health assessment.

Fig. 3.8 displays the correlation and Bland-Altman plots for RNN, SVR, and LR, illustrating the relationship between the ground truth and predicted values of EF. Fig. 3.9 shows the comparison between the utilization of LSTM as a regression technique and the direct application of Simpson's method for calculating EF.

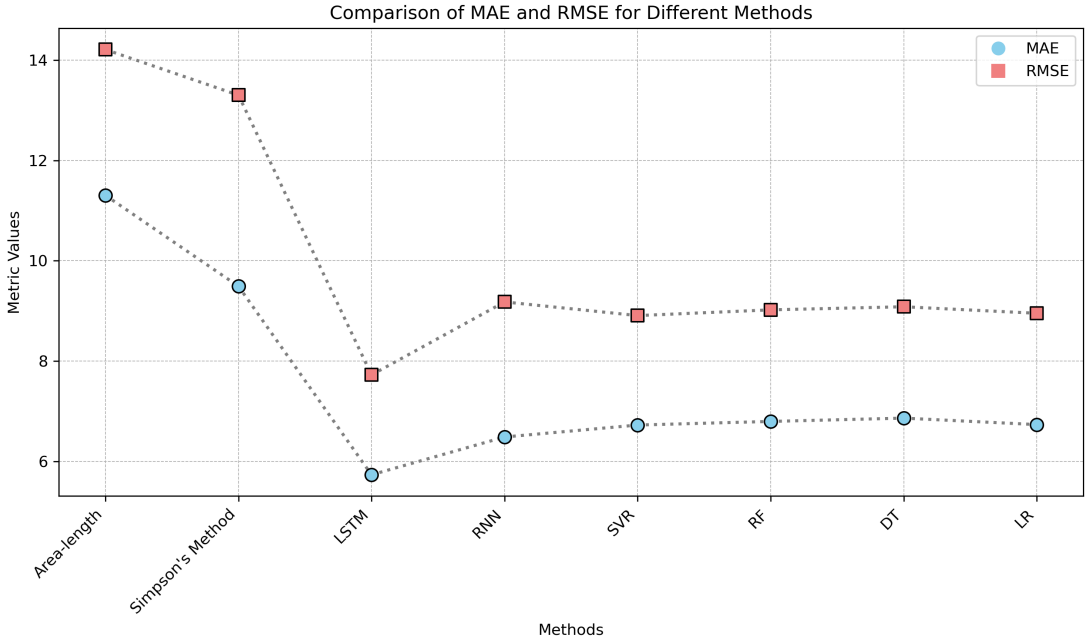
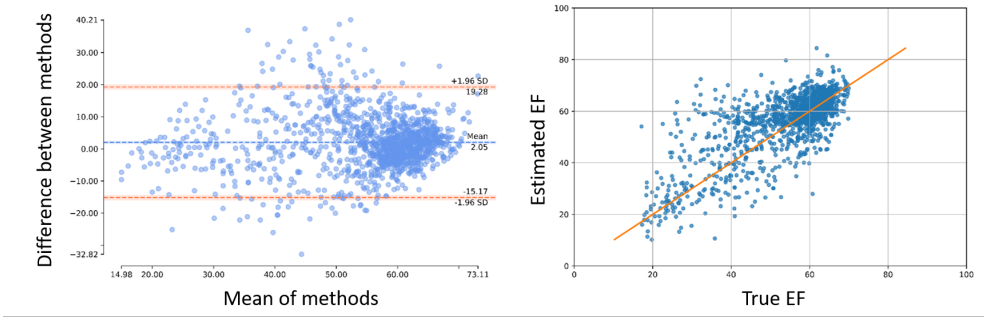
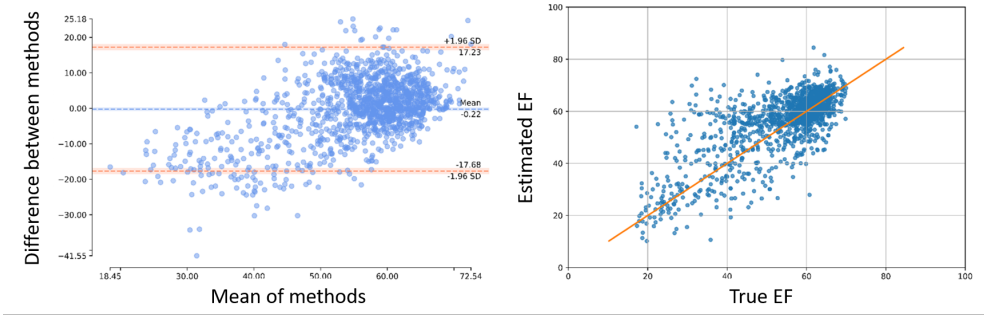


FIGURE 3.7: A comparison of MAE and RMSE of clinical methods with different ML methods.

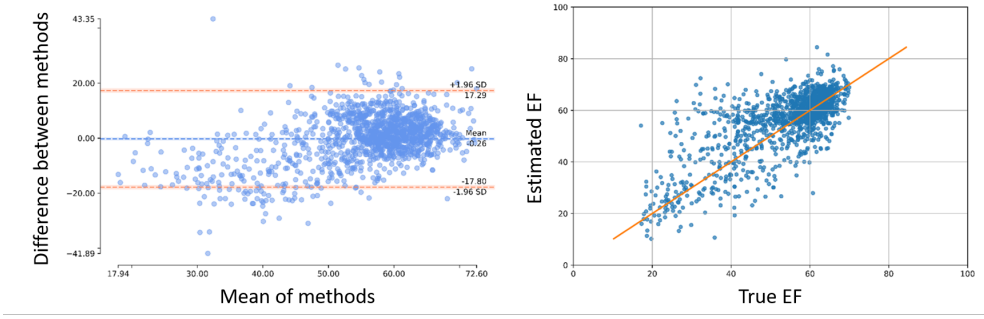
The Bland-Altman plot is a method used to assess the agreement between two different methods of measurement; the estimated and the ground truth EF values in our case. It is often employed in medical research and clinical studies to compare the agreement between two techniques that measure the same quantity. In a Bland-Altman plot, the differences between the two measurements are plotted against the averages of the measurements. This graphical representation assesses whether there is any systematic bias between the two methods and whether the differences between them are consistent across the range of measurements. Additionally, it provides insights into the limits of agreement, which represent the range within which 95% of the differences between the two measurements are expected to fall. As shown in Fig. 3.9; the EF data within 95% confidence interval for clinical Simpson's method is spread more than the data estimations obtained from LSTM. The correlation plot reinforces this observation, with EF values estimated from LSTM more concentrated around the line of perfect correlation as compared to the estimations obtained from the direct application of Simpson's method. Similarly, Fig. 3.8, which presents the plots for other ML methods such as RNN, SVR, and LR, demonstrates a better agreement between the estimated EF and the



(**a**) RNN



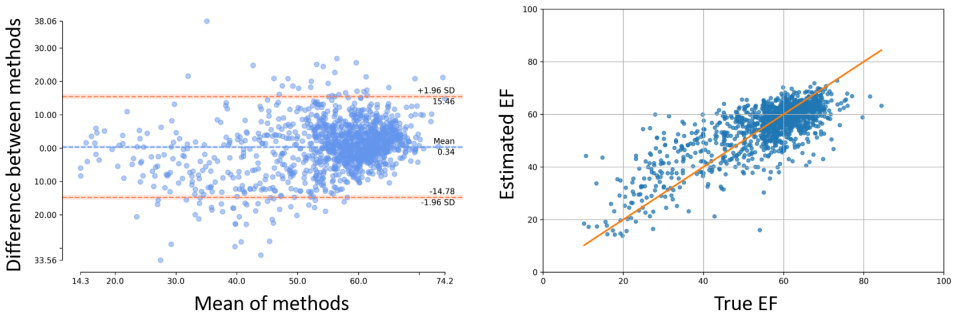
(**b**) SVR



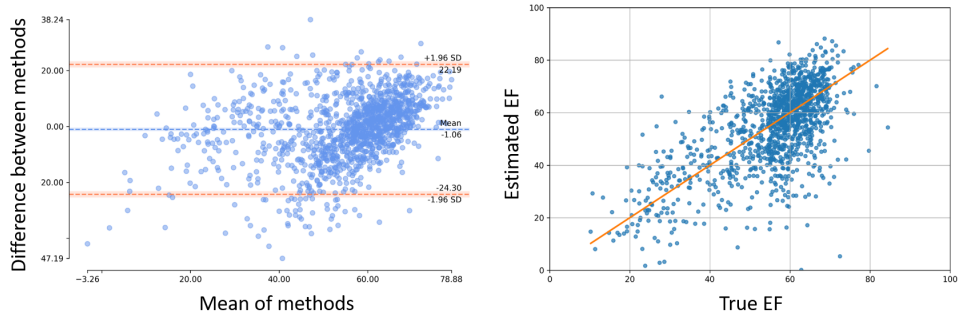
(**c**) LR

FIGURE 3.8: Left: Bland–Altman plot for RNN, SVR, and LR - The blue line shows the line of perfect average agreement, and the red lines show the limit of agreement bounds at ±1.96 standard deviation. Right: Correlation plot - The red line shows the line of perfect fit.

ground truth EF in comparison to the results from the Simpson's method. These findings highlight the superior accuracy and reliability of ML based approaches over traditional clinical methods for EF estimation.



(**a**) LSTM



(**b**) Simpson's method

FIGURE 3.9: Left: Bland–Altman plot for LSTM and Simpson's method - The blue line shows the line of perfect average agreement, and the red lines show the limit of agreement bounds at ±1.96 standard deviation. Right: Correlation plot - The red line shows the line of perfect fit.

3.5 Discussion

We initiated this research by investigating simpler and computationally less intensive methods, which are grounded in clinical principles. This initial exploration aimed to grasp the process of feature extraction and the relationship between features and various structural and functional parameters of the LV. As part of this exploration, we introduced two feature functions for use in polynomial regression. These functions not only streamlined the regression process but also yielded results that were comparable to more complex techniques.

In this part of the research, when EF was estimated using direct clinical methods on both EchoNet-Dynamic and CAMUS datasets, the monoplane Simpson's method

produced the best results; as indicated in Table 3.3. The proposed feature functions were then used to perform polynomial regression on the EchoNet-Dynamic data, resulting in the two models for EF estimation presented in Eqs. (3.9) and (3.10). The best results were achieved using the model of Eq. (3.10) based on the feature $SDCR_{simp}$. The models developed in Eqs. (3.9) and (3.10) were also applied on CAMUS to evaluate EF. The effectiveness of the suggested models is demonstrated by the comparable MAE achieved for CAMUS, even though the models had never been exposed to this data previously.

Building upon this foundation, the research progressed to employ ML and NN techniques on the extracted features for estimating EF. This proposed method once again aligns with the clinical workflow of the EF estimation and gives insight into the intermediate steps undertaken by cardiologists in the process of EF calculation, including LV segmentation and finding volume tracings. By incorporating these essential components, the proposed method not only ensures interpretability but also maintains consistency with the established clinical workflow.

In clinical practice, LV segmentation is typically obtained from the end-diastole and end-systole frames of an echocardiogram. The LV is divided into disks, which are used to calculate the left ventricular end-diastolic and end-systolic volumes. EF is then derived using these volumes obtained from both the A2C and A4C views. In this part of the study, the main focus was on the limitations caused by the unavailability of orthogonal views of data. This limitation restricts the use of Simpson's biplane method, which relies on having orthogonal views. In such a case, using ML regression algorithms for EF estimation provided improved outcomes as compared to clinical Simpson's method.

In this study, both NNs and traditional ML algorithms have been utilized. A key consideration in this research revolves around the intrinsic need to leverage the temporal information embedded within the data. The EF, which quantifies the difference between the volume of blood pumped into the LV during systole and the volume pumped out during diastole, heavily relies on the temporal information encapsulated between these two consecutive frames. The utilization of RNN

and LSTM models enables the exploitation of this information between systolic and diastolic frames. These network architectures are specifically designed to handle sequential data and capture the temporal dependencies between consecutive inputs, enabling them to learn from the contextual information across time.

TABLE 3.9: Comparative analysis with other studies for EF estimation.

Model	**MAE**	**RMSE**	**Corr**
MC3 * [33]	5.91	6.80	-
R2 + 1D * [33]	6.87	7.55	-
LSTM [75]	8.08	11.98	0.35
DL Based Workflow [76]	6.50	-	0.76
Proposed	5.73	7.72	0.78

* MC3: Mixed Convolution 3, R2 + 1D: Spatiotemporal Convolutional block.

The minimum MAE for estimating EF was obtained when LSTM was employed for regression with Simpson's diameters and LV length taken as a set of features. In Table 3.9, a comparative analysis was conducted for EF estimation using various existing methods on the EchoNet-Dynamic dataset. The proposed method outperformed the estimates by other studies, demonstrating its superior performance in EF estimation.

3.6 Summary

The objective of this chapter was to estimate the LV function utilizing echocardiographic frames from the EchoNet-Dynamic dataset, comprising the A4C views. The initial phase entailed the segmentation of the LV from the videos, followed by the extraction of relevant features from the obtained segmentation results. Subsequently, various regression algorithms were employed to estimate the left ventricular EF based on these extracted features.

Among the regression techniques employed, the LSTM network gave the best results with the least MAE. Additionally, the SVR algorithm demonstrated comparable outcomes while offering the advantage of lower computational complexity.

The evaluation of EF is limited when relying solely on the A4C view, as its accuracy is heavily dependent on the selection of a single imaging plane. If the chosen plane does not adequately represent the entire LV volume, the measurements may not accurately reflect the true LV function. However, tracing the endocardial border in multiple phases of the cardiac cycle can be a time-consuming and labor-intensive process, especially in challenging scenarios such as poor image quality or fast heart rates. As a result, the proposed techniques offer a valuable alternative by yielding results of acceptable accuracy even when measurements are based on a single imaging plane. This simplifies the analysis process and provides a practical solution for clinical applications.

Chapter 4

Quantification of LV Structure and Function from Echocardiogram Videos

The accurate quantification of both LV structure and function directly from the echocardiogram videos is important for the diagnosis and management of various cardiovascular conditions. Numerous studies have utilized DL techniques in the quest to develop automated methods that can reliably and precisely quantify both the structural and functional aspects of the LV. For the quantification of LV structure, Leclerc et al. [15] used variations of a UNet fine-tuned for LV segmentation on their publicly released dataset, CAMUS. Later, they [31] developed LU-Net, influenced by Mask R-CNN principles, which first predicts an ROI around the heart and then accurately segments within the ROI. This approach demonstrated improved results over their previous UNet architecture. Moradi et al. [32] proposed MFP-Unet, inspired by feature pyramid networks (FPN), incorporating dilated convolutions to expand the receptive field and upscale feature maps, enhancing 2D echocardiographic segmentation performance. Ouyang et al. [14] who released the EchoNet-Dynamic [33] dataset for assessing EF and segmenting the LV in A4C sequences, introduced a DL system for weak supervised learning segmentation of the LV and EF estimation across cardiac cycles. Their results showed that ML

techniques outperformed human expertise in chamber segmentation. They highlighted the importance of preprocessing to improve data quality and avoid biases. Ghorbani et al. [46] extended this work by introducing a model capable of classifying cardiac structures and estimating volumetric measures. Their model also predicted demographic information from echocardiography images. Zhang et al. [5] developed an automated echocardiography interpreting process that involved preprocessing, view classification, segmentation, and cardiac cycle detection using CNNs, estimating LV length, area, volume, mass, EF, and longitudinal strain, followed by disease detection. Liu et al. [48] proposed PLANet, a DL based segmentation approach that improves low-contrast regions and reduces noise impact by considering neighboring pixel results. They introduced a deep pyramid local attention neural network to learn pairwise label interdependencies, which was tested on CAMUS and EchoNet-Dynamic subsets.

Typically, these algorithms perform these tasks in a pipeline, i.e., performing segmentation initially, and identifying ES and ED frames based on segmentation outcomes, upon which EF estimation is conducted. Extending the work presented in Chapter 3, a single fully automated multitask network, the EchoFused Network (EFNet) is introduced that simultaneously addresses both LV segmentation and EF estimation tasks through cross-module fusion. The proposed approach makes use of semi-supervised learning to estimate the EF from the entire cardiac cycle, yielding more dependable estimations and obviating the need to identify specific frames. To facilitate joint optimization, the losses from task-specific modules are combined using a normalization technique, ensuring commensurability on a comparable scale.

4.1 EF Quantification from Cardiac Cycle

The current recommended clinical method to find EF estimates i.e. biplane method of disks (modified Simpson's rule), requires the extraction of ES and ED frames from an echocardiogram video. The LV tracings are derived from these

frames through manual labeling, which leads to the computation of EF. This procedure is time-consuming and prone to variability due to human involvement. The ground truth labeling for LV segmentation is typically limited, often available only for ES and ED frames due to the impracticality of labeling the entire video sequence encompassing numerous frames. However, accurate estimation of EF relies on capturing temporal information from a video sequence containing at least one complete cardiac cycle.

In the majority of existing studies on echocardiograms, the tasks of LV segmentation and regression of EF have been treated as independent tasks. Simultaneous feature learning from segmentation and regression models is a relatively novel concept that hasn't been extensively explored before. By training these tasks simultaneously, the models can exploit the mutually shared information, leading to improved performance and more accurate outcomes. In this study, we attempt to address two important tasks: LV segmentation and EF estimation. Despite their distinct nature and varying outcome requirements, these tasks are intricately interconnected and rely heavily on each other. Exploiting the interdependencies between them holds great potential for achieving our research objectives efficiently. Therefore, we aim to use cross-module learning through multitask optimization to utilize this shared information. In multitask optimization, task-specific models with distinct weights are employed, and a combined cost function is utilized. This allows the models to jointly optimize a single objective function while ensuring similarity in their parameters. Multitask optimization provides various benefits, including efficient data utilization, accelerated model convergence, and mitigation of model overfitting through shared representations.

The proposed multitask model EFNet, enables concurrent segmentation and regression from echocardiogram videos, employing joint optimization of the objective function and leveraging their interconnectedness to enhance overall performance. The integration of objective functions from two distinct tasks, each with varying scales, is carefully examined and addressed systematically. Effective strategies are devised to seamlessly combine these objective functions, ensuring coherence and

consistency in the model's training process. The proposed model undergoes training and evaluation using a larger dataset, enabling robust learning from a diverse range of samples. Furthermore, the model is fine-tuned on a smaller dataset, investigating the potential benefits of leveraging DL techniques to train effectively on limited data resources. This additional step aims to explore the model's adaptability and performance optimization when dealing with smaller datasets. Data augmentation methods are utilized to enhance the dataset's size and diversity.

4.2 Background on ML Techniques: Segmentation

Semantic segmentation creates a pixel-wise mask that identifies and delineates different objects or regions in the image. Delineating the LV through semantic segmentation can be achieved through DL models that generate pixel-wise masks of the LV region, enabling quantitative analysis of its size, shape, and function. However, it is a challenging task due to variations in heart anatomy, image quality, and potential artifacts. The different DL architectures used for segmentation in this study include DeepLabv3, DeepLabv3+, FCN, and UNet models, which are state-of-the-art models. Different versions of the ResNet architecture, ranging from ResNet18 to ResNet101, have been employed as the backbone architecture for these models. Varied depths were explored to determine the optimal network configuration that exhibits the minimum depth required while maintaining an acceptable level of accuracy.

4.2.1 DeepLabv3

The fundamental element of the DeepLab model is the utilization of atrous convolutions, also known as dilated convolutions. By adopting DeepLabv3 with ResNet as its foundation, the approach incorporates atrous convolution in a parallel fashion. This enables the model to capture contextual information across multiple

scales by employing different atrous rates [72]. Moreover, the model's performance is augmented through the inclusion of the Atrous Spatial Pyramid Pooling module in conjunction with image-level features. In DeepLabv3+, the inclusion of the Feature Pyramid Pooling (FPP) module enhances segmentation accuracy by integrating feature maps from multiple levels of the network hierarchy. This module specifically addresses the challenge of capturing fine-grained details, resulting in improved segmentation performance.

4.2.2 Fully Connected Neural Network

The fundamental concept behind Fully Convolutional Networks (FCNs) is to retain the spatial information across the network, allowing for pixel-wise predictions in semantic segmentation tasks. FCN architectures commonly integrate skip connections to merge features from multiple levels of the network hierarchy. These connections play a crucial role in capturing both local and global contexts, thereby improving the accuracy of segmentation results.

4.2.3 UNet

UNet was originally introduced for biomedical image segmentation by Olaf Ronneberger et al. [28]. In the UNet architecture, the encoder plays a crucial role in capturing high-level features from the input data. It consists of multiple downsampling blocks, comprising convolutional layers followed by max-pooling. On the other hand, the decoder part, known as the expanding path, consists of upsampling blocks that utilize up-convolutional layers. These blocks are then concatenated with skip connections originating from the corresponding contracting path. This mechanism facilitates accurate object localization by combining both low-level and high-level features. UNet++ is an extension of the original UNet architecture, proposed by Zongwei Zhou et al. [77]. It further improves the performance of UNet by capturing comprehensive contextual information at multiple scales through nested skip connections, enhancing the segmentation results.

4.3 Background on ML Techniques: Regression

For quantification of the LV function, the regression module utilized in this study is the R2Plus1D model available in the torchvision library. This model is a variant of the two-stream 3D CNN architecture, originally designed for video action recognition tasks. The (2+1)D convolution, also referred to as spatiotemporal convolution, plays a crucial role in capturing both spatial and temporal features in video data. In our case, the temporal information is contained in the echocardiogram multiple-frame video sequence, which captures the variation in left ventricular volumes. During the (2+1)D convolution operation, filters are applied across the spatial dimensions of each frame as well as the temporal dimension of the sequence. This allows the filters to capture spatial patterns within individual frames and temporal patterns across the cardiac cycle. The resulting feature maps are then flattened and fed into fully connected layers to learn complex relationships between the extracted features and the regression target. The final layer of the model produces the regression output, representing the predicted values for the EF estimation. By leveraging the spatiotemporal capabilities of the R2Plus1D model, we can achieve a more accurate and robust assessment of LV function. This approach enhances the precision of EF estimations, ultimately contributing to improved diagnostic capabilities in clinical practice.

4.4 Proposed Model: EFNet

The proposed EFNet simultaneously performs LV segmentation and EF estimation on echocardiogram videos by performing cross-module fusion. EFNet comprises two modules: a segmentation module and a regression module. The objective of the segmentation module is to delineate the LV's boundary by creating a pixel-wise mask. The regression module aims to estimate the EF from the video. The cross-module fusion integrates embeddings extracted from the segmentation module into the regression module, allowing for simultaneous training of both modules

through joint loss optimization. Fig. 4.1 illustrates an example input consisting of a sequence of frames from a cardiac cycle.

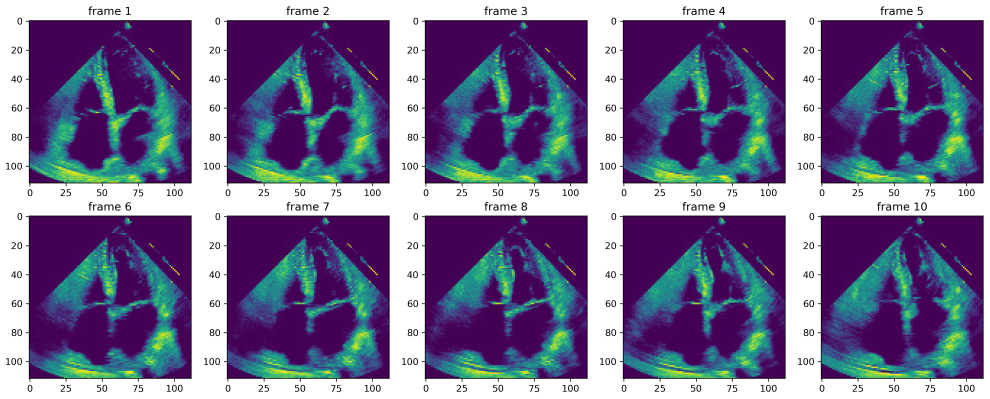


FIGURE 4.1: Example video frames extracted from a cardiac cycle.

Before providing data to EFNet, it undergoes a few preprocessing steps to enhance model performance. Data augmentation is employed to enhance the diversity of the dataset and to add robustness to the model. The transformations that are employed include rotation, shearing, translation, and composite transformations (sequence of translation, rotation, and shear). Details of data augmentation techniques are provided in section 4.4.3. After augmentation, the data is normalized to ensure consistent scales and facilitate convergence during training.

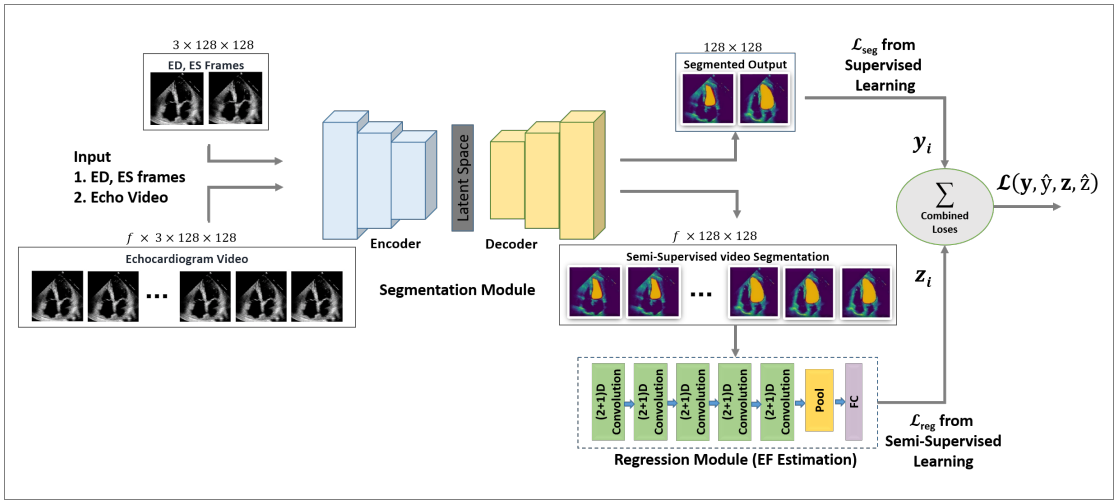


FIGURE 4.2: Proposed model: The EchoFused Network (EFNet). Input is echocardiogram videos with f frames, $\hat{z}_i$ is the EF estimate. $\hat{y}_i$ is the segmentation estimate. $\mathcal{L}$ is the loss function

The proposed methodology of the EFNet is illustrated in Fig. 4.2. An echocardiogram video sequence consisting of f frames is given as an input to the EFNet. The

dataset, as described in section 1.7, has limited labels for the segmentation of LV, available only for the ES and ED frames; no labels are available for the remaining frames in the video sequence. Due to the limited availability of labels, segmentation training is conducted in a semi-supervised manner to obtain embeddings from the video data. While the sparse ground truth labels are utilized to assess the losses in the segmentation module, the segmentation embeddings derived from the entire video sequence are employed in the training of the regression module. The UNet++ encoder is utilized as a common encoder to extract features from the input data. Skip paths are used to connect the encoder and decoder in UNet++ [77] to minimize the gap between the feature maps from the two in advance of their fusion. The ResNet50 architecture is used as the backbone for feature extraction. Fig. 4.3 gives the detailed architecture of the segmentation module.

The skip connection between the encoder and the decoder consists of a dense convolutional block. Let $x^{m,n}$ represent the output from the node $X^{m,n}$ where m represents the downsampled layer along the encoder and n represents the layer of the dense block along the skip connection. The feature maps are combined to obtain an estimate $\hat{y}_i$ according to Eq. (4.1);

$$\hat{y}_i = f_{ReLU}\left(conv\left(concat(x^{m,0}, x^{m,1}\ldots, x^{m,n}), E(x^{m+1,n-1})\right)\right), \tag{4.1}$$

where:

$$E(x^{m+1,n-1}) = f_{ReLU}\left(upconv\left(x^{m+1,n-1}\right)\right). \tag{4.2}$$

Here $\hat{y}_i$ is the output feature map, $f_{ReLU}(\cdot)$ is the non-linear activation function, $concat(\cdot)$ represents the concatenation layer and $upconv(\cdot)$ represents the transposed convolutional layer for upsampling. The loss function for segmentation is computed from the dissimilarity between the predicted masks and the ground truth masks obtained from ES and ED frames.

The embeddings obtained from the segmentation module on the video sequence are processed by the regression module to estimate the EF. This regression module is

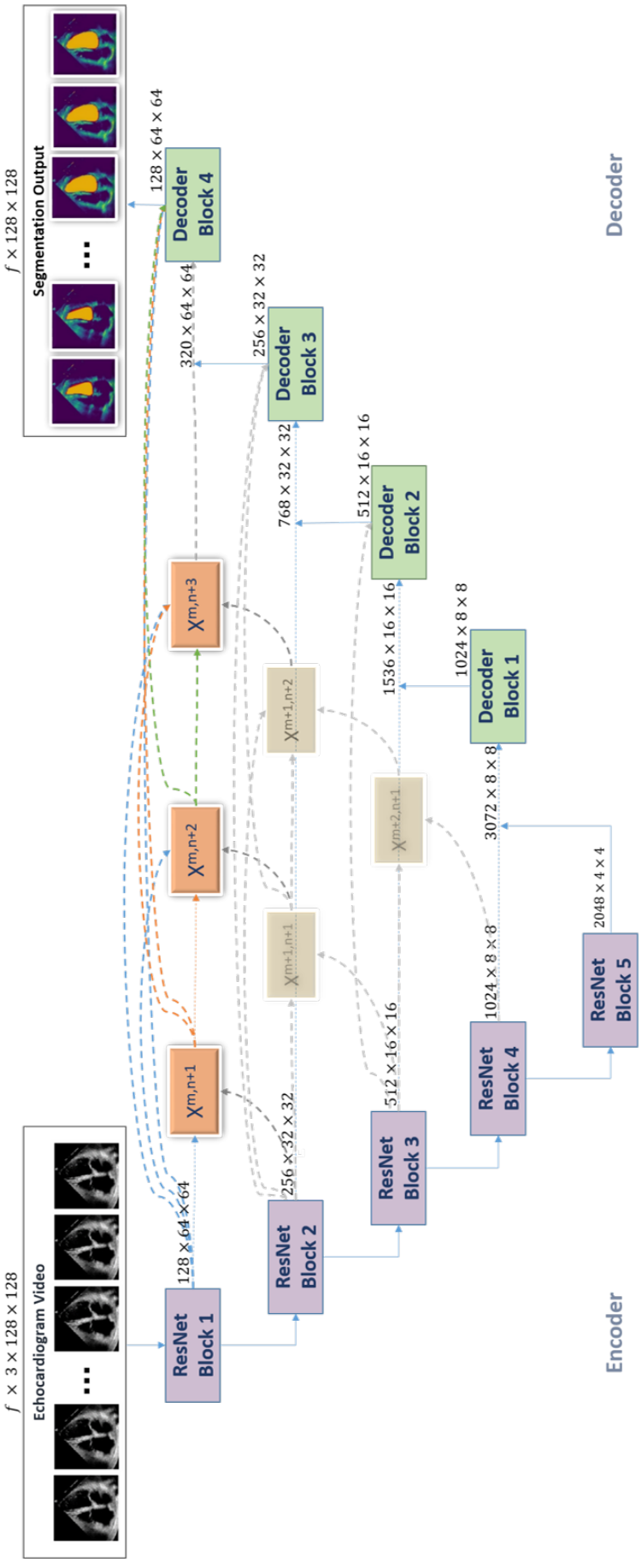


FIGURE 4.3: Encoder-decoder based segmentation module

the R2Plus1D model, which is a variant of the two-stream 3D convolutional neural network architecture, originally designed for video action recognition tasks.

The R2Plus1D spatiotemporal convolutions in the regression module decompose 3D convolutions into separate spatial and temporal components. The first convolutional layer performs 2D spatial convolution on each frame embedding $\hat{y}_i$ of the video sequence obtained from the segmentation module, given by Eq. (4.3). The 1D temporal convolution is then applied to the spatial tensors z_s of the entire video, yielding the output tensor z_t, which is given by Eq. (4.4).

$$z_s = conv2d\left(\hat{y}_i, W_s, b_s\right), \tag{4.3}$$

$$z_t = conv1d\left(z_s, W_t, b_t\right). \tag{4.4}$$

Here, W_s, b_s, W_t, and b_t are weights and biases for spatial and temporal convolutions, respectively. The output from each residual block of R2Plus1D, represented by z_r, is given by Eq. (4.5);

$$z_r = f\left(conv\left(z_t, W, b\right)\right). \tag{4.5}$$

Finally, the output z_i from the fully connected layer is given by Eq. (4.6);

$$z_i = f\left(\sum_j \left(W_{ij} z_r + b_i\right)\right). \tag{4.6}$$

4.4.1 Joint Loss Function

The training of cross-module EFNet depends on effectively combining and optimizing the segmentation and the regression loss functions, which are binary cross-entropy (BCE) and mean squared error (MSE), respectively. The BCE loss function computes the binary cross-entropy loss for each element in the input tensor and then calculates the mean of these individual losses. This ensures that the returned loss value represents the average loss across all elements in the input tensor. The reduction operation aggregates these individual losses into a single

scalar value, reflecting the overall loss of the entire batch or dataset. By taking the mean, the loss value becomes normalized and independent of the batch size or dataset size, facilitating easier comparison and interpretation. This normalization is crucial for maintaining consistency and comparability across different training iterations and data samples. On the other hand, the MSE loss function assigns a higher penalty to large differences between predicted and true values by squaring the differences. Consequently, larger errors contribute more significantly to the overall loss. By taking the mean, the MSE loss yields a scalar value representing the average squared difference between the predicted and true values. In our case, since the EF value is a percentage, the MSE range lies between 0 and 100. This range provides a clear and interpretable measure of the model's predictive accuracy, making it easier to assess performance improvements during the training process.

To enable simultaneous learning, joint optimization is employed using various strategies to ensure that losses from both networks are on a compatible scale. These strategies are thoroughly investigated and discussed in detail in section 4.4.2, to find the most effective combination of loss functions. The combined loss function from the segmentation and regression modules, given by Eq. (4.7), is backpropagated to update both the models' weights. This approach leverages shared features and encourages the model to find a balance between both tasks.

$$\mathcal{L}(y, \hat{y}, z, \hat{z}) = \frac{1}{N\rho}\sqrt{\sum_i (z_i - \hat{z}_i)^2} - \sum_i y_i \log(\hat{y}_i), \tag{4.7}$$

where

$$\rho = z_{max} - z_{min}. \tag{4.8}$$

Here, z and $\hat{z}$ are ground truth and estimated values from the regression module, while y and $\hat{y}$ are ground truth and estimated values from the segmentation module, respectively. N represents the sample size.

The training steps for our proposed model, EFNet, are outlined in Algorithm 1.

Algorithm 1 Cross-Module Regression and Segmentation Training

Input:
Video Data: $\{\{X_{ij}\}_{i=1}^{f}\}_{j=1}^{N}$, (f frames, N samples)
Ground truth for regression: $\{EF\}_{i=1}^{N}$
Pairs of frames (Segmentation): $\{F_s\}_{i=1}^{N}, \{F_d\}_{i=1}^{N}$
Ground truth Masks (Segmentation): $\{M_s\}_{i=1}^{N}, \{M_d\}_{i=1}^{N}$
Other Parameters: Number of epochs; n_{epochs}, Learning rate for Regression, Learning rate for Segmentation.
Initialization:
Select backbone ResNet architecture for both regression and segmentation modules.
Initialize both modules with pretrained ImageNet weights.
Training Loop:
for $epoch = 1$ to n_{epochs} **do**
    Sample input and ground truth video data for both regression and segmentation modules.
    Sample F_s, F_d along with M_s and M_d for segmentation loss
    Segmentation:
    Generate Segmentation feature maps $\hat{y}_i$ from F_s and F_d frames using Eq. (4.1).
    Compute Segmentation loss by comparing with M_s and M_d.
    Cross-Module Fusion:
    Pass video sequence $\{\{X_{ij}\}_{i=1}^{n}\}_{j=1}^{N}$ through segmentation module and obtain resulting embedding z.
    Pass this embedding to the regression module.
    Regression:
    Obtain spatial and temporal embeddings; z_s and z_t on z using Eqs. (4.3) and (4.4). Find output z_i from the fully connected layer using Eq. (4.6).
    Compute regression loss from the ground truth EF and estimated EF.
    Combine Losses:
    Normalize the regression loss.
    Combine the normalized Regression and Segmentation losses using Eq. (4.7).
    Backpropagation:
    Backpropagate the combined loss to update the weights of both modules simultaneously.
end for

4.4.2 Loss Function Normalization Techniques

To combine loss functions, it becomes essential to normalize the MSE loss obtained from the regression module to make it comparable with the segmentation loss. The methods explored and experimented in this study to normalize the regression loss include normalization of the RMSE by i) the standard deviation, ii)

the interquartile range, iii) the difference between the maximum and minimum values, and iv) the mean of the ground truth values [78], which are given as under;

$$NRMSE = \frac{RMSE}{\sigma}, \tag{4.9a}$$

$$NRMSE = \frac{RMSE}{Q3 - Q1}, \tag{4.9b}$$

$$NRMSE = \frac{RMSE}{z_{max} - z_{min}}. \tag{4.9c}$$

$$NRMSE = \frac{RMSE}{\bar{z}}. \tag{4.9d}$$

where NRMSE is the normalized root mean squared error, however, determining the most suitable normalization method is not straightforward as there is no evident superiority of one over the other. There are a few implications that we need to consider when computing normalized RMSE. While scaling by mean and standard deviation is commonly used, it may not always be the optimal approach. Normalization by standard deviation alters the original data scale and range, making it susceptible to the influence of outliers and skewness. Consequently, this method can lead to misleading results if the data follows a different distribution or if outliers hold significance in the analysis.

In our study, outliers are typically undesirable and arise due to cardiac chamber foreshortening or certain abnormalities in the heartbeat. When dealing with such cases, utilizing the interquartile range also tends to obscure the underlying patterns, making it necessary to opt for robust alternatives. On the other hand, max-min normalization maintains the relative order and distance of data points, in contrast to mean, standard deviation, and interquartile range normalization. This was also validated through experiments conducted on these normalization schemes, and it was determined that the most accurate results were obtained by

normalizing the RMSE by the difference in maximum and minimum EF values as given in Eq. (4.9c).

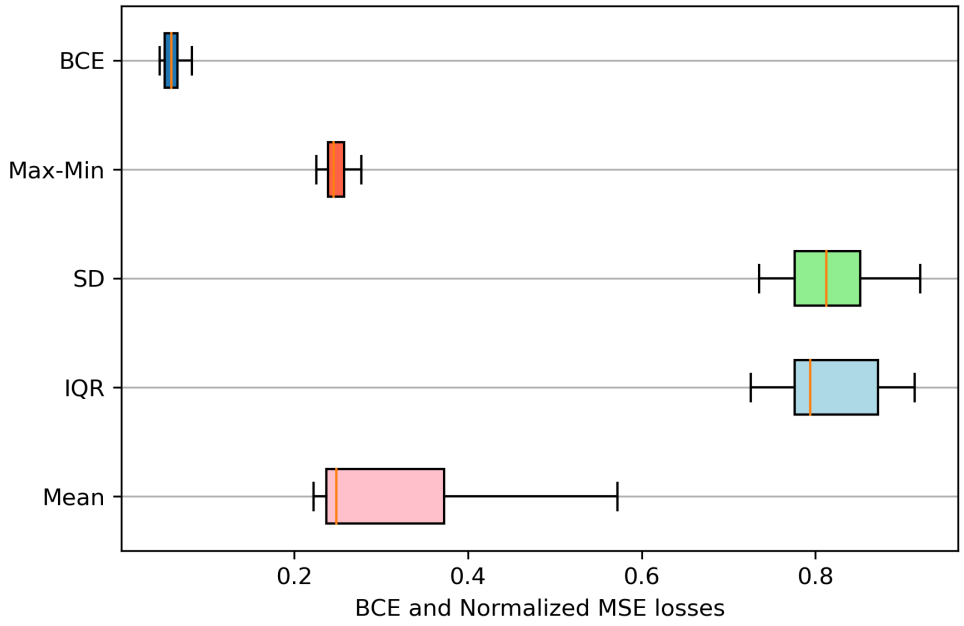


FIGURE 4.4: Comparison of different RMSE normalization methods. Loss samples are obtained on the validation set from EchoNet-Dynamic with a batch-size of 20. SD: Standard deviation, IQR: Interquartile range

A comparison of NRMSE using different normalization techniques is illustrated in Fig. 4.4. The losses are obtained on the validation set with a batch size of 20. The model used is FCN with resnet50 backbone architecture. Notably, normalization by mean and interquartile range has spread the range of losses, preventing their direct comparison with the BCE losses from segmentation.

4.4.3 Data Augmentation

Image augmentation techniques are used in this study to expand the training dataset in order to address the challenge of limited data availability. We performed an in-depth analysis by applying commonly used augmentation techniques to our training data and selected the most suitable ones based on both qualitative and quantitative analyses. It was found that geometric transformations produced better results as compared to intensity-based transformations. The techniques

applied include rotation, translation, shear, and composite transformations. The values for different parameters required for respective augmentation techniques are chosen based on experimental analysis. For rotation, a random rotation angle ranging from 0 to 25 degrees is selected. The translation operation is applied on both the horizontal and vertical axes using values sampled randomly from the range of 0 to 0.15, which is the maximum absolute fraction for horizontal and vertical translations. The horizontal shift is randomly sampled within the range [-image width * 0.15, -image width * 0.15] and the vertical shift is randomly sampled within the range [-image height * 0.15, -image height * 0.15]. The shear transformations can be applied both horizontally and vertically. They help improve the model's ability to recognize objects from different viewpoints and orientations. In this method, a shearing angle, which is selected randomly within the range of 0 to 10 degrees, is used to apply the shearing operation along the x and y axes. Lastly, a composite transformation comprising a sequence of translation, shearing, and rotation is applied to the data. The translation was carried out as previously described. For shearing, we selected shear angles from the range of -5 to 5 degrees. As for rotation, we randomly chose angles from -15 to 15 degrees. These combined transformations allowed us to create diverse and augmented datasets for our experiments. Each augmentation technique replicated the entire dataset once; hence, the total dataset size was increased five times after applying augmentation, bringing the training size from 7,460 to 37,300 videos.

The augmentation is performed simultaneously on the ES and ED frames along with their respective masks, in order to train the segmentation module. It is also performed on the sequence of video frames selected for training the regression module. Fig. 4.5 shows an example frame for each augmentation technique.

4.4.4 Model Evaluation

The input frames to EFNet from the echocardiogram videos are scaled to 128 $\times$ 128 pixels. The model was implemented using Python version 3.10.12 and PyTorch version 2.0.0. The experiments were performed on a server featuring an NVIDIA

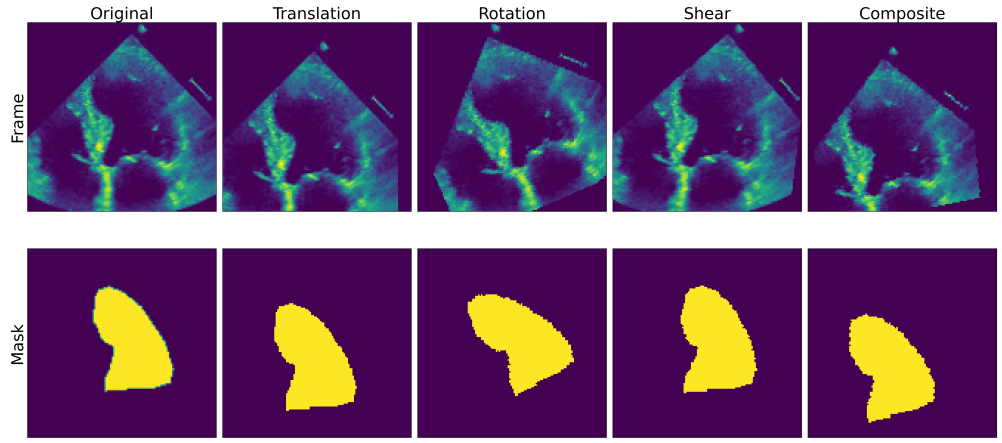


FIGURE 4.5: Example frames with their respective masks showing results of various augmentation techniques applied to EchoNet-Dynamic

Tesla P100 GPU. The proposed model undergoes training and evaluation using a larger dataset known as EchoNet-Dynamic. The model is initially trained on this sizable dataset, enabling it to capture a comprehensive range of cardiac dynamics and variations. Subsequently, the model is fine-tuned using a smaller dataset called CAMUS and subjected to evaluation. This approach leverages the benefits of both datasets, harnessing the richness and diversity of EchoNet-Dynamic during initial training and refining the model's performance by adapting it to the specific characteristics of the CAMUS dataset. Through this two-step process, the model achieves a robust and adaptable performance across these two different datasets.

4.5 Results

4.5.1 Evaluation Metrics

The DSC is used to evaluate the accuracy of segmentation performance using the formula given in Eq. 3.4.

Other metrics used to evaluate the performance of segmentation models are Pixel Accuracy and Intersection over Union (IoU). Pixel Accuracy quantifies the ratio

of accurately categorized pixels in the predicted output and is given in Eq. 4.10.

$$\text{Pixel Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4.10}$$

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

To find the overlap between the ground truth and the estimated segmentation masks, the IoU metric has also been utilized. It is given by Eq. 4.11;

$$IoU = \frac{|\hat{y} \cap y|}{|\hat{y} \cup y|}. \tag{4.11}$$

where $\hat{y}$ and y represent the estimate of the segmented mask and the ground truth, respectively.

The metrics used to access the estimation of EF include MAE, RMSE, r-squared (R^2), and Pearson correlation coefficient. The Receiver Operating Characteristic (ROC) curve provides insight into the discriminatory ability of our model across various decision thresholds for the detection of cardiomyopathy. The (ROC) curve is a fundamental tool for evaluating the performance of binary classification models.

4.5.2 Quantitative Analysis

4.5.2.1 EF Estimation

The experimental results for EF estimation obtained from video-based joint learning performed on EchoNet-Dynamic and CAMUS datasets through EFNet are given in Table 4.1.

In an independent test dataset from EchoNet-Dynamic unseen during model training, the EFNet demonstrated an EF prediction with an MAE of 4.35%, an RMSE of 5.83%, and an R^2 value of 0.88 when compared to human expert annotations

TABLE 4.1: EF estimation

Dataset	**MAE***	**RMSE***	R^2	**Corr**
EchoNet-Dynamic	4.35	5.83	0.78	0.88
CAMUS	5.69	6.99	0.73	0.87

*MAE and *RMSE results are presented in percentage (%) values

(ground truth values). These metrics comfortably align with the usual measurement discrepancies seen among different clinicians, known as inter-observer variation, which may reach up to 13.9% [8].

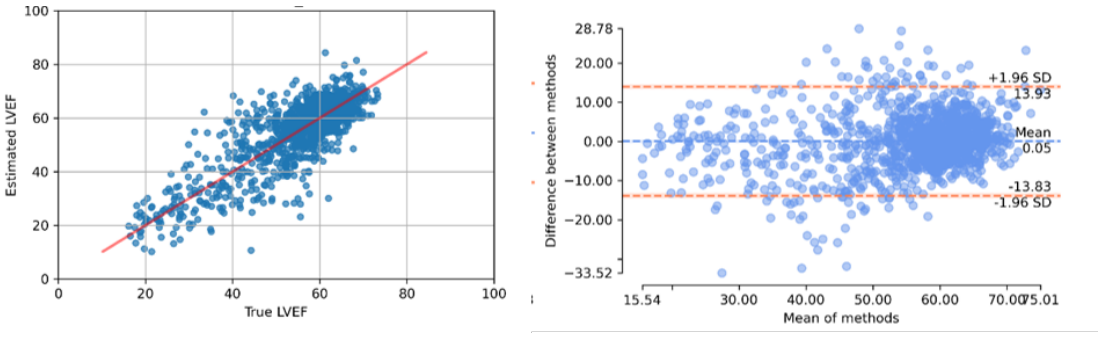


FIGURE 4.6: Bland-Altman and EF correlation plot (EchoNet-Dynamic)

The Bland-Altman plot and correlation graphs illustrating the true and predicted EF for EchoNet-Dynamic are depicted in Fig. 4.6. A comparison between these graphs and those presented in Fig. 3.8 and Fig. 3.9, which display the EF estimations derived from the method introduced in Chapter 3, reveals evident improvements resulting from cross-module training on video data. In the correlation plot shown in Fig. 4.6, the EF estimations exhibit a more uniform distribution around the ground truth as compared with the distribution seen in Fig. 3.9(a). Similarly, for the Bland-Altman plot in Fig. 4.6, the data is comparatively more concentrated within the 95% confidence interval.

The bar graph in Fig. 4.7 depicts the comparison between actual and predicted values categorized into different ranges of EF as suggested by ASE and EACVI and given in Table 1.1. It offers a visual representation of how accurately the model predicts EF across various ranges. The x-axis represents the EF ranges, while the y-axis indicates the frequency of samples falling within each range. The graph

reveals insights into the model's performance, highlighting areas of accurate prediction and those where discrepancies occur. This analysis aids in evaluating the model's effectiveness in capturing the subtleties of EF estimation across different ranges.

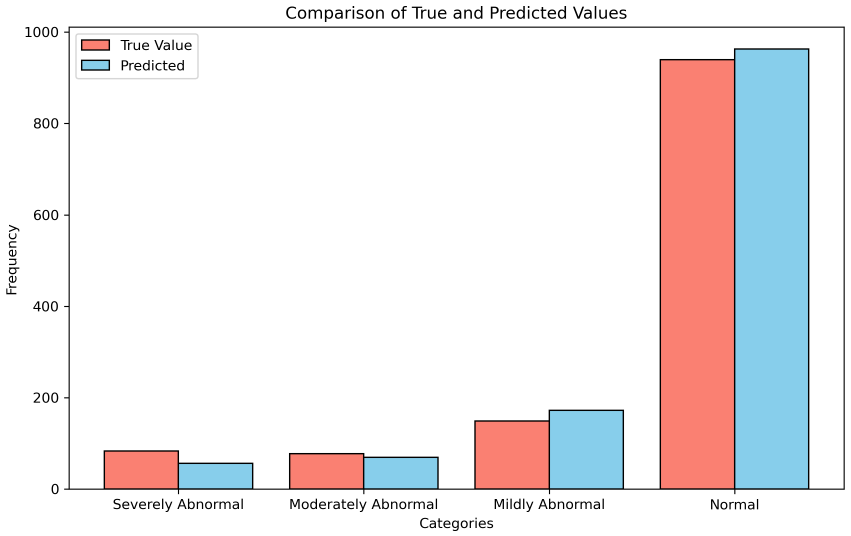


FIGURE 4.7: Comparison of ground truth and predicted EF categorization in different ranges of EF given by [8]

4.5.2.2 LV Segmentation

TABLE 4.2: Segmentation (EchoNet-Dynamic)

DL Model	DSC		IoU		Pixel Accuracy	
	ED	ES	ED	ES	ED	ES
DeepLabv3	0.9254	0.9012	0.8612	0.8202	0.9854	0.9824
DeepLabv3+	0.9190	0.9006	0.8501	0.8192	0.9843	0.9800
FCN	0.9233	0.9041	0.8575	0.8250	0.9859	0.9821
UNet	0.9163	0.8978	0.8455	0.8146	0.9841	0.9795
EFNet	0.9309	0.9135	0.8707	0.8408	0.9878	0.9844

The results for LV segmentation on EchoNet-Dynamic for ES and ED frames are given in Table 4.2. We also compared the results of segmentation obtained from EFNet with existing state-of-the-art segmentation networks which include DeepLabv3, DeepLabv3+, FCN, and UNet. We replicated the model proposed

in [14] for LV segmentation, employing DeepLab with a ResNet50 backbone. We independently implemented their model with our selection of hyperparameters. Notably, our proposed multitask network with cross-module fusion surpassed these state-of-the-art segmentation networks and yielded enhanced results. Across all utilized metrics, EFNet consistently demonstrated improved performance.

The DSC was computed individually for both ES and ED frames, providing insights into the segmentation accuracy at different phases of the cardiac cycle. Aggregating these DSC values allowed for the determination of an overall performance metric as depicted in Fig. 4.8. The histogram in the figure illustrates a predominance of DSC values ranging from 0.8 to 1.0, indicating a consistently high level of segmentation accuracy across the dataset.

Furthermore, the histogram reveals that approximately 80% of the segmentation predictions achieved a DSC greater than 90%, underscoring the model's proficiency in delineating the left ventricular boundaries accurately. A closer examination of the segmentation accuracy between ES and ED frames reveals notable differences, Table 4.3 further confirms this, illustrating that the accuracy for ED frames exceeds that of ES frames.

This discrepancy in segmentation accuracy between ES and ED frames suggests potential challenges or complexities associated with segmenting the LV during different phases of the cardiac cycle. The higher accuracy observed for ED frames may be attributed to factors such as clearer delineation of anatomical structures or reduced motion artifacts compared to ES frames. During end-systole, the LV contracts, leading to more complex motion patterns and deformation of the ventricular walls. This motion can result in blurring or distortion of the LV boundaries, making segmentation more challenging. Similarly, end-systolic frames may have lower image quality compared to end-diastolic frames due to increased motion artifacts or reduced contrast between the LV and surrounding tissues. This poorer image quality can make it harder for segmentation algorithms to accurately delineate the LV boundaries.

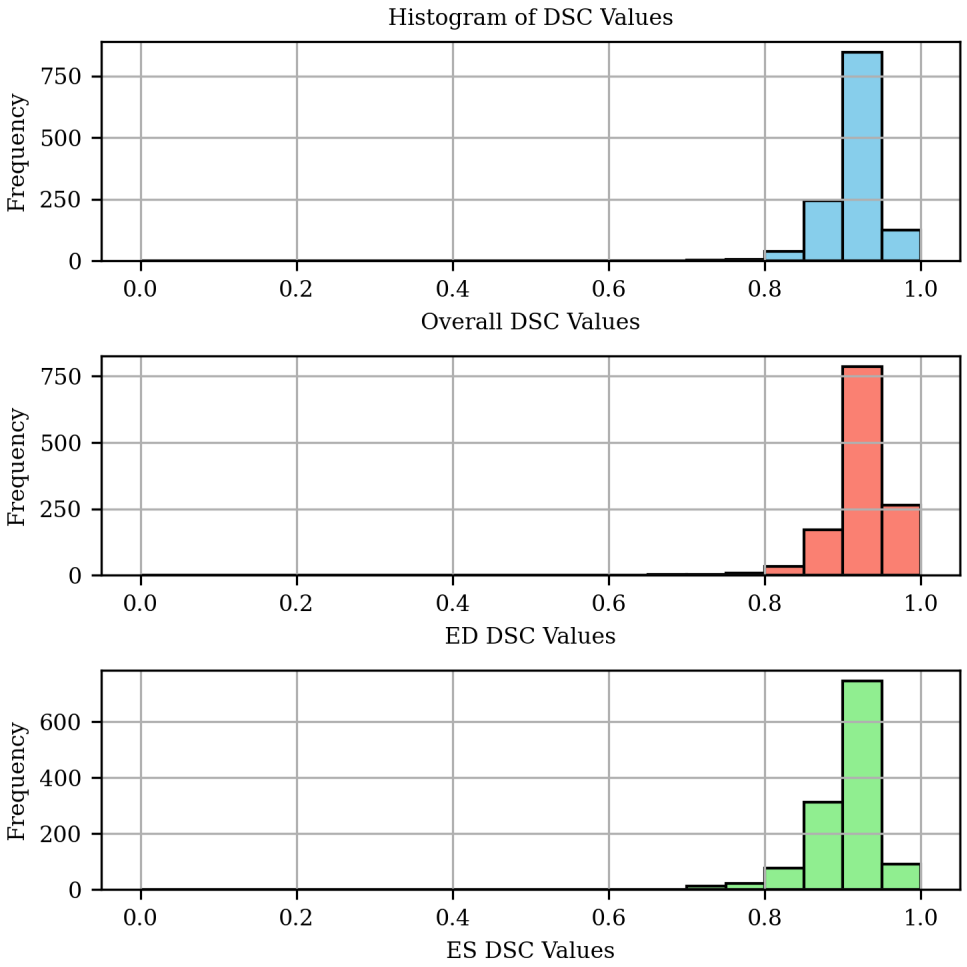


FIGURE 4.8: Histogram showing DSC values obtained for ES and ED frames, along with the overall DSC for EchoNet-Dynamic.

For evaluation on the CAMUS dataset, EFNet pre-trained on a larger dataset i.e. EchoNet-Dynamic was fine-tuned for several epochs. The results for LV segmentation on this dataset are given in Table 4.3.

TABLE 4.3: Segmentation (CAMUS)

DL Model	**DSC**		**IoU**		**Pixel Accuracy**	
	ED	**ES**	**ED**	**ES**	**ED**	**ES**
DeepLabv3	0.9275	0.8866	0.8648	0.7962	0.9881	0.9874
DeepLabv3+	0.9079	0.8700	0.8313	0.7699	0.9843	0.9832
FCN	0.9062	0.8897	0.8285	0.8013	09879	0.9841
UNet	0.9247	0.8968	0.8600	0.8128	0.9885	0.9866
EFNet	0.9366	0.9154	0.8807	0.8440	0.9905	0.9886

Another comparative analysis was conducted between the EF estimation results obtained by EFNet and those derived without cross-module fusion. The same regression model was used in both for comparison. It became evident that integrating cross-modules within EFNet led to a substantial 36.7% enhancement in EF estimation accuracy. Comparison of EFNet results with independent regression and segmentation results are given in Table 4.4.

TABLE 4.4: Comparison of EFNet with segmentation and regression networks trained without joint optimization

Model	**MAE**	**RMSE**	**ED**	**ES**
Segmentation (UNet++)	-	-	0.91	0.89
Regression (R2Plus1D)	6.87	7.55	-	-
EFNet	4.35	5.83	0.93	0.91
Improvement	36.6%	22.8%	2.19%	2.24%

4.5.2.3 Cardiomyopathy Detection

Fig. 4.9 illustrates receiver-operating characteristic curves for diagnosing heart failure with reduced ejection fraction using the test dataset.

In our analysis, we iterated over different threshold values, ranging from 35 to 50, representing the cutoff points at which the model classifies instances as positive or negative for cardiomyopathy. For each threshold value, the True Positive Rate (TPR) and False Positive Rate (FPR) are computed.

The TPR, also known as sensitivity, measures the proportion of actual positive cases correctly classified by the model as positive. Conversely, the FPR represents the proportion of actual negative cases incorrectly classified as positive by the model. By plotting these rates against each other, the ROC curve illustrates the trade-off between sensitivity and specificity (1 - FPR) across different threshold values.

Additionally, the Area Under the Curve (AUC) is calculated for each threshold, providing a single metric to summarize the model's discriminatory performance. The AUC quantifies the probability that the model will rank a randomly chosen

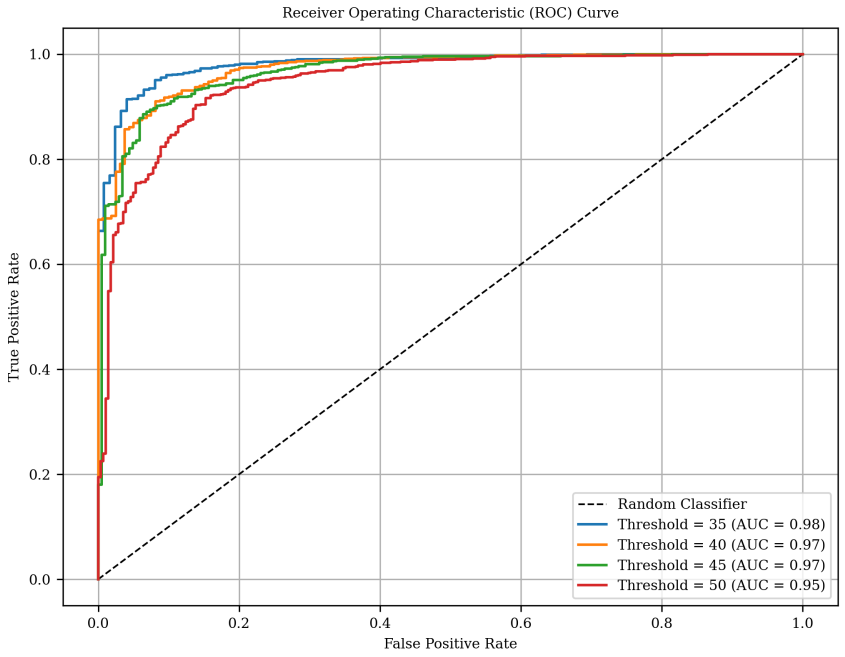


FIGURE 4.9: Receiver operating curve for the diagnosis of cardiomyopathy based on different thresholds for detection boundary.

positive instance higher than a randomly chosen negative instance. Higher AUC values indicate better discriminatory ability of the model.

Each curve on the ROC plot corresponds to a specific threshold value, with the label for each curve indicating the threshold value along with the corresponding AUC score. By analyzing the ROC curve and AUC scores, we gain insights into how well the model distinguishes between positive and negative cases of cardiomyopathy across different decision thresholds.

4.5.3 Qualitative Analysis

Fig. 4.10 shows a few samples of ground truth segmentation masks compared to the predicted segmentation masks obtained when EFNet was trained on EchoNet-Dynamic. Visually, it is evident that the results are of high quality, indicating the effectiveness of the models in accurately predicting the segmentation masks. The close alignment between the ground truth and predicted masks demonstrates the

robustness of EFNet in handling real-world echocardiographic data, showcasing its potential for reliable clinical application.

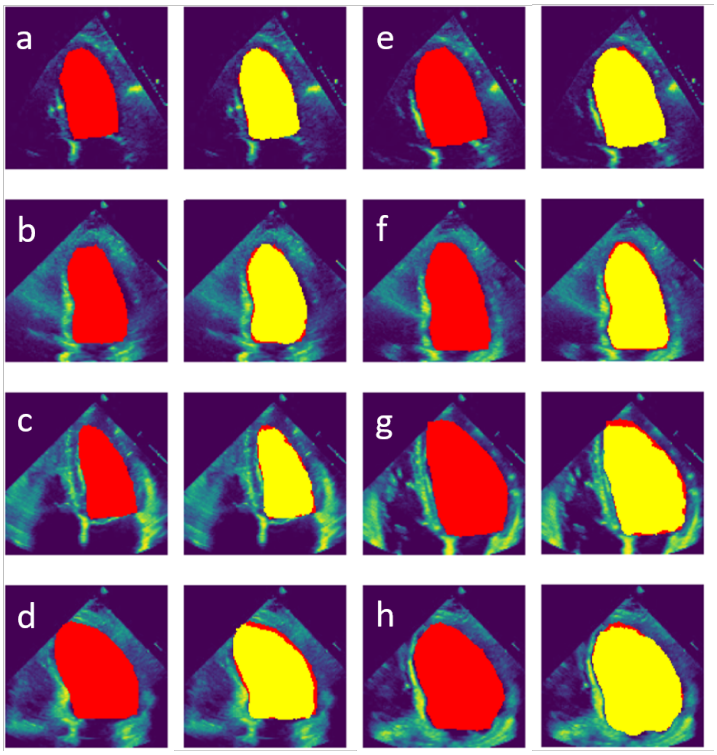


FIGURE 4.10: Predicted segmentation masks from EFNet. The red masks show the ground truth. Yellow masks show the predictions obtained from EFNet. (a-d) ES frames. (e-h) ED frames.

In our analysis, we conducted a qualitative comparison between the segmentation results obtained from EFNet, which incorporates cross-module fusion, and the results obtained when segmentation was carried out as an independent task without fusion with a regression module. The comparative analysis, as illustrated in Fig. 4.11, demonstrates the superior performance achieved when segmentation is integrated into the EFNet framework compared to its independent implementation. The DSC values are also provided with each frame, which further supports our assertion.

Furthermore, there were instances where the ground truth labels had discrepancies due to incorrect labeling, chamber foreshortening, or poor image quality [14]. EFNet exhibited the ability to accurately delineate the LV boundary in these cases too, demonstrating its robustness. A few such cases are shown in Fig. 4.12.

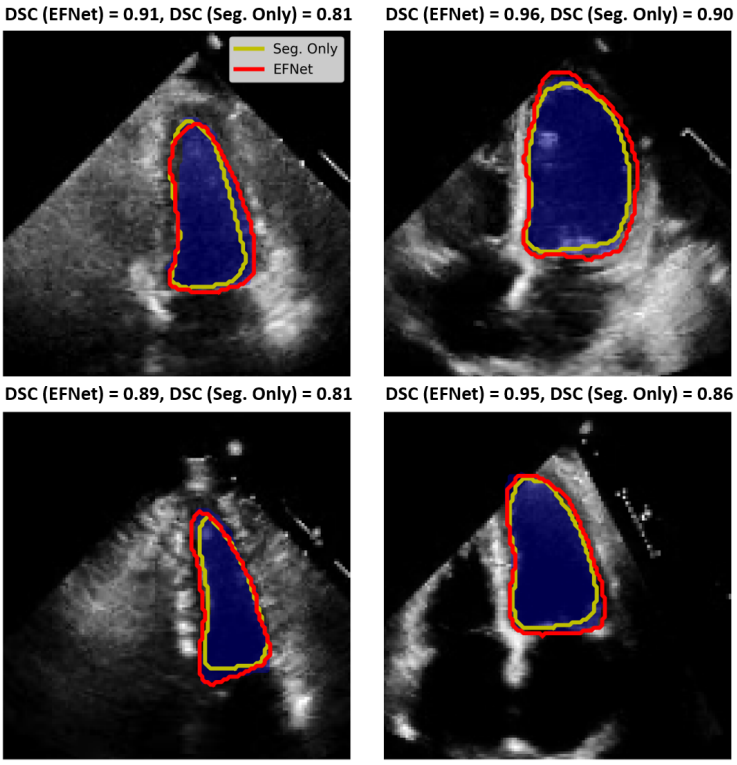


FIGURE 4.11: Illustrative frames with ground truth: EFNet vs. segmentation model predictions without cross-module fusion. Blue masks represent ground truth segmentation, with EFNet's segmentation boundary in red and without multitasking in yellow

4.6 Discussion

Due to inherent limitations in the imaging principle of echocardiography, speckle noise is often prominent, leading to blurred boundaries and the presence of artifacts in the cardiac tissue. These factors make it challenging for physicians to accurately trace the left ventricular endocardium. As a result, the clinical calculation of the left ventricular EF is highly dependent on empirical assumptions, leading to substantial errors and compromising the reliability of the results. Automating the procedure of EF calculation using AI can offer significant advantages in such scenarios. By leveraging AI algorithms, it becomes possible to overcome the limitations of manual tracing and improve the accuracy of EF estimation.

To achieve this objective, a multitask network employing cross-module fusion has been devised in this chapter, enabling simultaneous training of segmentation and regression tasks through joint optimization. By incorporating a cross-module

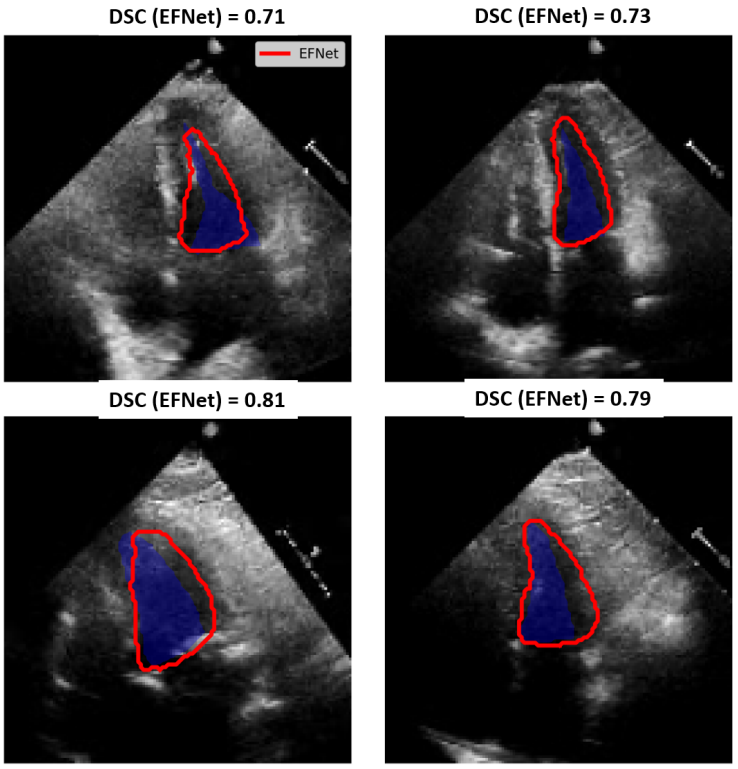


FIGURE 4.12: Illustrative frames with erroneous ground truth: EFNet predictions outperform ground truth. Blue masks show the ground truth segmentation. The segmentation boundary obtained through EFNet is shown in red.

framework, the model ensures compatibility with the clinical workflow while effectively integrating and leveraging the interdependencies between segmentation and regression processes. Moreover, within our proposed model, the prior knowledge obtained from the segmentation task enables the regression model to focus specifically on the delineated candidate areas for feature generation.

The simultaneous training of multiple tasks in multitask learning can sometimes result in a decline in overall performance. Smaller independent networks often outperform the multitask approach in such cases. This decrease in performance can be attributed to various factors. One possible reason is that the tasks being learned may require different rates of learning, making it challenging to find an optimal balance. Additionally, one task may dominate the learning process, leading to limited progress on the other tasks and an imbalance in performance. The gradients of the different tasks can also interfere with each other during training,

further affecting the overall performance. Moreover, combining multiple loss functions can create a more complex optimization problem, posing challenges for the learning process. To address these issues, we have focused on finding appropriate ways to combine the losses and optimize the objective function jointly. Equal weighting is assigned to both losses, ensuring that both tasks are optimized with equal importance. To ensure that the tasks are learned at similar rates, the losses are scaled to the same range. The two tasks at hand here are disjoint but dependent on each other and by leveraging their shared characteristics we were able to optimize the loss function and achieve satisfactory results. The overall results obtained through EFNet demonstrated improved accuracy as compared to previous studies that performed these two tasks independently.

In Tables 4.5 and 4.6, a comparative analysis is conducted for EF estimation using various existing methods on both EchoNet-Dynamic and CAMUS datasets, respectively. The proposed method of jointly optimizing segmentation with a regression module in a cross-module fusion model outperformed the estimates by other studies and yielded significantly improved results.

TABLE 4.5: Performance of EFNet against existing methods for EF estimation (EchoNet-Dynamic)

Model	**MAE**	**RMSE**	**Corr**
R3D [33]	5.44	6.16	-
MC3 [33]	5.91	6.80	-
R2+1D [33]	6.87	7.55	-
LSTM [75]	8.08	11.98	0.348
DL Based Workflow [76]	6.5	-	0.76
UVT (R*) [79]	6.76	8.70	0.48
UVT (M*) [79]	5.95	8.38	0.52
UltraSwin-small [80]	5.72	7.63	0.58
UltraSwin-base [80]	5.59	7.59	0.59
EchoGNN [81]	4.45	0.76	-
DL, LSTM [16]	5.73	7.72	0.78
MAEF-Net [50]	6.29	-	-
Proposed Method	**4.35**	**5.83**	**0.879**

R3D: 3D ResNet, MC3: Mixed Convolutional Networks, R2+1D: ResNet 2+1D, LSTM: Long Short-Term Memory, UVT: Ultrasound Vision Transformer, *R indicates random sampling, *M indicates mirror sampling of video sequences.

TABLE 4.6: Performance of EFNet against existing methods for EF estimation (CAMUS)

Model	**MAE**	**RMSE**	**Corr**
SRF [15]	12.8	-	0.465
BEASM-fully [15]	10.7	-	0.731
BEASM-semi [15]	10.0	-	0.790
UNet [15]	5.6	-	0.791
ACNN [15]	5.7	-	0.799
SHG [15]	5.7	-	0.770
UNet++ [15]	5.6	-	0.789
Automated EF [82]	6.7	-	-
Proposed Method	**5.54**	**8.02**	**0.822**

SRF: Structured Random Forest, BEASM: B spline Explicit Active Surface Model, ACNN: Anatomically Constrained Neural Networks, SHG: Stacked Hourglass

In this study, EF estimation played a crucial role in the classification of cardiomyopathy based on different thresholds. By applying various thresholds to the EF values obtained from echocardiographic data, we were able to classify patients into different categories, ranging from severely abnormal to normal cardiac function. This classification was instrumental in identifying individuals with cardiomyopathy, allowing for early intervention and appropriate management strategies.

However, it's essential to acknowledge that while this study presents a method for diagnosing cardiomyopathy based on specific criteria, the final decision regarding the classification of cardiomyopathy is often dependent on the expertise and judgment of cardiologists. Cardiologists consider a multitude of factors, including but not limited to ejection fraction, symptoms, medical history, and additional diagnostic tests, to make an accurate diagnosis and classification. Therefore, while the criteria presented in this study may aid in the diagnostic process, they should be interpreted in conjunction with clinical expertise and patient-specific information for accurate classification and treatment planning.

4.6.1 Time-Space Complexity Analysis

Table 4.7 gives the time-space complexity of EFNet and compares it with other semantic segmentation models. EFNet exhibits the highest number of parameters

(80.28 million) among all models listed in the table. Despite its complexity, EFNet maintains a reasonable model size of 374.48 MB, indicating efficient parameter utilization and potentially robust feature representation.

TABLE 4.7: Time-Space complexity analysis.

	EFNet	**DLv3**	**DLv3+**	**FCN**	**UNet**
Parameters (m)↓	80.28	73.30	57.98	85.62	63.82
Model Size↓	374.48	321.13	204.21	415.60	248.78
GFLOPs↓	131.01	123.95	148.12	130.51	155.57
MACs↓	65.41	61.88	73.91	65.15	77.63
FPS↑	17.32	15.23	17.08	13.31	16.42
Time (sec)↓	0.0578	0.0657	0.0585	0.0752	0.0609
Memory (GB)↓	0.95	0.75	1.37	0.78	1.32
DSC	0.9309	0.9254	0.9190	0.9233	0.9163

DLv3 - DeepLabv3. DLv3+ - DeepLabv3+. GFLOPs - Giga-floating-point operations. MACs - Multiply-accumulate operations. FPS - Frames per second. m - millions. Time and Memory per prediction are given.

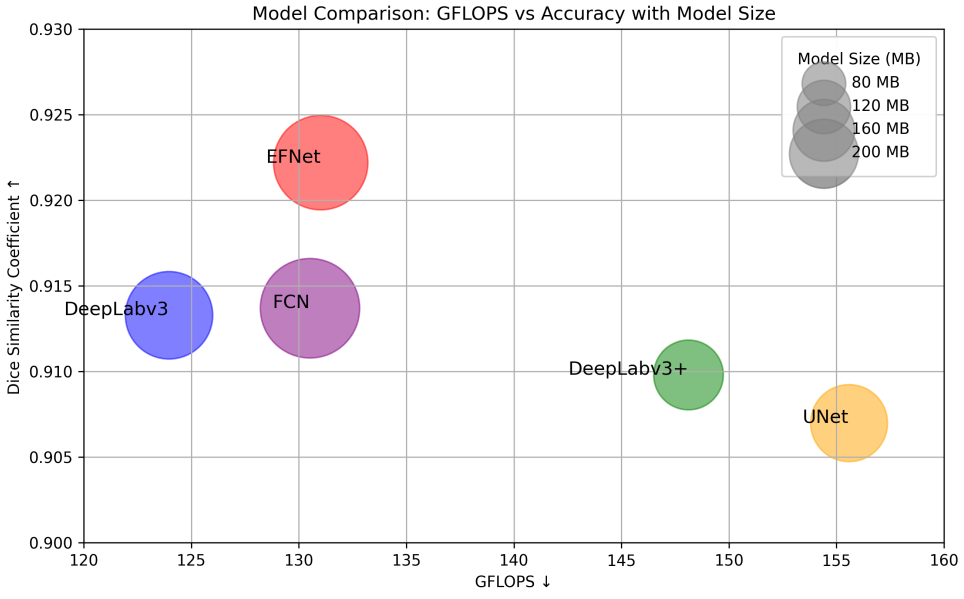


FIGURE 4.13: Trade-off between accuracy and computational efficiency based on GLOPs

When considering computational efficiency, EFNet demonstrates a lower or a comparable number of floating-point operations per second (GFLOPs) at 131.01 and multiply-accumulate operations (MACs) at 65.41 as compared to DeepLabv3+, FCN and UNet. DeepLabv3, on the other hand, has the lowest values of GFLOPs

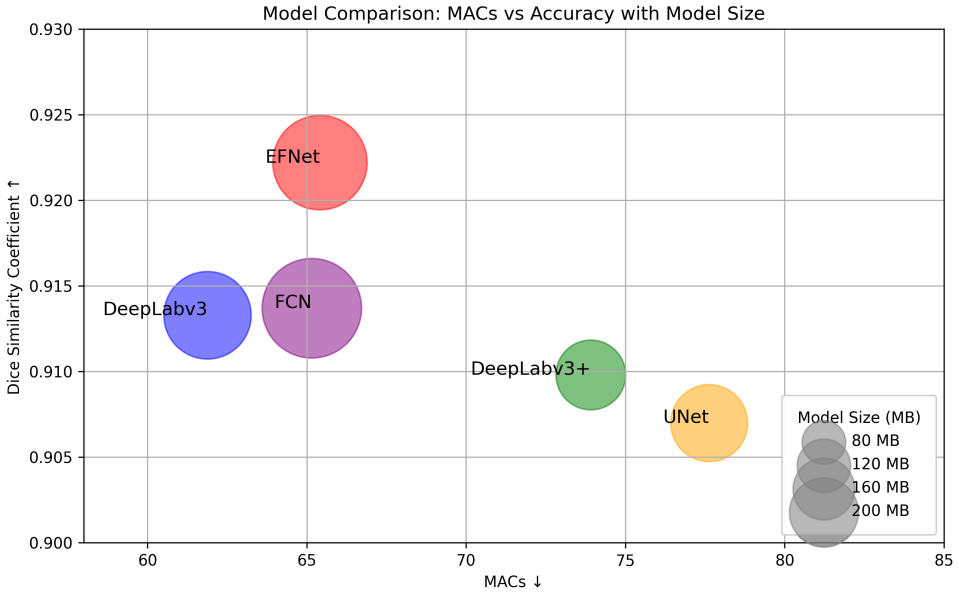


FIGURE 4.14: Trade-off between accuracy and computational efficiency based on MACs

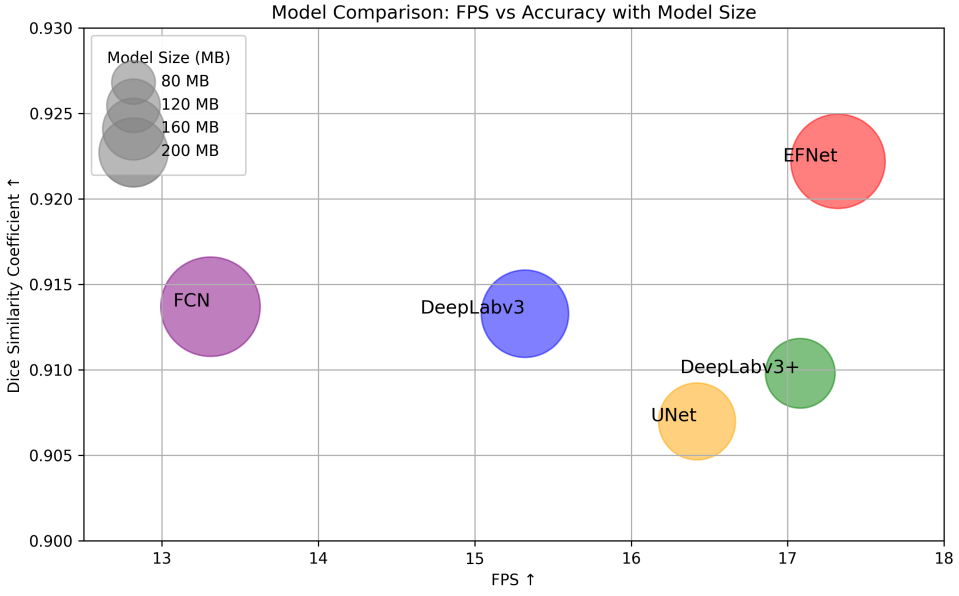


FIGURE 4.15: Trade-off between accuracy and processing speed measured in FPS

and MACs. In terms of inference speed, EFNet achieves a competitive Frames per Second (FPS) rate of 17.32 and a relatively low time per prediction of 0.0578 seconds. These values indicate EFNet's ability to process predictions swiftly, contributing to efficient real-time performance. Although EFNet requires slightly

more memory per prediction (0.95 GB) compared to some models, it remains within reasonable limits, ensuring compatibility with various hardware configurations.

The figures 4.13, 4.14 and 4.15 illustrate the efficiency of EFNet by comparing its DSC performance against MACs, GFLOPs, and FPS, respectively. The size of the circles in the plot represents the model sizes. From the plots, it is evident that EFNet achieves the highest accuracy among all models while maintaining a balanced tradeoff in terms of speed and computational efficiency.

4.7 Summary

In this study, the proposed model; EFNet, enables simultaneous quantification of cardiac chamber structure and function, resulting in a workflow consistent with the clinical evaluation of EF. It automates the procedure of volume tracing to obtain EF, which reduces the burden of human labeling. The evaluation in this work relies on a single imaging plane i.e. the A4C view whereas, in clinical practice, both A4C and A2C views are utilized for EF calculation. However, the process of manually tracing the endocardial border during various phases of the cardiac cycle can be demanding and time-consuming, especially in challenging situations characterized by inadequate image quality or elevated heart rates. Hence, the method proposed in this chapter offers an alternative by providing acceptable accuracy even with measurements from a single imaging plane. This simplifies analysis, making it a practical solution for clinical applications.

Chapter 5

Improved Quantification of LV Structure Through a Decoupled Edge Guided Module

Building upon the work laid out in the previous chapter, the significance of precise LV delineation in enhancing LV quantification remains paramount. This prompted an investigation into methods aimed at refining LV segmentation. This chapter delves deeper into improving LV segmentation techniques, leveraging insights gained from the joint EF estimation and LV segmentation performed previously.

Various segmentation algorithms utilizing DL techniques have been developed for the quantification of the LV structure. These algorithms focus on pixel classification within the object's body, emphasizing high-level features while overlooking low-level details such as edges and boundaries of the objects in context, leading to less precise detection of LV borders. Moreover, this task remains challenging due to the low signal-to-noise ratio, unclear borders, and organ variability in echocardiogram data. Precise delineation of the LV is essential in carrying out an accurate diagnosis of various cardiac conditions, highlighting the need for advanced techniques to improve segmentation accuracy.

This chapter introduces a multitask network designed to improve the quality and accuracy of LV delineation by including boundary information. The network employs a common encoder for shared feature extraction from echocardiogram data. It utilizes separate decoder modules for semantic segmentation and edge prediction, each with its individual cost function, combined to perform joint optimization within the network. Our proposed method exhibits enhanced accuracy across multiple metrics compared to existing state-of-the-art semantic segmentation models that do not include edge prediction. This improvement demonstrates the effectiveness of our approach in overcoming the challenges associated with LV delineation.

5.1 Limitations of Semantic Segmentation

Recently, there has been considerable work done in the field of semantic segmentation, specifically on the use of encoder-decoder-based neural networks to predict labels for each pixel within an input image. Semantic segmentation involves identifying specific image class pixels and isolating them from other image classes by applying a segmentation mask overlay. Employing classification architectures for pixel-level categorization has several limitations [83–85]. An encoder typically comprises a backbone network like VGG [86], ResNet [87], and MobileNet [88], among others. A widely used encoder-decoder architecture, UNet [28] is known for its efficiency in producing semantic segmentations. UNet was originally introduced for biomedical image segmentation by Olaf Ronneberger et al. [28]. In the UNet architecture, the encoder generates low-level features from the input data by convolving it through layers of filters. The downsampled feature maps are upsampled by the decoders to the original size of the input data. These reconstructed features provide pixel-level labels that provide semantic segmentation. However, during the process, important spatial information that is necessary for segmentation is lost. Fully Convolutional Networks (FCNs) are also considered effective for semantic segmentation; however, downsampling and upsampling operations in FCNs may also lead to the loss of fine-grained details, which might cause FCNs to

struggle to precisely delineate object boundaries. This can result in fuzzy or imprecise object boundaries in the segmentation output. These existing methods with reduced spatial resolution emphasize solely low-level semantic attributes like color, shape, and texture within a single deep architecture. As a result, these methods may lack understanding of attributes extending beyond pixel-level classification, particularly key low-level information like boundary details.

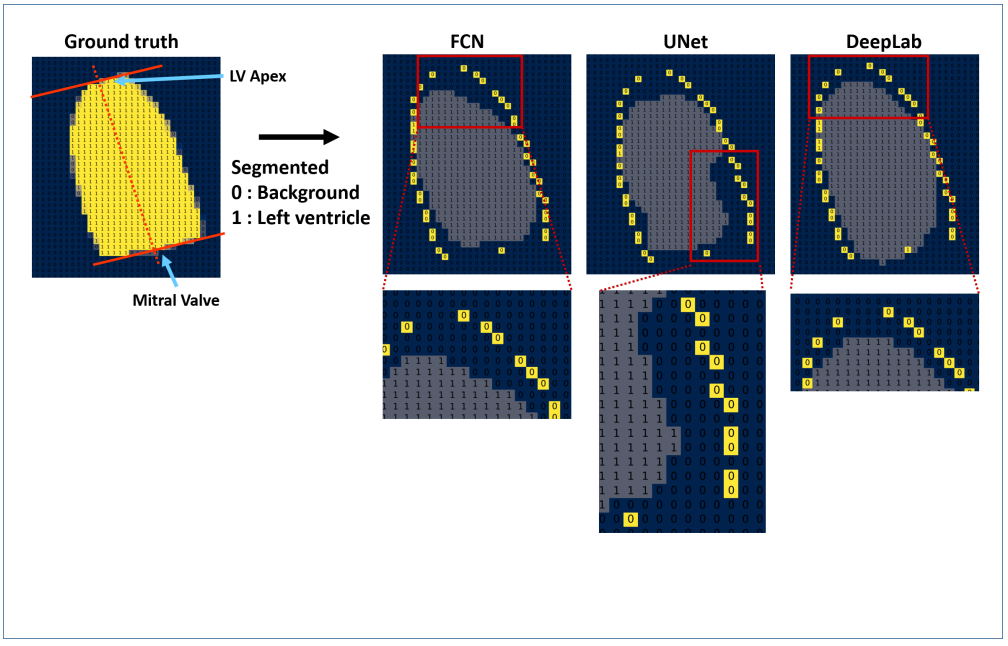


FIGURE 5.1: LV segmentation outcomes derived from SOTA semantic segmentation models. The images display pixel values, where 0s signify detected background and 1s represent detected LV. The yellow background pixels within the segmented images indicate the ground truth boundary. The enlarged regions highlight some of the undetected LV areas.

In LV delineation, it is important to establish distinct demarcations between adjacent structures by minimizing pixel ambiguity near the LV border. The inclusion of particular boundary points is important, particularly the LV apex and the mitral valve annulus [8]. Moreover, echocardiogram data encounter issues like low signal-to-noise ratio, noise, low contrast, and organ variation. Ensuring a precise LV boundary is crucial for cardiologists to derive accurate clinical insights [36, 89, 90]. Illustrated in Fig. 5.1 are outcomes of LV segmentation obtained from state-of-the-art (SOTA) semantic segmentation models. These images depict pixel values, where 0s indicate the detected background (blue-colored region)

and 1s signify the detected LV (gray-colored region). The yellow background within the segmented images represents the ground truth border. Observing the segmented images reveals that a considerable portion of border pixels, including adjacent areas, primarily register as 0s, indicating undetected segments within the LV region. The image size has been reduced to 50x50 to enhance pixel-level visibility.

5.2 Decoupled Mask and Edge Processing Techniques

The precision and quality of LV delineation can be enhanced by integrating boundary information through the decoupled mask and edge processing techniques. By decoupled processing, we mean separately carrying out the prediction of object masks (semantic segmentation) and the prediction of object boundaries (edge prediction) and fusing them in a way that enhances the performance of each other to obtain better accuracy. In contrast to semantic segmentation, which offers high-level details, the edge prediction component provides low-level information. Decoupling facilitates the development and integration of dedicated modules specifically designed for edge processing. This approach extends beyond conventional segmentation networks, enabling more flexibility and diversity in network design. Separate learning objectives and loss functions for edges and masks might also reduce overfitting as the model learns distinct features for different tasks.

In order to achieve these objectives, a multitask DL model featuring a common encoder for shared feature extraction from input data is introduced. This model also includes two distinct modules; the Mask Generation Decoder for mask segmentation and the Edge Predictor for boundary prediction. By incorporating edge supervision from the Edge Predictor the network's ability to preserve spatial boundary details is significantly improved, resulting in enhanced semantic segmentation performance. The multitask model optimizes through joint training by combining losses from both the Mask Generation Decoder and Edge Predictor

Thorough research on the structure of the Edge Predictor led to the proposal of the one that exhibited the best performance in the regression of edge coordinates.

5.3 Proposed Decoupled Edge Guided Module

We introduce a multitask model aimed at: i) conducting semantic segmentation to derive LV masks, and ii) performing regression for edge prediction. By leveraging both tasks collaboratively, our approach optimizes the multitask model to enhance segmentation accuracy, exploiting the combined information from these tasks. A common encoder is created for the two tasks in order to extract common features and share parameters between them. The inputs to the encoder are the ES and ED echocardiogram frames. The encoder output is used to perform semantic segmentation through the Mask Generation Decoder and boundary estimation through the Edge Predictor. The losses from both heads are combined to generate one combined loss to update the weights of the model during training. The proposed model is illustrated in detail in Fig. 5.2.

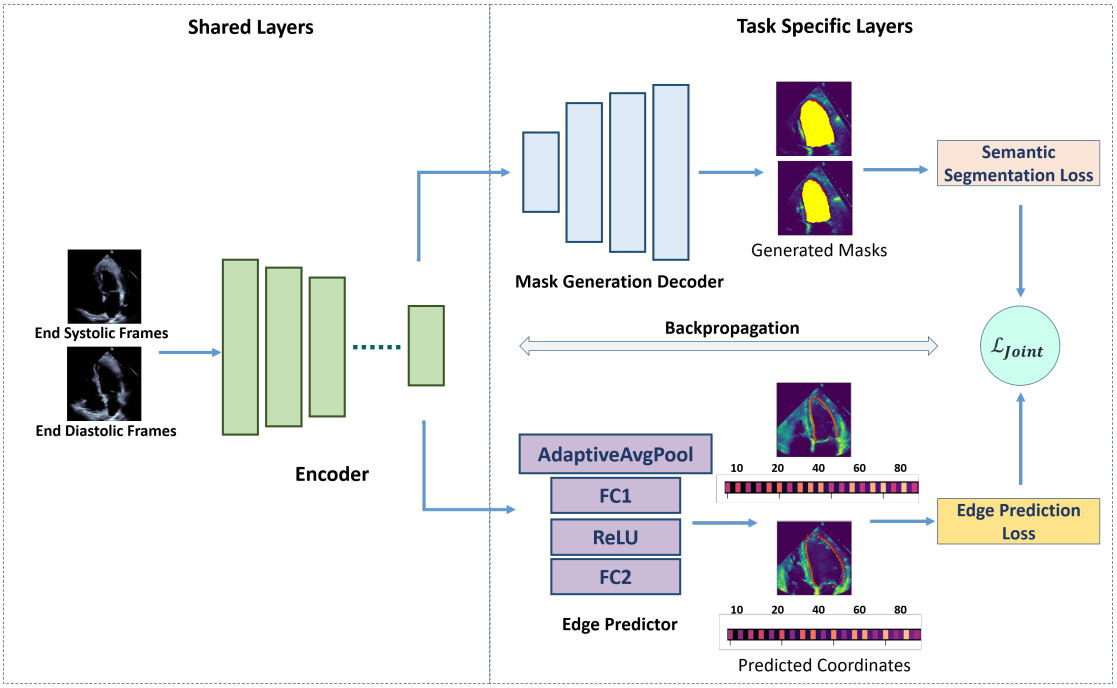


FIGURE 5.2: The proposed encoder-decoder based model. Input comprises ES and ED frames from echocardiogram data. Mask Generation Decoder produces semantic segmentation masks, Edge Predictor produces coordinates of edges.

The model architecture adopts the UNet framework combined with the ResNet34 backbone, which is integrated into the encoder to enhance feature extraction capabilities. The U-Net architecture consists of two main parts: the contracting path (encoder) and the expansive path (decoder). The contracting path extracts features through convolutional layers while reducing spatial dimensions. The expansive path gradually upsamples the features to produce a segmentation map. The ResNet34 architecture employs residual blocks consisting of skip connections to mitigate vanishing gradient problems and facilitate deeper network training. The architecture of the proposed model is illustrated in Fig. 5.3.

The encoder based on the ResNet34 backbone comprises five convolutional blocks (ConvBlock^k, k is the layer number). Each ConvBlock comprises M feature maps. The feature maps extracted from each layer of a ConvBlock are created according to the equation given by Eq. (5.1);

$$y_i^k = f(b_i^k + W_i^k * x) \text{ where } i = 1, \ldots, M. \tag{5.1}$$

Where W_i^k is the weight, b^k is the bias, and f is a non-linear activation function.

5.3.1 Mask Generation Decoder

The Mask Generation Decoder mirrors the encoder's architecture but performs upsampling operations to recover the spatial resolution by utilizing skip connections from the encoder to maintain fine-grained details. The outputs from the decoder are the segmentation masks for both ES and ED frames. For the mask-based segmentation head, BCE loss is used, which measures the dissimilarity between the predicted probabilities and the true binary labels. It is mathematically defined in Eq. (5.2) as;

$$\mathcal{L}_{mask} = -\left(y \log(\rho) + (1 - y) \log(1 - \rho)\right). \tag{5.2}$$

In the equation, y represents the true binary label (0 or 1), ρ denotes the predicted probability of the positive class, and log denotes the natural logarithm.

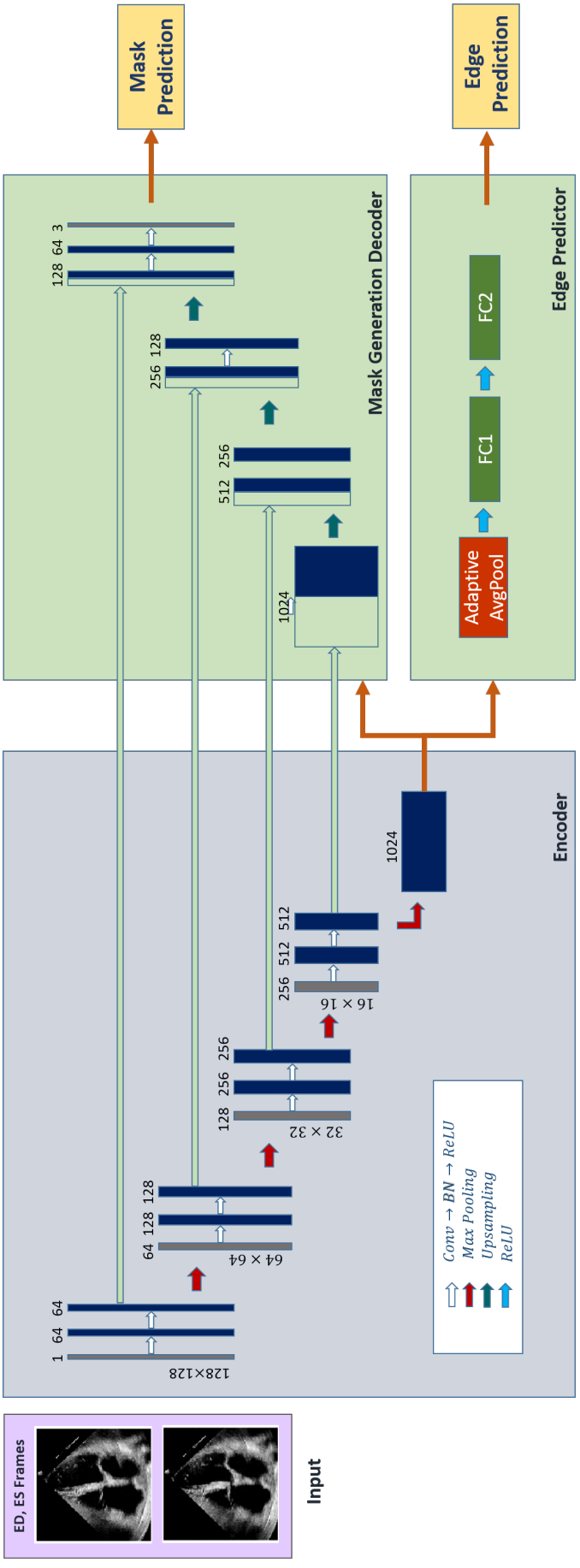


FIGURE 5.3: Architecture for joint training of Mask Generation Decoder and Edge Predictor. A UNet based common encoder extracts features from the input frames. Mask Generation Decoder upsamples extracted features to produce estimated segmentation masks. Edge Predictor performs regression on the extracted features to provide estimated edge coordinates.

5.3.2 Edge Predictior

The data for regression comprises coordinate pairs representing the edge points of the LV boundary. The coordinate pairs are transposed to convert them into a linear array, as shown in Fig. 5.4. In order to perform regression on these points,

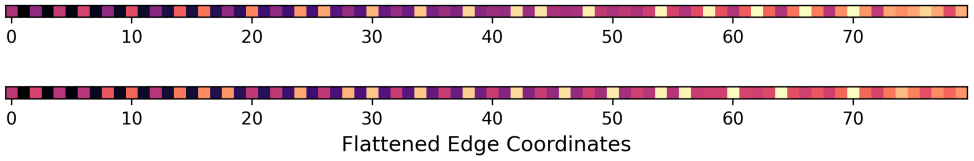


FIGURE 5.4: Range of ground truth values of boundary points for regression of edge prediction for ES and ED frames, respectively.

a regression head is designed. The regression head comprises fully connected (FC) layers designed to transform the encoded features into predictions for the regression task. The feature maps from the encoder of UNet are flattened out using the Adaptive Average Pooling layer to reduce the spatial dimensions of the input feature maps to a predefined output size of 1x1. This is followed by connecting two FC layers. The first FC layer has 512 input features and 256 output features. The last FC layer has the number of linear activations equal to the number of dimensions of the target space, allowing it to produce continuous values for regression tasks. The two FC layers create a deeper network structure. The first FC layer (512 to 256) extracts and transforms the features from a high-dimensional space to an intermediate space, reducing the dimensionality. The second FC layer (256 to 80) further processes these intermediate features to generate the final output. Using two FC layers allows for hierarchical feature extraction, potentially enabling the network to capture more intricate and abstract relationships between the input and output spaces. The outputs from the regression head are predicted continuous values for the coordinates of the edge of the LV.

For the Edge Predictor, the MSE loss function is used in the regression head. This loss function measures the average squared difference between the predicted values ($\hat{x}$) and the true values (x). The MSE loss is calculated by taking the mean of the

squared differences across the dataset, as given in Eq. (5.3);

$$\mathcal{L}_{edge} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2. \tag{5.3}$$

In the equation, N represents the total number of samples in the dataset, and the summation is performed over all the samples.

5.3.3 Joint Loss Function

During training, a combined loss function is formulated to jointly train both the segmentation and regression components of the model. The joint loss function is formed by a combination of losses from both the segmentation and regression tasks, given by Eq. (5.4);

$$\mathcal{L}_{Joint} = \mathcal{L}_{mask} + \mathcal{L}_{edge}. \tag{5.4}$$

The joint loss function, combining both segmentation and regression losses, is backpropagated through the network during training. The model's parameters are updated to minimize the overall loss, ensuring that both the segmentation and regression components are optimized simultaneously.

5.4 Results

5.4.1 Evaluation Metrics

To have a comprehensive evaluation of the segmentation model's performance, the metrics used include the DSC, IoU, F1-score, F2-score, Accuracy, and Recall.

To assess the accuracy of binary segmentation masks, the DSC is used. The formula for calculating the DSC is provided in Eq. (3.4), as previously discussed. In addition to DSC, Accuracy, and Recall are also used to evaluate the segmentation

performance. Accuracy evaluates the proportion of correctly classified pixels to the total number of pixels in the image and is given previously in Eq. (4.10). In the context of binary segmentation, Recall or Sensitivity measures the proportion of true positive predictions of the target class out of all actual positive instances present in the ground truth data, given by Eq. (5.5).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{5.5}$$

Here, TP represents the number of correctly predicted positive instances. FN represents the number of positive instances that were incorrectly predicted as negative and FP represents the number of negative instances that were incorrectly predicted as positive.

F1 score, taken as the harmonic mean of precision and recall, balancing both measures; is given by Eq. (5.6);

$$\text{F1 score} = \frac{\text{TP}}{\text{TP} + 0.5(\text{FP} + \text{FN})}. \tag{5.6}$$

F2 measure is obtained by using weighted mean, given in simplified form in Eq. (5.7). The F2 score places more emphasis on Recall as compared to the F1 score.

$$\text{F2 score} = \frac{TP}{\text{TP} + 0.2\text{FP} + 0.8\text{FP}}. \tag{5.7}$$

To evaluate the accuracy of edge prediction, the HD and MAE between ground truth and the predicted edge coordinates are used.

5.4.2 Quantitative Results

The outcomes for both semantic segmentation and edge prediction are detailed in Table 5.1. The results are given for both the ES and ED frames. The first two rows depict the outcomes of mask prediction and edge prediction obtained using our proposed multitask network, which integrates a decoupled edge detection module. In contrast, the last two rows present the results of semantic segmentation without

the inclusion of the edge detection module. The comparison highlights that the incorporation of edge information improves segmentation accuracy.

TABLE 5.1: Mask segmentation and edge prediction.

	Mask Prediction	**Edge Prediction**	
	DSC	**HD (mm)**	**MAE (mm)**
Diastolic	0.910	6.33	2.05
Systolic	0.930	5.88	1.72
Diastolic (w/o Edge)	0.894	–	–
Systolic (w/o Edge)	0.912	–	–

For comparison with various SOTA segmentation models, we utilized well-known segmentation networks: FCN with a ResNet50 backbone, UNet with ResNet34 and ResNet50 backbones, and UNet++ with a ResNet50 backbone. We have also replicated the model proposed in [14] for LV segmentation, employing DeepLabv3 with a ResNet50 backbone.

TABLE 5.2: Comparison of the proposed model with SOTA semantic segmentation models.

Models	**DSC**	**IoU**	**F1**	**F2**	**Accuracy**	**Recall**
UNet_34	0.905	0.818	0.900	0.892	0.982	0.886
UNet_50	0.903	0.818	0.900	0.893	0.983	0.889
UNet++_50	0.909	0.826	0.905	0.896	0.983	0.890
FCN	0.910	0.813	0.897	0.871	0.981	0.854
DeepLabv3	0.914	0.823	0.903	0.882	0.982	0.869
Proposed	0.920	0.834	0.910	0.930	0.984	0.942

We independently implemented their model with our selection of hyperparameters. Notably, our proposed network, which integrates a decoupled edge prediction module, surpassed these SOTA networks and yielded enhanced results. The evaluation encompassed multiple metrics outlined in section 5.4.1. The comprehensive outcomes for an aggregate of ES and ED frames are summarized in Table 5.2. Across all utilized metrics, our proposed model consistently demonstrated improved performance. The comparison is also illustrated in Fig. 5.5.

The validation loss curve, shown in Fig. 5.6, illustrates the trend of the model's loss function on a validation dataset across multiple epochs during the training

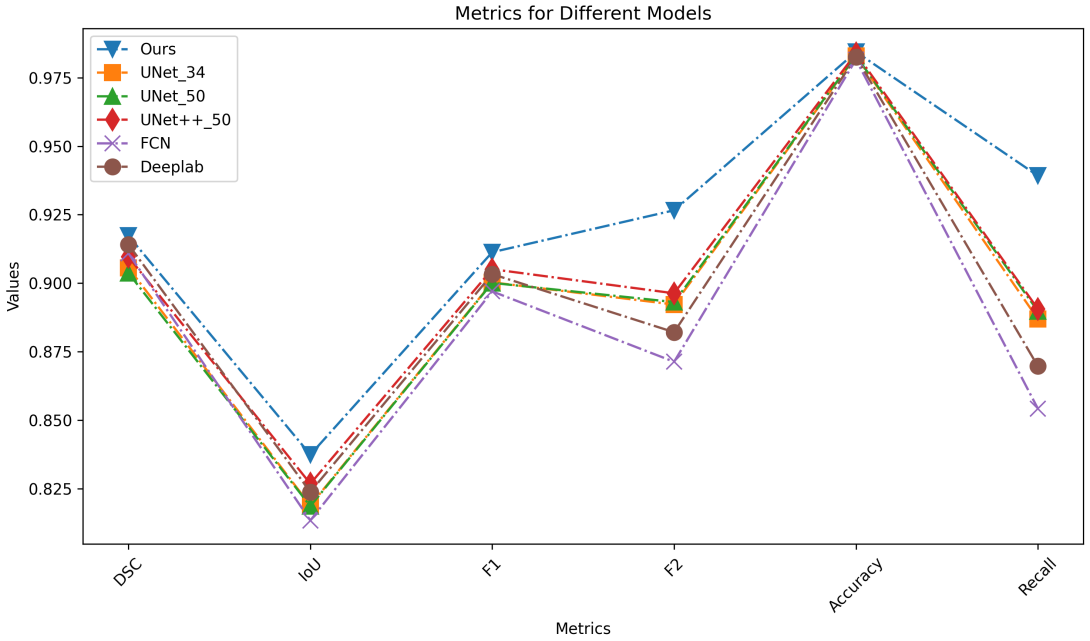


FIGURE 5.5: Comparison of the proposed model results with SOTA segmentation models.

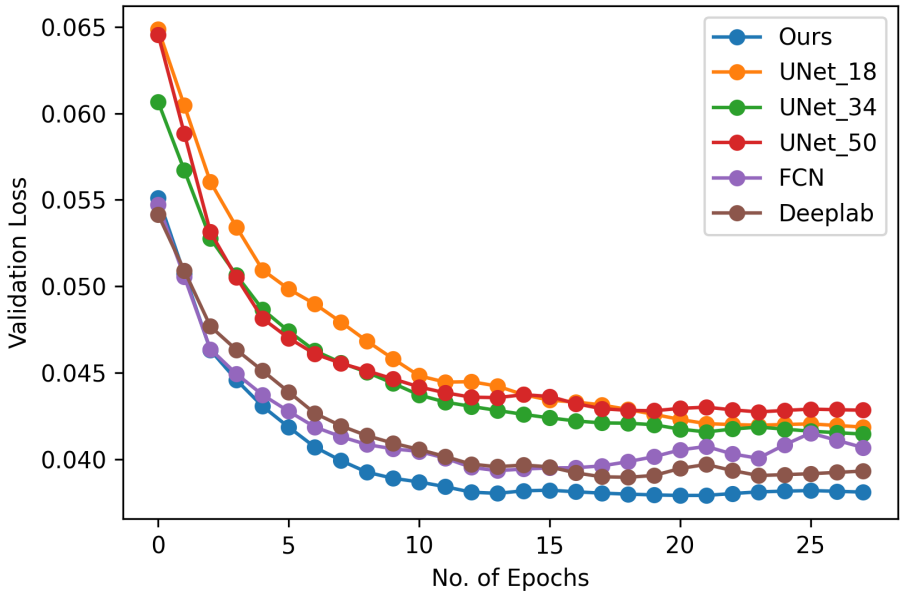


FIGURE 5.6: Validation loss curve for the proposed model and other SOTA segmentation models.

process. This curve shows the proposed model's superior performance compared to that of other SOTA models.

Detailed results for both the ES and ED frames are shown in Table 5.3.

TABLE 5.3: Comparison of LV Segmentation results for ES and ED frames.

Models	DSC		IoU		F1		F2		Accuracy		Recall	
	ED	**ES**	**ED**	**ES**	**ED**	**ES**	**ED**	**ES**	**ED**	**ES**	**ED**	**ES**
UNet_34	0.894	0.912	0.801	0.836	0.889	0.911	0.891	0.893	0.985	0.980	0.892	0.881
UNet_50	0.892	0.911	0.798	0.839	0.887	0.912	0.889	0.896	0.984	0.981	0.892	0.887
UNet++_50	0.896	0.918	0.807	0.846	0.893	0.916	0.890	0.902	0.985	0.981	0.887	0.893
FCN	0.894	0.921	0.795	0.831	0.886	0.907	0.865	0.877	0.984	0.979	0.848	0.859
DeepLabv3	0.897	0.924	0.809	0.838	0.894	0.911	0.880	0.883	0.985	0.980	0.871	0.868
Proposed	0.910	0.930	0.808	0.861	0.894	0.925	0.921	0.934	0.985	0.983	0.942	0.942

5.4.3 Qualitative Results

To offer a qualitative analysis, we've shown segmentation results obtained on a few samples of both ES and ED frames in Fig. 5.7 and Fig. 5.8. The ground truth segmentation is depicted in red, while the segmentation results from the proposed model are shown in yellow. The yellow segmentation overlays the red ground truth to facilitate comparison. We have also presented outcomes derived from various SOTA models on these frames for comparison. The visual depictions indicate that models lacking boundary information struggle to achieve accurate segmentation near the edges. In contrast, our proposed approach offers precise boundary delineation, as the visual comparisons show. These representations highlight the superior performance of the proposed method compared to alternative approaches.

The edge prediction results obtained from these frames are also shown in Fig. 5.9. In the figure, the positions of the ground truth coordinates are marked by red dots, while the yellow marks show the positions of the predicted coordinates. The visual results illustrate the close alignment between the predicted coordinate values and the ground truth coordinates.

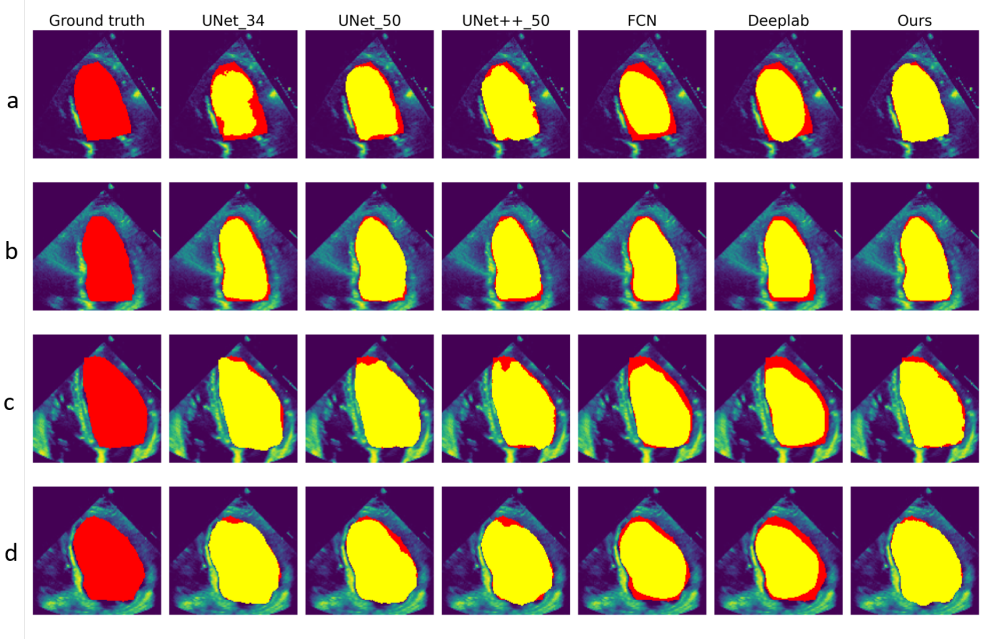


FIGURE 5.7: Visual depiction illustrating the qualitative comparison between our proposed model and other SOTA segmentation models for ED frames. Rows (a-d) represent the evaluation obtained on different input samples.

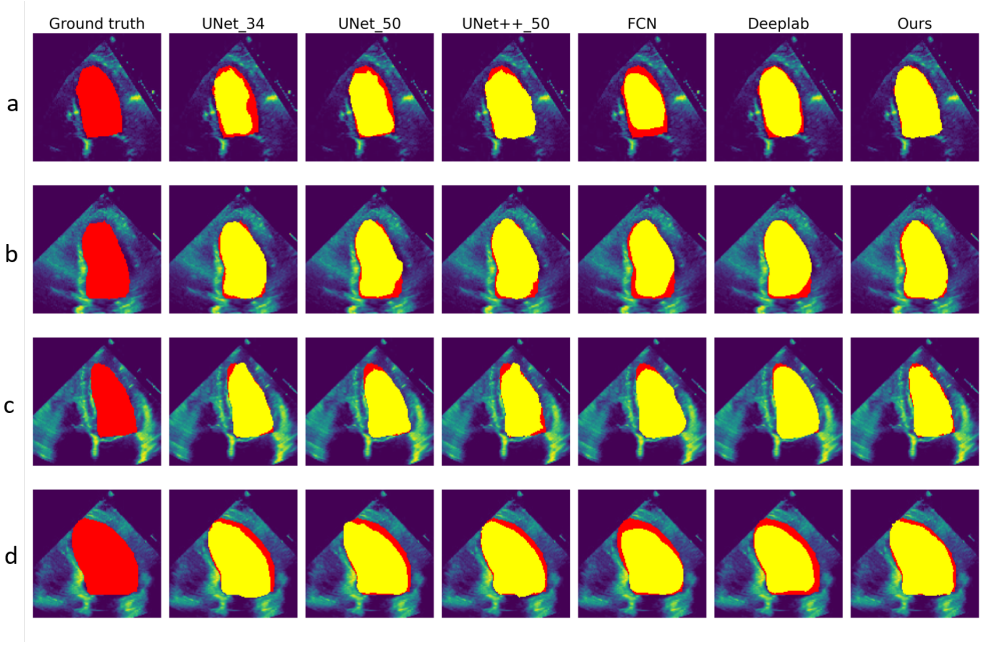


FIGURE 5.8: Visual depiction illustrating the qualitative comparison between our proposed model and other SOTA segmentation models for ES frames. Rows (a-d) represent the evaluation obtained on different input samples.

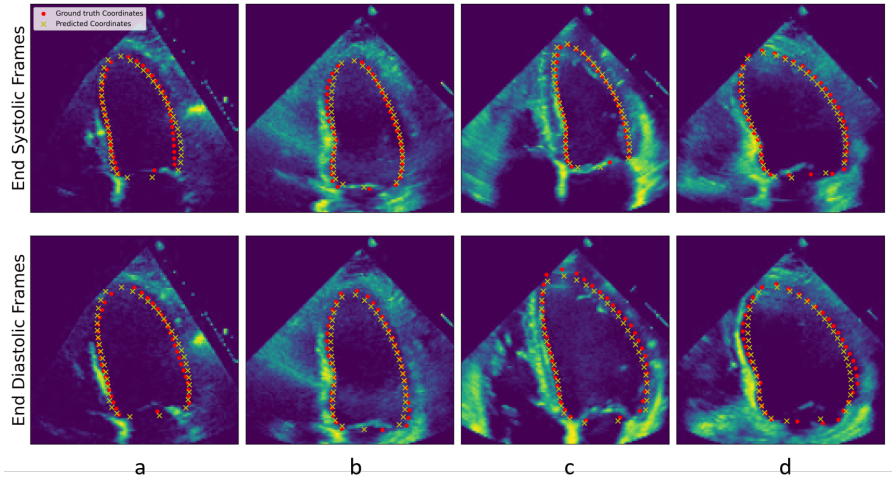


FIGURE 5.9: Edge prediction on ES and ED frames respectively. Red dots show the positions of ground truth coordinates, and yellow crosses show the positions of predicted coordinates. Columns (a-d) represent evaluations obtained on different samples.

5.4.4 Ablation Experiments

The ablation study systematically investigated the impact of various components and configurations within the proposed multitask architecture. Our experimentation involved utilizing UNet-based encoders—specifically, UNet18, UNet34, UNet50,

and UNet50++—each paired with different loss functions, including DICE and BCE losses. For the regression ablation experiments, we explored different configurations of FC layers. This exploration aimed to assess how the depth of the FC layers affected the model's regression performance. Furthermore, we experimented with different loss backpropagation techniques within our model. We conducted experiments by simply adding losses from the Mask Generation Decoder and Edge Predictor, evaluating their combined impact on the overall performance. Additionally, we explored another approach involving loss backpropagation. This method involved generating masks from the Edge Predictor and combining them with the output from the segmentation mask through operations such as ANDing, ORing, and averaging. We then calculated the loss on this combined output, subsequently backpropagating it to understand its influence on the model's performance. Following extensive experimentation, it was concluded that the optimal results were achieved by combining the individual losses derived from the Mask Generation Decoder and Edge Predictor rather than the loss from the combined masks.

Through these rigorous ablation experiments, we aimed to understand the contributions and effects of different architectural components and loss propagation methods on the overall performance and functionality of our proposed model. The results of several ablation experiments are given in Table 5.4.

TABLE 5.4: Ablation experiment results.

Baseline	$\mathcal{L}_{\mathcal{BCE}}$	$\mathcal{L}_{\mathcal{DICE}}$	FC_2	FC_4	**DSC**	**IoU**	**F1**	**F2**	**Accuracy**	**Recall**
UNet_34	✓	×	✓	×	0.920	0.834	0.910	0.930	0.984	0.942
UNet_34	✓	×	×	✓	0.900	0.812	0.896	0.900	0.982	0.913
UNet_34	×	✓	✓	×	0.917	0.836	0.910	0.931	0.984	0.965
UNet_34	×	✓	×	✓	0.914	0.830	0.907	0.928	0.983	0.947
UNet_50	✓	×	✓	×	0.917	0.836	0.911	0.926	0.984	0.939
UNet_50	✓	×	×	✓	0.903	0.817	0.899	0.899	0.982	0.905
UNet_50	×	✓	✓	×	0.918	0.839	0.912	0.933	0.984	0.949
UNet_50	×	✓	×	✓	0.910	0.821	0.901	0.923	0.982	0.942

$\mathcal{L}_{\mathcal{BCE}}$ - BCE loss for segmentation training. $\mathcal{L}_{\mathcal{DICE}}$ - DICE loss for segmentation training. FC_2 - Two FC layers in Edge Predictor. FC_4 - Four FC layers in Edge Predictor.

After examining various ablation experiments, it is evident that utilizing two FC layers yielded the most favorable outcomes. The increase in the number of FC

layers led to overfitting, resulting in a degradation in performance. Despite experimenting with deeper ResNet architectures, there was no substantial enhancement in accuracy. Consequently, UNet with ResNet34 backbone was selected as the proposed encoder due to its computational efficiency while fulfilling the intended purpose, in contrast to the ResNet50 backbone. The choice of loss function did not have any considerable impact on training.

5.4.5 Time-Space Complexity Analysis

Table 5.5 provides a comprehensive comparison of the time-space complexity of the proposed model with the SOTA models. In terms of the number of parameters, the proposed model stands out with 24.59 million parameters, slightly surpassing UNet_34 and falling short of UNet_50. Additionally, its model size of 93.96 MB places it between UNet_34 and UNet_50, making it relatively compact compared to UNet++_50, which has the largest size at 187.24 MB. Regarding computational requirements, the giga-floating-point operations per second (GFLOPs) required by each model during inference, are compared.

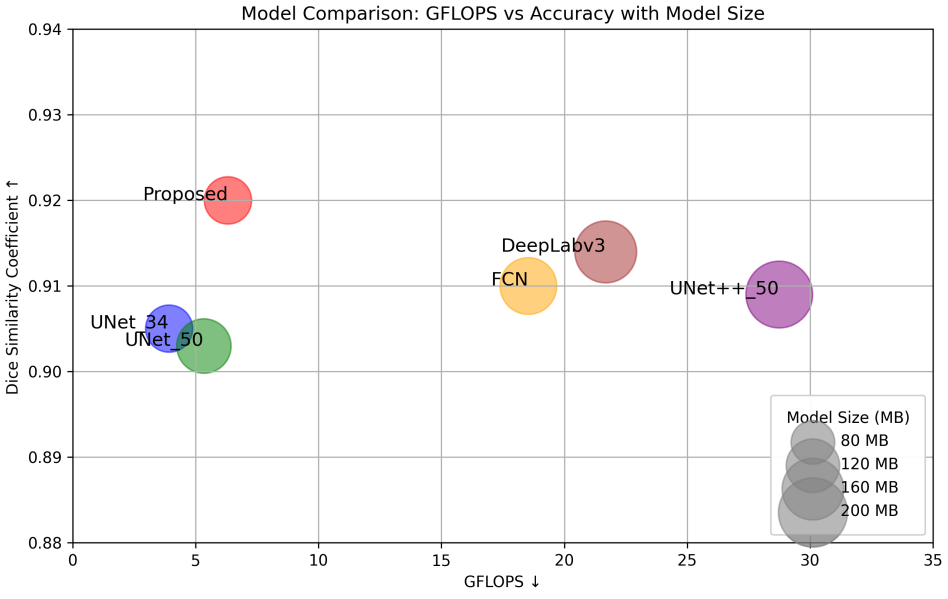


FIGURE 5.10: Trade-off between accuracy and computational efficiency based on GLOPs

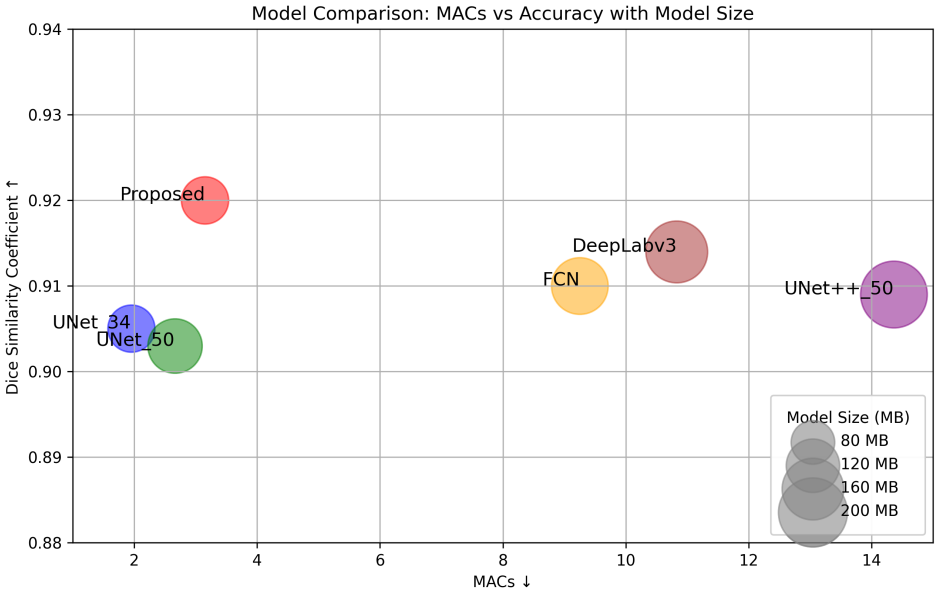


FIGURE 5.11: Trade-off between accuracy and computational efficiency based on MACs

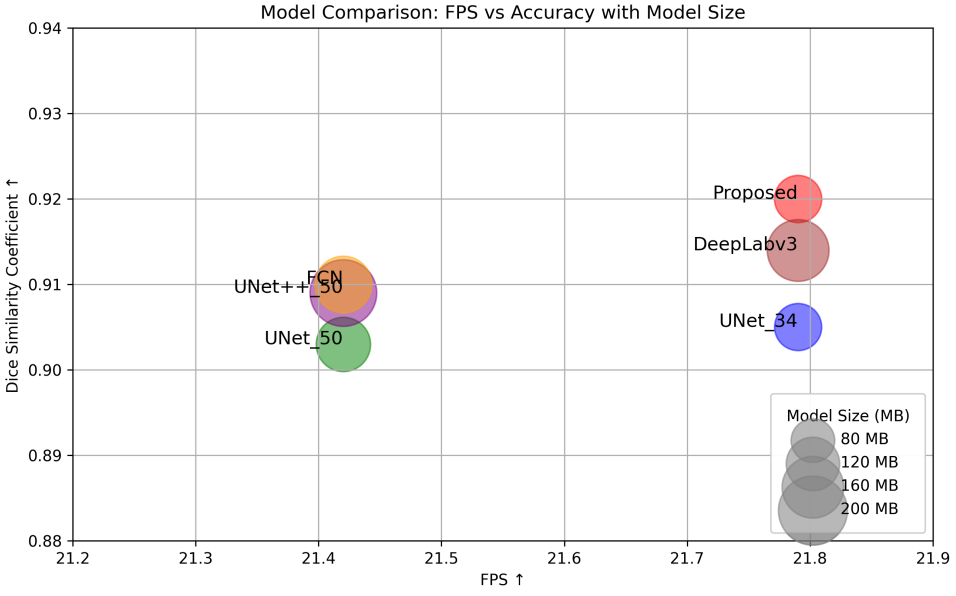


FIGURE 5.12: Trade-off between accuracy and processing speed measured in FPS

The proposed model demands 6.31 GFLOPs, outstripping UNet_34 and UNet_50 but substantially less than UNet++_50, which requires 28.74 GFLOPs. Similarly, its requirement of 3.15 MACs, which represents the number of multiply-accumulate operations required by each model during inference, places it ahead of UNet_34

TABLE 5.5: Comparison of Time-Space complexity of the proposed model with SOTA models.

	Proposed	**UNet_34**	**UNet_50**	**UNet++_50**	**FCN**	**DeepLabv3**
Parameters (m)↓	24.59	24.44	32.53	48.99	35.32	42.00
Model Size (MB)↓	93.96	93.38	124.39	187.24	135.06	160.57
GFLOPs↓	6.31	3.91	5.33	28.74	18.53	21.68
MACs↓	3.15	1.95	2.66	14.36	9.25	10.83
FPS↑	21.79	21.79	21.42	21.42	21.42	21.79
Time per prediction (sec)↓	0.0459	0.0459	0.0467	0.0467	0.0467	0.0459
Memory per prediction (GB)↓	0.20	0.27	0.39	0.76	0.52	0.38

and UNet_50 but below UNet++_50. The proposed model achieves frames per second (FPS) of 21.79, matching the performance of UNet_34 and DeepLabv3. The time per prediction is 0.0459 seconds, similar to UNet_34 and DeepLabv3. Furthermore, its memory requirement per prediction is the most efficient among all models listed, at 0.20 GB. These findings suggest that the proposed model strikes a balance between complexity and efficiency, performing competitively in terms of time-space complexity compared to established SOTA models. The figures 5.10, 5.11 and 5.12 illustrate this by plotting DSC performance against MACs, GFLOPs, and FPS, respectively. The size of the circles in the plot represents the model sizes.

5.5 Summary

In this chapter, we introduced a novel multitask learning framework designed to enhance the segmentation of the LV from echocardiogram data. Our architecture involved extracting semantic features using shared layers, which were subsequently decoupled into task-specific layers comprising a Mask Generation Decoder and an Edge Predictor. By integrating boundary information with mask segmentation by training the network through a joint objective function, we achieved an enhancement in overall segmentation accuracy. The inclusion of an edge predictor within the multitask network refined the mask segmentation process and helped in the detection of near-border regions as well. Our experimental findings demonstrated the method's robustness and its substantial ability to improve segmentation performance when compared to other SOTA segmentation models.

A limitation of the proposed approach is its sensitivity to variations in data quality or imaging conditions. Differences in echocardiogram resolution or artifacts may impact the model's performance. To enhance robustness for real-world applications, using a more diverse dataset with various echocardiogram perspectives, beyond the A4C view used in this study, could be beneficial.

Chapter 6

Conclusion and Future Work

6.1 Conlusion

AI has the potential to expedite the delivery of care by accelerating diagnosis, aiding healthcare systems in proactive population health management, and directing resources to areas where they can make the most significant difference. This heightened efficiency enables healthcare systems to offer improved care to a larger population, enhancing the experience of healthcare professionals. With AI's assistance, practitioners can allocate more time to direct patient care, thereby reducing burnout.

DL and ML have the potential to provide automatic measurements that are not only consistent and accurate but also less time-consuming. Specifically, DL algorithms, which support both supervised and semi-supervised learning, can be particularly valuable in scenarios where data availability is limited. In fields like echocardiography, where certain relationships or vital features remain undiscovered, these self-learning methods hold the potential to uncover insights beyond current knowledge. However, it's essential to ensure that these algorithms remain interpretable. Therefore, transparency in their design and decision-making process is crucial for understanding and ultimately trusting the results they provide.

Recent research in echocardiography has stressed the utilization of independent DL models for tasks such as view classification, LV segmentation, and EF estimation. However, the accuracy of these methods is contingent upon various factors including data clarity, volume, and the accuracy of clinical ground truth. Notably, the accuracy of each task in automated pipelines is heavily influenced by the accuracy of the preceding one, as seen in the reliance of EF estimation on precise LV boundary delineation.

Existing limitations include a lack of exploration of Simpson's method for LV feature extraction and a dearth of interpretability and alignment with clinical workflow in current methodologies. Studies often rely on the area-length method to determine LV volumes and EF from segmented LVs, despite Simpson's method being the preferred technique in clinical practice for LV EF evaluation. However, the area-length method's assumption of a bullet-shaped LV is not always accurate.

In studies that estimate EF directly using DL methods, the process often entails identifying systolic and diastolic frames within the cardiac cycle, which is prone to inaccuracies and time constraints. However, the exploration of the entire echocardiographic cine for both LV segmentation and EF estimation remains relatively unexplored. Furthermore, the simultaneous training of LV segmentation and EF regression, which are traditionally treated as separate tasks, has not been extensively explored.

Additionally, prevalent LV segmentation algorithms often overlook high-level details like edges and boundaries, leading to inaccurate delineations. Addressing these gaps necessitates the development of robust, standardized, and clinically validated automated methods for EF estimation from echocardiogram videos.

This study was conducted in two phases to achieve the aims and objectives stated in the first chapter of this thesis. Initially, we explored clinical methods for quantifying LV structure and function. LV segmentation was performed using a deep neural network, followed by feature extraction from the segmented LV based on

Simpson's method, a widely recommended clinical approach. Various ML techniques and neural networks were then applied to these features for ejection fraction estimation. After rigorous experimentation, we concluded that LSTM, with its temporal memory capabilities, produced the most accurate results.

To fully automate the process, EFNet was proposed in Chapter 4, which aimed to eliminate the need for manually crafting features from the segmented LV. EFNet performed LV segmentation and EF estimation directly from echocardiogram videos, taking advantage of the interconnectedness between these processes. Doing so eliminated the need to extract ED and ES frames from the video. EFNet not only streamlined the entire process but also yielded improved results compared to those obtained previously in Chapter 3 based on clinical methods.

During the training phase of the model introduced in Chapter 4, features are derived from the embeddings acquired through the LV segmentation module. The regression network then trains on these features to perform EF estimation. Consequently, the accuracy of the segmentation network is conclusive for achieving precise EF estimation. Hence, in Chapter 5, the work was further extended to explore the possible improvement in the segmentation process. This was achieved by proposing the decoupling of edge and mask segmentation processes, which could potentially provide better detection of edge coordinates.

To accomplish this, an encoder-decoder based architecture was employed. The encoder resembled that of a UNet, while the decoder consisted of two modules: a mask generation decoder that performed semantic segmentation based on mask detection and an edge predictor module that performed boundary prediction. By fusing the outcomes from both modules, their respective performances were complemented, resulting in enhanced accuracy of LV segmentation. These conclusions indicate that the aims outlined in section 1.8 have been addressed, aligning with the initial objectives set at the beginning of this work.

6.2 Implications in Clinical Practice

Using DL methods to estimate left ventricular ejection fraction brings numerous benefits to everyday clinical practice. DL algorithms have the capability to automate LV segmentation and EF estimation, resulting in reduced time and effort for clinicians. This streamlined process allows for faster and more efficient assessment of cardiac function, enabling timely decision-making and patient management. Additionally, DL based approaches provide uniform and consistent EF measurements, significantly mitigating the inter-observer variability often encountered in manual assessments. Consequently, the evaluation of cardiac function becomes more accurate and reliable, fostering consistency across diverse healthcare settings and among clinicians.

Automated EF estimation using ML and DL techniques also has significant applicability in point-of-care ultrasound (POCUS) devices. These devices are typically compact and portable, making them suitable for use in ambulances, remote clinics, and far-flung areas. POCUS devices equipped with DL-based EF estimation algorithms can be readily used by non-expert healthcare providers, such as paramedics or clinicians in remote areas and enable real-time EF estimation, providing immediate feedback to aid in diagnostic decision-making and patient management. This enables remote consultation, where acquired ultrasound images and EF estimations can be shared with experts located in urban centers, facilitating expert guidance.

It is important to note that while these advantages present promising prospects, the integration of ML into daily clinical practice requires careful validation, standardization, and regulatory considerations. Moreover, the effectiveness and applicability of the DL based methods heavily rely on the quality and diversity of the training data in order to generalize across different populations and imaging conditions. Certain factors such as variations in imaging protocols, equipment and anatomical variances among individuals can influence the accuracy and reliability of the results. These limitations present opportunities for future research and improvements. Nonetheless, the potential benefits of DL based EF estimation make

it an exciting and promising avenue for enhancing clinical decision-making and patient care.

6.3 Limitations

There are several limitations to consider in this work. The performance and generalizability of the proposed method heavily rely on the availability and quality of the training data. Limited access to diverse and representative datasets may impact the model's ability to generalize to different populations and imaging conditions. The proposed method may also be sensitive to variations in imaging protocols and equipment. Factors such as image resolution, image quality, and anatomical variations among individuals can affect the accuracy and reliability of the LV segmentation and EF estimation. These limitations provide opportunities for future research and improvements to enhance the accuracy, interpretability, and clinical applicability of the proposed EF estimation methods.

6.4 Future Work

There are various aspects that could be explored further to enhance the reliability and applicability of the proposed algorithms to provide valuable directions for future research, encompassing both technical advancements and clinical applications. By addressing these areas, the proposed algorithms can be further refined and validated, ultimately contributing to advancements in cardiovascular imaging and patient care.

- **Incorporation of A2C View Data:** Expanding the study to incorporate data from the A2C view alongside the A4C view presents an opportunity to improve the accuracy and reliability of EF estimation. This expansion allows

for a more comprehensive assessment of cardiac function, considering different imaging perspectives and potentially capturing additional information that may complement the A4C view.

- **Personalized EF Assessment:** Utilizing ML techniques to incorporate patient-specific information for EF estimation is a promising avenue for personalized medicine. By considering factors such as age, gender, comorbidities, and medical history, the algorithms can adapt to individual characteristics, leading to more tailored treatment plans and improved patient outcomes. This approach aligns with the growing trend towards precision medicine and could significantly enhance the clinical utility of EF estimation.

- **Data Augmentation and Diversification:** Given the data-intensive nature of DL techniques, incorporating more diverse datasets can enhance the reliability and generalizability of the proposed algorithms. By including data from different healthcare settings and diverse patient populations, the algorithms can better accommodate variations in echocardiogram characteristics across different demographics and ethnicities, leading to more robust models with broader applicability.

- **Integration of Improved Segmentation Module:** Integrating the improved segmentation module from Chapter 5 into the EFNet proposed in Chapter 4 is a logical step towards enhancing the overall accuracy of LV structure and function quantification. The improved segmentation module is expected to provide a more precise delineation of LV boundaries, which can directly contribute to more accurate EF estimation. This integration ensures that advancements made in one aspect of the algorithm are effectively transferred to the overall system, maximizing performance gains.

- **Use of Efficient Backbone Architectures:** One potential avenue for extending this work involves exploring the use of computationally more efficient backbone architectures, such as EfficientNet. While the current work has effectively utilized ResNet as the backbone, adopting a more efficient architecture like EfficientNet could enhance performance by offering

a better trade-off between accuracy and computational cost. EfficientNet's compound scaling method enables it to maintain high accuracy with fewer parameters and lower computational demands compared to ResNet. This improvement would be particularly advantageous for deploying the model in resource-constrained environments or when processing large-scale datasets, ultimately leading to more practical and scalable solutions.

- **Extension to Disease Classification:** Beyond EF estimation and LV segmentation, there is potential to extend the proposed algorithms to the classification of various cardiovascular diseases. By leveraging the insights gained from EF estimation and LV segmentation, the algorithms can contribute to more comprehensive diagnostic workflows, aiding clinicians in accurately identifying and classifying different cardiac pathologies. This extension aligns with the broader goal of improving clinical decision-making and patient care in the field of cardiology.

Bibliography

[1] World Health Organization (WHO), "Global status report on noncommunicable diseases 2010 Geneva, Switzerland," *World Health*, 2010.

[2] G. R. Dagenais, D. P. Leong, S. Rangarajan, F. Lanas, P. Lopez-Jaramillo, R. Gupta, R. Diaz, A. Avezum, G. B. Oliveira, A. Wielgosz, S. R. Parambath, P. Mony, K. F. Alhabib, A. Temizhan, N. Ismail, J. Chifamba, K. Yeates, R. Khatib, O. Rahman, K. Zatonska, K. Kazmi, L. Wei, J. Zhu, A. Rosengren, K. Vijayakumar, M. Kaur, V. Mohan, A. H. Yusufali, R. Kelishadi, K. K. Teo, P. Joseph, and S. Yusuf, "Variations in common diseases, hospital admissions, and deaths in middle-aged adults in 21 countries from five continents (PURE): a prospective cohort study," *The Lancet*, vol. 395, no. 10226, pp. 785–794, 2020.

[3] K. Seetharam, N. Kagiyama, and P. P. Sengupta, "Application of mobile health, telemedicine and artificial intelligence to echocardiography," 2019.

[4] M. A. Chamsi-Pasha, P. P. Sengupta, and W. A. Zoghbi, "Handheld Echocardiography: Current State and Future Perspectives," *Circulation*, vol. 136, no. 22, 2017.

[5] J. Zhang, S. Gajjala, P. Agrawal, G. H. Tison, L. A. Hallock, L. Beussink-Nelson, M. H. Lassen, E. Fan, M. A. Aras, C. R. Jordan, K. E. Fleischmann, M. Melisko, A. Qasim, S. J. Shah, R. Bajcsy, and R. C. Deo, "Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy," *Circulation*, vol. 138, no. 16, pp. 1623–1635, 2018.

[6] Q. Ciampi and B. Villari, "Role of echocardiography in diagnosis and risk stratification in heart failure with left ventricular systolic dysfunction," 2007.

[7] C. Mitchell, P. S. Rahko, L. A. Blauwet, B. Canaday, J. A. Finstuen, M. C. Foster, K. Horton, K. O. Ogunyankin, R. A. Palma, and E. J. Velazquez, "Guidelines for Performing a Comprehensive Transthoracic Echocardiographic Examination in Adults: Recommendations from the American Society of Echocardiography," *Journal of the American Society of Echocardiography*, vol. 32, no. 1, 2019.

[8] R. M. Lang, L. P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova, P. Lancellotti, D. Muraru, M. H. Picard, E. R. Rietzschel, L. Rudski, K. T. Spencer, W. Tsang, and J. U. Voigt, "Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the American society of echocardiography and the European association of cardiovascular imaging," *European Heart Journal Cardiovascular Imaging*, vol. 16, no. 3, 2015.

[9] C. W. Yancy, M. Jessup, B. Bozkurt, J. Butler, D. E. Casey, M. H. Drazner, G. C. Fonarow, S. A. Geraci, T. Horwich, J. L. Januzzi, M. R. Johnson, E. K. Kasper, W. C. Levy, F. A. Masoudi, P. E. McBride, J. J. McMurray, J. E. Mitchell, P. N. Peterson, B. Riegel, F. Sam, L. W. Stevenson, W. H. Tang, E. J. Tsai, and B. L. Wilkoff, "2013 ACCF/AHA guideline for the management of heart failure: A report of the american college of cardiology foundation/american heart association task force on practice guidelines," *Circulation*, vol. 128, no. 16, 2013.

[10] S. M. Dunlay, V. L. Roger, and M. M. Redfield, "Epidemiology of heart failure with preserved ejection fraction," 2017.

[11] P. A. Heidenreich, B. Bozkurt, D. Aguilar, L. A. Allen, J. J. Byun, M. M. Colvin, A. Deswal, M. H. Drazner, S. M. Dunlay, L. R. Evers, J. C. Fang, S. E. Fedson, G. C. Fonarow, S. S. Hayek, A. F. Hernandez, P. Khazanie, M. M. Kittleson, C. S. Lee, M. S. Link, C. A. Milano, L. C. Nnacheta, A. T.

Sandhu, L. W. Stevenson, O. Vardeny, A. R. Vest, and C. W. Yancy, "2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines," 2022.

[12] J. E. Wilcox, J. C. Fang, K. B. Margulies, and D. L. Mann, "Heart Failure With Recovered Left Ventricular Ejection Fraction," *Journal of the American College of Cardiology*, vol. 76, no. 6, 2020.

[13] S. P. Murphy, N. E. Ibrahim, and J. L. Januzzi, "Heart Failure with Reduced Ejection Fraction: A Review," 2020.

[14] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, "Video-based AI for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, 2020.

[15] S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. M. Jodoin, T. Grenier, C. Lartizien, J. Dhooge, L. Lovstakken, and O. Bernard, "Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography," *IEEE transactions on medical imaging*, vol. 38, no. 9, 2019.

[16] S. Batool, I. A. Taj, and M. Ghafoor, "Ejection Fraction Estimation from Echocardiograms Using Optimal Left Ventricle Feature Extraction Based on Clinical Methods," *Diagnostics*, vol. 13, no. 13, 2023.

[17] S. Batool, I. Taj, and M. Ghafoor, "Efnet: A multitask deep learning network for simultaneous quantification of left ventricle structure and function," *Physica Medica*, vol. 125, p. 104505, 2024.

[18] A. Madani, J. R. Ong, A. Tibrewal, and M. R. K. Mofrad, "Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease," *npj Digital Medicine*, vol. 1, no. 1, dec 2018.

[19] S. R. Snare, H. Torp, F. Orderud, and B. O. Haugen, "Real-time scan assistant for echocardiography," in *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 59, no. 3, 2012.

[20] A. H. Abdi, C. Luong, T. Tsang, G. Allan, S. Nouranian, J. Jue, D. Hawley, S. Fleming, K. Gin, J. Swift, R. Rohling, and P. Abolmaesumi, "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, jun 2017.

[21] A. H. Abdi, C. Luong, T. Tsang, J. Jue, K. Gin, D. Yeung, D. Hawley, R. Rohling, and P. Abolmaesumi, "Quality assessment of echocardiographic cine using recurrent neural networks: Feasibility on five standard view planes," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10435 LNCS. Springer Verlag, 2017, pp. 302–310.

[22] M. Razaak and M. G. Martini, "CUQI: cardiac ultrasound video quality index," *Journal of Medical Imaging*, vol. 3, no. 1, p. 011011, mar 2016.

[23] X. Gao, W. Li, M. Loomes, and L. Wang, "A fused deep learning architecture for viewpoint classification of echocardiography," *Information Fusion*, vol. 36, 2017.

[24] H. Vaseli, Z. Liao, A. H. Abdi, H. Girgis, D. Behnami, C. Luong, F. Taheri Dezaki, N. Dhungel, R. Rohling, K. Gin, P. Abolmaesumi, and T. Tsang, "Designing lightweight deep learning models for echocardiography view classification," Tech. Rep., 2019.

[25] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, "Fast and accurate classification of echocardiograms using deep learning," jun 2017.

[26] S. A. MeloJúnior, B. Macchiavello, M. M. Andrade, J. L. Carvalho, H. S. Carvalho, D. F. Vasconcelos, P. A. Berger, A. F. da Rocha, and F. A. Nascimento,

"Semi-automatic algorithm for construction of the left ventricular area variation curve over a complete cardiac cycle," *BioMedical Engineering Online*, vol. 9, 2010.

[27] A. John and K. B. Jayanthi, "Extraction of cardiac chambers from echocardiographic images," in *Proceedings of 2014 IEEE International Conference on Advanced Communication, Control and Computing Technologies, ICACCCT 2014*, 2015.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, 2015.

[29] Z. Yue, W. Li, J. Jing, J. Yu, S. Yi, and W. Yan, "Automatic segmentation of the Epicardium and Endocardium using convolutional neural network," in *International Conference on Signal Processing Proceedings, ICSP*, vol. 0, 2016.

[30] J. F. Silva, J. M. Silva, A. Guerra, S. Matos, and C. Costa, "Ejection Fraction Classification in Transthoracic Echocardiography Using a Deep Learning Approach," in *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, vol. 2018-June, 2018.

[31] S. Leclerc, E. Smistad, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. Rye Berg, M. Belhamissi, S. Israilov, T. Grenier, C. Lartizien, P. M. Jodoin, L. Lovstakken, and O. Bernard, "LU-Net: A Multistage Attention Network to Improve the Robustness of Segmentation of Left Ventricular Structures in 2-D Echocardiography," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, 2020.

[32] S. Moradi, M. G. Oghli, A. Alizadehasl, I. Shiri, N. Oveisi, M. Oveisi, M. Maleki, and J. Dhooge, "MFP-Unet: A novel deep learning based approach for left ventricle segmentation in echocardiography," *Physica Medica*, vol. 67, 2019.

[33] D. Ouyang, B. He, A. Ghorbani, M. P. Lungren, E. A. Ashley, D. H. Liang, and J. Y. Zou, "EchoNet-Dynamic: a Large New Cardiac Motion Video Data Resource for Medical Machine Learning," *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, no. NeurIPS 2019, 2019.

[34] M. J. Mortada, S. Tomassini, H. Anbar, M. Morettini, L. Burattini, and A. Sbrollini, "Segmentation of Anatomical Structures of the Left Heart from Echocardiographic Images Using Deep Learning," *Diagnostics*, vol. 13, no. 10, 2023.

[35] M. Liao, Y. Lian, Y. Yao, L. Chen, F. Gao, L. Xu, X. Huang, X. Feng, and S. Guo, "Left Ventricle Segmentation in Echocardiography with Transformer," *Diagnostics*, vol. 13, no. 14, 2023.

[36] R. Ge, G. Yang, Y. Chen, L. Luo, C. Feng, H. Zhang, and S. Li, "PV-LVNet: Direct left ventricle multitype indices estimation from 2D echocardiograms of paired apical views with deep neural networks," *Medical Image Analysis*, vol. 58, 2019.

[37] M. Li, C. Wang, H. Zhang, and G. Yang, "MV-RAN: Multiview recurrent aggregation network for echocardiographic sequences segmentation and full cardiac cycle analysis," *Computers in Biology and Medicine*, vol. 120, 2020.

[38] H. Wei, H. Cao, Y. Cao, Y. Zhou, W. Xue, D. Ni, and S. Li, "Temporal-Consistent Segmentation of Echocardiography with Co-learning from Appearance and Shape," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12262 LNCS, 2020.

[39] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, 2017.

[40] S. Liu, W. Ding, C. Liu, Y. Liu, Y. Wang, and H. Li, "ERN: Edge loss reinforced semantic segmentation network for remote sensing images," *Remote Sensing*, vol. 10, no. 9, 2018.

[41] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-Octob, 2019.

[42] R. S. Zimmermann and J. N. Siems, "Faster training of Mask R-CNN by focusing on instance boundaries," *Computer Vision and Image Understanding*, vol. 188, 2019.

[43] T. Cheng, X. Wang, L. Huang, and W. Liu, "Boundary-Preserving Mask R-CNN," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12359 LNCS, 2020.

[44] X. Zuo, H. Lin, D. Wang, and Z. Cui, "A Method of Crop Seedling Plant Segmentation on Edge Information Fusion Model," *IEEE Access*, vol. 10, 2022.

[45] B. Sui, Y. Cao, X. Bai, S. Zhang, and R. Wu, "BIBED-Seg: Block-in-Block Edge Detection Network for Guiding Semantic Segmentation Task of High-Resolution Remote Sensing Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, 2023.

[46] A. Ghorbani, D. Ouyang, A. Abid, B. He, J. H. Chen, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, "Deep learning interpretation of echocardiograms," *npj Digital Medicine*, vol. 3, no. 1, 2020.

[47] F. M. Asch, N. Poilvert, T. Abraham, M. Jankowski, J. Cleve, M. Adams, N. Romano, H. Hong, V. Mor-Avi, R. P. Martin, and R. M. Lang, "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, 2019.

[48] F. Liu, K. Wang, D. Liu, X. Yang, and J. Tian, "Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography," *Medical Image Analysis*, vol. 67, 2021.

[49] M. Tokodi, B. Magyar, A. Soós, M. Takeuchi, M. Tolvaj, B. K. Lakatos, T. Kitano, Y. Nabeshima, A. Fábián, M. B. Szigeti, A. Horváth, B. Merkely, and A. Kovács, "Deep Learning-Based Prediction of Right Ventricular Ejection Fraction Using 2D Echocardiograms," *JACC: Cardiovascular Imaging*, vol. 16, no. 8, 2023.

[50] Y. Zeng, P. H. Tsui, K. Pang, G. Bin, J. Li, K. Lv, X. Wu, S. Wu, and Z. Zhou, "MAEF-Net: Multi-attention efficient feature fusion network for left ventricular segmentation and quantitative analysis in two-dimensional echocardiography," *Ultrasonics*, vol. 127, 2023.

[51] K. Y. Leung and J. G. Bosch, "Localized shape variations for classifying wall motion in echocardiograms," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4791 LNCS, no. PART 1, 2007.

[52] M. Qazi, G. Fung, S. Krishnan, R. Rosales, H. Steck, R. B. Rao, D. Poldermans, and D. Chandrasekaran, "Automated heart wall motion abnormality detection from ultrasound images using Bayesian networks," in *IJCAI International Joint Conference on Artificial Intelligence*, 2007.

[53] A. Shalbaf, H. Behnam, Z. Alizade-Sani, and M. Shojaifard, "Automatic classification of left ventricular regional wall motion abnormalities in echocardiography images using nonrigid image registration," *Journal of Digital Imaging*, vol. 26, no. 5, 2013.

[54] T. Araki, N. Ikeda, D. Shukla, N. D. Londhe, V. K. Shrivastava, S. K. Banchhor, L. Saba, A. Nicolaides, S. Shafique, J. R. Laird, and J. S. Suri, "A new method for IVUS-based coronary artery disease risk stratification: A link between coronary & carotid ultrasound plaque burdens," *Computer Methods and Programs in Biomedicine*, vol. 124, 2016.

[55] S. G. Mougiakakou, S. Golemati, I. Gousias, A. N. Nicolaides, and K. S. Nikita, "Computer-aided diagnosis of carotid atherosclerosis based on ultrasound image statistics, laws' texture and neural networks," *Ultrasound in Medicine and Biology*, vol. 33, no. 1, 2007.

[56] U. R. Acharya, M. R. K. Mookiah, S. Vinitha Sree, D. Afonso, J. Sanches, S. Shafique, A. Nicolaides, L. M. Pedro, J. Fernandes E Fernandes, and J. S. Suri, "Atherosclerotic plaque tissue characterization in 2D ultrasound longitudinal carotid scans for automated classification: A paradigm for stroke risk assessment," *Medical and Biological Engineering and Computing*, vol. 51, no. 5, 2013.

[57] U. Raghavendra, H. Fujita, A. Gudigar, R. Shetty, K. Nayak, U. Pai, J. Samanth, and U. R. Acharya, "Automated technique for coronary artery disease characterization and classification using DD-DTDWT in ultrasound images," *Biomedical Signal Processing and Control*, vol. 40, 2018.

[58] B. Smitha and K. P. Joseph, "A new approach for classification of atherosclerosis of common carotid artery from ultrasound images," *Journal of Mechanics in Medicine and Biology*, vol. 19, no. 1, 2019.

[59] G. N. Balaji, T. S. Subashini, and N. Chidambaram, "Detection and diagnosis of dilated cardiomyopathy and hypertrophic cardiomyopathy using image processing techniques," *Engineering Science and Technology, an International Journal*, vol. 19, no. 4, 2016.

[60] S. Narula, K. Shameer, A. M. Salem Omar, J. T. Dudley, and P. P. Sengupta, "Machine-Learning Algorithms to Automate Morphological and Functional Assessments in 2D Echocardiography," *Journal of the American College of Cardiology*, vol. 68, no. 21, 2016.

[61] S. Ruder, "An Overview of Multi-Task Learning for Deep Learning," 2017.

[62] Y. Zhang, J. W. Liu, and X. Zuo, "Survey of Multi-Task Learning," *Jisuanji Xuebao/Chinese Journal of Computers*, vol. 43, no. 7, 2020.

[63] A. Amyar, R. Modzelewski, H. Li, and S. Ruan, "Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation," *Computers in Biology and Medicine*, vol. 126, 2020.

[64] S. El-Sappagh, T. Abuhmed, S. M. Riazul Islam, and K. S. Kwak, "Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data," *Neurocomputing*, vol. 412, 2020.

[65] N. Seshadri, D. S. McCalla, and R. Shah, "Early Prediction of Alzheimer's Disease with a Multimodal Multitask Deep Learning Model," *Journal of Student Research*, vol. 10, no. 1, 2021.

[66] S. Tabarestani, M. Eslami, M. Cabrerizo, R. E. Curiel, A. Barreto, N. Rishe, D. Vaillancourt, S. T. DeKosky, D. A. Loewenstein, R. Duara, and M. Adjouadi, "A Tensorized Multitask Deep Learning Network for Progression Prediction of Alzheimer's Disease," *Frontiers in Aging Neuroscience*, vol. 14, 2022.

[67] M. E. Hsieh and V. S. Tseng, "Boosting Multi-task Learning Through Combination of Task Labels - with Applications in ECG Phenotyping," in *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, vol. 9A, 2021.

[68] S. Vesal, M. Gu, A. Maier, and N. Ravikumar, "Spatio-Temporal Multi-Task Learning for Cardiac MRI Left Ventricle Quantification," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, 2021.

[69] J. Torres-Soto and E. A. Ashley, "Multi-task deep learning for cardiac rhythm detection in wearable devices," *npj Digital Medicine*, vol. 3, no. 1, 2020.

[70] C. Yu, Z. Gao, W. Zhang, G. Yang, S. Zhao, H. Zhang, Y. Zhang, and S. Li, "Multitask Learning for Estimating Multitype Cardiac Indices in MRI and CT Based on Adversarial Reverse Mapping," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, 2021.

[71] W. Xue, A. Islam, M. Bhaduri, and S. Li, "Direct Multitype Cardiac Indices Estimation via Joint Representation and Regression Learning," *IEEE Transactions on Medical Imaging*, vol. 36, no. 10, 2017.

[72] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Rethinking Atrous Convolution for Semantic Image Segmentation Liang-Chieh," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, 2018.

[73] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10553 LNCS, 2017.

[74] S. F. Qadri, H. Lin, L. Shen, M. Ahmad, S. Qadri, S. Khan, M. Khan, S. S. Zareen, M. A. Akbar, M. B. Bin Heyat, and S. Qamar, "CT-Based Automatic Spine Segmentation Using Patch-Based Deep Learning," *International Journal of Intelligent Systems*, vol. 2023, 2023.

[75] M. Blaivas and L. Blaivas, "Machine learning algorithm using publicly available echo database for simplified "visual estimation" of left ventricular ejection fraction." *World journal of experimental medicine*, vol. 12, no. 2, pp. 16–25, mar 2022.

[76] J. Tromp, P. J. Seekings, C. L. Hung, M. B. Iversen, M. J. Frost, W. Ouwerkerk, Z. Jiang, F. Eisenhaber, R. S. Goh, H. Zhao, W. Huang, L. H. Ling, D. Sim, P. Cozzone, A. M. Richards, H. K. Lee, S. D. Solomon, C. S. Lam, and J. A. Ezekowitz, "Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study," *The Lancet Digital Health*, vol. 4, no. 1, 2022.

[77] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Lecture Notes*

in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11045 LNCS, 2018.

[78] S. A. Otto, M. Kadin, M. Casini, M. A. Torres, and T. Blenckner, "A quantitative framework for selecting and validating food web indicators," *Ecological Indicators*, vol. 84, 2018.

[79] H. Reynaud, A. Vlontzos, B. Hou, A. Beqiri, P. Leeson, and B. Kainz, "Ultrasound Video Transformers for Cardiac Ejection Fraction Estimation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12906 LNCS, 2021.

[80] L. Fazry, A. Haryono, N. K. Nissa, Sunarno, N. M. Hirzi, M. F. Rachmadi, and W. Jatmiko, "Hierarchical Vision Transformers for Cardiac Ejection Fraction Estimation," in *IWBIS 2022 - 7th International Workshop on Big Data and Information Security, Proceedings*, 2022.

[81] M. Mokhtari, T. Tsang, P. Abolmaesumi, and R. Liao, "EchoGNN: Explainable Ejection Fraction Estimation with Graph Neural Networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13434 LNCS, 2022.

[82] E. Smistad, A. Østvik, I. M. Salte, D. Melichova, T. M. Nguyen, K. Haugaa, H. Brunvand, T. Edvardsen, S. Leclerc, O. Bernard, B. Grenne, and L. Løvstakken, "Real-Time Automatic Ejection Fraction and Foreshortening Detection Using Deep Learning," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, 2020.

[83] Z. Zhang, S. Fidler, and R. Urtasun, "Instance-level segmentation for autonomous driving with deep densely connected MRFs," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, 2016.

[84] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, 2017.

[85] J. Ji, X. Lu, M. Luo, M. Yin, Q. Miao, and X. Liu, "Parallel Fully Convolutional Network for Semantic Segmentation," *IEEE Access*, vol. 9, 2021.

[86] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[87] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, 2016.

[88] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[89] S. Dong, G. Luo, C. Tam, W. Wang, K. Wang, S. Cao, B. Chen, H. Zhang, and S. Li, "Deep Atlas Network for Efficient 3D Left Ventricle Segmentation on Echocardiography," *Medical Image Analysis*, vol. 61, 2020.

[90] S. S. Ahn, K. Ta, S. Thorn, J. Langdon, A. J. Sinusas, and J. S. Duncan, "Multi-frame Attention Network for Left Ventricle Segmentation in 3D Echocardiography," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12901 LNCS, 2021.