Classification of Selected Cardiac Abnormalities through Machine Learning

by
Kalayvaani Murugan

Thesis presented in partial fulfilment of the requirements for the degree of Master of Engineering (Mechanical) in the Faculty of Engineering at Stellenbosch University

Supervisor: Dr Gareth Erfort

April 2022

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: April 2022

Abstract

Classification of Selected Cardiac Abnormalities through Machine Learning

K. Murugan

Department of Mechanical and Mechatronic Engineering,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MScEng (Biomedical)

April 2022

Cardiovascular diseases contribute to a large number of deaths worldwide per year. From an engineering perspective, an opportune point of intervention is the examination phase of a patient where the equipment and supporting software is concerned. This study aims to develop a prototype supervised machine learning algorithm that can be used as a diagnostic tool in medical practise. Four hundred and six (406) Echocardiography examinations were collected containing six (6) different cardiac abnormalities associated with the left ventricle and aortic valve. Data was considerably insufficient thus augmentation techniques were required to generate synthetic samples. Image processing techniques and various calculations were used to derive measurements and features to be suitable input for the machine learning models. Random Forest and Neural Network models with a variety of dimensions were developed and trained in 3 different tests. The first 2 tests investigated the value of engineering (measurement-derived) and medical (patient information) features to model outputs. Test 3 investigated the effect of various training set sizes. Both models were better informed by medical features than those extracted geometrically or calculated. This was found due to the effect of noise distorting measurements extracted for features. Models also performed better on the largest training set size (90% of data). All models were evaluated by selected performance metrics and/or learning curves (where applicable). The most suitable model selected was a Random Forest instance, as Neural Networks were prone to overfitting training data. These results were not true reflections of either model's capabilities due to the underlying data representativeness issue.

Uittreksel

Classification of Selected Cardiac Abnormalities through Machine Learning

K. Murugan

Department of Mechanical and Mechatronic Engineering,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Tesis: MScEng (Biomedical)

April 2022

Kardiovaskulêre siektes dra by tot 'n groot aantal sterftes wêreldwyd per jaar. Vanuit 'n ingenieursperspektief is die ondersoekfases van pasiëntbehandeling die venster vir intervensie. Die doel van die studie is om 'n prototipe toesig masjienleer-algoritme te ontwikkel wat as 'n diagnose hulpmiddel in die mediese praktyk gebruik kan word. Vierhonderd-en-ses (406) eggokardiogram toetse is ingesamel vir ses (6) verskillende kardiale abnormaliteite wat verband hou met die linker ventrikel en aortaklep. Uiters onvoledoende data is versamel, en daarom was dit noodsaaklik om van aanvullingstegnieke gebruik te maak om sintetiese monsters te genereer. Beeldverwerkingstegnieke en verskeie berekeninge is gebruik om afleidings kenmerke en metings te identifiseer om geskikte insette vir die masjienleermodelle te bied. Random Forest en Neurale Netwerk modelle met 'n verskeidenheid dimensies is ontwikkel en opgelei in 3 verskillende toetse. Die eerste twee toetse het die waarde van ingenieurswese (meting-afgeleide) en mediese (pasiëntinligting) kenmerke ondersoek om uitsette te modelleer. Die derde toets het die effek van verskeie opleidingstel groottes ondersoek. Beide modelle is beter ingelig deur mediese kenmerke as die wat meetkundig uitgewerk of bereken is. Die rede hiervoor is later geïndetifiseer as geraas metings wat as kenmerke onttrek is en gebruik is. Modelle het beter gevaar met die grootste opleidingstel grootte (90% van die data). Alle modelle is geëvalueer deur prestasiemaatstawwe of leerkurwes (waar van toepassing). Die mees geskikte gekose model, was 'n Random Forest geval, aangesien Neurale netwerke geneig was om opleidingsdata te oorpas. Hierdie resultate was nie ware weerspieëling van enige van die modelle se vermoëns nie as gevolg van die teenwoordigheid van onderliggende data kwessies en geraas.

Acknowledgements

I would like to express my gratitude to my academic leaders and advisors.

To my initial supervisor, Prof. Pieter Fourie, thank you for the idea and inspiration behind this project, and for connecting me to helpful people and resources.

To Dr. Van der Bijl, thank you for your time, expertise and willingness to advise me at key points in this project, in both data acquisition and writing stages.

Most importantly, to my current supervisor, Dr. Gareth Erfort, thank you for taking me and this project on despite the circumstances around the time and your current load. I cannot thank you enough for your guidance, questions and the much-needed push to get to the finish line.

I would like to acknowledge my support system, without whom, I would have given up a long time ago. To the friends I count as family, thank you for helping me keep my sanity and hope alive.
Special mention of Jessica de Villiers and Francé Bresler for being extra sets of eyes in these closing stages.

To my amazing parents and siblings, thank you for your support and prayers that have been the wind beneath my wings more times than I care to count.

Last, but not least, to my phenomenal husband. Thank you for being my northern star, voice of reason and biggest supporter.

Soli Deo Gloria!

Dedication

To my parents, Desireé and Kovilan Murugan; and all those before me that could only dream of the things our family is achieving today. It was an honour to be the first, but I know I will not be the last...

Contents

Declaration i

Abstract ii

Uittreksel iii

Acknowledgements iv

Dedication v

Contents vi

List of Figures ix

List of Tables xi

Nomenclature xiii

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Motivation . . . 2
- 1.3 Problem Statement . . . 2
- 1.4 Aim and Objectives . . . 3
- 1.5 Document Structure . . . 3

2 Literature Review 5
- 2.1 Fundamentals of Cardiology . . . 5
  - 2.1.1 The Cardiovascular System . . . 5
  - 2.1.2 Anatomical Structures . . . 7
  - 2.1.3 The Conduction System . . . 9
  - 2.1.4 Basic Cardiac Physiology . . . 10
  - 2.1.5 Echocardiography . . . 14
  - 2.1.6 Cardiac Abnormalities . . . 14
- 2.2 Applied Machine Learning . . . 17
  - 2.2.1 Medical Image Processing . . . 17

2.2.2 Machine Learning 18
2.2.3 Data Augmentation Techniques 21
2.2.4 Feature Selection 22
2.2.5 Evaluation Methods 23

3 Methodology **28**
3.1 Sample Selection 28
3.2 Previous Work: Machine Learning in Cardiology 29
3.3 Experimental Work 31
3.3.1 Image Pre-processing 32
3.3.2 Model Input Preparation 33
3.3.3 Model Development 36
3.3.4 Model Performance Evaluation 37
3.3.5 Final Model Selection 39

4 Results and Discussion **40**
4.1 Image Pre-processing 40
4.2 Model Input Preparation 41
4.3 Model Development 44
4.4 Model Performance Evaluation 46
4.4.1 Engineering Tests 47
4.4.2 Medical Tests 56
4.4.3 Data Tests 65
4.5 Final Model Selection 73
4.6 Conclusion 77

5 Limitations & Recommendations **78**

6 Conclusion **81**

A Appendix 1: Engineering Tests Results **82**
A.1 Mutual Information Scores 82
A.2 Random Forest Results 85
A.3 Neural Networks Results 91

B Appendix 2: Medical Tests Results **95**
B.1 Mutual Information Scores 95
B.2 Random Forest Results 98
B.3 Neural Networks Results 103

C Appendix 3: Data Tests Results **108**
C.1 Shortlisted Models: Previous Results 108
C.2 Random Forest Results 108
C.3 Neural Networks Data tests Results 112

D Appendix 4: Additional Information **114**
D.1 Research Protocol Content 114
D.1.1 Randomisation, Confidentiality & Bias 114
D.1.2 Data Collection & Management 115
D.1.3 Project Commencement Plan 116
D.2 Continuous Ranked Probability Score 116
D.3 FIJI Image Processing Schematic 116

List of References **118**

List of Figures

2.1 Labelled diagram of the internal structures of the human heart (Marieb, 2015) . . . 6
2.2 Interaction of the pulmonary and systemic circuits (Burkhoff, 2002) 7
2.3 Arrangement of the 4 valves of the heart (Anatomy, 2017) . . . 8
2.4 Layers of the heart wall (Marieb, 2015) . . . 9
2.5 Ring-like arrangement of cardiac muscles . . . 9
2.6 Components of the Conduction System of the Heart (Marieb, 2015) 10
2.7 Key Pressure and Volume curves for 1 Cardiac Cycle . . . 11
2.8 Pressure-Volume Loop of the Left Ventricle for the Cardiac Cycle (Burkhoff, 2002) . . . 12
2.9 Standard chest windows for transducer positioning (left) and a labelled apical 2-chamber view (right) . . . 15
2.10 Example of Binarization, Erosion and Dilution effects on an image (OpenCV, 2021) . . . 18
2.11 Intermediary steps of the Watershed algorithm (OpenCV, 2021) . . 19
2.12 Inspiration of Neurons for Perceptrons of a Neural Network (Portilla, 2018) . . . 20
2.13 Confusion Matrix layout for Binary Classification . . . 24
2.14 Loss Learning Curves with differing Learning Rates (Venugopal and Ramaswamy, 2015) . . . 26
2.15 Degrees of Overfitting observed from Accuracy Learning Curves (Venugopal and Ramaswamy, 2015) . . . 26

3.1 Solution Pipeline . . . 32

4.1 Individually applied Data Augmentations . . . 41
4.2 Input and Output of FIJI Image Processing . . . 42
4.3 Visual representation of the effect of inherent and added noise on FIJI output . . . 42
4.4 Visual representation of area exclusion Criteria 1 & 2 . . . 43
4.5 Visual example of geometric parameters derived, with the system centreline (blue), centroid (green) and c and a parameters (red) . . 44
4.6 Mutual Information of all Features . . . 45

4.7 Schematic showing Random Forest instances that output the highest scores per Engineering test 48
4.8 Test E1 - Learning Curves of model instance D3W256 53
4.9 Test E12345 - Learning Curves of model instance D3W64 54
4.10 Schematic showing Random Forest instances that output the highest scores per Medical test 57
4.11 Test M123456: Learning Curves of model instance D3W256 62
4.12 Test M4: Learning Curves of model instance D2W256 63
4.13 Learning Curves of Random Forest instances 67
4.14 Learning Curves for Test-Train ratios 0.1-0.9 (left) and 0.9-0.1 (right) 70
4.15 D2W64 Learning Curves for all Test-Train ratios 72
4.16 Mutual Information Scores for Train-Test Split 0.9-0.1 74

C.1 Performance metric plots of shortlisted models for Engineering Tests 110
C.2 Performance metric plots of shortlisted models for Medical Tests . . 111

D.1 Flow diagram of FIJI image processing steps implemented 117

List of Tables

3.1 Exclusion-Inclusion Criterion 29
3.2 Input Features to Models with Variable Names in parentheses . . . 35
3.3 Pathologies, Abbreviations and Unaugmented case Totals 35
3.4 Engineering and Medical Tests: Test names and associated features removed 38
3.5 Data Tests Details 39

4.1 Highest scoring Random Forest instances per Engineering test with output Performance metrics. Minimums of each test are included in parentheses below 49
4.2 Confusion Matrices of highest (left) and lowest (right) scoring models for each pathology in tests EM0 and E12345 50
4.3 Comparison of performance metrics for highest and lowest scoring model instances for Tests EM0 and E12345 50
4.4 Best performing Neural Networks per Engineering Test with Validation performance metrics 52
4.5 Summary of highest scoring Engineering tests and model instances 55
4.6 Highest scoring Random Forest instances per Medical test with associated Performance metrics. Minimums of each test are included in parentheses below 59
4.7 Confusion Matrices of highest (left) and lowest (right) scoring models for each pathology in tests M123456 and M4 60
4.8 Performance metrics of highest and lowest scoring model instances for Tests M123456 and M4 60
4.9 Best performing Neural Networks per Medical Test Validation performance metrics 61
4.10 Summary of key Medical tests and model instances 64
4.11 Confusion Matrices for Random Forest models at a 0.9-0.1 train-test split 67
4.12 Averaged performance metrics from 0.9-0.1 train-test ratio 68
4.13 Confusion Matrices of E100D16 and D2W64 74
4.14 Performance Metrics of E100D16 and D2W64 74
4.15 Tallies and Total of correctly predicted Multi-label samples per model 75

4.16 Resulting measurements extracted from unaugmented and noised frame of the Heart Failure case 76

A.1 Mutual Information Scores of all individual feature deletions for Engineering Tests 82
A.2 Mutual Information Scores of all cumulative feature deletions for Engineering Tests 83
A.3 Best performing Random Forest models based on performance metrics for each pathology, listed by architectural descriptors 85
A.4 Averaged Performance Metrics for Best Performing Random Forest Models per Engineering Test 87
A.5 Averaged Performance Metrics for Best Performing Neural Network Models per Engineering Test 91

B.1 Mutual Information Scores of all individual feature deletions for Medical Tests 95
B.2 Mutual Information Scores of all cumulative feature deletions for Medical Tests 96
B.3 Best performing Random Forest models based on performance metrics for each pathology, listed by architectural descriptors 98
B.4 Averaged Performance Metrics for Best Performing Random Forest Models per Medical Test 100
B.5 Averaged Performance Metrics for Best Performing Neural Network Models per Medical Test 103

C.1 Performance metrics output for each Random Forests instance for each train-test ratio tested 108
C.2 Train and Validation set performance metrics for Neural Network model instances in all Data tests 112

Nomenclature

Abbreviations

AoP	Aortic Pressure
AR	Aortic Regurgitation (Pathology)
AS	Aortic Stenosis (Pathology)
AV	Atrioventricular (node/valves)
avg	Average/Mean
CNN	Convolutional Neural Network
CO	Cardiac Output
CRPS	Continuous Ranked Probability Score
D	Depth of Model Instance
DBP	Diastolic Blood Pressure
E	Engineering Test Number of Estimators (Random Forest Models)// Elastance (Physiology)
ECG	Electrocardiogram
EDP	End-Diastolic Pressure
EDPVR	End-Diastolic Pressure-Volume Relationship
EDV	End-Diastolic Volume
EF	Ejection Fraction
EMR	Electronic Medical Records
ESPVR	End-Systolic Pressure-Volume Relationship
ESV	End-Systolic Volume
FN	False Negative (Confusion Matrix)
FP	False Positive (Confusion Matrix)
GAN	Generative Adversarial Network
HF	Heart Failure (Pathology)
HR	Heart Rate
IQR	Interquartile range
LAP	Left Atrial Pressure

LV	Left Ventricle
LVH	Left Ventricular Hypertrophy (Pathology)
M	Medical Test
max	Statistical Maximum
min	Statistical Minimum
MI	Myocardial Infarction (Pathology)
N	Normal Cardiac Functioning (Pathology)
NN	Neural Network
P	Pressure
PR	Peripheral Resistance
ReLU	Rectified Linear Unit (Activation Function)
RF	Random Forest
ROI	Region of Interest
SBP	Systolic Blood Pressure
SA	Sinoatrial (node)
SD	Standard Deviation
SV	Stroke Volume
TN	True Negative (Confusion Matrix)
TP	True Positive (Confusion Matrix)
VAR	Variance

Symbols

A	Area
b	Centreline y-intercept
a	Ellipsoidal short axis
b	Bias term
c	Ellipsoidal long axis
C	Centroid
CO_2	Carbon Dioxide
n	Total number of composite areas
m	Centreline gradient
O_2	Oxygen
Q_3	Upper quartile
P_{es}	End-Systolic Pressure
V/Vol	Internal Volume
w	Learned Weights of Neural Networks
x	Input Model Features

Xc	Centroid x-coordinate
Yc	Centroid y-coordinate

Units of Measurement

bpm	Heart Rate in beats per minute
L	Litres

1 Introduction

The study presented in this document, entitled *Classification of Cardiac Abnormalities using Machine Learning*, is introduced in this chapter. Some background (in §1.1) is provided to set the context for the Motivation (§1.2) and Problem Statement (§1.3) that follow. Aims and Objectives (§1.4) are highlighted before the chapter closes with a brief description of the Document Structure (§1.5).

1.1 Background

For both animals and humans alike, the heart is among the five organs considered essential for survival together with the brain, lungs, liver and kidneys (Marieb, 2015). It is a key component of the cardiovascular system of the body, responsible for maintaining the continuous circulation of blood. This provides a means of transport for (i) oxygen and nutrients and/or carbon dioxide and cell metabolic products, and (ii) for endocrine hormones to specific sites in the body. The heart can be described as a double pump, with left and right sides functioning in parallel; each structurally suited to their specific purposes. It beats roughly 115 000 times per day, pumping more than 7000 L of blood for the average human being (Marieb, 2015). When the heart is not functioning normally, the entire body is starved of oxygen; preventing cells from working optimally before eventually dying.

The health of the heart affects every component of the body, and is usually a function of an individual's diet, lifestyle (especially physical health) choices and even psychological well-being. Diseases of the heart or cardiovascular diseases involve disorders with any part of the cardiovascular system. Cardiovascular diseases are the leading cause of deaths with recent statistics being roughly $\pm 33\%$ worldwide and $\pm 17.3\%$ for South Africa, according to the World Heart Federation (2017). In addition to an individual's lifestyle, contact points with health care professionals regarding any heart-related issues are of immense importance. Efficient and effective tests and practises, correct diagnoses and successive treatments are key factors in the care and mortality of heart-patients specifically.

The incorporation of Artificial Intelligence (AI) into the medical field specifically, has boosted the overall effectiveness of the practises and supporting technologies offering improved ways of identifying disorders, diagnosing and treating patients. Not only has it contributed to more sophisticated systems and enhanced patient care, but has saved - and is projected to save - millions (monetarily) in the healthcare industry (Greenlight Medical, 2020). This becomes more apparent as the influx of patients and challenges continuously increase, AI proves to ease many of the strains on health care professionals and administrative workers. With the increased productivity and more regulated incorporation of AI into the medical field, emerging technologies promise to consistently improve patient care.

1.2 Motivation

The motivation behind this research endeavour is to develop a diagnostic tool to aid or assist health care professionals in diagnostic processes by introducing suitable machine learning algorithms. Machine learning algorithms are capable of successfully processing both large and/or limited data sets, and are superb in recognising patterns to make accurate classifications, predictions, or in this case diagnoses (Greenlight Medical, 2020). Their addition also serves to improve the speed, efficiency and reliability of diagnoses as they continuously learn from the continuous stream of incoming data. Deep learning methods, specifically may further provide cardiologists new insights into correlations previously unknown. This project serves to contribute to the collective attempt to develop smart tools to be incorporated into routine medical practises and increase reaction windows and improve service delivery.

1.3 Problem Statement

The study explores the use of simple machine learning techniques in the development of a diagnostic tool to aid cardiologists in clinical practise. Echocardiogram data of 1183 patients will be used to develop a method of feature extraction before the most appropriate algorithm is selected based on key performance metrics. Six cardiac abnormalities were selected to be classified using information extracted from apical 2-chamber views from a standard transthoracic echocardiogram. Traditional means are preserved by means of key visual indicators (to cardiologists) when translated to engineering features to maintain relevance and/or readability to health care professionals.

1.4 Aim and Objectives

The aim of the study is to develop a prototype diagnostic tool that employs relevant features and appropriate machine learning models to correctly classify selected cardiac abnormalities. Cardiovascular diseases contribute to a significant fraction of annual deaths and is noted as an area where machine learning methods can aid health care professionals by improving the speed and reliability of diagnoses. Correct and efficient diagnoses with better supporting infrastructure may contributes to decreasing cardiovascular-related fatalities.

The objectives for the project can thus be stated as follows:

1. To collect a sufficiently high quality, diverse and unbiased dataset to be used to sufficiently train the machine learning algorithms.
2. To develop an effective, generalised pipeline for pre-processing input video data incorporating applicable computer vision techniques and key visual indicators of routine echocardiography practises.
3. To extract relevant measurements from input images and translate them to appropriate model features.
4. To compare appropriate machine learning models for the classification task.
5. To evaluate model performances using meaningful evaluation metrics that can be comprehended for the intended end-use.
6. To identify the best performing machine learning model for the classification of cardiac abnormalities.
7. To scrutinize all the findings to present germane recommendations for further improvements to the selected model and its corresponding performance.

1.5 Document Structure

The overall layout and content of successive chapters with each of their main sections outlined in this section. Following this introductory chapter is a thorough **Literature Review**, in Chapter 2, containing topics aimed at providing a theoretical foundation for the project. A discussion on the Fundamentals of Cardiology (§2.1) is presented that covers: the anatomical structures, circulatory and conduction systems, selected physiological aspects of the heart, working principles of Echocardiography practises and descriptions of the cardiac abnormalities selected for the study. The next section covers theory associated with Applied Machine Learning (§2.2). Special needs of medical (image)

data are discussed, such that they can be made appropriate for the algorithms selected and detailed thereafter. Topics such as data augmentation, feature selections and evaluation methods are included thereafter.

The **Methodology**, presented in Chapter 3, comprises of 3 main sections. First, Ethical Considerations (§3.1) discuss the aspects most fair and appropriate to data (patient) selections for the study. Previous work (§3.2) is then presented to set the context for recent work done in the respective fields. The chapter concludes with details of all practical work (§3.3), from data preprocessing to model development, training and evaluations.

Results and Discussion follow in Chapter 4, where the output of all practical steps are presented (§4.1 - §4.3). Model evaluations (§4.4) present detailed summaries of all findings, with many comparisons before a suitable model is identified. This final model is then discussed hypothetically in its desired context (§4.5). The **Limitations and Recommendations** outlined in Chapter 5, highlight areas of improvement considering all practical steps and methods, with alternative suggestions presented in some areas. The document closes with a **Conclusion** in Chapter 6, of all the work done and documented in this report.

2 Literature Review

The literature review presented in this chapter covers the most relevant aspects of Cardiology and Applied Machine Learning techniques needed for successive chapters, in Sections 2.1 and 2.2, respectively. The Fundamentals of Cardiology (§2.1) outlines the main subsystems, anatomy and applicable physiology of the heart (§2.1.1 - §2.1.3). With echocardiography (§2.1.5), this theory sets the backdrop for the discussion of the cardiac abnormalities selected for the study (§2.1.6). The Applied Machine Learning section includes considerations for medical image processing (§2.2.1) before describing supervised machine learning algorithms used (§2.2.2). Data augmentation and feature selection methods are presented before an evaluation methods discussion concludes the chapter (§2.2.3 - §2.2.5).

2.1 Fundamentals of Cardiology

The human heart can be described as a muscular pump, approximately the size of a clenched fist that rests atop the diaphragm, between the lungs. The function of the heart and the accompanying blood vessels is to transport nutrients, oxygen and metabolic products to and from specific places in the body. The heart's functioning is facilitated by its anatomical structures and conduction system. A labelled diagram of the internal structures and adjoining blood vessels of the heart can be seen in Figure 2.1.

2.1.1 The Cardiovascular System

The cardiovascular (or *circulatory*) system is responsible for the circulation of vital substances throughout the human body in a closed circuit. It comprises of the heart, blood vessels, lymphatic vessels / glands, blood and lymph. The blood vessels (or *vasculature*) consists of 5 classes: arteries, arterioles, capillaries, venules and veins - all structurally suited for their respective functions. Arteries transport blood out of heart, while veins transport blood toward the heart. The arteries are large, elastic blood vessels with more muscular walls for withstanding high pressures of blood pumped from the heart. Arteries subdivide into arterioles; smaller, finer branches with a thinner muscular layer, that

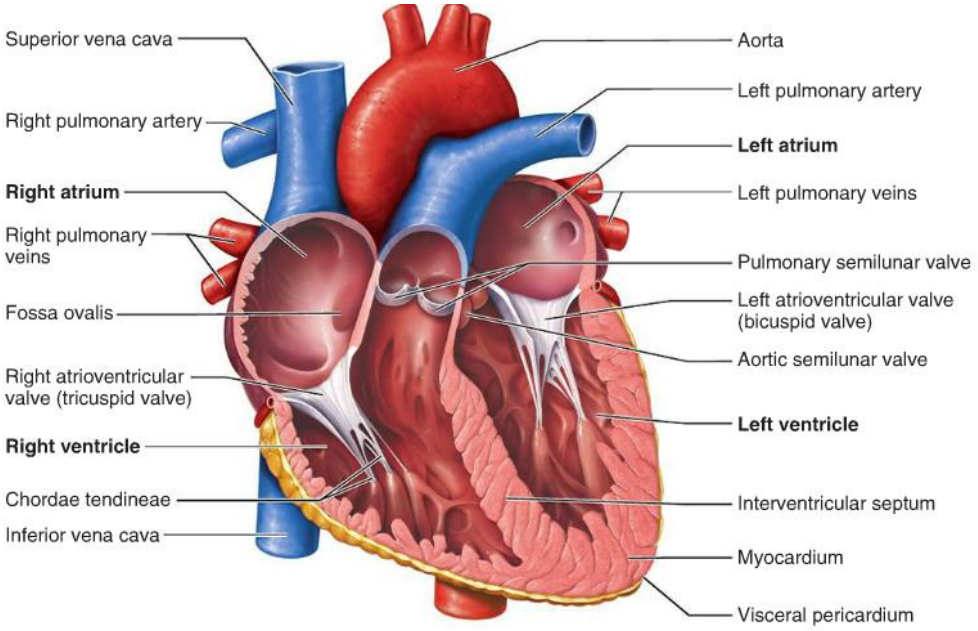

Figure 2.1: Labelled diagram of the internal structures of the human heart (Marieb, 2015)

further divide into capillaries. Capillary walls are thin and semipermeable to allow for diffusions to occur. Capillaries join arterioles to venules, which enlarge to form veins. Veins and venules have thinner, less elastic muscular walls (compared to arteries/arterioles) and valves to ensure one-way flow (Marieb, 2015).

The cardiovascular system can be divided into 2 sub-circuits; the pulmonary and the systemic circuit. The pulmonary circuit, or *venous circulation*, consists of the blood vessels that transport deoxygenated blood to the lungs where carbon dioxide (CO_2) is exchanged for oxygen (O_2). The systemic circuit, or *arterial circulation*, consists of the vessels that transport oxygenated blood to the rest of the body. The arterial system's secondary function is to smooth fluid flow oscillations (pulses) of blood pumped from the heart. Blood flow is thus non-pulsatile as it enters the capillaries and venous circuits thereafter (Walley, 2016).

The interaction of the pulmonary and systemic circuits can be seen in Figure 2.2: arrows indicate the direction of blood flow, red representing oxygenated blood and blue, the deoxygenated blood. With reference to Figure 2.1, the pulmonary circuit begins in the right ventricle, which pumps blood into the pulmonary artery towards the lungs. Within the lungs, arteries and arterioles divide into capillary networks to facilitate gas exchange (near the alveolar walls). Capillaries become venules and veins, returning blood from each lung

to the left atrium via pulmonary arteries. The systemic circuit begins as blood moves into the left ventricle and out the *aorta* (largest artery). The aorta branches to coronary arteries (that feed cardiac muscle), arteries and arterioles as it moves toward capillary networks near body tissues. At the capillary networks, O_2 and nutrients are exchanged for cellular metabolic products. Capillaries become arterioles and arteries, returning blood to the right atrium via the superior and inferior *vena cavae* (largest veins) (Marieb, 2015).

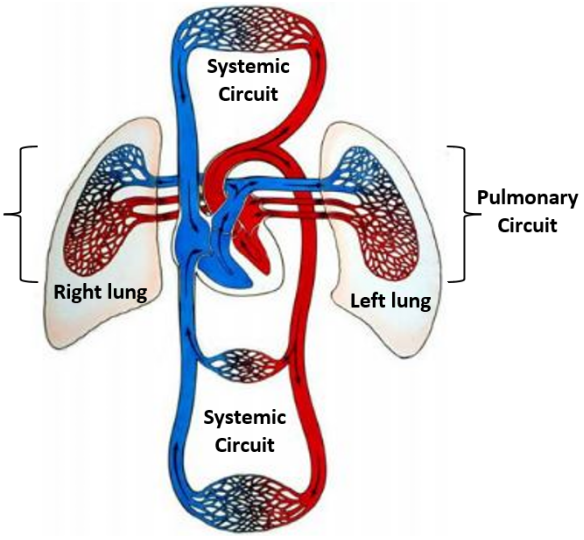

Figure 2.2: Interaction of the pulmonary and systemic circuits (Burkhoff, 2002)

2.1.2 Anatomical Structures

Anatomy of the heart refers to the study of its sub-structures (viz. the chambers, valves and papillary muscles) in the context of their functions. Figure 2.1 shows the 4 main divisions (*chambers*) of the heart; 2 smaller, upper (left and right) atria, and 2 larger, lower (left and right) ventricles. The role of the atria is to collect blood returned to the heart, acting as an intermediary reservoir before blood enters the ventricles. Ventricles collect and pump blood at high pressures out of the heart through the large arteries. All chambers of the heart differ significantly based on their required pump capacities. Due to low pressure blood returned to the heart, both atria have thin muscular walls compared to ventricles. The crescent-shaped right ventricle muscular wall is roughly $\pm$3-5 mm thick, as it pumps blood to the lungs in a shorter circuit. The asymmetric ellipsoidal left ventricle is distinguished by its thick muscular wall ($\pm$1 cm); necessary for generating high pressures to pump blood with enough force to reach parts of the body further away (against peripheral resistance caused by arterial walls and decreasing diameters) (Marieb, 2015).

The left and right chambers are separated by the muscular *septum* that prevents blood on either side from intermingling. Between each atrium and ventricle is an *atrioventricular* (AV) valve; a *tricuspid* valve on the right and *mitral* valve on the left. AV valves are bound to papillary muscle projections (*chordae tendineae*) that extend into the ventricles from their walls. Papillary muscles contract with the ventricles and prevent valves from swinging back into the atria. *Semi-lunar* valves occur between ventricles and arteries, and open during ventricular contraction to allow blood into the aorta and pulmonary artery. The pulmonary valve occurs between the right ventricle and pulmonary artery, and the aortic valve is between the left ventricle and aorta. AV valves prohibit backflow of blood into the atria from ventricles during ventricular contraction. Semi-lunar valves prohibit backflow into the ventricles from arteries during ventricular relaxation (Anatomy, 2017). The valvular arrangement can be seen in Figure 2.3.

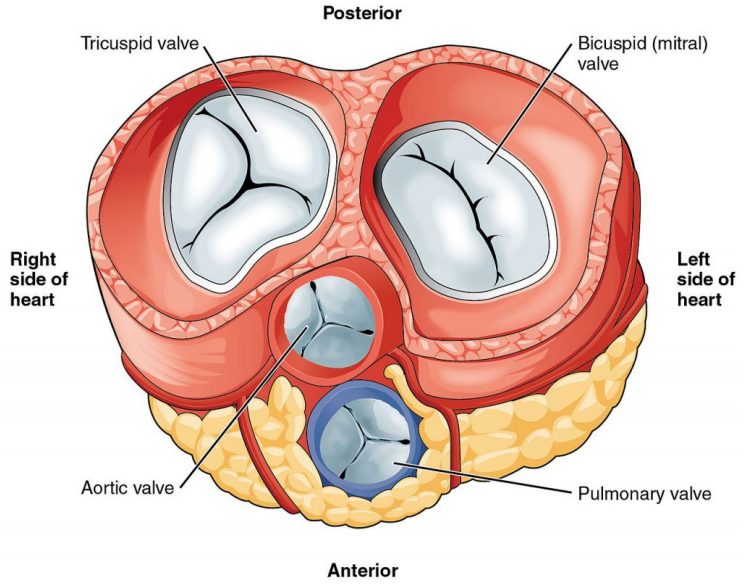

Figure 2.3: Arrangement of the 4 valves of the heart (Anatomy, 2017)

The walls of the heart consist of 3 functionally unique layers, as per Figure 2.4, described as follows:

1. The **pericardium** is a double-layered sac that encloses the heart and attaches it to the vasculature leaving the heart. At the attachment sites, the pericardium folds or *reflects* back on itself and continues along the muscular surface of the heart.

2. The **myocardium** is the thickest layer, consisting primarily of cardiac muscle tissue organised in ring-like arrangement, as per Figure 2.5 - such that ventricles twist upon contraction.

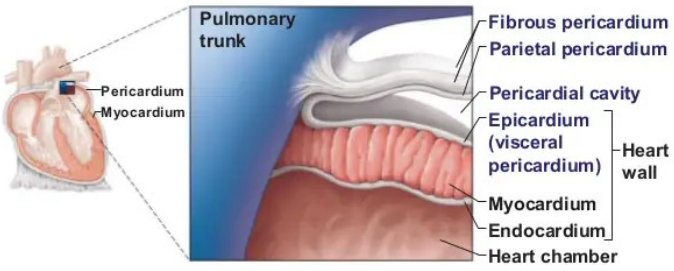

Figure 2.4: Layers of the heart wall (Marieb, 2015)

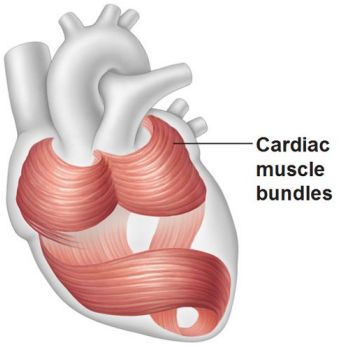

Figure 2.5: Ring-like arrangement of cardiac muscles

3. The **endocardium** forms the inner lining of the chambers, continuous with the attached blood vessels.

Cardiac muscle cells are cushioned by fibrous connective tissue in branched networks. When any portion of the network is stimulated by an impulse, contractions across chamber/s occurs as a unit (Marieb, 2015). Cardiac muscle cells are joined by junctions that facilitate impulse conductions that bring about coordinated movements regulated by the conduction system. Cardiac muscle cells are either myogenic or contractile. Myogenic cells are *autorhythmic* and are where impulses (spontaneous depolarisation) originate that aid rhythmic functioning. Contractile cells bring about muscle contractions in response to impulse stimuli (Biga *et al.*, 2019).

2.1.3 The Conduction System

The conduction (or *electrical*) system of the heart refers to the internal circuitry responsible for coordinating rhythmic contractions. This system comprises of nodes and/or strands of autorhythmic cardiac tissue located at specific regions of the heart, seen in Figure 2.6. These tissues initiate and distribute electrical impulses through cardiac muscle cells of the myocardium (Marieb, 2015).

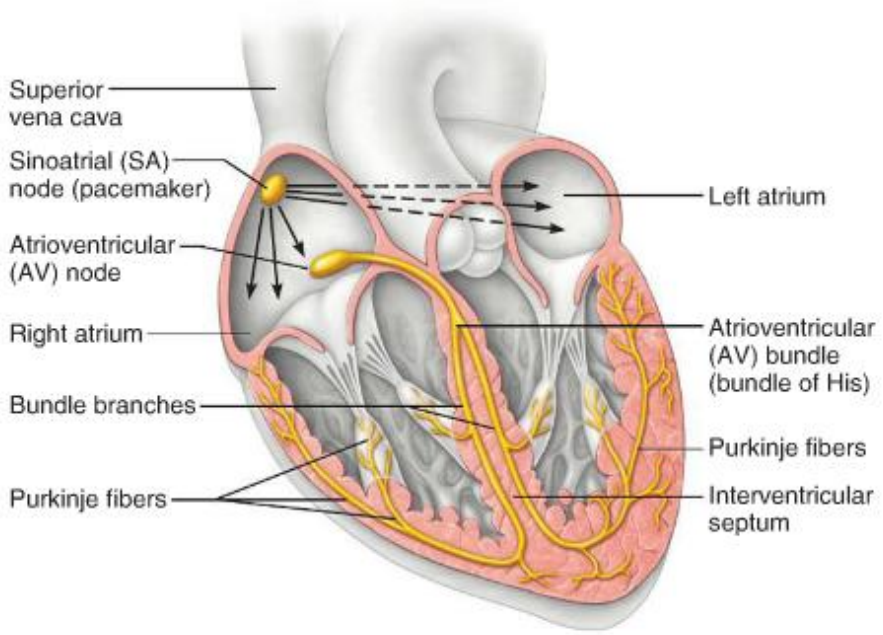

Figure 2.6: Components of the Conduction System of the Heart (Marieb, 2015)

The sinoatrial (SA) and atrioventricular (AV) nodes are small masses of autorhythmic cardiac tissue located beneath the epicardium of the right atrium. Impulses initiated from the SA node are conducted through the atrial *syncytium* (circuitry in the myocardium) bringing about atrial contractions. The atrial and ventricular syncytium are connected by the AV node, through which impulse conduction can continue. Impulse delayed (due to narrow fibres) at the AV node allows time for the atria to (actively) discharge blood into the ventricles. After the AV node, impulses are conducted through the *bundle of His* (in the interventricular septum) and *Purkinje fibres* that spread throughout cardiac muscles of ventricular walls and papillary muscles. Impulse stimulation brings about ventricular contractions (Burkhoff, 2002).

2.1.4 Basic Cardiac Physiology

Physiology of the heart includes a vast collection of knowledge encompassing all mechanisms involved in its functioning. The theory presented in this section summarises facets necessary for this project, and is tailored based on the relationships / calculations used. The cardiac cycle is first outlined (§2.1.4.1) before selected pressure-volume relationships are explored (§2.1.4.2). Most of the information in this section comes from Burkhoff (2002), unless stated otherwise.

2.1.4.1 The Cardiac Cycle

The cardiac cycle refers to the series of events that occur within the heart for every heartbeat. Each cycle comprises the following phases: *systole* (contraction and ejection) and *diastole* (relaxation and filling) - mostly used to describe the ventricles. Systole occurs when cardiac muscles transform from maximally relaxed to maximal mechanical activation. Diastole occurs when cardiac muscles transform from maximally activated to maximally relaxed. Pressure and volume changes for the left atrium, left ventricle and aorta during the cardiac cycle are included in Figure 2.7 to aid event descriptions below:

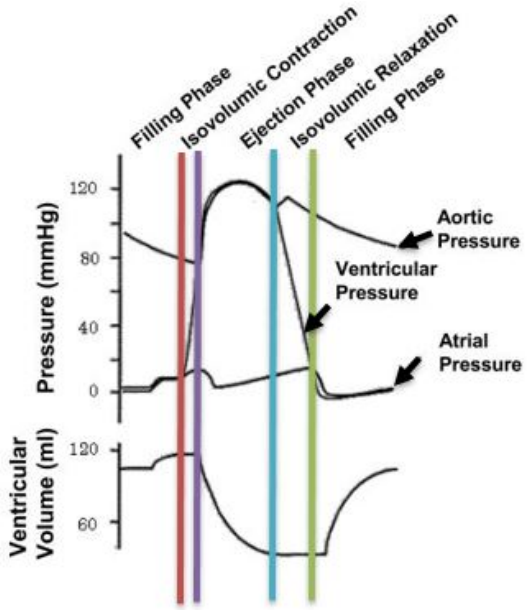

Figure 2.7: Key Pressure and Volume curves for 1 Cardiac Cycle

- Pressure in the heart is low as it becomes completely relaxed (ventricular diastole). Both AV valves are open, and blood actively fills the ventricles from the atria. Approximately 70% of atrial reserves enter the ventricles before atrial contraction begins, forcing out much of the remainder. Semi-lunar valves are closed during ventricular *Filling*.

- Contraction (ventricular systole) follows and ventricular pressure rapidly increases. Once atrial pressure is exceeded, AV valves close and ventricles begin *Isovolumic Contraction* (as both valves are closed) and ventricular pressure increases with no change in volume.

- When ventricular pressure exceeds pressure in the large arteries, semi-lunar valves open and *Ejection* begins. Low atrial pressures allow atrial

filling to begin for the next cardiac cycle. Ventricular pressure continues to rise (as the volume now decreases), reaching a maximum before declining (onset of ventricular diastole).

- When ventricular pressure is below that of major arteries, semi-lunar valves close and *Isovolumic Relaxation* occurs. The next cardiac cycle begins as ventricular pressure fall below atrial pressure and AV valves re-open.

2.1.4.2 Pressure-Volume Loops & Relationships

The cardiac cycle for the left ventricle can be represented as a Pressure vs. Volume plot. This *PV-Loop*, seen in Figure 2.8, is read chronologically in an anti-clockwise direction. The phase labels and coloured dots on the loop coincide with the coloured lines in Figure 2.7, (except the pink dot). The abbreviated labels are explained as follows:

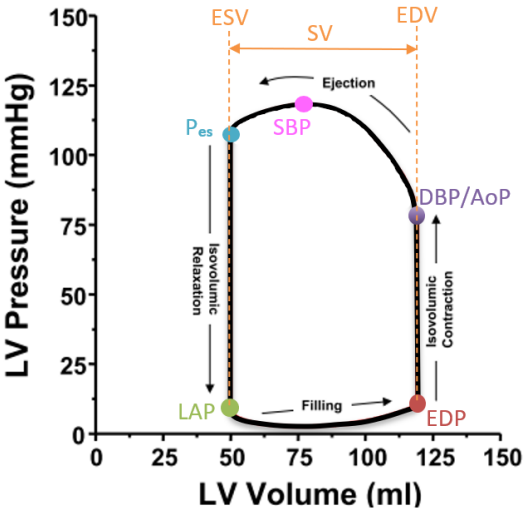

Figure 2.8: Pressure-Volume Loop of the Left Ventricle for the Cardiac Cycle (Burkhoff, 2002)

- P_{es}: End-systolic pressure of the ventricle
- LAP: Left atrial pressure; below which AV valves open and ventricular filling begins.
- EDP: End-diastolic pressure; the point before ventricular systole where there is maximal volume at minimal pressure.
- DBP/AoP: Diastolic blood pressure; occurs when ventricular pressure equals aortic pressure (AoP) where semi-lunar valves open and ejection begins.

- SBP: Systolic blood pressure; the highest pressure attained during the cardiac cycle (crest of ventricular and aortic pressure curves in Figure 2.7).
- ESV: End-systolic volume or minimum ventricular volume attained by the end of ejection.
- EDV: End-diastolic volume or maximum ventricular volume as filling ends.
- SV: Stroke volume; the volume of blood ejected per cardiac.

A host of properties / relationships can be derived from the PV-loops about the heart. Some examples include elastance, compliance, end-diastolic and end-systolic pressure volume relationships (EDPVR and ESPVR, respectively). Many of these relationships require pressure readings not included in the data acquired, thus only those utilised for later calculations are discussed. These include stroke volume, cardiac output and ejection fraction - all considerably important parameters in practise and among other studies (Bizopoulos and Koutsouris, 2019).

Blood Pressure:
Blood pressure is formally defined as the force exerted by blood against the interior walls of blood vessels; usually referring to pressure in the arteries - where it is highest (Marieb, 2015). Arterial pressures mimic those of the aorta and heart as it moves through phases of the cardiac cycle. The arterial walls distend upon blood inflow but recoil quickly thereafter. Blood pressure is affected by both blood volume and peripheral resistance. Pressure gradients are responsible for driving passive blood flow through the circulatory system (Walley, 2016).

Stroke Volume (SV):
Arterial blood pressure is a function of heart rate (HR), volume of blood, peripheral resistance (PR), blood viscosity and stroke volume. As seen in Figure 2.8, stroke volume is calculated by the difference between the maximum (EDV) and minimum (ESV) volumes of the left ventricle, as per Equation 2.1.1:

$$SV = EDV - ESV \tag{2.1.1}$$

Cardiac Output (CO):
The cardiac output of a heart refers to the volume of blood discharged from the left ventricle per minute; calculated by the product of SV and HR, as per Equation 2.1.2:

$$CO = SV * HR \tag{2.1.2}$$

Ejection Fraction (EF):
In clinical practice, one of the older and most employed measures of contractility is *ejection fraction*. Ejection fraction is the ratio between the blood volume ejected to the maximum ventricular volume; defined by Equation 2.1.3:

$$EF = \frac{SV}{EDV} * 100 \tag{2.1.3}$$

As per the American Heart Association (2017), normal ranges of ejection fractions lie between 55-70%; estimated by techniques such as nuclear imaging, cardiac catheterization, CT scans or echocardiography. Certain abnormalities such as Heart Failure result in ejection fractions below 40%. Ejection fraction is still the much-preferred contractility index used in practice today as a result of its firm knowledge foundations and long use history (Silva *et al.*, 2018).

2.1.5 Echocardiography

Echocardiography is the practise of utilising an ultrasound imaging modality to non-invasively observe functionality of the human heart (Bizopoulos and Koutsouris, 2019). When an echocardiogram is performed, the internal structures, functioning and size of the heart can be viewed in real-time for the purpose of examination (The Heart Foundation, 2017). One of the most effective tools for accurate diagnoses (Jeanrenaud *et al.*, 2015), standard 2D echocardiograms are routinely performed and are the preferred option for identification of cardiovascular diseases or normal functioning of all structural components (Mandes *et al.*, 2020).

Standard transthoracic echocardiograms are conducted by cardiac sonographers and involve moving a probe, over the specific windows of the chest's surface, shown left in Figure 2.9 (Lohr and Sivanandam, 2015). Complete echocardiogram examinations involve capturing around 12-16 different views of various anatomical structures; based on particular probe orientations at the chest windows. Cardiologists require a combination of technical skill, attention to detail, and holistic understanding of cardiac physiology, anatomy, and physiopathology to diagnose a patient (Jeanrenaud *et al.*, 2015). For this project, *apical 2-chamber* views from standard transthoracic echocardiograms were used to extract information of left heart structures, as per Figure 2.9 (right).

2.1.6 Cardiac Abnormalities

Many cardiovascular diseases (CVDs) that occur frequently today are due to poor lifestyle choices or aging, may be hereditary, or a result of other diseases. CVDs related to the heart specifically can occur in any of the subsystems previously discussed. For example, electrical disorders (*arrhythmias*) are

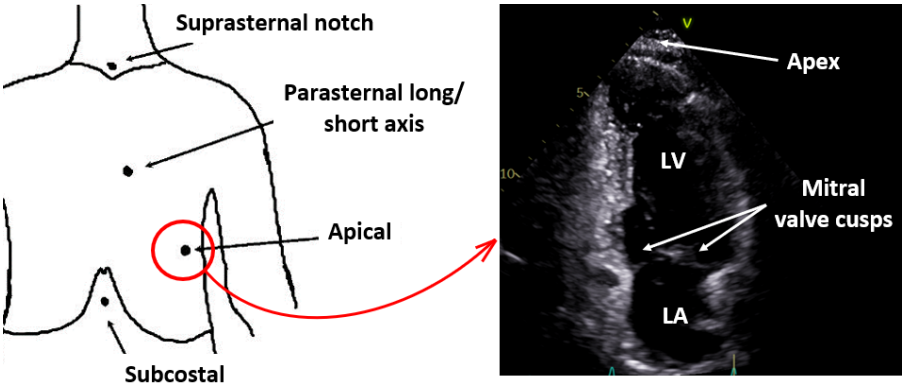

Figure 2.9: Standard chest windows for transducer positioning (left) and a labelled apical 2-chamber view (right)

due to issues of the conduction system that result in irregular, disorganized heartbeats. In this section, special attention is given to the cardiac abnormalities included in the study. All content of previous sections described normal functioning hearts, against which the selected abnormalities are contrasted (American Heart Association, 2017):

1. **Heart Failure (HF):**
   Heart failure occurs when cardiac muscles are not strong enough to efficiently pump blood to meet the demand of the body. It is a chronic condition that continually progresses, forcing the heart and body to compensate in other ways. The heart attempts to increase cardiac output by enlarging (chamber volume), increasing muscle mass or simply beating faster. The body compensates by narrowing blood vessels (to increase pressure) or diverting blood away from non-vital organs. As the heart and body cope less over time, the following mild symptoms surface: wet coughs, shortness of breath during activity, swelling of lower limbs, rapid weight gain, abdominal swelling/discomfort and trouble sleeping - all of which intensify as the condition worsens. Heart failure with preserved EF occurs when EF is normal; but can occur with reduced EF, where EF is around 30-40%.

2. **Left Ventricular Hypertrophy (LVH):**
   Hypertrophy refers to the enlargement/thickening of cardiac muscle cells (of the left ventricle) which cause inefficient pump functioning. Similar to heart failure, the muscles of the left ventricle expand to cope with the experienced demand. This works temporarily, but as the ventricle thickens it becomes progressively weak and less elastic/stiff. The symptoms of LVH, such as shortness of breath, fatigue or chest pain (in severe cases) overlap with other heart conditions; and is accompanied by heart palpitations or dizziness. LVH typically occurs due to high blood pressures, diabetes, or heart valve issues.

3. **Myocardial Infarction (MI):**
   Myocardial Infarction (heart attack) is caused by narrowed coronary arteries (due to blood clots or other causes) that restrict blood flow / oxygen supply to cardiac muscles. The muscles are damaged or killed, thus causing a heart attack. The extent of damage is a function of how long the vessel was blocked, time between the attack and medical intervention, and the size of the area affected. Heart attacks are the result of circulation issues, however can seldom be attributed to coronary artery spasms or spontaneous tearing. Signs of an impending heart attack include periodic chest pains; pain or discomfort in one/both arms, neck, back, stomach or jaw; shortness of breath; sporadic cold sweats; lightheadedness or nauseousness.

4. **Heart Valve Complications:**
   Typically hereditary or self-developing, issues of any valve affects cardiac output. Treatments of valve abnormalities usually involve blood-thinning medications or valve replacement surgery. In this study, 2 aortic valve instances were included:

   - **Aortic Regurgitation (AR):**
     Aortic (valve) regurgitation refers to a leaking aortic valve, in which blood pumped out of the left ventricle leaks back in during diastole. Since less blood (thus less oxygen) is pumped to the body, the heart compensates for the difference by increasing work done. Ventricle walls may thicken (as with LVH) resulting in ineffective pump functioning as cardiac muscles become fatigued and stiffen over time. It is mostly caused by high blood pressure, aging valve tissues, injury, untreated syphilis or a bacterial infection; resulting in symptoms that overlap with HF and LVH.
   - **Aortic Stenosis (AS):**
     Aortic stenosis refers to atypical narrowing of the aortic valve opening. Consequently, blood flow and oxygen supply to the body are restricted, and left atrial pressure may be affected. It is often the result of bicuspid aortic valves; a congenital heart defect that develops with age. Valve cusps may experience scarring or calcium accumulation that narrows the valve opening. Symptoms overlap with HF, with fluttering heartbeats and diminishing ability to do daily tasks. AS can develop further by muscular thickening of the left ventricle as a compensation measure to increase cardiac output - ultimately leading to HF.

2.2 Applied Machine Learning

This section includes literature required for the machine learning aspects of the study. The theory summarises processing steps implemented from medical data handling to the input requirements and evaluation methods of the selected models. For the remainder of this section, Müller and Guido (2016) was the main reference unless stated otherwise.

2.2.1 Medical Image Processing

Medical data can be collected from a multitude of (primary or secondary) sources in both qualitative and quantitative formats (NIH, 2021). Some of these formats include measurements, visual or audio formats, electrical signals or physical samples. Echocardiogram data specifically, is output in the form of images and videos. To be used as appropriate input data to any machine learning algorithm, various pre-processing steps must occur - with specifics being a function of the algorithm used. In the case of Random Forests and Neural Networks, features from the images must be extracted and input as 1D arrays - necessitating the use of computer vision techniques.

Computer Vision is a branch of computer science that aims to create systems that mimic the manner in which humans analyse images. These systems enable computers to analyse images at pixel-level before they can *understand* the content, and thus recognise or interpret graphic data (Babich, 2020). Common computer vision applications can be categorised into object detection, object classification or object tracking; with examples in the works of Chen *et al.* (2014), (Panda, 2018), and Jiang *et al.* (2020), respectively. Based on the underlining objective, many methods and platforms exist for image processing. Commonly applied steps include binarization, morphological operations and edge detections; all adopted in this project and described as follows:

Binarization

As per Thapliyal *et al.* (2017), binarization was a crucial first step for successive edge detection steps. Binarizing an image involves converting the pixel values of an image to be 1 of 2 values; namely 255 (white) or 0 (black). The degree of binarization is a function of a predefined threshold applied to image pixels. An example of a binarized image can be found in Figure 2.10.

Morphological Operators

Morphological operations are applied sets of kernels (matrices) that achieve various image effects by manipulating pixel gradients. Common examples include embossing, altering contrast or sharpening effects; less common examples include noise injection, dilation and erosion. Noise injection involves adding

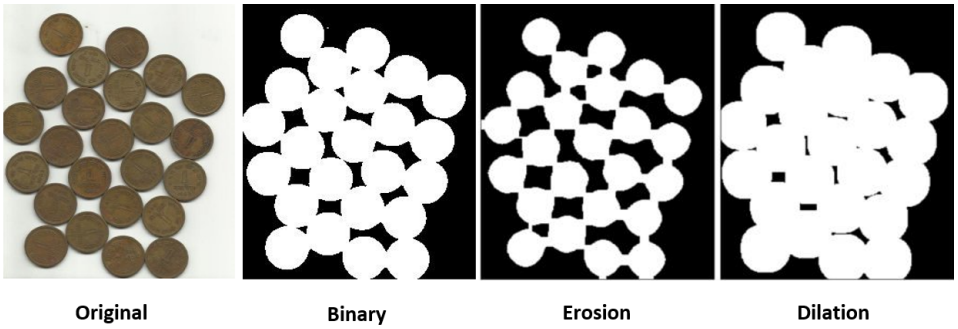

Figure 2.10: Example of Binarization, Erosion and Dilution effects on an image (OpenCV, 2021)

a randomly initialized 2D matrix (containing values in the range of pixel values) to an image, introducing haphazard colour/brightness variations (Portilla, 2018). Dilations and erosions are opposing functions; dilation serves to extend or thicken the boundaries of image features, while erosion shrinks boundaries, as seen in Figure 2.10. Dilations are performed by increasing pixel values of a neighbourhood (of some kernel size) to the highest pixel value within the neighbourhood; whereas erosions assign the lowest pixel values.

Edge Detection: Watershed Algorithm

The Watershed algorithm is a method of image segmentation, whereby objects in an image are distinguished from one another. The algorithm is based on the Watershed Transformation: Gradients of greyscale images are considered to resemble a topographic map. The map consists local minima (dark) and maxima (light), defined by pixel values. Assuming the topographic region floods - where hypothetical water fills the region submerging the global minimum first - until *sure* local maxima are left; providing information of where boundaries exist. An example of this is taken from OpenCV (2021) documentation, using the original image of Figure 2.10, various stages of the Watershed algorithm can be seen in Figure 2.11.

2.2.2 Machine Learning

Machine learning entails obtaining knowledge from data and is the meeting point of fields such as mathematics, statistics and computer science. There are countless applications of machine learning, limited only by human imagination and resources. In this technological era, machine learning is present in many facets of everyday life. Some examples encountered daily are embedded in cell phone applications, search engines, social media, production lines and telecommunication systems (Vashistha, 2019). Data is harvested through these platforms to improve the machine learning algorithms operating behind the scenes or to drive new research and developments.

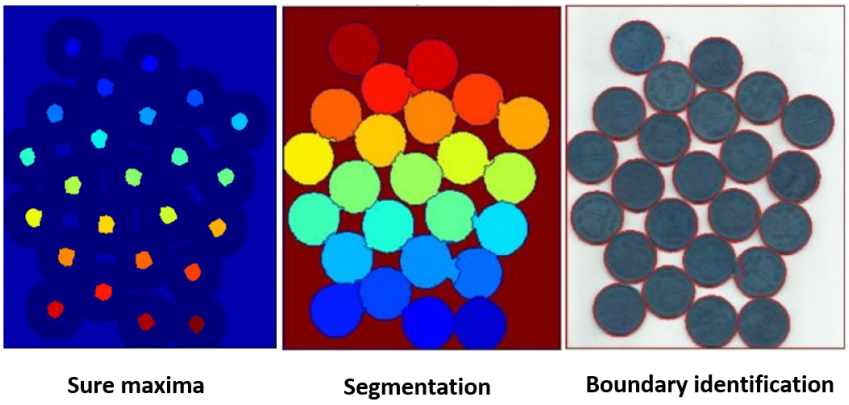

Figure 2.11: Intermediary steps of the Watershed algorithm (OpenCV, 2021)

2.2.2.1 Supervised Machine Learning

Machine learning algorithms are either *unsupervised* or *supervised*, based on the manner in which they *learn*. Unsupervised learning involves providing an algorithm with input data in which patterns must be found without prior knowledge of what to look for. Supervised learning involves model learning guided by expected outputs provided for each sample (Bizopoulos and Koutsouris, 2019). Supervised machine learning problems can be categorised as either *regression* or *classification*. Regression tasks are associated with predictions of continuous or real numbers, such as the prediction of housing prices (Kaggle, 2016). Classification tasks are those where the correct label, or *class*, must be predicted from a list of predefined classes. Examples include a *binary* (2 classes) classification task for emails identified as *spam* or *not spam*, respectively (Awad, 2011).

Many classification problems are *multi-class*; having more than 2 output classes. Each sample fits into 1 of many classes; such as classification of fruit, where each sample can only have one label. Other classification problems are *multi-label*, whereby samples in the data belong to more than 1 class; such as topic prediction for text or video sample, with many applicable labels like *religion* and *politics* (Scikit-learn, 2007). *Multi-output* problems are both multi-class and multi-label; as is with the classification task of this research where a fraction of the input data contains 2/3 labels (some patients have more than one abnormality).

The supervised machine learning models selected for this study include Random Forests and Neural Networks. Both methods are among the supervised machine learning algorithms in Python's *Scikit-learn* package capable of handling multi-output classification problems.

2.2.2.2 Random Forests

Random Forests (RF) is an ensemble method whereby multiple models (viz. Decision Trees) are combined to improve overall performance of the single model. Decision Tree models make decisions based on a learned hierarchy but tend to overfit the training data (explained in §2.2.5.1). Random Forests circumvent this tendency by its internal structure. For the specified number of trees (*estimators*) per forest, each tree is built with randomness injected to force variation among all trees. Each tree in the forest makes its own prediction on a random subset of the input data (known as "bagging" or Bootstrap Aggregation). With each individual tree slightly overfitting the data in differing ways, their averaged results serve to diminish overfitting in the forest. This is based on the concept of the "wisdom of crowds", where a committee of uncorrelated trees outperform individual trees (Yui, 2019).

Random Forests have various strengths and weaknesses identified in practise. One of their major advantages is they make up for the deficits of individual Decision Trees, with regards to overfitting while expanding overall capacity. They are also among the most popular machine learning methods; requiring minimal hyperparameter tuning or data scaling and operating well on default parameters. Some disadvantages include tendencies to perform poorly on sparse or considerably high dimensional datasets, they are more computationally expensive (due to size), and results are more difficult to interpret.

2.2.2.3 Neural Networks

Neural Networks (NN) refer to a collection of deep learning models inspired by biological principles and how the human brain learns. The building block of Neural Networks is the perceptron. They are comprised of numerous perceptrons in multiple layers that perform different steps of processing before the network arrives at a decision. Neural Network internal structures imitate neurons and neural pathways of the brain, seen in Figure 2.12.

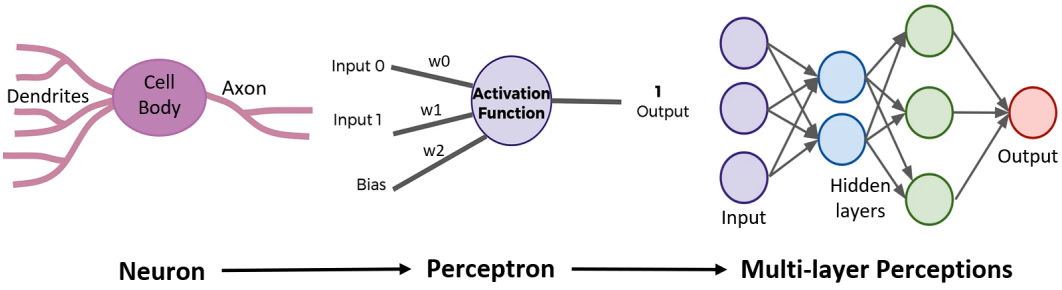

Figure 2.12: Inspiration of Neurons for Perceptrons of a Neural Network (Portilla, 2018)

Each perceptron applies some nonlinear (*activation*) function to its learned weights (w), bias term (b) and data variables (x); as per Equation 2.2.1 (Portilla, 2018). There is one weight between every unit (perceptron) in one layer and every unit in the next immediate layer. These models, thus, have large computational loads given the number of weights to learn. The activation functions applied at each unit is usually the rectified linear unit (ReLU) or sigmoid/hyperbolic tangent (tanh). In this project, and typically for classification tasks, both activation functions are used; ReLU for all hidden layer units except the second last layer where the sigmoid function is used (Brownlee, 2016).

$$\sum_{i=0}^{n} w_i x_i + b \tag{2.2.1}$$

One of the advantages of Neural Networks is that they are highly customisable with many parameters to tune to improve model performances - with many defaults also working relatively well. Another advantageous property is the random initialisation of weights at the onset of training and the inclusion of bias - facilitating better fits as training progresses. The disadvantages are that Neural Network training is computationally expensive, they are very sensitive to data scales and require relatively homogeneous data (where features have similar value).

2.2.3 Data Augmentation Techniques

Large datasets are one of the most effective and reliable ways to improve the performance of any machine learning algorithm; especially for complex (inherent patterns) datasets or complex (large) models. However, since ideal data amounts may be in the order of thousands, data acquisition is capped under generally limited conditions (Ng, 2015). One solution is *data augmentation*; whereby synthetically modified data is generated from original data to increase the amount of samples (m) available for model training and testing (Nolen, 2019). For image data, one way to synthesize data is to apply transformations (translations, rotations, reflections) or inject noise. Data augmentations ensure sufficient data for model development, equalise slightly skewed datasets, or aid in improving model robustness to outliers (Raj, 2018) and noise (Litjens *et al.*, 2019). For this research, the data was augmented by the following means using *OpenCV* functions:

- Clockwise and counter-clockwise **rotations** were applied to input images about image centres.
- **x- and y-translations** involve offsetting all pixels by variables specified for either direction.

- **Black noise injections** involve the addition of randomly located black pixels to the image, distorting the lighter (grey/white) foreground.

Suitable augments were chosen to synthesize data while preserving the essence of the original images in latter steps.

2.2.4 Feature Selection

Features are simply all variables present in a dataset. Feature selection, therefore, refers to techniques of evaluating all features of a dataset to identify the most informative ones. Features are regarded as informative based on the degree to which they inform the model about the desired output. Careful selection involves removing or creating features that improve model performance (Guhanesvar, 2021). There are 3 strategies used in practise for feature selection:

1. **Univariate Statistics**:
   Univariate statistical methods use confidence values to determine correlations between individual features and target variable/s. One example is *Mutual Information* (Information Theory), defined as the measure of uncertainty (entropy) between 2 variables. It evaluates the degree to which the (known) feature reduces uncertainty about the (unknown) target. For multi-class classification, mutual information scores are determined between each feature and each class. Mutual information scores are non-negative: higher values (never greater than 2) indicate stronger dependency, lower values indicate lower dependency and 0 implying mutual independence. They are simple to implement and interpret, computationally efficient, and robust to relationships of higher orders (Holbrook, 2021). Score are obtained from training sets and may inform successive decisions on model selection/parameters.

2. **Model-based Feature Selection**:
   This method entails using another supervised learning model to assess the importance of each feature before making a selection; thus more powerful than univariate tests. All features are considered altogether with their respective inter-feature relationships. Further details on this method are excluded as it was not used in the study.

3. **Iterative Feature Selection**:
   Multiple models are built with differing combinations of features included from their training datasets. For this approach, 1 of the following 2 iterative methods are undertaken until some predefined criteria is satisfied:

   - Either begin with 0 features and adding 1 at a time, or
   - Begin with all features included and remove 1 at a time

Features are only as informative to the extent that the model implemented can learn its relation to the target. In some cases, features may need further engineering to reveal associations learnable to a model. Feature engineering or selections must be guided by specific domain knowledge. Appropriate consideration must be given to features best for the model and those most relevant to the domain. This ensures models are effective when deployed into industry/practice.

2.2.5 Evaluation Methods

A vital part of model development is keeping in mind the overall goal (*business metric*) and the consequences of the model outputs (*business impact*). Model parameters, features and architectures should be decided based on what is most favourable in these two aspects. Another key aspect is robust evaluation; to ensure model performances are meaningful for their intended context. The errors/performances produced must be properly represented and accounted for with respect to the context. For medical practise, the type of errors that occur hold varying implications as human lives and resources are at stake.

2.2.5.1 Performance Metrics

Performance metric scores are some of the many ways to evaluate the predicted outputs of a model. Different scores reveal different characteristics / capabilities of a model; valuable for assessing key aspects of its behaviour. These assessments direct further improvements and/or give indications about how the model will perform in its desired context. From the consortium of performance metrics available, the Confusion Matrix and some of its derived metrics were utilized. Confusion Matrices are among the most comprehensive approaches when evaluating binary classification results with a *positive* and *negative* class. It is constructed by comparing each prediction of the model against the true labels (for each class in multi-class instances) and classifying the prediction in one of the following categories:

1. True Negative (TN): correctly predicted negative sample
2. False Positive (FP): incorrectly predicted positive (true label is negative)
3. False Negative (FN): incorrectly predicted negative (true label is positive)
4. True Positive (TP): correctly predicted positive sample

Categories are then tallied and represented in a Confusion Matrix, as seen in Figure 2.13. From different combinations of these tallies, the following performance metrics are defined:

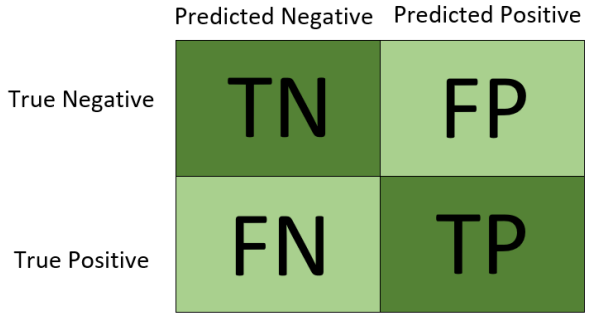

Figure 2.13: Confusion Matrix layout for Binary Classification

- **Accuracy**: a measure of all correct predictions as a fraction of the total number of predictions; defined by Equation 2.2.2

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.2.2}$$

- **Precision**: correctly predicted positives as a fraction of all (correctly and incorrectly) predicted positives; defined in Equation 2.2.3

$$Precision = \frac{TP}{TP + FP} \tag{2.2.3}$$

  Precision is used to assess the number of false positives predicted with the goal of reducing them.

- **Recall**: correctly predicted positives as a fraction of all true positives; defined by Equation 2.2.4

$$Recall = \frac{TP}{TP + FN} \tag{2.2.4}$$

  Recall is used to monitor the occurrences of false negative predictions.

- **F1-score**: a measure of the trade-off between Recall (false negatives) and Precision (false positives), providing insight about a model's false predictions. The F1-score refers to the F_β score in Equation 2.2.5 when $\beta = 1$; representing the harmonic mean between Recall and Precision.

$$F_\beta = (1 + \beta^2)\frac{Precision * Recall}{(\beta^2 Precision) + Recall} \tag{2.2.5}$$

$$F_{\beta=1} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.2.6}$$

For multi-class classification problems, all performance metrics are obtained by treating the problem as a collection of binary classification problems (equal to the number of output classes) and averaging across all classes. Average weightings are varied based on how imbalanced the dataset is.

The information present in all datasets can be classified as *signal* (informative) or *noise* (non-informative). The noise present in a dataset can either assist in building robust models or be detrimental by obscuring underlying patterns or contributing to model error. Error that measures how much the output differs from the target value is referred to as *Loss*. For each epoch (iteration) during model training, loss can be calculated as a quantitative measure (Jose, 2019). For this project, *binary cross-entropy* was used as the loss function, as it accommodates multi-output classification (TensorFlow, 2021).

2.2.5.2 Learning Curves

A Learning Curve is a mathematical representation of a model's performance over time (*experience*) (Brownlee, 2019). Before being adopted into machine learning, Learning Curves were used as to measure the effect on production when (i) engineering changes or (ii) workforce training was introduced (Adler and Clark, 1991). Learning curves were thus created as tools to assess the performance of staff over time exposed to a new variable in their daily tasks (Anzanello and Fogliatto, 2011). As repetitions continued, workers took less time to perform the new tasks and production increased with familiarity. Learning curves are used in a similar way for diagnosing machine learning models. During training, performance metrics are tracked per iteration as the training subset increases incrementally (per *epoch*). Loss and Accuracy are most commonly used to construct the Learning curves of the model and diagnose learning behaviours (Jose, 2019), such as the following:

Learning Rates

The Learning Rate refers to extent to which weights of a model are adjusted per epoch. They are identified by the rate of change of the (training) Loss curve gradient. Some examples of learning rates deduced from Loss curves are seen in Figure 2.14. The lower learning rates appear more linear, but become exponential with further increases. Tuning the learning rate forms part of the optimisation strategy for a given model and can positively affect performance if selected carefully (Peace *et al.*, 2015).

Model generalisation

Poor predictive abilities arise when models *overfit* or *underfit* their training data; reducing their ability to generalise new data. Overfitting occurs when the model, having excess complexity (model size), learns the noise or particularities of the training data. The performance on the training set is typically much better in comparison to that of the test set. Underfitting occurs when the model is not sufficiently complex to capture inherent patterns of the data, performing poorly for both the train and test sets. The ideal model generalisation is the sweet-spot between these two extremes.

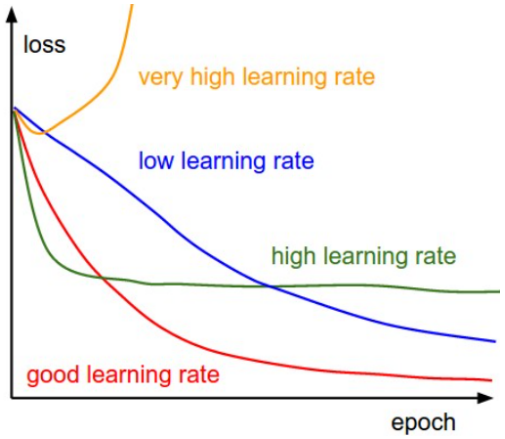

Figure 2.14: Loss Learning Curves with differing Learning Rates (Venugopal and Ramaswamy, 2015)

Model generalisation is diagnosed by the shape of the performance metric curves of both training and validation subsets. For example, Figure 2.15 shows sketches of hypothetical accuracy curves. The gaps present between the training and validation curves is indicative of the degree of overfitting present (Jose, 2019). Ideally, accuracy (and other performance metrics) training curves are minimally higher than their corresponding validation curves; although for loss curves, training curves are ideally minimally lower than validation curves. Underfit models display training and validation curves at low values that (may temporarily increase but) usually continue decreasing as learning progresses (Muralidhar, 2021).

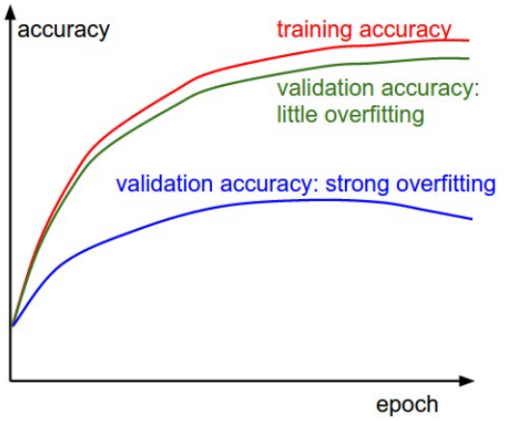

Figure 2.15: Degrees of Overfitting observed from Accuracy Learning Curves (Venugopal and Ramaswamy, 2015)

Data Representativeness

Data representativeness refers to how well a data subset (training and/or validation) captures the statistical attributes present in another subset drawn from the same domain (Brownlee, 2019). Unrepresentative data is usually a case of insufficient samples in one subset compared to the other, such as in the case where data classes are imbalanced. This can be seen when comparing training and validation curves for the same metric. Referring to Loss curves, unrepresentative data can be distinguished as follows:

- **Training data**: When the training subset is unrepresentative, it does not contain enough information (samples or learnable patterns) for the model to learn or generalise for the given validation subset. The loss learning curves for this instance contain a consistent considerable gap between the training and validation curves despite both showing improvement over time.
- **Validation data**: When the validation subset is unrepresentative, the model's ability to generalise cannot be properly evaluated. The validation data is either too easy or not sufficiently related (to the training data) for the model to predict. This occurs when the validation subset contains too few samples compared to the training set. On learning curves, this may be identified by (i) the presence of a lower validation than training loss; or (ii) if the validation loss is consistently noisy relative to the training loss.

3 Methodology

The following chapter details the approach executed to meet the aims and objectives of this research project. The most important factors regarding the study population chosen are discussed in §3.1 as per internal (Stellenbosch University) and external (sourcing hospital) institutional requirements for the use of human subjects in research. Previous work (§3.2), thereafter, contains a discussion on other endeavours investigating the use of machine learning in cardiology, specifically those used to identify pathologies. Lastly, Experimental Work (§3.3) includes a detailed discussion on practical steps involving all work with acquired data and the various machine learning models. The methods for each phase from input preparations to model development and evaluations is detailed in this section. The following terms are used interchangeably for the remainder of this report: (i) *target variables* and *labels* and (ii) *test* and *validation*. As per request of the sourcing hospital and associated staff, any identifying information was not to be included in this report.

3.1 Sample Selection

Since data of human subjects were used in this study, Ethical Clearance was required both internally and externally. In the proposal submitted to both Ethics committees, assurance of involved persons identity protection and demographic diversity / bias negation were among the most important considerations before final clearance could be granted. Details on (i) Randomisation, Confidentiality and Bias measures, (ii) Data Collection and Management and (iii) Project Commencement Plan can be found in Appendix D; in §D.1.1, §D.1.2 and §D.1.3, respectively. The remainder on this section discusses how the study population was selected with the removal of bias by means of an Exclusion-Inclusion criterion.

The Study Population is a target group of patients that satisfy a unique criterion that qualify them as observable subjects in answering a research question (Garg, 2016). The patients included were sampled from a larger population from the sourcing hospital's echocardiogram archives. Generally, for clinical trials, sampling is restricted to what is available from the source, some pre-

defined criteria or desired sample size (determined statistically). However, in machine learning applications, there are generally no upper bounds on the amount of data to use (Ng, 2015). The sample size needed depends on the complexity of the model or problem. The sample size must be sufficiently large enough to represent the population. The higher the quality and wider the variety of data available to train a model, the greater the corresponding performance with regard to predictive capabilities.

Patients were selected for the study by means of an Exclusion-Inclusion criterion - seen in Table 3.1. An Exclusion-Inclusion criterion is used as part of a clinical trial to identify eligible patients who can/cannot be considered for the study in an objective and consistent manner. Patients recruited in this manner ensure suitability (to the study) and minimal bias of the study population. Further details regarding randomisation measures, confidentiality considerations and bias mitigations can be found in Appendix D (§D.1.1).

Table 3.1: Exclusion-Inclusion Criterion

Exclusion	**Inclusion**
Patient cases lacking apical 2-chamber views	Patient cases with apical 2-chamber views
Patients tested before 2016	Patients tested between 2016-2020
Patient cases where noise heavily distorts video frames	Patients diagnosed with a shortlisted pathology

3.2 Previous Work: Machine Learning in Cardiology

Previously, diagnosis of a patient rested solely on the shoulders of a medical professional or physician. Ultimately, diagnosis is a function of their assessment method, training, experience, access to medical history and suitable equipment (Bizopoulos and Koutsouris, 2019). Physicians proceed to subjectively interpret and match the patient's information to some traditional taxonomy of medical conditions. Coupled with earlier, less sophisticated imaging techniques and significant manual tuning, these methods collectively serve to exacerbate errors. Before the advent of advanced imaging modalities in the domain of cardiology, relevant clinical indicators were obtained off cardiovascular images in the same error-prone manner. Since the introduction of machine learning into specialised medical and routine clinical practices, modern systems allow physicians to capture information more accurately (Litjens *et al.*, 2019).

The remainder of this subsection summarises a range of recent works with respect to cardiology, echocardiography and automated diagnostics research.

Venugopal and Ramaswamy (2015) sought to investigate a method hypothesized to allow early detection of heart disease. The dataset, sourced during the 2015 Kaggle Data Science Bowl (Kaggle, 2015), consisted of 500 anonymised MRI's with 30 time series images each. This data was fed into convolutional neural network (CNN) models of differing architectures to estimate and predict volumes associated with different stages of the cardiac cycle. The output was then used to assess how likely a patient was to experience heart disease, evaluated using the prescribed metric of Continuous Ranked Probability Score (CRPS) - description in Appendix D.2. Their best model was a 7-layered network that resulted in a 0.032 CRPS (where smaller scores are desired), earning the duo a place in the Top 20%.

Walley (2016) investigated left-ventricular functioning from a physiological point of view by measuring and characterising time-varying elastances. Incorporating physiological relationships, such as those defined in §2.1.4, new insights were uncovered about ventricular function and its role in regulating cardiac output. The match or mismatch of ventricular and aortic elastances were key in qualifying mechanical loads on the heart resulting from vasculature interactions. Another study by Bozkurt (2019) also used mathematical modelling to assess cardiac function by making use of heart chamber geometries. Inclusion of physiological relationships further allowed haemodynamic indicators to be incorporated. LV ejection fraction, end-diastolic and end-systolic volumes and sphericity indices could also be estimated. This study successfully proved a feasible model for healthy and dilated cardiomyopathy (DCM) cases in both adults and children.

In 2018, much research was published investigating automated echocardiogram diagnoses in efforts to further improve predictive accuracy of machine learning models. Three such examples follow:

- Two such studies were undertaken by Madani *et al.* (2018) and Zhang *et al.* (2018). Madani *et al.* (2018) sought to investigate a deep learning solution addressing the issue of unannotated data and the lack of accessible databases - typical of medical image data. Using supervised and semi-supervised learning approaches, they developed 2 models to perform view classification and LV hypertrophy classification on echocardiogram images. They conclude their study with deep learning solutions for cardiac assessments made from medical imaging.

- Zhang *et al.* (2018) designed a similar model to include cardiac chamber segmentation and detection of 3 additional pathologies. Segmentation

results were used to estimate LV volume, mass and ejection fractions. Additionally, longitudinal strains were deduced using speckle tracking techniques. The model built was of a CNN architecture trained on >14 000 echocardiograms for all the various tasks with impressive results obtained.

- A study by Silva *et al.* (2018) used a 3D-CNN to classify ejection fractions into 4 classes; unhealthy, intermediate, healthy and abnormally high. A dataset of 5600 transthoracic exams was used to train and test the model. The exams consisted of apical 4-chamber view cine-loops consisting of 30 sequential frames each. Their study investigated interesting architecture adjustments, exploring the effects of asymmetric convolution filters and residual learning blocks.

Ghorbani *et al.* (2020) stretched the capabilities of their deep learning model, EchoNet, in the identification of cardiac structures, cardiac function estimation and prediction of systemic phenotypes, exceeding the scope of expert interpretations. The model could predict certain abnormalities, LV end-systolic and end-diastolic volumes, and systemic phenotypes of sex, age, height and weight from echocardiogram information only. An interpretation analysis validated that EchoNet could identify regions of interest (ROIs) coinciding with those observed in practice by cardiologists.

3.3 Experimental Work

This section focuses on the practical work-flow of the project, design decisions and methodology of individual steps. It documents the Image Pre-processing (§3.3.1) steps followed prior to Model Input Preparation (§3.3.2), which expands on model features extracted. Model Development details model parameters and iterative training procedures. Lastly, Model Performance Evaluation (§3.3.4) explains the approach used to assess the model performances and predictions on test data. Finally, considering all tests, the most successful model instances are compared before a Final Model Selection (§3.3.5) concludes the practical workflow.

A schematic representation of the solution pipeline utilised in this project can be seen in Figure 3.1. The main work order is denoted by the bold chain (text and arrows) in the diagram, from *Start* to *End*. The flow of information and related inputs/outputs of each step are included in side branches. The groupings (colour blocks) in the schematic coincide with successive sections for the remainder of this chapter and Chapter 4.

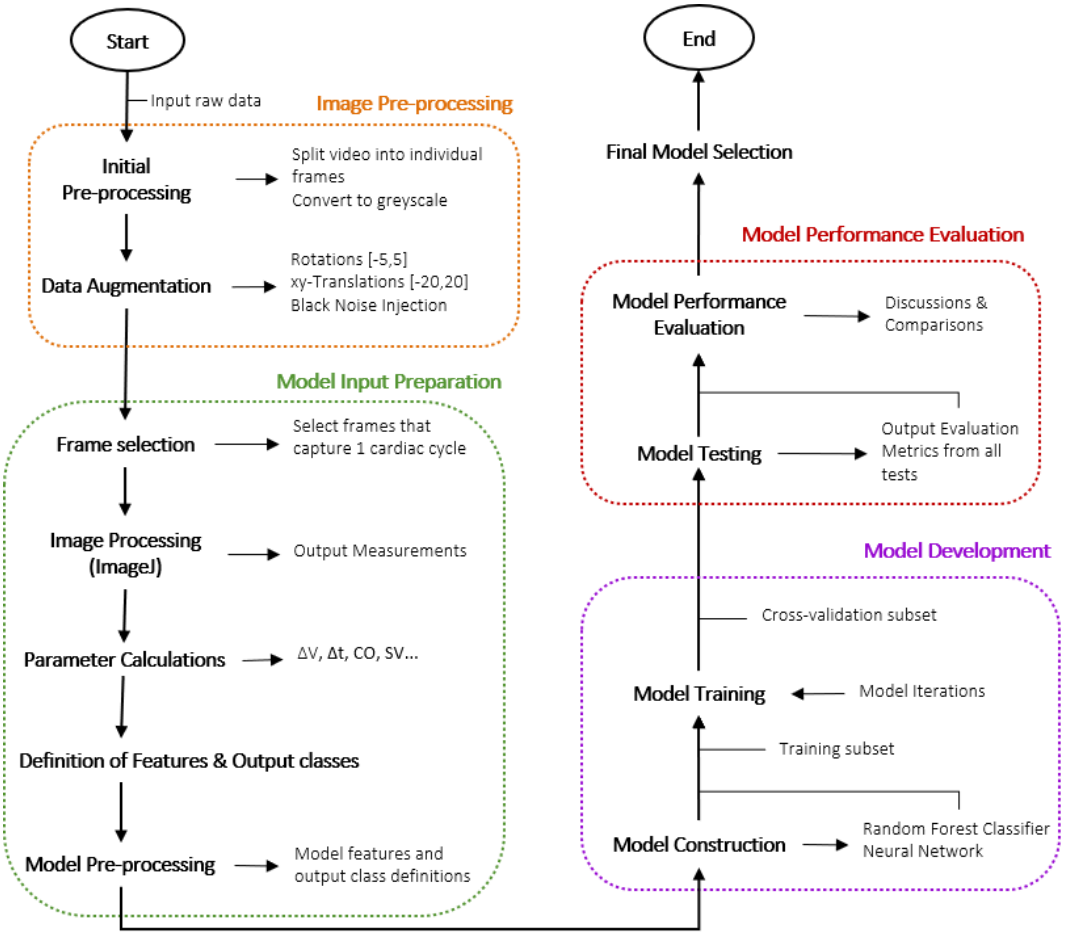

Figure 3.1: Solution Pipeline

3.3.1 Image Pre-processing

The pre-processing procedures involved with "neatening" the raw data derived from the videos, before being prepared for the model, have been divided into the following steps:

1. **Initial Processing**
   - Video Frame Isolation: all video data was separated into individual frames before successive augmentations could take place.
   - Conversion to Greyscale: this was done in preparation for the binarization step that forms part of later image processing.
2. **Data Augmentation**
   Data obtained from the sourcing hospital was considered insufficient when compared to the typical amounts of data required to train simple-to-complex machine learning algorithms (as per examples in §3.2). In this case, only 406 patient cases (break down in Table 3.3) were able to

be collected in the allocated time for data acquisition. Therefore, for realistic synthetic data, subtle (small ranged) augmentations were made to the existing cases, extending the data to 1 183 cases (200 per pathology with some multi-label samples). The methods and ranges of augmentations chosen, listed below, were applied in random combinations to all cases per pathology until the target number (200) was met:

- Rotations in a degree range of [-5,5] were randomly selected and applied using *SciPy*'s *ndimage.rotate* to an input image.
- Differing translations in both x- and y-directions (denoted t_x and t_y respectively) were randomly selected from a pixel range of [-20,20]. Translations were applied to individual frames using *OpenCV*'s *warpAffine* function by means of the translation matrix in Equation 3.3.1.

$$T_{mat} = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \end{bmatrix} \tag{3.3.1}$$

- Black noise was introduced by the addition of a dense matrix (of 0's and -255's for black and white, respectively) using *Numpy*'s function *random.randint*.

3.3.2 Model Input Preparation

This phase includes the preparation of data from augmented inputs to suitable model inputs. This involves further processing of the images for measurement extractions. These measurements allow for calculations of geometric parameters that define model input features.

1. **Frame Selection**
   Frames that most clearly showed one cardiac cycle were selected manually; all frames between two consecutive maximally relaxed (diastole) states were included. This facilitated the identification of key temporal attributes associated with maximum contraction (minimal volume) and maximum relaxation (maximum volume) for parameter calculations/features.

2. **Image Processing (FIJI)**
   FIJI was used to identify boundaries on the images by applying a range of standard plug-ins. All images in a case were processed as single image stacks. Whole stacks were converted to HSB stacks, before binarizing and denoising processes followed. The *Watershed* plug-in was applied to the images, demarcating the foreground in smaller shapes from which measurements (such as centroids or areas) could be extracted. This process was automated and applied as a FIJI macro (script file); a flow diagram of which can be seen in Appendix D.3, Figure D.1.

3. **Parameter Calculations & Feature Definitions**
   Many models in reviewed literature specify the need for consistency between the input variables, such as an equal number of frames per image stack/video processed, images of the same dimensions (Bozkurt, 2019), or in the case of echocardiograms, videos of a single cardiac cycle (Ouyang *et al.*, 2020). Since Random Forest Classifiers and simple Neural Networks were used, whole image stacks could not be processed. Therefore, (pixel) measurements were geometrically derived for feature definitions/calculations.

   Parameter calculations are based on the equations and relationships defined in §2.1.4; such as for cardiac output, stroke volume and ejection fraction. Dynamic volumes were calculated under the assumption that the left ventricle shape could be approximated as half an ellipsoid (Marieb, 2015). Other parameters included were from patient electronic medical records (EMR); such as age, gender and heart rate during testing. Certain features, although clinically significant, were excluded as key values for calculations were unavailable; such as pressure as it is not routinely (dynamically) tracked.

   Post FIJI processing, data for each case existed as 2D arrays; with measurements extracted from each slice representing a specific timestamp. The mean (avg), standard deviation (SD) and variance (VAR) of all time-varying parameters were recorded to capture dynamic behavioural aspects during each cardiac cycle. All dynamic characteristics associated with maximum and minimum volumes of the cardiac cycle were included. These summarised dynamic, statistic and static (EMR) parameters comprise the final 48 features; listed in Table 3.2.

4. **Definition of Output Classes**
   The output classes are based on the 6 pathologies listed in Table 3.3. Their abbreviated names will be used to refer to the pathologies for the remainder of this report.

Table 3.2: Input Features to Models with Variable Names in parentheses

Summarised Dynamic Features:	
Minimum Volume [V_{min}]	Maximum Volume [V_{max}]
Area @ min Vol [A_vmin]	Area @ max Vol [A_vmax]
Centroid coordinate @ V_{min} [Xc_min]	Centroid coordinate @ V_{max} [Xc_max]
Centroid coordinate @ V_{min} [Yc_min]	Centroid coordinate @ V_{max} [Yc_max]
Centerline gradient @ V_{min} [m_min]	Centerline gradient @ V_{max} [m_max]
Centerline intercept @ V_{min} [b_min]	Centerline intercept @ V_{max} [b_max]
Minor axis radius @ V_{min} [a_min]	Minor axis radius @ V_{max} [a_max]
Major axis radius @ V_{min} [c_min]	Major axis radius @ V_{max} [c_max]
Timestamp @ V_{min} [t_min]	Timestamp @ V_{max} [t_max]
Static Features	**Statistical Features**
Gender	Volume: V_avg, V_SD, V_VAR
Age	Area: A_avg, A_SD, A_VAR
Test year	x-coordinate: Xc_avg
2C,3C,4C Heart Rates [HR2/3/4]	y-coordinate: Yc_avg
Stroke Volume [SV]	Gradient m: m_avg, m_SD, m_VAR
Cardiac Output [CO]	Intercept b: b_avg, b_SD, b_VAR
Ejection Fraction [EF]	Minor radius a: a_avg, a_SD, a_VAR
Time per cycle [1cycl_dur]	Major radius c: c_avg, c_SD, c_VAR

Table 3.3: Pathologies, Abbreviations and Unaugmented case Totals

Pathology	**Abbreviation**	**No. cases**
Normal	N	106
Heart Failure	HF	58
Left Ventricular Hypertrophy	LVH	41
Myocardial Infarction	MI	100
Aortic Regurgitation	AR	42
Aortic Stenosis	AS	59

5. **Model Pre-processing**
   Pre-processing of input features was necessary prior to model training. These steps involved imputations of missing values with a *Simple Imputer*, scaling numerical data (using a *Min-Max Scaler*), discretizing categorical data and target labels (whose entries were strings) by using a *One-hot encoder* and *Multi-label Binarizer* respectively. All functions are from *Scikit-learn*'s *Impute* and *Preprocessing* packages; using mostly de-

fault parameters except where otherwise required. Train and test subsets were processed separately to avoid information leakage across the subsets; as pre-processing steps were fit to training sets and labels before being used to transform the test set and labels, respectively.

3.3.3 Model Development

The Model Development phase involved the general steps associated with training, iterating and testing model instances. The three steps from Figure 3.1 can be described as follows:

1. **Model Construction**
   Two types of models were developed side-by-side to investigate the differences between an ensemble model and a deep learning model for the classification task at hand. The model types chosen were:

   - Random Forests adapted for multi-output classification
   - Neural Networks

   Multiple model instances of varying dimensions for each model type were created to be tested. The models were also made to output comparable scoring metrics; including Accuracy, Recall, Precision and (occasionally) F1-score for successive comparisons and evaluations.

   The **Random Forest** model instances were created for all combinations of the following parameters and scoring methods:

   - Number of estimators (E): 100, 150, 200, 250, 300, 350, 400, 450, 500
   - Depths (D): 4, 8, 16
   - Metrics: Multi-label Confusion Matrix: entries were used to calculate the Accuracy, Recall, Precision and F1-score for each target variable. Averaged metrics were used for all comparisons.

   The **Neural Network** model instances were created similarly, investigating the following parameters and scoring methods:

   - Depths/Hidden layers (D): 1, 2, 3
   - Widths/Neurons per hidden layer (W): 16, 64, 256
   - Additions to hidden layers: Batch Normalisation, Dropout (30%), and ReLU Activation
   - Addition to last layer: Sigmoid Activation
   - Layer connection types: Dense

- Metrics: The models were compiled to output Accuracy, Recall and Precision. F1-score was calculated according to the formulae in Equation 2.2.6 for comparisons (with Random Forests).
- Call backs: Models were compiled with Early Stopping where learning ceased after 15 epochs of no improvement (*Patience*) or if absolute change <0.001 (*Minimum delta*). The *restore_ best_ weights* argument was also activated; thus, the models were saved with the weights from epochs with the best scores.

2. **Model Training**
   This phase involved training all models through the respective tests with their associated features, discussed below (§3.3.4).

3.3.4 Model Performance Evaluation

This phase of the pipeline included final testing and outputting of performance metrics of the models. Individual steps as per Figure 3.1 thus follow:

1. **Model Testing**
   To investigate the capabilities and/or preferences of the models all underwent various tests to identify the best suited model for the classification task. In order of application, test details follow:

   - *Engineering* Tests: involved removing "medical" features considered to be more relevant to health care professionals. The aim of these tests was to investigate the predictive capabilities of the model as it relies increasingly on "engineering"/geometric features. Medical features were removed individually then cumulatively until only engineering features - those derived from the images - remained. All models used a train-test ratio of 0.8-0.2, and features removed per test are tabulated in Table 3.4.
   - *Medical* Tests: similar to the above approach, involved the removal of engineering features until only medical features remained. The aim of this test was to investigate the degree to which patient data assists the predictive abilities of a model. Models were developed with an 0.8-0.2 train-test ratio, with features removed specified in Table 3.4.

Table 3.4: Engineering and Medical Tests: Test names and associated features removed

Test	Number/s	Description
E		Denotes Engineering tests
M		Denotes Medical tests
EM	0	Baseline test where all features were included
E/M	12,123, 1234,12345, 123456	Combinations of test numbers indicate that features associated with individual numbers (below) were cumulatively removed
E	1	All Heart Rate features removed
E	2	Age removed
E	3	Gender removed
E	4	Test year removed
E	5	Single cycle duration removed
M	1	Y-intercept data removed
M	2	Centroid data removed
M	3	Center axis data removed
M	4	All Area features removed
M	5	Centerline gradient
M	6	Ellipsoidal radius

- *Data* Tests: were the last set of tests applied to the best performing Random Forest and Neural Network instances for the first two tests. In these tests, model instances were trained on differing train-test ratios to find the ideal split for the models before final selection. Feature selections for data tests are based upon both domain knowledge and previous results. Test details are included in Table 3.5 with respective reasoning for features deletions that follow:

  a) Test year: For the time period over which data was collected, there are no known correlations between any of the selected pathologies nor the frequency at which they occurred and the given year.

  b) Centreline y-intercept: Although this variable was useful in extracting other geometric information from the echocardiogram images, it is not medically informative. For data collected, none showed visual variation to the degree where this feature would prove valuable.

  c) Centroid data: Most useful for extracting other geometric characteristics, it is not relevant in practice.

Table 3.5: Data Tests Details

Train-Test Ratios	**Features Deleted**	**Tested Model Architectures**
0.1-0.9	Test year	**Random Forests:**
0.2-0.8		E100D16
0.3-0.7	Centreline y-intercept data [b variables]	E300D16
0.4-0.6		E500D16
0.5-0.5		
0.6-0.4		**Neural Networks:**
0.7-0.3	Centroid data [Xc and Yc variables]	D1W16
0.8-0.2		D2W64
0.9-0.1		D3W256

2. **Model Performance Evaluation**
   Assessments of the model types are based on applicable evaluations of performance metrics and/or learning curves for all tests. Discussions presented in Chapter 4 account for the effects of feature deletions and model architectures when comparing the results. Comparative discussions of the best performing model instances and feature combinations for both model types conclude all tests. Additionally, performances and feature combinations are discussed (i) in the context of their mutual information scores, and (ii) compared to results of the baseline test EM0.

3.3.5 Final Model Selection

The best performing model instances for Random Forests and Neural Networks were thoroughly compared before one was selected as best suited for the classification task. The section concludes by assessing the selected model's results in its desired context.

4 Results and Discussion

This chapter expands on the results of all phases of the solution pipeline outlined in §3.3; namely Pre-processing (§4.1), Model Input Preparation (§4.2), Model Development (§4.3), Evaluation (§4.4) and Final Selection (§4.5). The results will be discussed in one of the following ways, according to what is most applicable:

- Using an example case to display results associated with steps such as pre-processing, feature extraction and model input preparations.
- Considering the whole dataset in discussions of model performances and/or comparisons, where confusion matrices or learning curves are investigated.

For this chapter, some phrases or terminology are used interchangeably or are differentiated as follows:

1. *Centroid* refers to the centroid of individually demarcated area output from the Watershed plug-in. *Composite centroid* refers to the centroid of the system; considering all areas.
2. *Model* and *model instance* are used to refer to the specific dimensions a model *type* (either Random Forest or Neural Network).
3. *Epochs* and *experience* are used interchangeably when discussing the learning curves of the Neural Network instances.
4. *Echo* refers to echocardiogram.
5. *Dynamic*, *statistical* and *geometric* are used interchangeably to highlight various aspects of the engineering features.
6. *EMR* is sometimes used to refer to medical features.

4.1 Image Pre-processing

Data augmentations included combinations of noise, translations and rotations. In Figure 4.1, the (original) first frame of the example case can be seen

alongside its (individually applied) augments. The dataset acquired contained 34 samples with more than one pathology/label. For multi-label instances, caution was taken to ensure no repeats of originals or its augments existed to avoid data leakage between training and test subsets during model development.

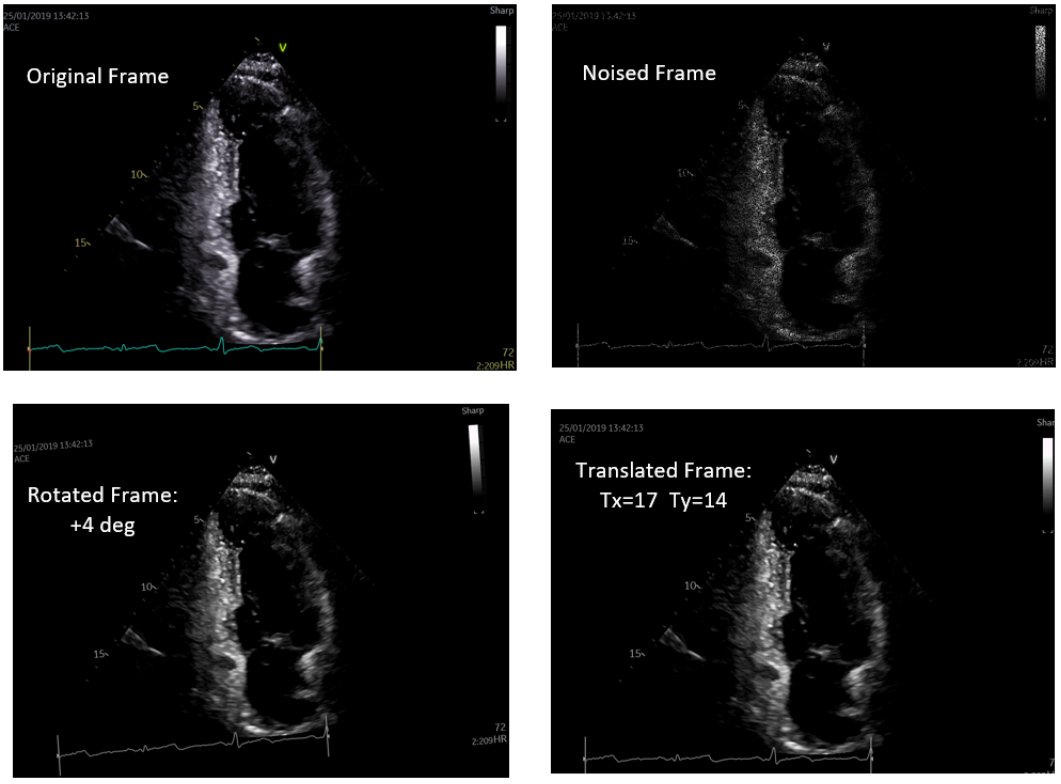

Figure 4.1: Individually applied Data Augmentations

4.2 Model Input Preparation

After applying the FIJI macro (as per §3.3.2), the output images contained outlines of the (white) foreground identified in the input images; as seen in Figure 4.2. The *Watershed* plug-in segmented the foreground into smaller areas/shapes sharing common boundaries. The *Analyse Particles* plug-in extracted measurements in pixel values thereafter - since no measurement scale was provided.

Unfortunately, many of the unaugmented cases contained noise; seen by the white speckles or blurs in areas where there is no cardiac tissue. This, inadvertently, had a detrimental effect on the FIJI output measurements, and thus features extracted. For example, in Figure 4.3, is a Heart Failure (HF)

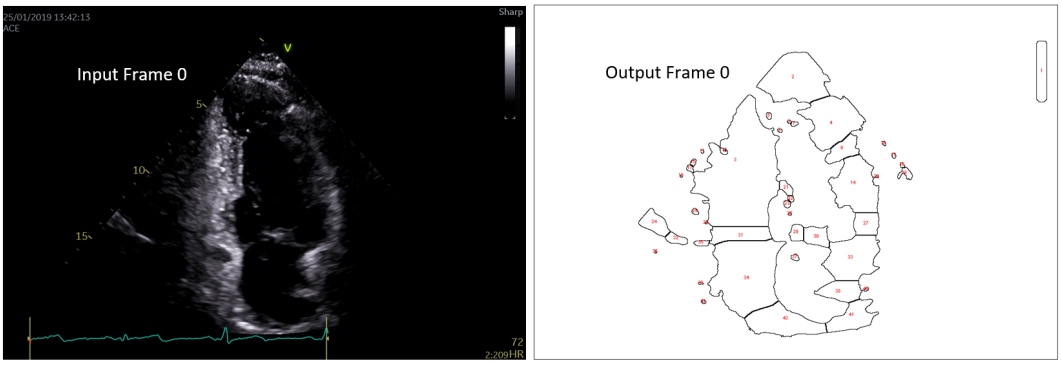

Figure 4.2: Input and Output of FIJI Image Processing

case where much noise is present in the unaugmented image (top left); resulting in poor FIJI output (top right). The bottom images to provide a visualisation of the exacerbated errors due to a further noised augment (bottom left) of the original image.

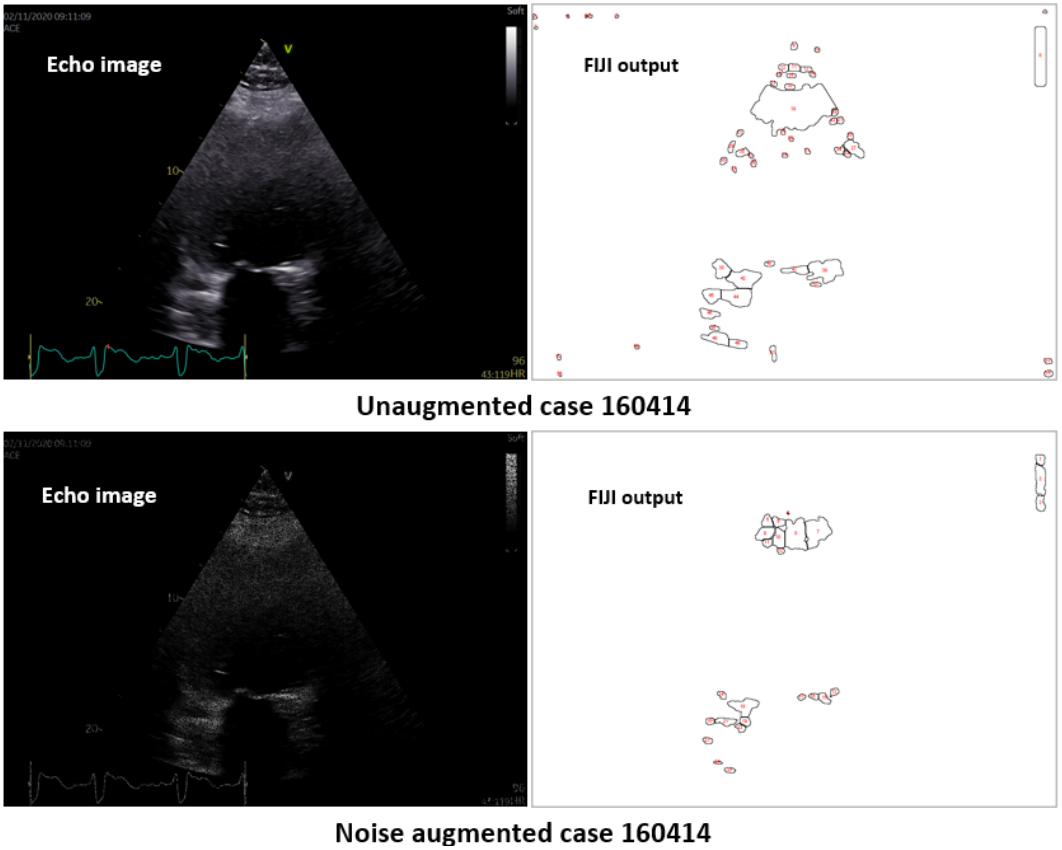

Figure 4.3: Visual representation of the effect of inherent and added noise on FIJI output

The area of interest in a 2-chamber view are those that coincide with the

anatomical structures represented in the foreground of the echo images. In each frame, centroids of the smaller shapes (composite centroids) were used to calculate the centroid of the whole system according to Equation 4.2.1:

$$C = (\frac{\sum x_n}{n}, \frac{\sum y_n}{n}) \tag{4.2.1}$$

Centroid calculations were heavily influenced by the presence of outliers and noise. Thus, an area exclusion criterion was defined and applied in the order below:

1. Exclude areas found outside a range of x-coordinates, as seen by the red boundaries in Figure 4.4 (left).

2. Exclude areas in the first 2 (of 100) bins of the Area histogram as smaller speckles were considered noise. The Histogram bins selected are shown by the red bracket in Figure 4.4 (right).

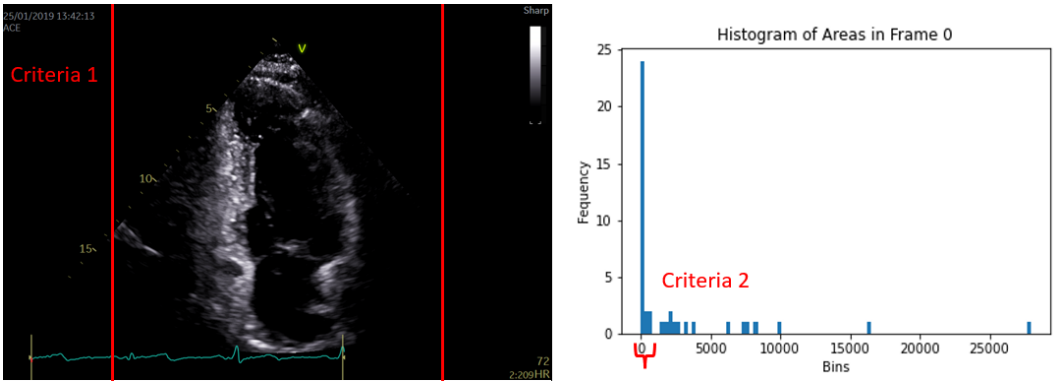

Figure 4.4: Visual representation of area exclusion Criteria 1 & 2

3. Exclude areas whose distance to the centroid is greater than some threshold defined by the Interquartile Range (IQR) rule in Equation 4.2.2. This was to exclude noise too far from the area of interest centrally located:

$$threshold = Q_3 + IQR \tag{4.2.2}$$

The composite centroid was used to define the centreline (or central axis) that passes through the apex of the ventricle and the mitral valve. The centreline, described by gradient m and y-intercept b, was used to estimate the long and short ellipsoidal radii, c and a respectively. In the frames, c is identified as the distance from the apex of the heart to the centroid, while a coincides with the maximum perpendicular distance from the centreline to the chamber walls. Figure 4.5 depicts the centreline (blue), centroid (green) and perpendicular distances calculated to find a (grey lines and text). For this particular frame, a was identified as 142.

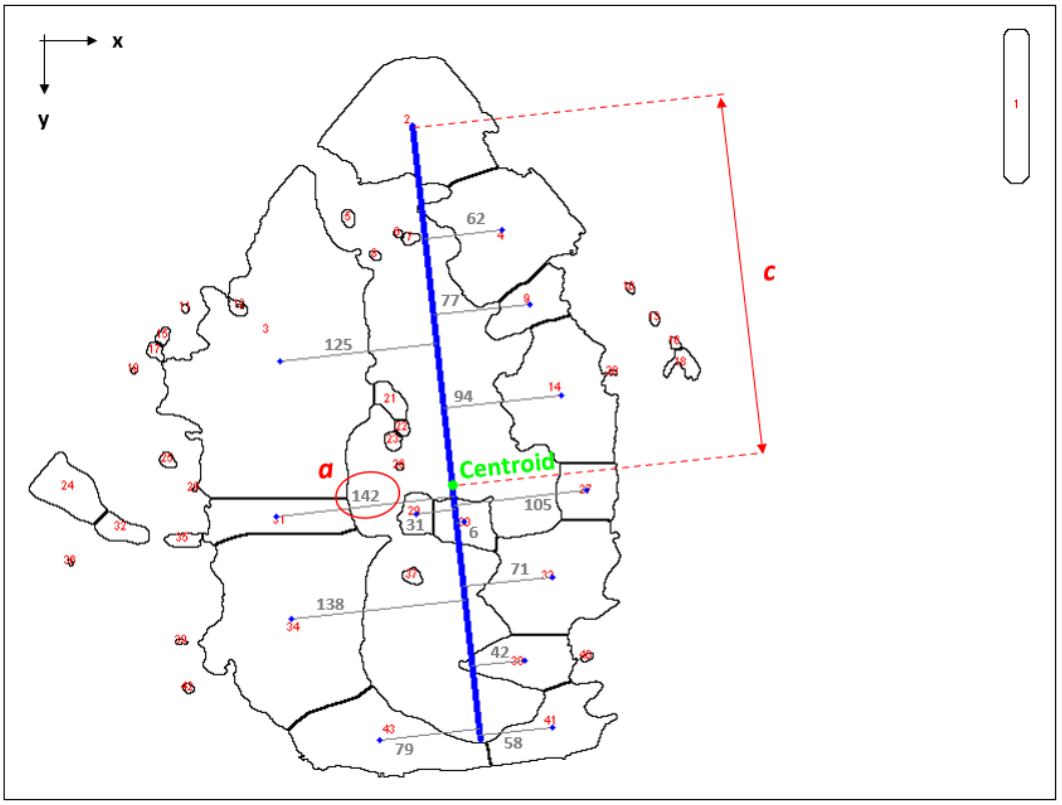

Figure 4.5: Visual example of geometric parameters derived, with the system centreline (blue), centroid (green) and c and a parameters (red)

The instantaneous ventricular volumes were calculated according to Equation 4.2.3. Thereafter, parameters such as stroke volume (SV), cardiac output (CO) and ejection fraction (EF) could be calculated as per Equations 2.1.1, 2.1.2 and 2.1.3, respectively.

$$V = \frac{2}{3}\pi a^2 c \tag{4.2.3}$$

Summarising dynamic data involved calculations of means, standard deviations and variances. This, in conjunction with other parameters at key volumes, EMR data and static variables were thus in a suitable features format to be used as model input.

4.3 Model Development

The Random Forest and Neural Network model instances were constructed as per the specifications outlined in Section 3.3.3. For all tests and model instances, data was pre-processed in the same manner. Prior to all testing, an investigation on the mutual information scores of all features were obtained

to gain insight on their value in relation to the target variables. For multi-class problems, mutual information scores are obtained for each target variable then summed together, constituting the "Final Contributions". The mutual information scores of all features (in baseline test EM0) can be seen in Figure 4.6; feature names as per Table 3.2.

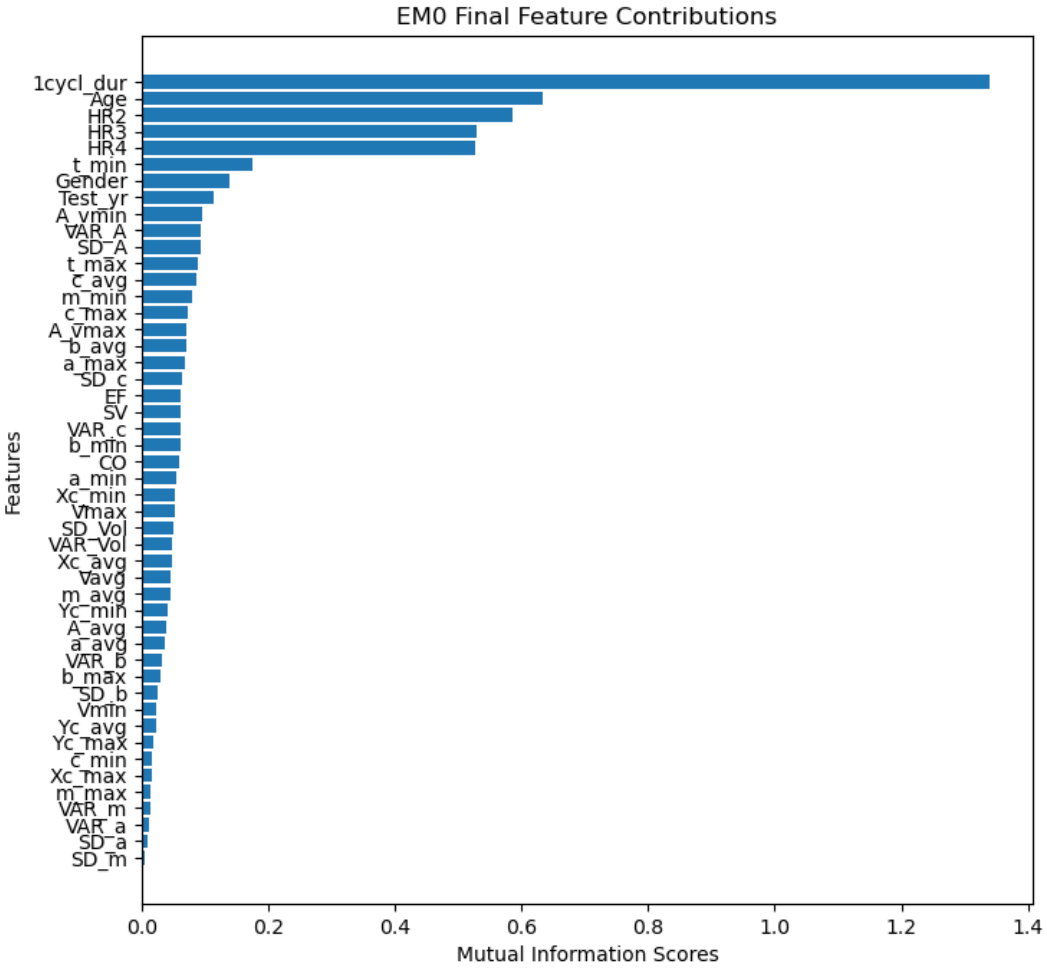

Figure 4.6: Mutual Information of all Features

Medical features have the highest mutual information scores while most other features all score poorly. Of the top 5 scores, heart rate information seems to be most important - as duration of a single cardiac cycle ($1cycl\_dur$) was calculated using heart rate too. This reinforces the importance of considering the patient's unique history and vitals in the process of diagnosis.

An alternative reason to the above observation could be due to the multiple duplicates within certain (pathology) categories from multi-label samples and

augmented data. For instance, 4 of the 6 pathologies selected have rare occurrences, thus as little as 41-59 samples were collected over the period (2016-2020) of interest. As a result, to reach the target number (200 per pathology), some cases were augmented 3-5 times. Despite all image augmentations being unique and providing the desired geometric variations, the EMR data remained unchanged. This was assumed to be acceptable as many pathologies are more common to certain patients (based on characteristics such as age or gender), thus augmented cases would remain on par with original cases.

Examining the dynamic and statistical features more closely, certain features expected to score higher, fall short. For instance, cardiologists visually observe pump functionality when inspecting an echocardiogram. The shape descriptors of the left ventricle, included in features such as centreline gradient (m) or minor ellipsoidal radius (a), were included to inform the model about the movement of the heart as it beats. The centreline gradient provides a measure of how much a particular heart oscillates in the view plane. The minor ellipsoidal radius was expected to be important as it experiences much variation as ventricular volume changes throughout the cardiac cycle.

Mutual Information scores were checked for every test to assess dependencies between the remaining features and the target variables. Mutual information scores of all Engineering tests can be found in Appendix A (Tables A.1 and A.2) and those of Medical tests in Appendix B (Tables B.1 and B.2).

4.4 Model Performance Evaluation

This section presents a discussion of the model performances for the Engineering, Medical and Data tests in §4.4.1, §4.4.2 and §4.4.3, respectively. Discussions within subsections are structured similarly; multi-output Random Forests are discussed first, then Neural Networks, before a comparison of the best instances of each type. Recall is considered with much importance in all discussions with F1-scores (where applicable). These comparisons aid in identifying the most suitable model types and instances for the Data tests (§4.4.3) and final model selection (§4.5). Due to the length of the results tables, most figures and tables in this section summarise selected instance findings only. Full tables for the Engineering, Medical and Data tests can be found in Appendices A, B and C, respectively. For the remainder of this chapter, abbreviated forms of pathologies (as per Table 3.3) and model instance descriptors (listed below) will be used:

- E: Number of estimators (Random Forests only)
- W: Width or neurons per hidden layer (Neural Networks only)

- D: Depth (Random Forests) / number of hidden layers (Neural Networks)

4.4.1 Engineering Tests

The Engineering Tests explore model dependence on engineering features as medical/ EMR features are removed. For this section, all test labels presented in tables or figures align with the names and descriptions in Table 3.4. The results of the Random Forest models (§4.4.1.1) are discussed prior to the Neural Networks (§4.4.1.2). The section closes with a comparative discussion on their respective performances (§4.4.1.3).

4.4.1.1 Random Forest Models

The Random Forest models exhibited much variation between which model architecture performed best across the different tests. However, there was some correlation between model dimensions and pathologies predicted. Tables A.3 and A.4, in Appendix A, lists the models that scored highest per test and pathology and their performance metrics, respectively. Summarising the most important aspects, the following points are listed:

- No D4 models performed well enough for any of the individual pathologies. D16 models occur most frequently across all tests. D8 models begin to perform on par with D16 models as more cumulative deletions occur.
- A range of correlations exist between model architectures and pathologies predicted. A strong correlation example includes E100D16; predicting N well for 7 of the 9 tests (EM0 excluded). The same model predicts AR and AS well only 5 out of 9 tests, but never predicts HF well for any test. These sorts of patterns are evident across all models; showing consistently strong / weak abilities in predicting specific pathologies despite feature variations.
- Continuing the above point; smaller (less estimators), deeper forests predict N, HF, AS and AR better, while larger (more estimators), deeper forests perform better for LVH and MI.
- HF is least predicted well; at best predicted by E200D16 in 3 tests.
- There seems to be no influence of the number of original cases on the models (dimensions) that predict certain pathologies better. For example, N (having 106 original cases) and AS (having only 59) are both predicted consistently well by E100D16. As with LVH (with 41 original cases) and MI (with 100) which are both predicted well by E500D16.

Figure 4.7 is a dot plot of the highest scoring models for all Engineering tests; where dots of the same colour represent equal (averaged) scores. From the figure, the most versatile model instances are E100D16 and E500D16; performing well for all tests. E150D16, E200D16 and E450D16 trail behind, performing well for 8 tests (excluding EM0). Additionally, all D16 models perform well for tests E2, E3, E5 and E123, while E1234 is predicted equally well by D8 and D16 models. D8 models begin to perform on par with D16 models when valuable features (referring to their respective mutual information scores) are deleted. This suggest data complexity decreases enough to be predicted comparably well by shallower models.

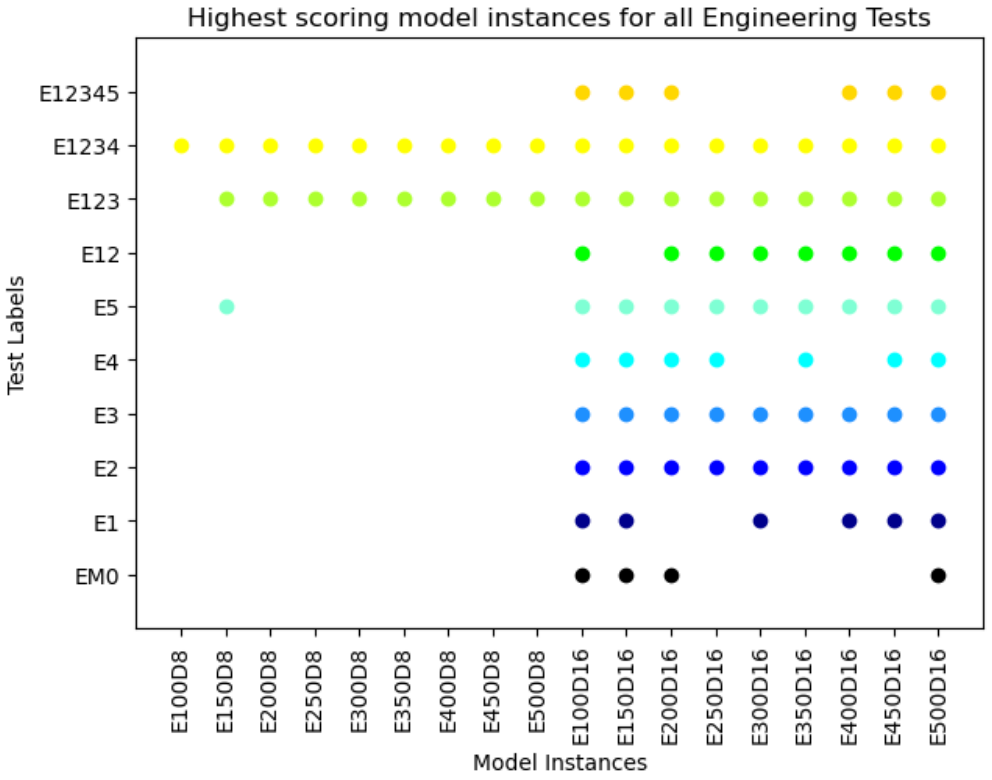

Figure 4.7: Schematic showing Random Forest instances that output the highest scores per Engineering test

To investigate the apparent convergence of D8 and D16 models, performance metrics of the highest and lowest scoring instances per test are tabulated in Table 4.1 - including those of D4 models (generally lowest). From the table, the following observations are made:

Table 4.1: Highest scoring Random Forest instances per Engineering test with output Performance metrics. Minimums of each test are included in parentheses below

Test	Best Model	Accuracy	Recall	Precision	F1-score
EM0	E150D16	0,837 (0,800)	0,226 (0,039)	0,950 (0,827)	0,352 (0,083)
E1	E500D16	0,820 (0,797)	0,139 (0,011)	0,933 (0,833)	0,233 (0,039)
E2	E500D16	0,828 (0,800)	0,196 (0,035)	0,931 (0,769)	0,304 (0,090)
E3	E100D16	0,832 (0,800)	0,208 (0,032)	0,942 (0,833)	0,327 (0,101)
E4	E500D16	0,833 (0,799)	0,205 (0,035)	0,956 (0,861)	0,320 (0,091)
E5	E250D16	0,832 (0,799)	0,209 (0,024)	0,958 (0,857)	0,322 (0,084)
E12	E200D16	0,814 (0,797)	0,127 (0,011)	1,000 (0,774)	0,250 (0,041)
E123	E150D16	0,814 (0,796)	0,118 (0,007)	0,935 (0,667)	0,193 (0,040)
E1234	E100D16	0,808 (0,796)	0,093 (0,007)	0,902 (0,667)	0,153 (0,040)
E12345	E100D16	0,806 (0,796)	0,077 (0,003)	0,886 (0,500)	0,134 (0,034)

- A notable decline in performance metrics ranges is evident for cumulative feature are deletions (E12 - E12345) compared to individual deletions (E1 - E5). As medical features are removed individually and (more so) cumulatively, the models perform worse than the baseline test EM0 for all Recalls and F1-scores. This is explained by the high mutual information scores of medical features.
- The highest scoring test was EM0 (the baseline test) where no medical features are removed, with model E150D16 scoring highest.
- The lowest scores were produced in E12345, where all medical features were removed, with E100D16 performing best.
- All model instances per test perform within narrow ranges of each other showing convergence of their predictive capacities as more medical features are excluded.

- Precision does not follow the trends of other metrics. For example, test EM0 produces the highest Recall, Accuracy and F1-score, while E12 has the highest Precision.

The internal workings of the model instances are investigated by means of confusion matrix outputs. The outputs of the highest and lowest scoring models (for tests EM0 and E12345) are tabulated in Table 4.2; on the left and right respectively. For multi-output Random Forests, confusion matrix values are output for each pathology. Their respective performance metrics are presented altogether in Table 4.3. From both tables, the following is noted:

Table 4.2: Confusion Matrices of highest (left) and lowest (right) scoring models for each pathology in tests EM0 and E12345

	E150D16 in EM0						**E450D4 in EM0**					
	N	**HF**	**LVH**	**MI**	**AR**	**AS**	**N**	**HF**	**LVH**	**MI**	**AR**	**AS**
TN	203	190	191	179	188	174	203	192	191	179	188	174
FP	0	7	0	0	0	0	0	5	0	0	0	0
FN	30	25	35	46	37	52	34	33	45	57	48	63
TP	4	15	11	12	12	11	0	7	1	1	1	0
	E100D16 in E12345						**E450D4 in E12345**					
	N	**HF**	**LVH**	**MI**	**AR**	**AS**	**N**	**HF**	**LVH**	**MI**	**AR**	**AS**
TN	202	192	190	179	188	174	203	196	191	179	188	174
FP	1	5	1	0	0	0	0	1	0	0	0	0
FN	32	33	42	55	47	60	34	40	46	57	49	63
TP	2	7	4	3	2	3	0	0	0	1	0	0

Table 4.3: Comparison of performance metrics for highest and lowest scoring model instances for Tests EM0 and E12345

Test	**EM0**		**E12345**	
Performance metrics	**E150D16**	**E450D4**	**E100D16**	**E450D4**
Accuracy	0,837	0,800	0,808	0,796
Recall	0,226	0,039	0,093	0,003
Precision	0,947	0,896	0,902	0,500
F1-score	0,352	0,096	0,153	0,034

- The best and worst model instances for each test are distinguished by their TP values. Models for each test have similar values for TN, FP and FN; all within small ranges of each other.
- These tests were performed on the same train-test split of data, thus all columns sum to 237. High Accuracies of all models are attributed to the high fraction of TN values.
- High Precision values noted previously are a result of to the low FP values of the model instances.
- Recalls across all Random Forest models are low due to the low fractions of FN and TP to the total amount of predictions (TN + FP + FN + TP). This is especially noteworthy as Recall is more valuable when assessing model performance for the healthcare context; as false negative diagnoses have more serious repercussions (depending on the abnormality).
- The low F1-scores indicate unsatisfactory balances between Recalls and Precisions for all Random Forest models, evident by their major differences in FN and FP values.

In conclusion, no Random Forest model clearly outperformed another across all Engineering Tests. With no clear winner regarding model instances or feature/s deleted, the models of interest identified are those which were most versatile; viz. E100D16 and E500D16. E100D16 performed best for tests E3, E1234 and E12345 while E500D16 performed best for E1, E2 and E4. E150D16 in EM0 was also considered the best performing model due to high metrics, though it is not as versatile as those aforementioned.

4.4.1.2 Neural Network Models

The best performing model instance per Engineering test and their performance metrics are summarised from Table A.5, in Appendix A. From Table 4.4, the following observations about the Neural Network models are made:

Table 4.4: Best performing Neural Networks per Engineering Test with Validation performance metrics

Test	Model	Accuracy	Recall	Precision	F1-score
EM0	D3W256	0,540	0,500	0,681	0,577
E1	D3W256	0,536	0,507	0,631	0,562
E2	D3W256	0,523	0,438	0,585	0,501
E3	D3W256	0,498	0,438	0,620	0,513
E4	D3W256	0,468	0,403	0,594	0,480
E5	D3W256	0,511	0,483	0,645	0,552
E12	D2W256	0,418	0,314	0,558	0,402
E123	D1W256	0,359	0,152	0,489	0,232
E1234	D1W256	0,304	0,093	0,500	0,157
E12345	D3W64	0,304	0,048	0,483	0,088

- There is a decrease in most performance metrics relative to the baseline test EM0, for every test performed - except test E1. Additionally, the scores decrease steadily for all cumulative medical features deletions (tests E12 - E12345).
- For single feature deletions, largest model D3W256 performs best. As the tests involved more cumulative feature deletions, the architectures required to fit the data become smaller; typically, shallower (D decreases) before narrowing (W decreases).
- Low performance metric scores and smaller models associated with cumulative feature deletions suggest that less complex relationships exist between the features and target variables. This is reinforced by the low mutual information scores (discussed further in §4.4.1.3) confirming the low contributions of remaining features to the targets.
- Another trend observed when investigating the results for all models (seen in Table A.5), is that the deep, narrow models perform similarly, if not worse, than shallower models of the same width. This confirms the sensitivity of Neural Network performances to model complexity and features deleted.

From Table 4.4, the best and worst results are associated with tests E1 and E12345, respectively. To explore the different model instance behaviours in these tests, their Learning Curves are analysed. The learning curves include Training and Validation curves for Loss, Accuracy and Recall to investigate the internal workings of the models. All dashed lines represent training curves, while solid lines represent validation curves, colour-coded per metric for easy

identification. The learning curves of model D3W256 , Test E1, in Figure 4.8 reveal and/or display very strong characteristics about the model and its response to the data:

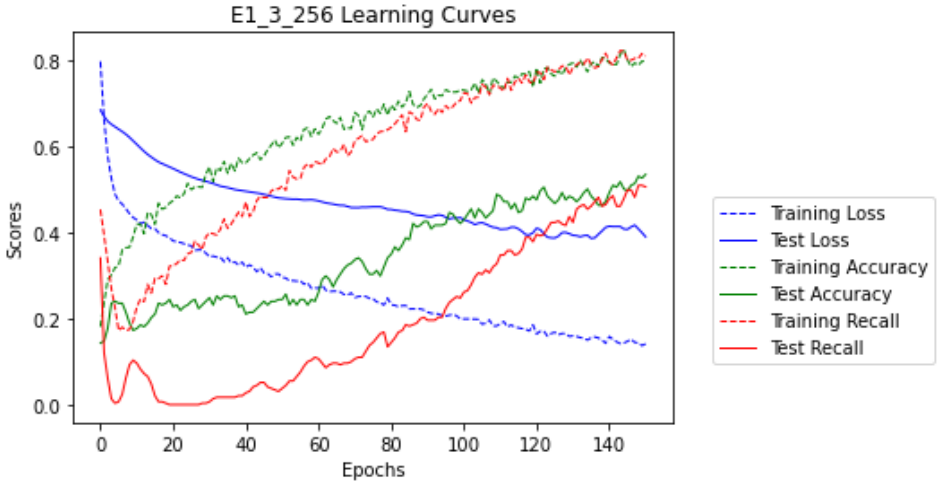

Figure 4.8: Test E1 - Learning Curves of model instance D3W256

- A high learning rate is present, seen by the initial gradient change in the Training Loss curve.
- Training and Validation Loss curves continue to decrease as epochs increase with neither stabilising. This is an indicator that the model's capacity is greater than required for the problem at hand. Continual decrease of both Loss curves indicate continued learning.
- The large gap between the Training and Validation Loss curves indicates strong overfitting. This is in agreement with the above points, as the model continually learns until the stopping criteria was met. The noise increasingly present in the Validation Loss curve further supports this conclusion.
- The other indicators of strong overfitting are the large gaps between the Training and Validation curves of both Accuracy and Recall. The noise in the Validation curves further support that the model has learnt particulars in the training dataset, thus unable to generalise well.
- The noise present in all Validation curves (especially Accuracy and Recall) and gap between all Training and Validation Loss curves suggest that that the model has been trained on an unrepresentative training dataset. For the model complexity, overfitting was thus an expected occurrence.

Observing D3W64 in Test E12345, the model at the opposite end of the performance spectrum, the following can be noted on model and data characteristics:

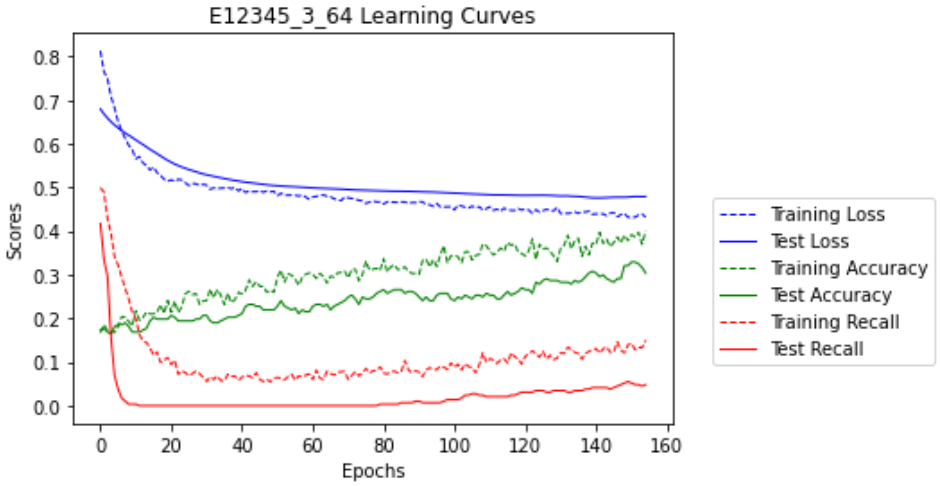

Figure 4.9: Test E12345 - Learning Curves of model instance D3W64

- Gradient change at the beginning of the Training Loss curve suggests the learning rate is considerably high.
- The maximum epochs (when the stopping criteria was reached) is similar to D3W256 despite the Loss of D3W64 remaining relatively high and the Accuracy/Recall remaining low. Training Loss continues to decrease to a lesser degree than that of D3W256, indicating some continued learning.
- The gap between the Loss curves widen with experience, indicating increasing presence of overfitting - also slight in comparison to D3W256 in test E1.
- Slight overfitting is confirmed by the gap between the Training and Validation Accuracy curves that increase with experience.
- The slight noise in both the Accuracy (more evidently) and Recall Validation curves are attributed to the degree of overfitting present. The degree of noise and overfitting are much less in D3W64 (than D3W256) as it is better suited to the data complexity in E12345.
- The noise in the Training and Validation Recall and Accuracy curves are a common symptom for model instances trained or tested with unrepresentative data, capping the overall performance of the model. Underfitting may have been a possibility considering both Loss curves continually decrease over time yet remain high or Accuracy/Recall curves that remain low.

Considering models D3W256 and D3W64 alongside one another in the context of E1 and E12345, both models may be improved by additional data. Both model instances experience different effects of unrepresentative data, exhibiting varying degrees of overfitting and noise due to their sizes. This was seen in all model instances for all Engineering tests.

4.4.1.3 Comparison of Best Models & Features

A summary of the best and worst Engineering tests (representing feature combinations) and model instances for the Random Forests and Neural Networks can be found in Table 4.5. For both model types, the highest scoring tests were EM0 and E1 (Heart Rate features deleted); while E12345 (all medical features deleted) was unanimously worst. Considering the mutual information scores of features in tests E1 and E12345, in Tables A.1 and Table A.2 respectively, insight into model performances can be further explained.

Table 4.5: Summary of highest scoring Engineering tests and model instances

	Random Forests	**Neural Networks**
Best Test	EM0	E1
Worst Test	E12345	E12345
Best Models	E100D16, E500D16 E150D16	D3W256

Tests EM0 and E1 result in the best performance outputs as they contain the most medical features - which maintain the highest mutual information scores. Apart from individual architectures, models depend mostly on features of *Age* and *single cycle duration* for E1. For test E12345, no feature has a significant relationship with the target variable. Despite both tests having features with scores that vary, the presence of the higher scores associated with medical features contribute to the performance differences. Cumulative deletions of medical features are the main contributors to the performance convergence of the D8 and D16 models (with D4 models trailing, seen in Table 4.1) for Random Forests, and the smaller/narrower models outperforming larger Neural Network models.

In general, the Random Forests produced much higher accuracies than the Neural Networks. Upon further investigation, the Random Forest accuracies were skewed on account of the TN. True predictive capacity of the model instances was revealed better when observing TP verses FN values represented in Recall scores. Conversely, Neural Networks generally had Recalls roughly double those of the Random Forests. No clear winner exists between the

Random Forest model instances (as compared to for the Neural Networks), although results were harmonious in reflecting the better performance of deeper forests. The same can be said for the Neural Networks; however, further investigation of the associated learning curves showed overfitting and high learning rates for deeper, wider instances.

From Table 4.5, the best model instances are listed and considered on the following basis:

- **Random Forests**:
  E100D16 and E500D16 represent the most versatile of the architectures based on their ability to predict more pathologies. E150D16 was listed as it produced the highest Recall in test EM0 and performs best for E123. The results show that small, medium and larger forests all exhibit different strengths and abilities to predict certain pathologies as these models are sensitive to the features selected and internal architectures.

- **Neural Networks**:
  The best performing and most versatile Neural Network model instance was D3W256; scoring highest for most tests. Further investigation into respective learning curves revealed the inevitable tendency of this model to overfit the training data. Neural Networks also showed sensitivity to features present, data representativeness and model dimensions.

4.4.2 Medical Tests

The Medical Tests explore the dependence of the models on the medical features as more engineering features are excluded. As was for the previous section, all test labels presented correspond with the names and descriptions in Table 3.4. Similarly, the results of the Random Forests are presented separately §4.4.2.1, before the Neural Network models (§4.4.2.2) and a comparison of their respective performances (§4.4.2.3).

4.4.2.1 Random Forest Models

Tables B.3 and B.4 in Appendix B, show all model instances that best predicted certain pathologies for all 11 Medical tests (excluding test EM0). Summarised observations made from these tables are listed below, with many similarities seen previously in Engineering test results:

- All of the best performing model instances listed are D16 forests.

- Most models, at best, predict 1 or 2 pathologies well. For example, model instance E100D16 predicts N and AR best, however does poorly for HF and LVH.

- As previously seen, there are no correlations between model dimensions and the number of unique cases in certain pathologies. For example, E100D16 best predicts N (having 106 original cases) and AR (which has 42 original cases) as specified in Table 3.3.
- The above points further confirm that specific model instances are considerably invariant to feature changes when predicting certain pathologies well - as with E100D16 predicting N well in most Engineering and Medical tests.
- Smaller forests (E100-E200) predict N, AR and AS better; middle-sized forests(E250-E350) do well for HF, and larger forests (E400-E500) do better for LVH. MI is an exception to this trend and the above point; as there was no clear pattern to identify which model architecture worked best - varying based on the features removed.

To better understand how model versatilities compare, Figure 4.10 shows the highest scoring models against all tests. Dots of the same colours indicate the same high score was achieved. The following observations are thus listed:

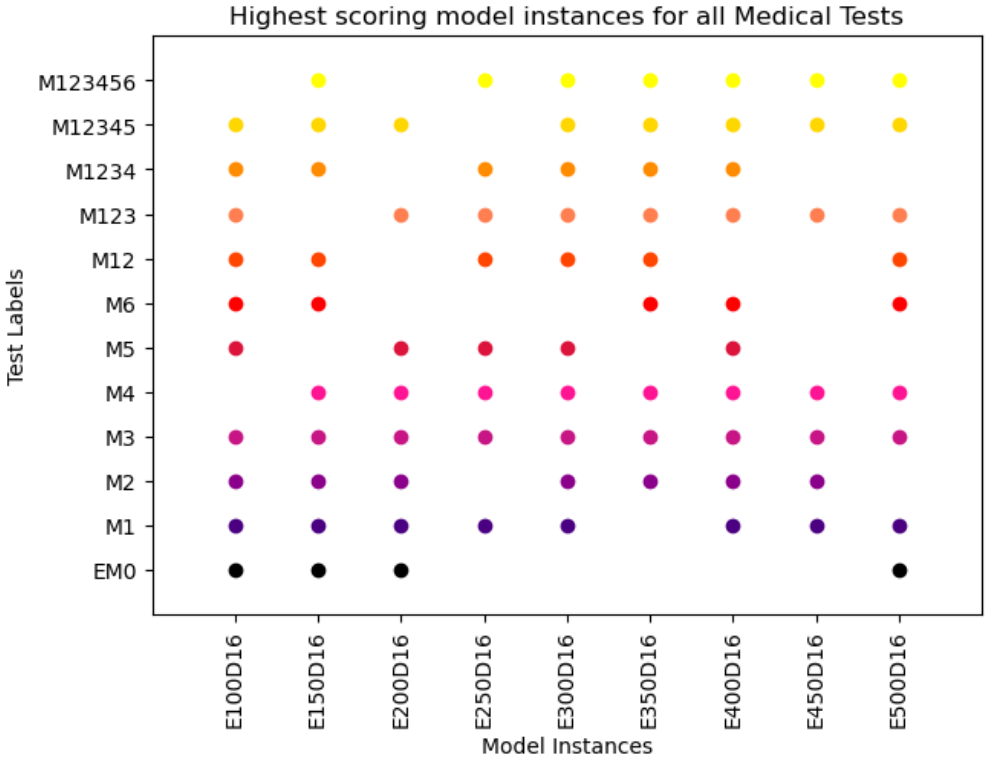

Figure 4.10: Schematic showing Random Forest instances that output the highest scores per Medical test

- Since all models are deep forests (D16), this preludes to the high value of the remaining features as the more complex models perform best for all tests.
- No model instances do well across all Medical tests. The closest instances are E100D16, E150D16 and E400D16 that perform well for 10 of the 11 tests.
- Test M3 (where major ellipsoid radius data was removed) resulted in all instances performing well and in close range of each other.

The range in which all models (including D4 and D8 forests) perform per test can be seen in Table 4.6. The performance metrics of the highest and lowest scoring models per test are tabulated for all Medical tests. The best performing model instances were identified based on Recall and F1-score values primarily. General observations were noted as follows:

- In general, the ranges of performance metrics across all tests increase as more engineering features are removed. This indicates that the models somewhat diverge in their predictive capacities as engineering features are deleted individually and cumulatively. This is expected given their low mutual information scores.
- As with the results of the engineering tests, there is no clear pattern as to which architectures perform best in terms of forest sizes.
- There is improvement of Recall and F1-scores in all tests compared to test EM0. The highest scores are achieved by E300D16 in M123456 - roughly doubled the Recall and F1-score of EM0. Recall and F1-score values for all tests including single feature deletions (M1-M6) remain close to those in EM0. Both metrics increase as features are cumulatively removed (M12-M123456); opposing the trend seen in Engineering tests.
- Precisions show no clear pattern with feature deletions across all tests.

Table 4.6: Highest scoring Random Forest instances per Medical test with associated Performance metrics. Minimums of each test are included in parentheses below

Test	Best Model	Accuracy	Recall	Precision	F1-score
EM0	E150D16	0,837 (0,800)	0,226 (0,039)	0,950 (0,827)	0,352 (0,083)
M1	E150D16	0,842 (0,797)	0,249 (0,028)	0,957 (0,750)	0,383 (0,077)
M2	E400D16	0,839 (0,800)	0,239 (0,039)	0,946 (0,792)	0,365 (0,098)
M3	E250D16	0,836 (0,799)	0,219 (0,028)	0,958 (0,800)	0,346 (0,125)
M4	E450D16	0,834 (0,803)	0,212 (0,049)	0,980 (0,918)	0,337 (0,099)
M5	E100D16	0,837 (0,800)	0,224 (0,036)	0,952 (0,900)	0,346 (0,076)
M6	E350D16	0,836 (0,797)	0,228 (0,028)	0,952 (0,815)	0,356 (0,082)
M12	E250D16	0,843 (0,798)	0,270 (0,039)	0,946 (0,808)	0,402 (0,080)
M123	E300D16	0,852 (0,798)	0,299 (0,035)	0,957 (0,761)	0,443 (0,088)
M1234	E100D16	0,855 (0,804)	0,313 (0,054)	0,969 (0,913)	0,462 (0,112)
M12345	E500D16	0,873 (0,804)	0,388 (0,048)	0,944 (0,873)	0,541 (0,126)
M123456	E300D16	0,887 (0,809)	0,458 (0,081)	0,947 (0,861)	0,613 (0,208)

To assess the quality of model predictions, the confusion matrices of the highest and lowest scoring models (from tests M123456 and M4) are investigated. All respective confusion matrix outputs are tabulated, per pathology, in Table 4.7: highest scoring models on the right and lowest on the left. The associated performance metrics of these models can be found in Table 4.8. Observing information from both tables, the following is noted:

- Despite the TP values of E300D16 and E150D4 being similar, the largest differences exist between the FN and TP values of the two models. In line with the trends noted for Random Forests, the shallower (D4) forests struggle to model more complicated relationships - noted by the differences in FN values.

- The considerable differences between the Recall and F1-score values are due to the FN values and its respective trade-off with FP values (Precision).
- Interestingly, the highest and lowest scoring models of test M4 have similar TN values. E450D16 has higher FP values for HF and LVH, resulting in the slightly lower Precision of this model compared to instances E200/E250D4. Their large differences in Recall values are attributed to the difference in TP values.

Table 4.7: Confusion Matrices of highest (left) and lowest (right) scoring models for each pathology in tests M123456 and M4

	E300D16 in M123456						**E150D4 in M123456**					
	N	**HF**	**LVH**	**MI**	**AR**	**AS**	**N**	**HF**	**LVH**	**MI**	**AR**	**AS**
TN	200	196	191	179	188	172	202	195	191	179	188	174
FP	3	1	0	0	0	2	1	2	0	0	0	0
FN	25	19	22	34	23	32	32	33	46	55	39	63
TP	9	21	24	24	26	31	2	7	0	3	10	0
	E450D16 in M4						**E200/250D4 in M4**					
	N	**HF**	**LVH**	**MI**	**AR**	**AS**	**N**	**HF**	**LVH**	**MI**	**AR**	**AS**
TN	203	192	190	179	188	174	203	194	191	179	188	174
FP	0	5	1	0	0	0	0	3	0	0	0	0
FN	29	27	35	48	37	54	34	32	45	57	47	62
TP	5	13	11	10	12	9	0	8	1	1	2	1

Table 4.8: Performance metrics of highest and lowest scoring model instances for Tests M123456 and M4

Test	**M123456**		**M4**	
Performance metrics	**E300D16**	**E150D4**	**E450D16**	**E200D4/ E250D4**
Accuracy	0,887	0,809	0,834	0,803
Recall	0,458	0,082	0,212	0,049
Precision	0,941	0,861	0,940	0,945
F1-score	0,613	0,208	0,337	0,100

In conclusion, the best performing Random Forest in the Medical Tests was E300D16 in Test M123456 where most engineering features have been removed and the model depended predominantly on medical features. As with the Engineering tests, there were no instances that outperformed the others with regards to performance metrics. However, a model of interest, due once again to versatility, was instances E100D16 performing well for many Medical Tests.

4.4.2.2 Neural Network Models

The best performing model instances per Medical test and their final performance metrics summarised in Table 4.9 are based on Table B.5 in Appendix B. The following is observed about the performances of Neural Network models instances and tests:

Table 4.9: Best performing Neural Networks per Medical Test Validation performance metrics

Test	Model	Accuracy	Recall	Precision	F1-score
EM0	D3W256	0,540	0,500	0,681	0,577
M1	D3W256	0,451	0,431	0,610	0,505
M2	D3W256	0,523	0,497	0,632	0,556
M3	D3W256	0,549	0,510	0,682	0,584
M4	D2W256	0,489	0,428	0,629	0,509
M5	D3W256	0,540	0,534	0,683	0,600
M6	D3W256	0,532	0,497	0,634	0,557
M12	D3W256	0,527	0,517	0,617	0,563
M123	D3W256	0,519	0,514	0,618	0,561
M1234	D3W256	0,553	0,476	0,627	0,541
M12345	D2W256	0,527	0,455	0,660	0,539
M123456	D3W256	0,570	0,538	0,678	0,600

- The widest models (W256) produce the highest scores for all Medical tests, based on their final performance metrics. Again, model D3W256 most frequently performs best (except in M4 and M12345).
- Performance metrics values typically increase as Engineering features are cumulatively deleted.
- Test M1, at best, produced performance metrics all lower than that of the baseline test EM0. However, based on Recall alone, the worst result was of model instance D2W256 in test M4.

- The best results are of test M123456, where almost all engineering features are removed with some static and EMR data remaining. However, the best Precision was associated test M5.
- Tests M123456 and M4 resulted in highest and lowest scores, respectively; similar to the Random Forest results previously.

Learning curves were used to further investigate model behaviours associated with tests M123456 (D3W256) and M4 (D2W256) in response to the data, features and relative model sizes. The following information can be deduced from the curves of D3W256, test M123456, seen in Figure 4.11:

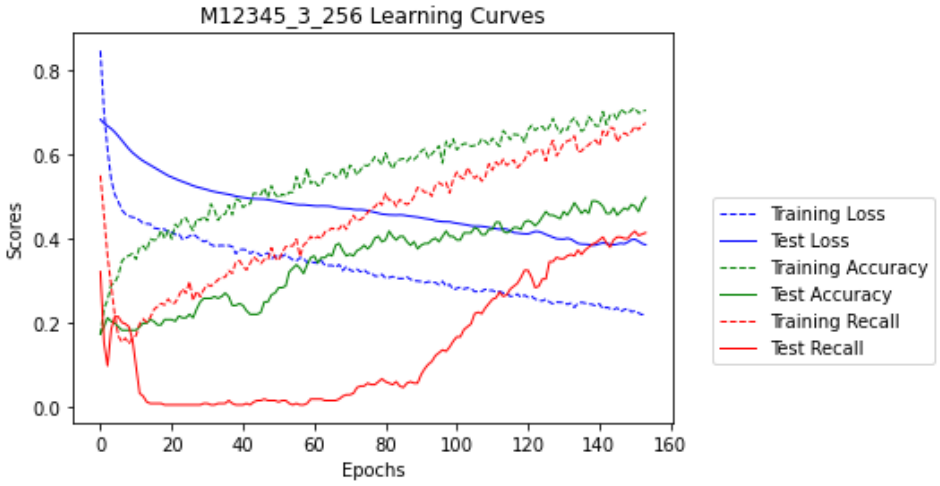

Figure 4.11: Test M123456: Learning Curves of model instance D3W256

- A higher learning rate is present; seen by the steep gradient in the Training Loss curve.
- Both the Training and Validation Loss curves continually decrease without stabilizing. This indicates the continued learning of the model until the stopping criteria was reached. This behaviour can be attributed to the model's capacity compared to the few remaining features; introducing the likelihood of overfitting.
- A relatively large gap between the Training and Validation Loss curves results from overfitting. Furthermore, the gap remains more-or-less constant throughout model learning.
- Large gaps between the Training and Validation Accuracy and Recall curves further confirms the overfitting present. The large gaps in the curves, most prominent for Recall curves.

- There is a significant amount of noise seen in the Accuracy and Recall curves over time suggesting that the model has been trained/tested on unrepresentative data.
- Overfitting is also the cause of the decrease in Validation Accuracy around ±130 epochs

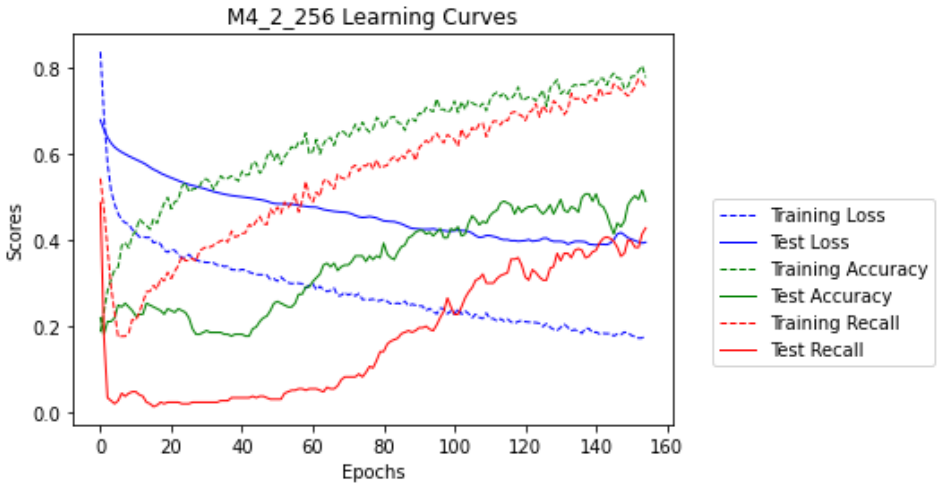

Figure 4.12: Test M4: Learning Curves of model instance D2W256

From Figure 4.12, the following can be noted about the performance of D2W256 in test M4:

- Training and Validation Loss curves continually decrease without stabilizing, while the gap between them gets larger despite ongoing learning. This suggests the model has overfit the data or that training halted prematurely as the curves did not stabilize before the stopping criteria was reached.
- Larger gaps (compared to D3W256) present between the Training and Validation curves for both Accuracy and Recall support that strong overfitting is present.
- Both Validation curves for Accuracy and Recall show a greater degree of noise compared to their Training counterparts; suggesting unrepresentative data.

Comparing the model architectures associated with tests M123456 and M4, both exhibit similar behaviours despite differences in final performance metrics values. The effects of unrepresentative data on both models support the need for more training data for more complex models. As with the Engineering tests, data representativeness is an issue common to all Neural Network models, presenting itself differently based on the architecture and features included.

4.4.2.3 Comparison of Best Models & Features

The highest and lowest scoring Medical tests, together with the highest scoring models are summarised in Table 4.10.

Table 4.10: Summary of key Medical tests and model instances

	Random Forests	**Neural Networks**
Best Test	M123456	M123456
Worst Test	M4	M4
Best Models	E300D16	D3W256

The best and worst tests are unanimous among both the Random Forest and Neural Network instances. To investigate further, mutual information scores of both tests, from Table B.1 and B.2 are compared. Test M123456 has the least features present among all tests; as most engineering features, where most variation exists, are removed. This set of features results in the best performance metrics for both Random Forests and Neural Network model instances, as opposed to their highest Engineering test results. The effect of less features (with low contributions) involved for Medical tests can be seen by the number of small-to-medium sized Random Forest models that perform best in Table 4.6.

Another possible explanation for this result may be in the unaugmented EMR data. Multiple samples containing similar (EMR) data exists in both training and test subsets. This implies that there may be a degree of data leakage between these subsets - whereby many of the EMR features, with the highest mutual information scores, are replicated. There may not be enough variation provided by non-EMR features remaining (such as EF or CO) to better distinguish one sample from another given their associated (lower) mutual information scores. Thus, higher model scores may be a result of learning similar data.

In test M4, model instance scores are lowest compared to other Medical tests due to the deletion of surface area (A) information. Mutual information scores of 4 of the 5 A features deleted are in the top 20 for test EM0, in Figure 4.6. This deletion is thus considered a loss for all models predictive capabilities. This can be seen by the difference in model instance dimensions that scored best (E450D16 and D2W256) that differed from respective trends observed.

The models in Table 4.10 were selected based on final Recall scores. Additionally, Random Forest instances E100D16, E150D16, E250D16 are also

worth noting. Similar to the Engineering tests, these model instances (including E300D16) were slightly more versatile, performing well for 2 (of 11) tests each. Larger forests do not repeatedly perform well among Medical tests. As previously noted, model instances of certain sizes perform better for certain pathologies (Tables B.3 and B.4). For example, E300D16 predicts HF and LVH well, but poorly for N, AS and AR; which E100D16 / E150D16 predicts better.

The highest scoring, most commonly occurring Neural Network model instance is D3W256. However, as before, this particular instance is prone to overfitting the training data. This is of greater concern in the Medical tests (especially M123456) where features do not differ to the degree of the engineering features removed; enabling overfitting and poor generalisation.

4.4.3 Data Tests

This test is performed on shortlisted Random Forest and Neural Network models from previous tests results. These models and feature deletions are discussed (§4.4.3.1) before Data tests results are reported for the Random Forest (§4.4.3.2) and Neural Networks (§4.4.3.3). The Data tests train-test ratios are specified in Table 3.5 (§3.3.4). The results of this section were used to select the most suitable model for the classification task at hand.

4.4.3.1 Shortlisted Features & Models

Based on both Engineering and Medical tests results, the following instances of each model type were selected:

- **Random Forests:** E100D16, E300D16, E500D16
  These models either had the highest performance metrics, were versatile (despite feature deletions) or consistent with the pathologies they predicted well. Since there was no distinct pattern, 3 model instances (all D16) across the range of forest sizes were included to investigate the effect of altered training set sizes on model performances. Other depths were not included as they did not have comparable results in previous tests.

- **Neural Network:** D1W16, D2W64, D3W256
  The pattern exhibited for the Neural Network performances was unanimous among all tests: D3W256 performed best. Due to the observed tendency to overfit, models across the dimension spectrum (both depths and widths) were included to investigate the effect of altered training set sizes on model performances.

The effects of each Engineering and Medical test on the models can be seen in Figures C.1 and C.2 in Appendix C. Associated observations are summarised as follows:

- **Random Forests:**
  For the various medical features deleted, all Random Forests instances respond similarly; unanimously increasing or decreasing for specific feature/s deletions. Model instances across all tests scored less than test EM0, however features removed in E1 (heart rates) and E3 (Gender) were most detrimental, producing the poorest Recalls. Cumulative medical feature deletions result in the gradual decline of Recall. An opposing response to engineering feature deletions was seen (in Figure C.2) as all instances improved for the individual deletion in test M1 (y-intercept data) and cumulatively for most tests thereafter.

- **Neural Network:**
  Each of the shortlisted Neural Network model instances exhibit vastly different behaviours to feature deletions. For both tests, there are no similar responses between model instances. For example, models D1W16, D2W64 and D3W256 decrease, increase and (slightly) increase, respectively, for test E1 (compared to EM0). These results, thus, contribute negligibly to the features selected for the Data tests. They do, however, provide insight as to what to expect with the inclusion/exclusion of certain features.

The final list of features excluded for the Data tests is based loosely on the Random Forests and on clinical relevance:

- Most medical features greatly contribute to the performance of the model, however in practice the year of testing (*Test_yr*) was not be considered relevant to disease occurences and thus excluded.

- Based loosely on the Random Forest results in Figure C.1; features of (test M12) *y-intercept* and *centroid* data don't have significance in practice and were excluded.

4.4.3.2 Random Forest Results

Figure 4.13 graphically presents learning curves for Accuracy, Recall and F1-score that result from the shortlisted Random Forest model instances. All models experience a slow, steady increase in performance as the training set size increases. The largest differences are seen in the Recall (orange) and F1-score (green) curves in the figure, as Accuracy (blue) improves only slightly.

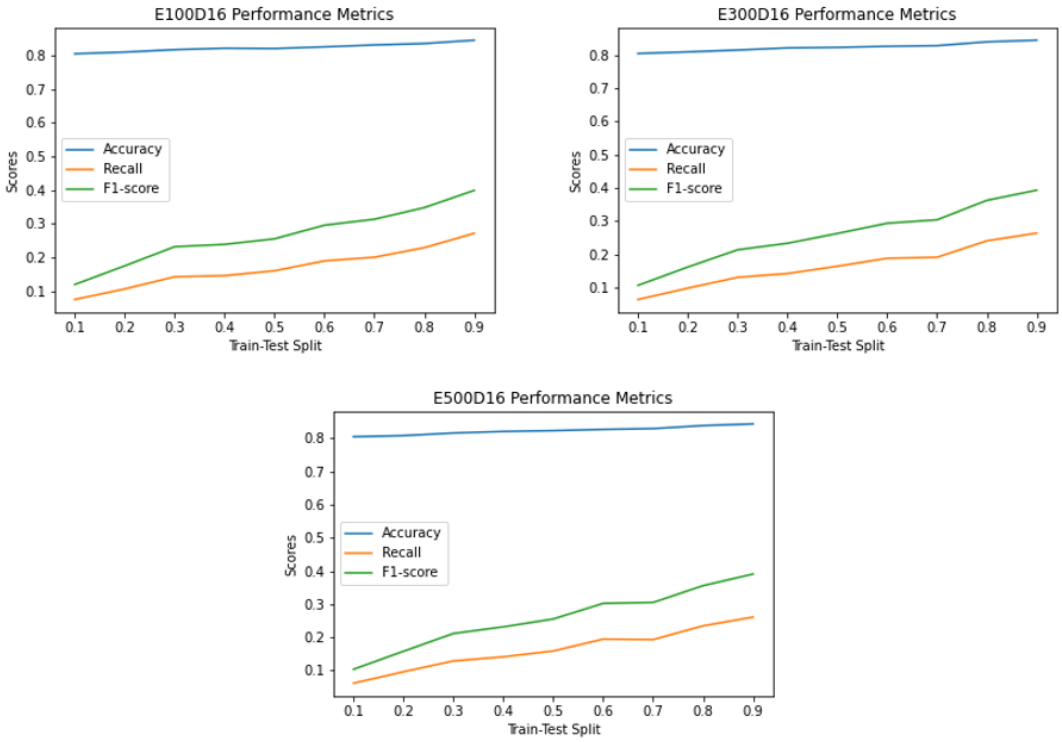

Figure 4.13: Learning Curves of Random Forest instances

The highest scores were seen for train-test split 0.9-0.1. Tables 4.11 and 4.12 show confusion matrices and corresponding performance metrics for all instances, respectively; from which the following is noted:

Table 4.11: Confusion Matrices for Random Forest models at a 0.9-0.1 train-test split

		N	HF	LVH	MI	AR	AS
E100D16	**TN**	102	94	96	86	95	89
	FP	0	3	0	1	0	1
	FN	15	14	20	26	14	17
	TP	2	8	3	6	10	12
E300D16	**TN**	102	95	96	87	95	89
	FP	0	2	0	0	0	1
	FN	14	14	21	27	14	18
	TP	3	8	2	5	10	11
E500D16	**TN**	102	95	96	86	95	89
	FP	0	2	0	1	0	1
	FN	14	15	21	26	14	18
	TP	3	7	2	6	10	11

- Model instances for Data tests show a range of predictive abilities with the pathologies included in the study; consistent with performances in previous (E/M) tests. Across all model instances, HF, AR and AS have high (similar) TP values - greatly improving for HF and worsening for N. Previously, model E100D16 best predicted N and AR (and occasionally well for MI and AS). Similarly, E300D16 and E500D16 predicted better for HF, LVH and MI.
- As seen previously, TN and FP values are minimal; resulting in the high Accuracies and Precisions.
- The models correctly identify small numbers of TP for this train-test ratio and more FN than previously - seen by the low Recalls. The low F1-scores reinforce the poor trade-off between Recall and Precision.

Table 4.12: Averaged performance metrics from 0.9-0.1 train-test ratio

	E100D16	**E300D16**	**E500D16**
Accuracy	0,845	0,845	0,843
Precision	0,918	0,953	0,925
Recall	0,272	0,263	0,261
F1-score	0,399	0,393	0,391

For the features selected, the shortlisted Random Forest instances perform in similar ranges to the Engineering tests despite the altered train-test ratios. This is attributed to the many features with inherently little mutual information with the targets. However, model instance E100D16 was identified as most successful, achieving the highest performance metric scores at train-test split 0.9-0.1.

4.4.3.3 Neural Network Results

The figures and tabulated results of the Neural Network instances in this section are representations of those in Appendix C, Table C.2. The behaviour of all models was analysed to identify the most successful model. The learning curves of the 3 model instances for train-test splits 0.1-0.9 and 0.9-0.1 can be found in Figure 4.14. These ratios were selected for each instance to represent the progression of learning behaviour for models from the minimum to the maximum training set size. The general trends are thus listed below:

- As model dimensions increase (going down a column), the noise present in the Training (dashed-lined) curves decrease.

- As the training set size increases (from left to right), the degree of noise in the Training curves decrease.
- Learning rates increase as model dimensions and training set size increase. This is seen by the rate of change of the gradients in the Training Loss curves of the models for split 0.1-0.9 (left) compared to 0.9-0.1 (right). For example, D1W16 (top left) has a low learning rate at 0.1-0.9 that improves by 0.9-0.1; whereas D3W256 has an increasingly high learning rate present from split 0.1-0.9.
- All models with train-test ratio 0.1-0.9 (left column) experience noise in their Training Loss curves. The training set is unrepresentative at this ratio. Unrepresentative training data prevents the model from learning the problem and thus making good predictions on validation sets - seen by the low validation scores.
- The noise in Validation Loss curves at larger train set sizes increases as model dimensions increase; opposite to that of Training Loss at lower train set sizes. Unrepresentative validation data does not provide enough information to properly assess the predictive capabilities of a model - also seen by the low validation scores produced.
- For all train-test ratios tested, the gap between the Training and Validation curves for Loss, Accuracy and Recall increases with model dimensions. Gaps in specific curves have unique interpretations:
  - For D1W16 in 0.1-0.9 and 0.9-0.1, a small gap is present between the Training and Validation Loss curves (usually ideal) which stabilize at a high values; indicating underfitting. Accuracy curves reveal underfitting as both Training and Validation curves show improvement but remain at low values. Model complexity of D1W16 may also be insufficient for the task.
  - For both D2W64 and D3W256, there is overfitting with unrepresentative training data in test split 0.1-0.9. These model complexities are more than what is required for the given data. Both models show improvements in all Training curves, however overfitting (unrepresentative training data) resulted in very large gaps with respective Validation curves. Validation curves are relatively flat in comparison, showing little to no improvement with experience. As the training set sizes increase (as in 0.9-0.1), the degree of overfitting (gap) decreases and Validation curves show improvements. The effects of unrepresentative data are seen in the poor final scores for all Validation sets relative to Training sets.

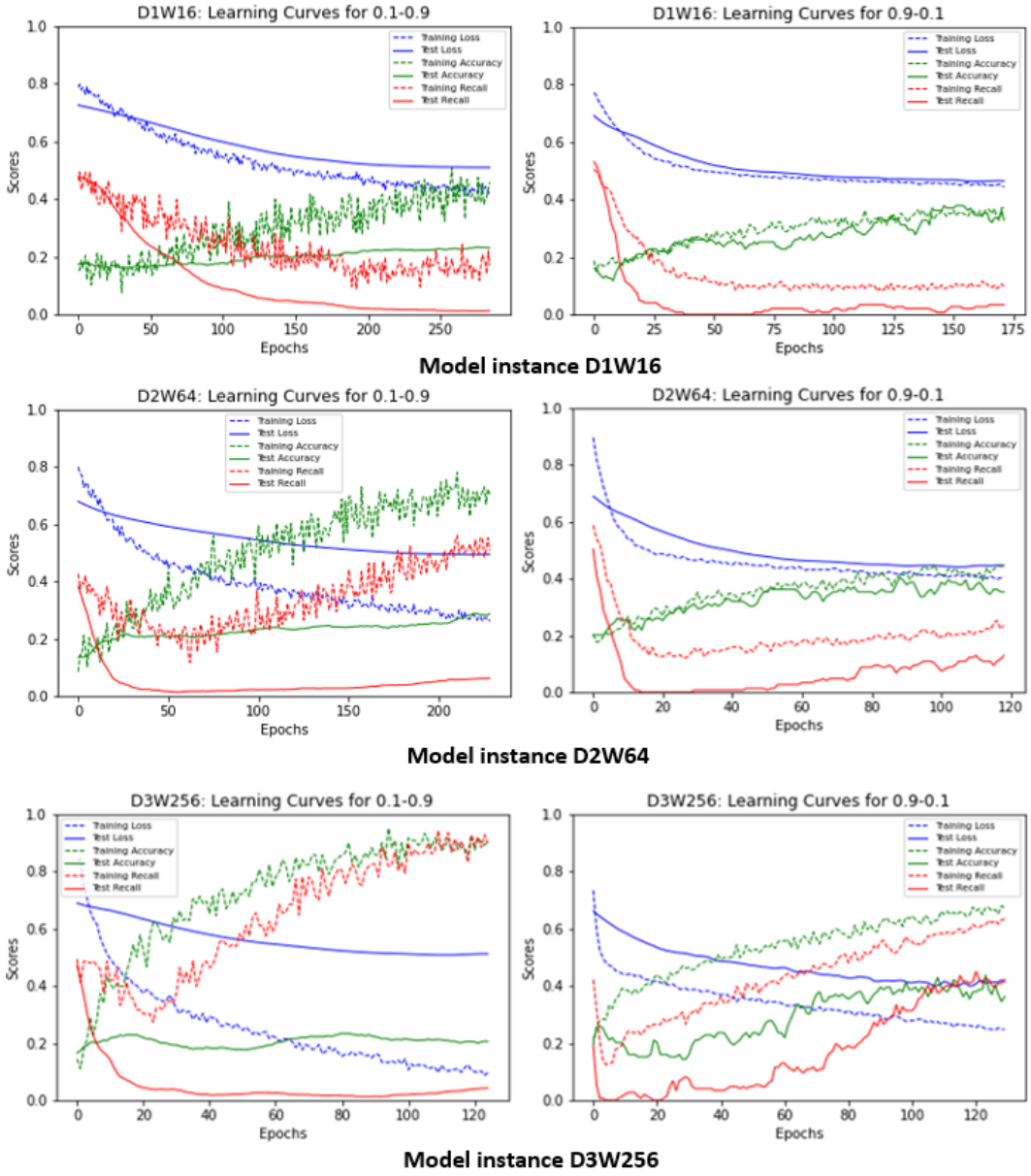

Figure 4.14: Learning Curves for Test-Train ratios 0.1-0.9 (left) and 0.9-0.1 (right)

Across all model instances and train-test ratios, performances are generally poor, exhibiting some degree of under- or overfitting. This implies that the dataset, in general, is unrepresentative in that no model instance could identify inherent patterns within it. As a result, increasing the training data did not have the desired effect of improving performances.

Based on the above observations, instances D1W16 and D3W256 cannot qualify as suitable models. Model D1W16 consistently exhibited underfitting and slow learning rates, irrespective of the train-test split used. Model D3W25 was too complex for the data and/or classification task at hand, consistently overfitting the data with undesirably high learning rates. As a result, model D2W64 can be identified as the best-suited instance. D2W64 learning curves of remaining train-test splits are seen in Figure 4.15. From the learning curves, and resulting performance metrics in Table C.2, Appendix C; the following points were considered in identifying the optimal train-test ratio:

- The learning rate steadily moves from relatively low for the smallest split (0.1-0.9) (from Figure 4.14), to relatively high as the training set size increases.
- The Training Loss stabilizes better as curves begin to flatten with increasing train-test ratios. The gap between the Training and Validation Loss curves continually decreases too.
- Training set sizes of 10-40% have the largest amount of noise present in their Training curves. These train-test ratios also result in larger gaps between the Training and Validation curves for Loss (in 0.1-0.9 and 0.2-0.8), Accuracy and Recall. Additionally, the final Validation scores for all metrics are among the poorest: showing low Recalls and Accuracies, and the high Losses.
- Training sizes of 50-80% contain less noise for the Training curves than those aforementioned. This indicates better learning from the training subset, although gaps between the Training and Validation curves are significant. Validation curves improve as training set sizes increase; however, the noise present increases too.
- For train-test split 0.9-0.1, although overfitting (gap) is minimal, the final scores for Accuracy and Recall are among the lowest.
- Looking at specific performance metrics on the validation subset, 3 train-test ratios stand out for D2W64. Split 0.5-0.5 results in the highest Validation Accuracy and the smallest gap between the Training and Validation Accuracy curves. Split 0.6-0.4 results in the highest Recall and lowest Loss. Split 0.9-0.1, however, results in the smallest overall degree of overfitting.

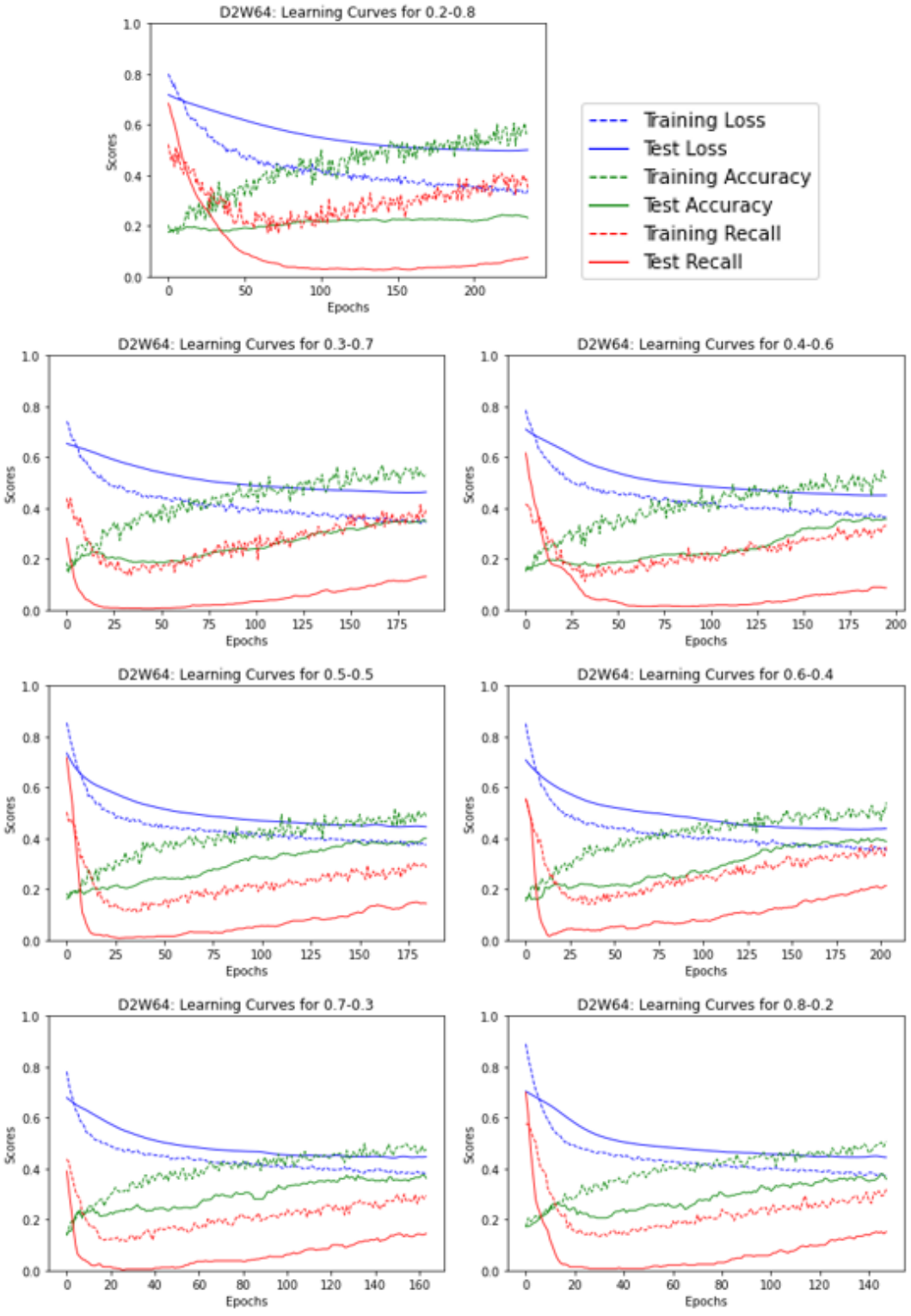

Figure 4.15: D2W64 Learning Curves for all Test-Train ratios

- Conversely, the lowest performance metrics on the Validation set are of splits 0.1-0.9 and 0.2-0.8. The smallest training size also results in the largest gaps between all Training and Validation curves.

Considering the above observations, the most promising training set sizes identified were 50%, 60% and 90%. Of these three training set sizes, split 0.6-0.4 resulted in the highest Recall while 0.9-0.1 exhibited the best learning behaviour (relatively). For final model comparisons, D2W64 was thus considered with the 0.9-0.1 train-test split. Even though noise in its Validation curves show that the validation dataset is unrepresentative; it is preferred over the degree of overfitting present for split 0.6-0.4.

4.5 Final Model Selection

This final section compares the most successful model instances, at their selected train-test splits, to identify which model best classifies selected cardiac abnormalities. Random Forest E100D16 and Neural Network D2W64 will be assessed on the quality of their predictions, before a discussion on the influence of model type, architecture and hyper-parameters is presented. The section concludes with commentary on the overall findings and clinical implications of the selected model.

Mutual Information Scores and Train-Test Ratios

The mutual information scores vary with the size of the training set from which they are calculated. Observing those of the 0.9-0.1 train-test split in Figure 4.16, more features decrease (21) than increase (17) relative to the scores test EM0. As a result, it was expected that both models would not perform better than test EM0, having mostly lower feature scores.

Predictions

Upon assessing model predictions against true labels, both models exhibited many similarities. For most multi-label samples, the models were only able to predict 1 of the 2/3 labels correctly. In the case of the Random Forest E100D16, there were much less FP compared to those of the Neural Network D2W64, as it would often misclassify a sample of one pathology class as another.

Comparing New Results

Models E100D16 and D2W64 were re-initialised and re-run on re-shuffled data at train-test split 0.9-0.1. As a common means of comparison, confusion matrices Table 4.13 were constructed from each model's predictions against the true labels. The resulting performance metrics found in Table 4.14. D2W64 has higher FP, FN and TP values compared to E100D16.

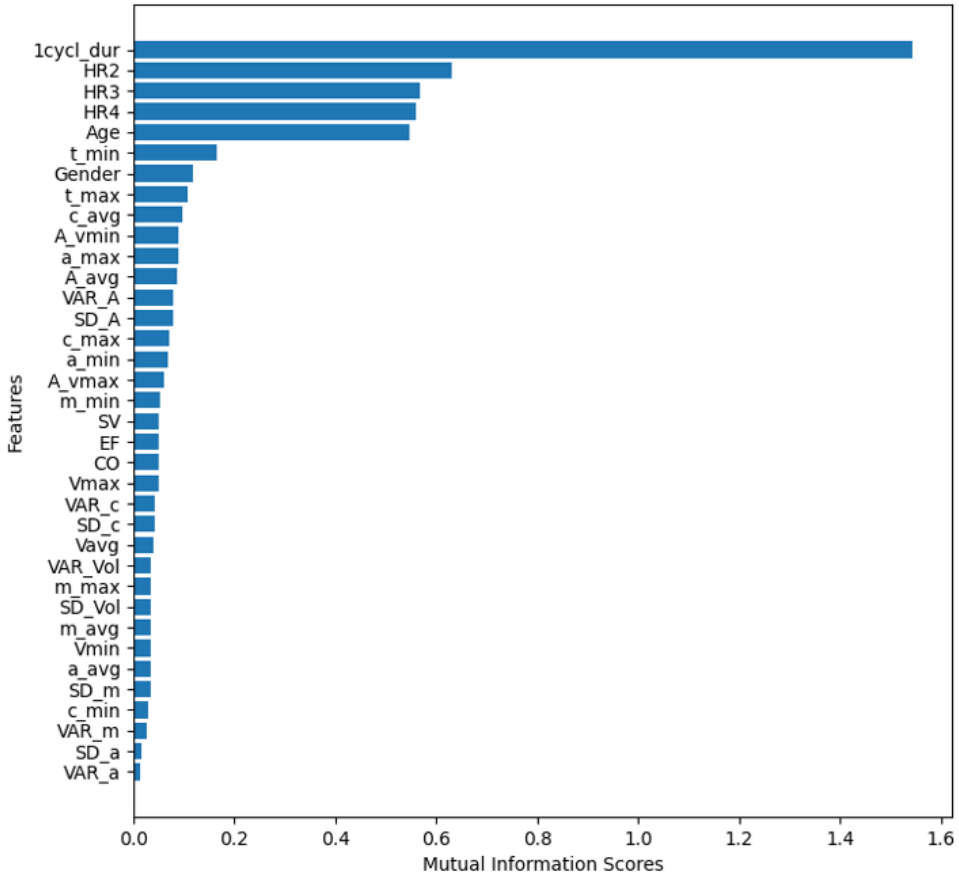

Figure 4.16: Mutual Information Scores for Train-Test Split 0.9-0.1

Table 4.13: Confusion Matrices of E100D16 and D2W64

		N	HF	LVH	MI	AR	AS
E100D16	**TN**	101	92	91	97	92	95
	FP	0	0	0	1	0	0
	FN	15	20	18	13	17	17
	TP	3	7	10	8	10	7
D2W64	**TN**	95	91	90	96	92	93
	FP	6	1	1	2	0	2
	FN	10	19	26	18	22	20
	TP	8	8	2	3	5	4

Table 4.14: Performance Metrics of E100D16 and D2W64

	E100D16	**D2W64**
Accuracy	0,859	0,822
Precision	0,981	0,732
Recall	0,304	0,218
F1-score	0,458	0,314

Label-swapping and Class Confusions

As previously noted, both models perform in small ranges of each other. The Random Forest model outputs higher performance metrics, mostly due to the differences in FN and FP values. Upon inspection of D2W64's 12 FP predictions, there were a range of classes confused (labels switched). The most common labels switched were (i) N and LVH, and (ii) N and MI (occuring 3 times each) - the single FP of E100D16 was as per (ii). Other classes confused by D2W64 included N-AS, N-HF, MI-AS, MI-AR, and LVH-AS. Confusions of aortic valve abnormalities (AR/AS) can be attributed to the lack of features that inform models on valve action more specifically, as 2-chamber views do not view the aortic valve directly.

Multi-label Predictions

The test subset contains 7 multi-label samples, as per Table 4.15. Both models recover 2 of these combinations at low rates. Table 4.15 columns show tallies of the multi-label samples discovered by each model compared to the Total in the test set. Missed combinations account for FN values of both models in Table 4.13; missing many HF, MI, AS and AR samples. Both models were able to predict 2 labels, however there was no discovery of samples with 3 labels.

Table 4.15: Tallies and Total of correctly predicted Multi-label samples per model

Class Combinations	E100D16	D2W64	Total
AS-AR	4	1	7
MI-AS-AR	0	0	2
MI-AR	0	0	1
MI-AS	0	0	2
HF-MI	4	1	7
HF-AR	0	0	1
HF-LVH	0	0	2

Best-suited Model Identified

Based on the relative quality of the predictions and resulting performance metrics, the Random Forest instance is the most suitable classifier for this application. E100D16 correctly predicts both single label and multi-label samples more frequently than D2W64, thus produces slightly better performance metrics, although both sets of results are relatively poor.

Input Data and Noise effects

Special mention of the feature quality is made to address the underlying issue of unrepresentative data, which ultimately caps the performance of all model instances in all tests. The consistently low performance metric scores despite the number of samples available suggests that the data has little discoverable patterns. As a result, the true predictive capacities of both model types tested are not represented well in the results.

A deeper exploration of the pipeline steps was done to identify the cause / degree of variation-correlation trade-off present in the input data. Figure 4.3 (§4.2) presents an example of a noisy HF case with the output from the image pre-processing (FIJI) step - during which measurements were acquired prior to geometry calculations. Comparing the unaugmented and augmented echocardiogram images, the naked eye can identify a rough shape of the heart. The unaugmented images were already considerably noisy, with the addition of black noise lessening the prominence/clarity of the foreground. The diminished (white) areas of interest were partially filtered and removed in image pre-processing steps. Noise was further removed using a histogram (of areas) when model input features were being prepared. These observations also apply to inherently noisy original cases augmented by other means.

The effect of noise on features can be seen in Table 4.16 where selected measurements are included to reveal output differences. Parameters calculated by multiplications, divisions, additions or subtractions of these measurements thus exacerbate errors further - seen in the difference of volume (V) values calculated from ellipsoidal radii (c and a) measurements. Volumes were then used to estimate stroke volume (SV), ejection fraction (EF) and cardiac output (CO). They were also used to identify measurements at maximal and minimal volumes - such as associated times (t) and surface areas (A). Lastly, due to the dynamic nature of the data, the statistical measures intended to summarise temporal variations do not represent original data patterns/relationships. As a result, inherent relationships cannot become apparent to the machine learning models. Measurement noise across all input data thus cause model instances to confuse to pathology classes.

Table 4.16: Resulting measurements extracted from unaugmented and noised frame of the Heart Failure case

	c	**a**	**A**	**V**	**m**	**b**
Unaugmented	243,14	137,67	24956	9650767	-8,722	4502,50
Augmented	247,81	145,97	10379	11059009	-5,777	3016,54

Clinical Implications

Considering E100D16 in the context of medical practice, the likelihood of FN and FP do not favour its implementation. For this project, there are a range of low-to-high risk cardiac abnormalities selected. Patients usually go for echocardiograms as a secondary investigation to further understand issues experienced pertaining to their heart or to monitor certain characteristics post some abnormal experience. Diagnoses are successful when information from patient history and practitioner's knowledge / experience are combined. Therefore, with no medical history or medical expert in tandem with the model, false negative / positive predictions could have serious implications if undetected. For example: a false negative classification for non-lethal abnormality may not be as serious. However, false negatives for more serious issues have mortal consequences. Alternatively, false positives may result in patients undergoing unnecessary treatments; wasting both time, money and resources. As a result, this model requires further development before clinical trials or any implementation can occur.

4.6 Conclusion

The chapter opened detailing the image pre-processing (FIJI) outputs and the feature definitions to prepare data as suitable model inputs. The issues associated with noisy measurement/feature extractions presented themselves early; in the FIJI outputs. This was then confirmed by the mutual information scores (§4.3) where most geometric and statistical features derived from images were poorly related to the targets. These insights remained consisted throughout all tests, as model performances improved when features with higher mutual information scores were included. Some of the more valuable features were not clinically relevant; thus, feature selection for Data tests were based more on medical practice. Performance did not improve much for the training set sizes tested, as the model instances were found to poorly identify innate relationships in the data. Predictions were closely scrutinised to highlight specific issues found to be common to both models tested - such as failure to identify most multi-label samples. Random Forest E100D16 was selected as the best suited model on the basis of relatively better detection of multi-label instances and higher performance metric scores.

5 Limitations & Recommendations

This chapter presents a discussion on the gaps/limitations on this project together and related recommendations. All practical work phases of project are considered, including the clinical and engineering aspects of all events in Figure 3.1.

The initial reasoning for conservative ranges in the data augmentations (applied to the echocardiogram video frames) was to maintain realistic synthetic samples. Based on the results for both models, more training data would have better contributed to improving overall performance, navigating the bias-variance (underfit-overfit) trade-off and balancing data in all classes. Therefore, the respective ranges for rotations and translations could be widened to provide more distinguishable variations in the extracted measurements from the images. Additionally, conservative augmentations could have been applied to the EMR data as these features were merely repeated for original and corresponding augmented cases. Minor adjustments to variables, such as *Age*, could have been reasonable - especially where cases were augmented more than twice.

Based on results from the tests and mutual information scores, medical history and specific patient data are major contributors to the predictive capabilities of machine learning models in this application. This lack of information ultimately caps model performances; compromising validity and effectiveness. Despite ethical regulations barring access to personal details, there is still room to collect other medically informative data; such as previous procedures, blood pressure (usually monitored before a procedure), or stress-strain information (available in more recent echocardiogram reports). Such information allows for other physiological attributes to be estimated and included as features; such as instantaneous pressure, elastance and compliance, or cardiac muscle characteristics. Research involving echocardiograms are normally integrated with supporting clinical or patient data. This presents a challenge in cases where conflicting information exists about prognosis (Mandes *et al.*, 2020).

Another limitation is the lack of medical data available over longer periods for more recently developed methods. Echocardiography, for example, is a newer technique that has undergone many additions and improvements for

its various imaging modalities, but lacks large public databases (Litjens *et al.*, 2019). As a result, there is high variability in existing literature from the models developed, architectural choices and evaluation methods; thus comparisons are difficult where there is little/no overlap (Bizopoulos and Koutsouris, 2019).

Of the many views and imaging modes available from an echocardiogram, only one mode and orientation was used to extract information in this project; viz. apical 2-chamber view. Including other views or modalities (with accompanying domain knowledge) would allow for more feature extractions. For example, information from a collection of views would provide more insightful features on global/localised behaviours of valves, papillary muscles or surround vasculature. More views/features about valves would have informed models better on abnormalities such as Aortic Regurgitation or Aortic Stenosis - not captured in a 2-chamber view.

Machine learning methods applied to any facet of healthcare require sufficiently large amounts of training data to achieve results of comparable quality. Working with medical data involves extensive manual labelling by healthcare specialists (Madani *et al.*, 2018). This is necessary for transdisciplinary research and development where experts in other fields attempt work on health care systems. Across hospital environments, not all data, apart from patient information, is stored in a standardized, well-labelled manner suitable for external research. Moreover, most medical data belongs to the wide *normal* category as opposed to the *abnormal* (Bizopoulos and Koutsouris, 2019); creating vastly unbalanced datasets. This was the case in the project at hand, which necessitated augmentations for sufficient training/testing data.

The complexity of the (relatively) larger Neural Networks included in the study proved to be, or quickly become, too complex for the classification task and data at hand. Other than collecting more training data and implementing early stopping; alternative approaches to minimize overfitting include increasing the dropout rate of hidden layer neurons or learning rate alterations (Zulkifli, 2018). A dropout rate of 30% was used for all Neural Network model sizes, however this can be increased to at least 50%. This would also negate the high learning rates typical of larger models applied to small / unrepresentative datasets. Learning rates themselves are affected by optimisers. The Adam (Adaptive Moment Estimator) optimiser was used in the compilation of the Neural Networks, however, model instances still tended to overfit often. Other optimisers such as AdaGrad (Adaptive Gradient Algorithm) or RMSprop (Root Mean Square Propagation) could be tried (Zulkifli, 2018); however, with the same dataset, improvements may plateau.

From a consultation with a collaborating doctor on the process of diagnosing any cardiac pathology, much of the decision-making is based on both years of training and experience in practice (Van der Bijl, 2021). In both ways, efforts toward transdisciplinary implementation of machine learning demands study or easy access (at least) to applicable domain knowledge. For this project, the intersection between characteristics observed in practice by experts and that which could be extracted from images alone, was the geometry. Although the video frames provided a means of extracting some dynamic / geometric information, it was not enough to inform the models (through better features) for good predictions. Furthermore, geometry was heavily affected by noise content of the images - compromising overall model performance.

Less explored domains were those of computer vision, statistics, mathematics and data science - which ultimately limit model quality and capability. The following areas of improvement were thus observed:

- Alternative computer vision techniques and appropriate quality control checks would optimise image processing steps.
- More complex statistics could have been applied to better capture the behaviour of the heart when used to summarise dynamic features.
- More sophisticated mathematical methods could be used to approximate / fit the extracted information (from images) to some derived function for volume (primarily). For example, fitting non-uniform rational B-splines (NURBS) could be further investigated as opposed to the geometric approach undertaken to approximate the asymmetric geometry of the left ventricle.
- Knowledge and training in the aforementioned fields are foundational in good data science practices; for better intuition about parameter tuning or evaluation methods, etc. Another direction to investigate would be the use of other deep learning architectures. Such alternatives include convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs are normally used for image classification, and RNNs are used for dynamic/temporal input data. Both conditions apply in this project and thus allow videos to be used as training data directly.
- Autoencoders and Generative Adversarial Networks (GANs) are other alternative deep learning architectures that could be used for dimension reductions (Raj, 2018) and realistic augmentations (Goodfellow *et al.*, 2014), respectively. However, successful application of these alternatives requires adequate training data.

6 Conclusion

This study endeavoured to develop a diagnostic tool that employs machine learning algorithms for the task of classifying selected cardiac abnormalities. In this final chapter, a review of all previous content, methods and conclusions are presented in relation to the aims and objectives outlined in Chapter 1.

For the context of this project, the fields of cardiology and supervised machine learning intertwine. Essential theory from both fields was collected and presented in a literature review in Chapter 2. The literature facilitated the translation of medically relevant information to measurable engineering features. Data sourced was in the form of echocardiogram videos, from which measurements were extracted for calculations and feature definitions, detailed with previous work in Chapter 3.

Upon close examination of all model instances in all tests (with associated feature combinations) in Chapter 4, clear model behaviours could be identified for both Random Forests and Neural Networks. In the Engineering and Medical tests, Random Forest architectures showed consistent strengths and weakness in predicting certain pathologies. Neural Network instances that scored highest usually overfit the training data. Both model types showed sensitivity to feature selections and/or data representativeness throughout all results. Upon altering training set sizes in Data tests, ideal train-test ratios were identified for both model type instances shortlisted. However, data representativeness hindered model performances due to the effect of inherent and/or injected noise in the echocardiogram video frames.

Limitations were discussed, in Chapter 5, from a variety of steps in the project, with recommendations or alternatives provided where possible; other solutions being available with more domain knowledge. Much of the latter steps in the pipeline were affected by noise present, and unfortunately the true potential of the models could not be explored. However, all outputs were analysed according to plan, providing valuable insights into model behaviour with regards to internal dimensions / parameters, value of the features included and response to the quality of data for training and testing.

A Appendix 1: Engineering Tests Results

A.1 Mutual Information Scores

The tables included in this section include the Mutual Information scores for all Engineering tests performed. They contain colour-coded blocks: green indicating increased scores compared to baseline EM0, and red indicating decreased scores. Table A.1 contains scores for Tests E1-E5; where single feature deletions were done. Table A.2 contain scores for Tests E12-E12345; where cumulative feature deletions were done. The baseline test results are included in both tables for comparison purposes.

Table A.1: Mutual Information Scores of all individual feature deletions for Engineering Tests

Features	EM0	E1	E2	E3	E4	E5
1cycl_dur	1,3407	1,3314	1,3353	1,3375	1,3353	
A_avg	0,0380	0,0380	0,0380	0,0380	0,0380	0,0380
a_avg	0,0359	0,0359	0,0359	0,0359	0,0359	0,0359
a_max	0,0675	0,0675	0,0675	0,0675	0,0675	0,0675
a_min	0,0548	0,0548	0,0548	0,0548	0,0548	0,0548
A_vmax	0,0707	0,0713	0,0711	0,0695	0,0711	0,0711
A_vmin	0,0964	0,0981	0,0979	0,0987	0,0979	0,0979
Age	0,6334	0,6170		0,6170	0,5913	0,5913
b_avg	0,0698	0,0698	0,0698	0,0698	0,0698	0,0698
b_max	0,0302	0,0302	0,0302	0,0302	0,0302	0,0302
b_min	0,0618	0,0618	0,0618	0,0618	0,0618	0,0618
c_avg	0,0873	0,0873	0,0870	0,0873	0,0870	0,0870
c_max	0,0737	0,0737	0,0737	0,0737	0,0737	0,0737
c_min	0,0157	0,0157	0,0157	0,0157	0,0157	0,0157
CO	0,0584	0,0584	0,0584	0,0584	0,0584	0,0584
EF	0,0623	0,0623	0,0623	0,0623	0,0623	0,0623
Gender	0,1375	0,0942	0,1323		0,1323	0,1323

Continued on next page

Table A.1 – *Continued from previous page*

HR2	0,5847		0,5951	0,5806	0,5951	0,5719
HR3	0,5285		0,5953	0,5338	0,5953	0,5229
HR4	0,5274		0,5745	0,5041	0,5745	0,6049
m_avg	0,0453	0,0453	0,0453	0,0453	0,0453	0,0453
m_max	0,0146	0,0146	0,0146	0,0146	0,0146	0,0146
m_min	0,0805	0,0805	0,0805	0,0805	0,0805	0,0805
SD_A	0,0927	0,0927	0,0927	0,0927	0,0927	0,0927
SD_a	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101
SD_b	0,0243	0,0243	0,0243	0,0243	0,0243	0,0243
SD_c	0,0646	0,0646	0,0646	0,0646	0,0646	0,0646
SD_m	0,0048	0,0048	0,0048	0,0048	0,0048	0,0048
SD_Vol	0,0494	0,0494	0,0494	0,0494	0,0494	0,0494
SV	0,0623	0,0623	0,0623	0,0623	0,0623	0,0623
t_max	0,0875	0,0778	0,0905	0,0855	0,0905	0,0905
t_min	0,1738	0,1836	0,1688	0,1683	0,1688	0,1688
Test_yr	0,1142	0,1308	0,1233	0,1308		0,0905
VAR_A	0,0932	0,0932	0,0932	0,0932	0,0932	0,0932
VAR_a	0,0109	0,0109	0,0109	0,0109	0,0109	0,0109
VAR_b	0,0310	0,0245	0,0345	0,0135	0,0345	0,0345
VAR_c	0,0621	0,0621	0,0621	0,0621	0,0621	0,0621
VAR_m	0,0140	0,0020	0,0113	0,0100	0,0113	0,0113
VAR_Vol	0,0483	0,0483	0,0483	0,0483	0,0483	0,0483
Vavg	0,0460	0,0460	0,0460	0,0460	0,0460	0,0460
Vmax	0,0520	0,0520	0,0520	0,0520	0,0520	0,0520
Vmin	0,0227	0,0227	0,0227	0,0227	0,0227	0,0227
Xc_avg	0,0483	0,0482	0,0482	0,0482	0,0482	0,0482
Xc_max	0,0148	0,0148	0,0148	0,0148	0,0148	0,0148
Xc_min	0,0522	0,0522	0,0522	0,0522	0,0522	0,0522
Yc_avg	0,0219	0,0219	0,0219	0,0219	0,0219	0,0219
Yc_max	0,0175	0,0175	0,0175	0,0175	0,0175	0,0175
Yc_min	0,0400	0,0400	0,0400	0,0400	0,0400	0,0400

Table A.2: Mutual Information Scores of all cumulative feature deletions for Engineering Tests

Features	EM0	E1	E12	E123	E1234	E12345
1cycl_dur	1,3407	1,3314	1,3377	1,3410	1,3356	
A_avg	0,0380	0,0380	0,0380	0,0380	0,0380	0,0380
a_avg	0,0359	0,0359	0,0359	0,0359	0,0359	0,0359
a_max	0,0675	0,0675	0,0675	0,0675	0,0675	0,0675

Continued on next page

Table A.2 – *Continued from previous page*

a_min	0,0548	0,0548	0,0548	0,0548	0,0548	0,0548
A_vmax	0,0707	0,0713	0,0703	0,0697	0,0696	0,0700
A_vmin	0,0964	0,0981	0,0999	0,0983	0,0990	0,0987
Age	0,6334	0,6170				
b_avg	0,0698	0,0698	0,0698	0,0698	0,0698	0,0702
b_max	0,0302	0,0302	0,0302	0,0302	0,0302	0,0302
b_min	0,0618	0,0618	0,0618	0,0618	0,0618	0,0618
c_avg	0,0873	0,0873	0,0873	0,0873	0,0873	0,0870
c_max	0,0737	0,0737	0,0737	0,0737	0,0737	0,0737
c_min	0,0157	0,0157	0,0157	0,0157	0,0157	0,0157
CO	0,0584	0,0584	0,0584	0,0584	0,0584	0,0584
EF	0,0623	0,0623	0,0623	0,0623	0,0623	0,0623
Gender	0,1375	0,0942	0,1003			
HR2	0,5847					
HR3	0,5285					
HR4	0,5274					
m_avg	0,0453	0,0453	0,0453	0,0453	0,0453	0,0453
m_max	0,0146	0,0146	0,0146	0,0146	0,0146	0,0146
m_min	0,0805	0,0805	0,0805	0,0805	0,0805	0,0805
SD_A	0,0927	0,0927	0,0927	0,0927	0,0927	0,0927
SD_a	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101
SD_b	0,0243	0,0243	0,0243	0,0243	0,0243	0,0243
SD_c	0,0646	0,0646	0,0646	0,0646	0,0646	0,0646
SD_m	0,0048	0,0048	0,0048	0,0048	0,0048	0,0048
SD_Vol	0,0494	0,0494	0,0494	0,0494	0,0494	0,0494
SV	0,0623	0,0623	0,0623	0,0623	0,0623	0,0623
t_max	0,0875	0,0778	0,0788	0,0955	0,0883	0,0928
t_min	0,1738	0,1836	0,1635	0,1647	0,1668	0,1837
Test_yr	0,1142	0,1308	0,1286	0,1550		
VAR_A	0,0932	0,0932	0,0932	0,0932	0,0932	0,0932
VAR_a	0,0109	0,0109	0,0109	0,0109	0,0109	0,0109
VAR_b	0,0310	0,0245	0,0314	0,0383	0,0318	0,0280
VAR_c	0,0621	0,0621	0,0621	0,0621	0,0621	0,0621
VAR_m	0,0140	0,0020	0,0020	0,0122	0,0108	0,0194
VAR_Vol	0,0483	0,0483	0,0483	0,0483	0,0483	0,0483
Vavg	0,0460	0,0460	0,0460	0,0460	0,0460	0,0460
Vmax	0,0520	0,0520	0,0520	0,0520	0,0520	0,0520
Vmin	0,0227	0,0227	0,0227	0,0227	0,0227	0,0227
Xc_avg	0,0483	0,0482	0,0482	0,0483	0,0483	0,0482
Xc_max	0,0148	0,0148	0,0148	0,0148	0,0148	0,0148
Xc_min	0,0522	0,0522	0,0522	0,0522	0,0522	0,0522
Yc_avg	0,0219	0,0219	0,0218	0,0219	0,0219	0,0219

Continued on next page

Table A.2 – *Continued from previous page*

Yc_max	0,0175	0,0175	0,0175	0,0175	0,0175	0,0175
Yc_min	0,0400	0,0400	0,0400	0,0400	0,0400	0,0400

A.2 Random Forest Results

The following tables are to be read in conjunction with one another, as the model entries in Table A.3 all coincide with their performance metrics in Table A.4. Coloured blocks in Table A.4 highlight the highest scores for the respective Engineering test.

Table A.3: Best performing Random Forest models based on performance metrics for each pathology, listed by architectural descriptors

Test	**N**	**HF**	**LVH**	**MI**	**AR**	**AS**
EM0	E100:D16				E100:D16	E100:D16
	E150:D16		E150:D16	E150:D16		
		E200:D16	E200:D16			
			E500:D16			
E1	E100:D16					
	E150:D16	E150:D16	E150:D16			
					E300:D16	
				E400:D16	E400:D16	E400:D16
				E450:D16		E450:D16
				E500:D16	E500:D16	E500:D16
E2					E100:D16	
			E150:D16		E150:D16	
				E200:D16	E200:D16	E200:16
				E250:D16		
			E300:D16	E300:D16	E300:D16	
				E350:D16	E350:D16	
			E400:D16	E400:D16	E400:D16	
			E450:D16	E450:D16	E450:D16	
	E500:D16	E500:D16	E500:D16	E500:D16	E500:D16	
E3						E100:D16
	E150:D16		E150:D16			
	E200:D16		E200:D16			
			E250:D16	E250:D16	E250:D16	
			E300:D16			
			E350:D16			

Continued on next page

Table A.3 – *Continued from previous page*

			E400:D16 E450:D16 E500:D16	E400:D16 E450:D16 E500:D16		
E4	E100:D16 E150:D16 E250:D16	 E500:D16	E100:D16 E200:D16	 E350:D16 E450:D16 E500:D16	E100:D16 E250:D16 E500:D16	 E450:D16
E5	E100:D16 E150:D8 E150:D16 E200:D16 E250:D16 E300:D16 E450:D16 E500:D16	 E200:D16	 E300:D16 E350:D16 E400:D16 E450:D16 E500:D16	 E200:D16 E250:D16 E300:D16 E350:D16 E400:D16 E450:D16 E500:D16	E100:D16	 E150:D16
E12	 E200:D16	 E200:D16	 E250:D16 E300:D16 E350:D16 E400:D16 E450:D16 E500:D16	E100:D16	 E200:D16	 E400:D16 E450:D16
E123	E150:D8 E200:D8 E250:D8 E300:D8 E350:D8 E400:D8 E450:D8 E500:D8 E100:D16 E150:D16 E200:D16 E250:D16 E300:D16 E350:D16	 E150:D16	 E350:D16	 E350:D16	 E150:D16	 E100:D16 E150:D16

Continued on next page

Table A.3 – *Continued from previous page*

	E400:D16		E400:D16			
	E450:D16		E450:D16	E450:D16		
	E500:D16		E500:D16	E500:D16		
E1234	E100:D8		E100:D8			
	E150:D8		E150:D8			
	E200:D8		E200:D8			
	E250:D8		E250:D8			
	E300:D8		E300:D8			
	E350:D8					
	E400:D8					
	E450:D8		E450:D8			
	E500:D8					
	E100:D16		E100:D16			E100:D16
			E150:D16	E150:D16	E150:D16	
			E200:D16		E200:D16	
			E250:D16			
			E300:D16			
		E350:D16	E350:D16			
		E400:D16	E400:D16			
			E450:D16			
			E500:D16			
E12345	E100:D16	NONE		NONE	E100:D16	E100:D16
	E150:D16				E150:D16	
			E200:D16		E200:D16	
					E400:D16	
					E450:D16	
					E500:D16	

Table A.4: Averaged Performance Metrics for Best Performing Random Forest Models per Engineering Test

Test	Model	Accuracy	Recall	Precision	F1-score
EM0	E100:D16	0,8333	0,2171	0,9309	0,3357
	E150:D16	0,8368	0,2264	0,9470	0,3523
	E200:D16	0,8354	0,2201	0,9493	0,3404
	E500:D16	0,8326	0,2063	0,9372	0,3243
	E200:D16	0,83544	0,22011	0,94928	0,34039
	E500:D16	0,83263	0,20626	0,93718	0,32433
E1	E100:D16	0,8136	0,1130	0,8512	0,1947
	E150:D16	0,8179	0,1325	0,8824	0,2222

Continued on next page

Table A.4 – *Continued from previous page*

	E300:D16	0,8172	0,1230	0,9167	0,2101
	E400:D16	0,8193	0,1348	0,9117	0,2276
	E450:D16	0,8193	0,1314	0,9213	0,2230
	E500:D16	0,8200	0,1390	0,9153	0,2327
E2	E100:D16	0,8221	0,1787	0,8442	0,2793
	E150:D16	0,8249	0,1830	0,8941	0,2862
	E200:D16	0,8256	0,1822	0,9179	0,2860
	E250:D16	0,8242	0,1713	0,9214	0,2692
	E300:D16	0,8242	0,1757	0,9179	0,2728
	E350:D16	0,8242	0,1749	0,8604	0,2781
	E400:D16	0,8256	0,1847	0,8699	0,2866
	E450:D16	0,8270	0,1874	0,9255	0,2918
	E500:D16	0,8284	0,1964	0,9281	0,3044
E3	E100:D16	0,8277	0,1319	0,8948	0,2192
	E150:D16	0,8319	0,1318	0,9335	0,2141
	E200:D16	0,8319	0,1153	0,9385	0,1916
	E250:D16	0,8298	0,1284	0,9258	0,2116
	E300:D16	0,8291	0,1165	0,9247	0,1941
	E350:D16	0,8291	0,1094	0,9265	0,1842
	E400:D16	0,8319	0,1115	0,9318	0,1897
	E450:D16	0,8298	0,1157	0,9265	0,1943
	E500:D16	0,8326	0,1115	0,9417	0,1883
E4	E100:D16	0,8312	0,1915	0,9510	0,3102
	E150:D16	0,8305	0,1943	0,9524	0,3083
	E200:D16	0,8270	0,1785	0,9444	0,2831
	E250:D16	0,8298	0,1909	0,9500	0,3045
	E350:D16	0,8312	0,1978	0,9470	0,3114
	E450:D16	0,8319	0,2005	0,9470	0,3152
	E500:D16	0,8326	0,2054	0,9493	0,3199
E5	E100:D16	0,8326	0,1970	0,9537	0,3153
	E150:D8	0,8165	0,1219	0,9249	0,2088
	E150:D16	0,8319	0,2044	0,9351	0,3186
	E200:D16	0,8326	0,2074	0,9414	0,3218
	E250:D16	0,8319	0,2091	0,9375	0,3218
	E300:D16	0,8319	0,2085	0,9354	0,3227
	E350:D16	0,8312	0,2071	0,9315	0,3163
	E400:D16	0,8291	0,1934	0,9318	0,3014
	E450:D16	0,8319	0,2051	0,9394	0,3195
	E500:D16	0,8312	0,1995	0,9348	0,3154
E12	E100:D16	0,8108	0,1105	0,8121	0,1868
	E200:D16	0,8136	0,1274	0,8732	0,2086
	E250:D16	0,8129	0,1183	0,9033	0,1937

Continued on next page

Table A.4 – *Continued from previous page*

	E300:D16	0,8122	0,1183	0,8699	0,1933
	E350:D16	0,8101	0,1066	0,8397	0,1774
	E400:D16	0,8122	0,1134	0,8984	0,1853
	E450:D16	0,8101	0,1022	0,8690	0,1702
	E500:D16	0,8136	0,1128	0,9097	0,1896
E123	E150:D8	0,7961	0,0070	0,6667	0,0402
	E200:D8	0,7961	0,0070	0,6667	0,0402
	E250:D8	0,7961	0,0070	0,6667	0,0402
	E300:D8	0,7961	0,0070	0,6667	0,0402
	E350:D8	0,7961	0,0070	0,6667	0,0402
	E400:D8	0,7968	0,0107	0,7778	0,0410
	E450:D8	0,7968	0,0107	0,7778	0,0410
	E500:D8	0,7975	0,0107	0,8333	0,0414
	E100:D16	0,8094	0,0988	0,8660	0,1646
	E150:D16	0,8143	0,1175	0,9352	0,1931
	E200:D16	0,8115	0,1031	0,9271	0,1728
	E250:D16	0,8108	0,1047	0,9259	0,1718
	E300:D16	0,8115	0,1000	0,9333	0,1673
	E350:D16	0,8115	0,1007	0,9333	0,1680
	E400:D16	0,8115	0,0966	0,9359	0,1638
	E450:D16	0,8115	0,0994	0,9286	0,1675
	E500:D16	0,8108	0,0926	0,9306	0,1580
E1234	E100:D8	0,8017	0,0520	0,8611	0,0923
	E150:D8	0,7996	0,0403	0,8083	0,0878
	E200:D8	0,7989	0,0374	0,8667	0,0809
	E250:D8	0,7996	0,0374	0,8750	0,0815
	E300:D8	0,7996	0,0374	0,8750	0,0815
	E350:D8	0,7989	0,0338	0,8750	0,0739
	E400:D8	0,7982	0,0338	0,8667	0,0734
	E450:D8	0,7996	0,0416	0,8800	0,0885
	E500:D8	0,7989	0,0380	0,8800	0,0809
	E100:D16	0,8080	0,0930	0,9021	0,1527
	E150:D16	0,8080	0,0883	0,8244	0,1501
	E200:D16	0,8066	0,0868	0,8160	0,1450
	E250:D16	0,8059	0,0821	0,8333	0,1381
	E300:D16	0,8073	0,0862	0,8438	0,1441
	E350:D16	0,8066	0,0875	0,8426	0,1432
	E400:D16	0,8066	0,0875	0,8426	0,1432
	E450:D16	0,8059	0,0834	0,8382	0,1383
	E500:D16	0,8045	0,0763	0,8333	0,1284
E12345	E100:D16	0,8059	0,0768	0,8417	0,1337
	E150:D16	0,8052	0,0664	0,8333	0,1188

Continued on next page

Table A.4 – *Continued from previous page*

	E200:D16	0,8052	0,0687	0,7778	0,1217
	E400:D16	0,8031	0,0651	0,7278	0,1150
	E450:D16	0,8024	0,0622	0,7333	0,1095
	E500:D16	0,8017	0,0622	0,6778	0,1093

A.3 Neural Networks Results

Each Neural Network model performance is summarised for all Engineering tests. Final scores of the training and validation subsets (differentiated by $v\_$) are presented in Table A.5, with blue coloured blocks highlighting the best performing model per test.

Table A.5: Averaged Performance Metrics for Best Performing Neural Network Models per Engineering Test

Test	Model	loss	accuracy	recall	precision	v_loss	v_accuracy	v_recall	v_precision
EM0	D1W16	0,4323	0,4034	0,1517	0,5768	0,4585	0,3840	0,0345	0,4545
	D1W64	0,3586	0,5333	0,3268	0,7109	0,4358	0,4135	0,1172	0,5000
	D1W256	0,2591	0,6663	0,5835	0,8344	0,4361	0,4093	0,2621	0,5891
	D2W16	0,2591	0,6663	0,5835	0,8344	0,4361	0,4093	0,2621	0,5891
	D2W64	0,3660	0,5048	0,3411	0,6972	0,4330	0,4219	0,1345	0,5909
	D2W256	0,1892	0,7635	0,7343	0,8758	0,4166	0,4388	0,3207	0,6039
	D3W16	0,4577	0,3590	0,0682	0,5468	0,4632	0,3038	0,0069	0,2222
	D3W64	0,3517	0,5533	0,3662	0,6974	0,4173	0,4304	0,2483	0,5669
	D3W256	0,1240	0,8163	0,8447	0,8945	0,3922	0,5401	0,5000	0,6808
E1	D1W16	0,4401	0,3928	0,1176	0,5928	0,4622	0,3544	0,0172	0,5000
	D1W64	0,3840	0,4857	0,2711	0,6331	0,4408	0,4093	0,0793	0,6389
	D1W256	0,2763	0,6526	0,5332	0,8060	0,4350	0,4051	0,2000	0,5631
	D2W16	0,4481	0,3442	0,0817	0,5322	0,4597	0,3376	0,0103	0,4286
	D2W64	0,3737	0,5016	0,3268	0,6642	0,4242	0,4135	0,1931	0,5895
	D2W256	0,1364	0,8237	0,8223	0,8998	0,4315	0,5274	0,4172	0,6471
	D3W16	0,4791	0,3031	0,0449	0,4464	0,4703	0,3544	0,0207	0,6667
	D3W64	0,3708	0,4879	0,3061	0,6713	0,4321	0,4135	0,1724	0,5102

Continued on next page

Table A.5 – *Continued from previous page*

	D3W256	0,1416	0,7994	0,8106	0,8923	0,3906	0,5359	0,5069	0,6309
E2	D1W16	0,4498	0,3622	0,0978	0,5648	0,4777	0,3080	0,0379	0,6875
	D1W64	0,3776	0,4974	0,2549	0,6961	0,4672	0,3586	0,1207	0,5833
	D1W256	0,4114	0,4182	0,1993	0,6510	0,5732	0,2025	0,0241	0,4118
	D2W16	0,4761	0,3094	0,0709	0,5163	0,4786	0,3038	0,0103	0,7500
	D2W64	0,3849	0,4583	0,2873	0,6584	0,4480	0,3544	0,1448	0,5000
	D2W256	0,1681	0,7899	0,7612	0,8843	0,4099	0,4641	0,3931	0,6064
	D3W16	0,4774	0,2872	0,0287	0,5161	0,4874	0,2869	0,0069	0,2500
	D3W64	0,3859	0,4667	0,2684	0,6458	0,4502	0,3882	0,1655	0,5106
	D3W256	0,1525	0,7888	0,7998	0,8752	0,4343	0,5232	0,4379	0,5853
E3	D1W16	0,4517	0,3611	0,0880	0,4804	0,4658	0,3080	0,0207	0,5000
	D1W64	0,3768	0,5090	0,2953	0,6985	0,4456	0,3797	0,1034	0,5263
	D1W256	0,2712	0,6737	0,5530	0,8031	0,4510	0,3924	0,2103	0,5083
	D2W16	0,4496	0,3474	0,0987	0,5473	0,4489	0,3671	0,0345	0,9091
	D2W64	0,3519	0,5312	0,3555	0,6655	0,4213	0,4388	0,2138	0,6392
	D2W256	0,1614	0,7909	0,7675	0,8824	0,4255	0,4557	0,3207	0,6078
	D3W16	0,4761	0,2851	0,0359	0,4301	0,4720	0,3122	0,0069	0,4000
	D3W64	0,3950	0,4593	0,2612	0,5988	0,4399	0,3755	0,1241	0,5455
	D3W256	0,1316	0,8036	0,8294	0,9041	0,4256	0,4979	0,4379	0,6195
E4	D1W16	0,4427	0,3728	0,1194	0,5473	0,4627	0,3460	0,0172	0,5000
	D1W64	0,3769	0,4921	0,2846	0,7060	0,4497	0,3755	0,1103	0,5424
	D1W256	0,2720	0,6790	0,5503	0,8024	0,4510	0,4135	0,1931	0,5437
	D2W16	0,4535	0,3495	0,0952	0,5550	0,4634	0,3207	0,0621	0,5806
	D2W64	0,3601	0,5079	0,3348	0,6907	0,4355	0,3755	0,1655	0,5647
	D2W256	0,1705	0,7951	0,7603	0,8888	0,4238	0,4641	0,4034	0,5969

Continued on next page

Table A.5 – *Continued from previous page*

	D3W16	0,4734	0,2946	0,0485	0,5000	0,4699	0,3038	0,0172	0,7143
	D3W64	0,3812	0,4752	0,2935	0,6462	0,4394	0,3797	0,1655	0,5106
	D3W256	0,1709	0,7730	0,7792	0,8637	0,4310	0,4684	0,4034	0,5939
E5	D1W16	0,4434	0,3759	0,1329	0,5175	0,4568	0,3629	0,0207	0,4000
	D1W64	0,3644	0,5143	0,2890	0,6736	0,4374	0,4093	0,1103	0,6154
	D1W256	0,3929	0,4784	0,2478	0,6781	0,5748	0,2236	0,1069	0,4306
	D2W16	0,4413	0,3759	0,1059	0,5673	0,4516	0,3291	0,0448	0,6190
	D2W64	0,3508	0,5143	0,3833	0,6789	0,4211	0,4008	0,2172	0,5780
	D2W256	0,1329	0,8416	0,8106	0,9177	0,4007	0,5401	0,4448	0,6355
	D3W16	0,4703	0,2957	0,0539	0,5455	0,4690	0,2911	0,0000	0,0000
	D3W64	0,3615	0,5153	0,3582	0,6763	0,4128	0,4262	0,2345	0,5574
	D3W256	0,1433	0,8046	0,8214	0,8815	0,3745	0,5105	0,4828	0,6452
E12	D1W16	0,4569	0,3664	0,0808	0,5882	0,4800	0,3165	0,0172	0,6250
	D1W64	0,4079	0,4488	0,2217	0,6730	0,4666	0,3502	0,0448	0,5417
	D1W256	0,3072	0,6241	0,4695	0,7877	0,4544	0,3502	0,1552	0,5357
	D2W16	0,4700	0,3126	0,0503	0,5000	0,4846	0,3207	0,0069	0,4000
	D2W64	0,3866	0,4762	0,2738	0,6408	0,4605	0,3418	0,1276	0,5781
	D2W256	0,1965	0,7687	0,7253	0,8587	0,4483	0,4177	0,3138	0,5583
	D3W16	0,4837	0,2756	0,0171	0,5000	0,4813	0,3333	0,0000	0,0000
	D3W64	0,3958	0,4752	0,2648	0,6330	0,4475	0,3207	0,1655	0,5581
	D3W256	0,3403	0,5364	0,4004	0,6778	0,5092	0,1857	0,0483	0,3889
E123	D1W16	0,4857	0,2619	0,0422	0,3507	0,4940	0,2363	0,0000	0,0000
	D1W64	0,4111	0,4245	0,1777	0,6535	0,4694	0,2954	0,0276	0,3200
	D1W256	0,3183	0,5660	0,4210	0,7363	0,4696	0,3586	0,1517	0,4889
	D2W16	0,4723	0,2988	0,0422	0,4608	0,4787	0,2785	0,0241	0,5833

Continued on next page

Table A.5 – *Continued from previous page*

	D2W64	0,4048	0,4277	0,2172	0,6302	0,4583	0,3249	0,0793	0,5610
	D2W256	0,3312	0,5913	0,4013	0,7500	0,5087	0,1814	0,0069	0,2857
	D3W16	0,4775	0,2883	0,0287	0,5614	0,4737	0,3122	0,0034	0,2500
	D3W64	0,4133	0,4161	0,1939	0,6067	0,4583	0,3502	0,1241	0,4557
	D3W256	0,3403	0,5322	0,4372	0,6850	0,5196	0,2321	0,0655	0,3393
E1234	D1W16	0,4746	0,2988	0,0359	0,4938	0,4971	0,2574	0,0034	0,5000
	D1W64	0,4167	0,4129	0,1598	0,6473	0,4833	0,2869	0,0276	0,2963
	D1W256	0,3378	0,5649	0,3770	0,7460	0,4795	0,3038	0,0931	0,5000
	D2W16	0,4785	0,2946	0,0305	0,5862	0,4983	0,2025	0,0000	0,0000
	D2W64	0,4108	0,4182	0,1966	0,6329	0,4748	0,3291	0,0655	0,4634
	D2W256	0,3486	0,5385	0,3680	0,7335	0,5073	0,1646	0,0000	0,0000
	D3W16	0,5012	0,2408	0,0251	0,3944	0,4979	0,2236	0,0000	0,0000
	D3W64	0,4230	0,4065	0,1688	0,6006	0,4783	0,3080	0,0552	0,4000
	D3W256	0,3405	0,5343	0,4210	0,6897	0,5047	0,2658	0,0034	0,3333
E12345	D1W16	0,4944	0,2555	0,0242	0,3462	0,5037	0,2152	0,0000	0,0000
	D1W64	0,4296	0,3939	0,1338	0,6032	0,4889	0,2700	0,0207	0,5455
	D1W256	0,4419	0,3664	0,1050	0,5442	0,5823	0,2068	0,0000	0,0000
	D2W16	0,4933	0,2693	0,0233	0,2955	0,5012	0,2447	0,0000	0,0000
	D2W64	0,4258	0,4192	0,1697	0,6097	0,4783	0,3586	0,0483	0,5000
	D2W256	0,3463	0,5396	0,3636	0,7155	0,5062	0,1688	0,0034	0,1667
	D3W16	0,4933	0,2376	0,0171	0,4419	0,5002	0,2194	0,0000	0,0000
	D3W64	0,4318	0,4002	0,1499	0,5986	0,4792	0,3038	0,0483	0,4828
	D3W256	0,3543	0,5417	0,3842	0,6971	0,5290	0,2405	0,0034	0,2000

B Appendix 2: Medical Tests Results

B.1 Mutual Information Scores

The tables included in this section include the Mutual Information scores for all Engineering tests performed. They contain colour-coded blocks: green indicating increased scores compared to baseline EM0, and red indicating decreased scores. Table B.1 contains scores for Tests M1-M6; where single feature deletions were done. Table B.2 contain scores for Tests M12-M123456; where cumulative feature deletions were done. The baseline test results are included in both tables for comparison purposes.

Table B.1: Mutual Information Scores of all individual feature deletions for Medical Tests

Features	EM0	M1	M2	M3	M4	M5	M6
1cycl_dur	1,3407	1,3446	1,3664	1,3446	1,3446	1,3446	1,3446
A_avg	0,0380	0,0380	0,0380	0,0380		0,0380	0,0380
a_avg	0,0359	0,0359	0,0359	0,0359	0,0359	0,0359	
a_max	0,0675	0,0675	0,0675	0,0675	0,0675	0,0675	
a_min	0,0548	0,0548	0,0548	0,0548	0,0548	0,0548	
A_vmax	0,0707	0,0695	0,0703	0,0695		0,0695	0,0695
A_vmin	0,0964	0,0995	0,0970	0,0995		0,0995	0,0995
Age	0,6334	0,6413	0,6035	0,6413	0,6413	0,6413	0,6413
b_avg	0,0698		0,0702	0,0698	0,0698	0,0698	0,0698
b_max	0,0302		0,0302	0,0302	0,0302	0,0302	0,0302
b_min	0,0618		0,0618	0,0618	0,0618	0,0618	0,0618
c_avg	0,0873	0,0870	0,0873		0,0870	0,0870	0,0870
c_max	0,0737	0,0737	0,0737		0,0737	0,0737	0,0737
c_min	0,0157	0,0157	0,0157		0,0157	0,0157	0,0157
CO	0,0584	0,0584	0,0584	0,0584	0,0584	0,0584	0,0584
EF	0,0623	0,0623	0,0623	0,0623	0,0623	0,0623	0,0623
Gender	0,1375	0,1106	0,1638	0,1106	0,1106	0,1106	0,1106

Continued on next page

Table B.1 – *Continued from previous page*

HR2	0,5847	0,6077	0,5857	0,6077	0,6077	0,6077	0,6077
HR3	0,5285	0,5546	0,5684	0,5546	0,5546	0,5546	0,5546
HR4	0,5274	0,5816	0,5727	0,5816	0,5816	0,5816	0,5816
m_avg	0,0453	0,0453	0,0453	0,0453	0,0453		0,0453
m_max	0,0146	0,0146	0,0146	0,0146	0,0146		0,0146
m_min	0,0805	0,0805	0,0805	0,0805	0,0805		0,0805
SD_A	0,0927	0,0927	0,0927	0,0927		0,0927	0,0927
SD_a	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101	
SD_b	0,0243		0,0243	0,0243	0,0243	0,0243	0,0243
SD_c	0,0646	0,0646	0,0646		0,0646	0,0646	0,0646
SD_m	0,0048	0,0048	0,0048	0,0048	0,0048		0,0048
SD_Vol	0,0494	0,0494	0,0494	0,0494	0,0494	0,0494	0,0494
SV	0,0623	0,0623	0,0623	0,0623	0,0623	0,0623	0,0623
t_max	0,0875	0,0960	0,0884	0,0960	0,0960	0,0960	0,0960
t_min	0,1738	0,1713	0,1633	0,1713	0,1713	0,1713	0,1713
Test_yr	0,1142	0,1595	0,1594	0,1595	0,1595	0,1595	0,1595
VAR_A	0,0932	0,0932	0,0932	0,0932		0,0932	0,0932
VAR_a	0,0109	0,0109	0,0109	0,0109	0,0109	0,0109	
VAR_b	0,0310		0,0421	0,0361	0,0286	0,0286	0,0361
VAR_c	0,0621	0,0621	0,0621		0,0621	0,0621	0,0621
VAR_m	0,0140	0,0057	0,0155	0,0057	0,0094		0,0057
VAR_Vol	0,0483	0,0483	0,0483	0,0483	0,0483	0,0483	0,0483
Vavg	0,0460	0,0460	0,0460	0,0460	0,0460	0,0460	0,0460
Vmax	0,0520	0,0520	0,0520	0,0520	0,0520	0,0520	0,0520
Vmin	0,0227	0,0227	0,0227	0,0227	0,0227	0,0227	0,0227
Xc_avg	0,0483	0,0482		0,0482	0,0482	0,0482	0,0482
Xc_max	0,0148	0,0148		0,0148	0,0148	0,0148	0,0148
Xc_min	0,0522	0,0522		0,0522	0,0522	0,0522	0,0522
Yc_avg	0,0219	0,0219		0,0219	0,0219	0,0219	0,0219
Yc_max	0,0175	0,0175		0,0175	0,0175	0,0175	0,0175
Yc_min	0,0400	0,0400		0,0400	0,0400	0,0400	0,0400

Table B.2: Mutual Information Scores of all cumulative feature deletions for Medical Tests

Features	**EM0**	**M1**	**M12**	**M123**	**M1234**	**M12345**	**M123456**
1cycl_dur	1,3407	1,3446	1,3270	1,3648	1,3486	1,3660	1,3316
A_avg	0,0380	0,0380	0,0380	0,0380			
a_avg	0,0359	0,0359	0,0359	0,0359	0,0359	0,0359	
a_max	0,0675	0,0675	0,0675	0,0675	0,0675	0,0675	

Continued on next page

Table B.2 – *Continued from previous page*

a_min	0,0548	0,0548	0,0548	0,0548	0,0548	0,0548	
A_vmax	0,0707	0,0695	0,0707	0,0706			
A_vmin	0,0964	0,0995	0,0984	0,0982			
Age	0,6334	0,6413	0,5976	0,5823	0,5713	0,5852	0,6210
b_avg	0,0698						
b_max	0,0302						
b_min	0,0618						
c_avg	0,0873	0,0870	0,0870				
c_max	0,0737	0,0737	0,0737				
c_min	0,0157	0,0157	0,0157				
CO	0,0584	0,0584	0,0584	0,0584	0,0584	0,0584	0,0584
EF	0,0623	0,0623	0,0623	0,0623	0,0623	0,0623	0,0623
Gender	0,1375	0,1106	0,1297	0,1201	0,1199	0,1038	0,0869
HR2	0,5847	0,6077	0,6025	0,5882	0,5722	0,5799	0,5962
HR3	0,5285	0,5546	0,5063	0,5450	0,5466	0,5630	0,5129
HR4	0,5274	0,5816	0,5854	0,5318	0,5611	0,5960	0,6279
m_avg	0,0453	0,0453	0,0453	0,0453	0,0453		
m_max	0,0146	0,0146	0,0146	0,0146	0,0146		
m_min	0,0805	0,0805	0,0805	0,0805	0,0805		
SD_A	0,0927	0,0927	0,0927	0,0927			
SD_a	0,0101	0,0101	0,0101	0,0101	0,0101	0,0101	
SD_b	0,0243						
SD_c	0,0646	0,0646	0,0646				
SD_m	0,0048	0,0048	0,0048	0,0048	0,0048		
SD_Vol	0,0494	0,0494	0,0494	0,0494	0,0494	0,0494	0,0494
SV	0,0623	0,0623	0,0623	0,0623	0,0623	0,0623	0,0623
t_max	0,0875	0,0960	0,0828	0,0851	0,0958	0,0835	0,0908
t_min	0,1738	0,1713	0,1654	0,1669	0,1651	0,1610	0,1693
Test_yr	0,1142	0,1595	0,0833	0,0937	0,1083	0,1243	0,1963
VAR_A	0,0932	0,0932	0,0932	0,0932			
VAR_a	0,0109	0,0109	0,0109	0,0109	0,0109	0,0109	
VAR_b	0,0310						
VAR_c	0,0621	0,0621	0,0621				
VAR_m	0,0140	0,0057	0,0060	0,0082	0,0094		
VAR_Vol	0,0483	0,0483	0,0483	0,0483	0,0483	0,0483	0,0483
Vavg	0,0460	0,0460	0,0460	0,0460	0,0460	0,0460	0,0460
Vmax	0,0520	0,0520	0,0520	0,0520	0,0520	0,0520	0,0520
Vmin	0,0227	0,0227	0,0227	0,0227	0,0227	0,0227	0,0227
Xc_avg	0,0483	0,0482					
Xc_max	0,0148	0,0148					
Xc_min	0,0522	0,0522					
Yc_avg	0,0219	0,0219					

Continued on next page

Table B.2 – *Continued from previous page*

Yc_max	0,0175	0,0175					
Yc_min	0,0400	0,0400					

B.2 Random Forest Results

The following tables are to be read in conjunction with one another, as the model entries in Table B.3 all coincide with their performance metrics in Table B.4. Coloured blocks in Table B.4 highlight the highest scores for the respective Medical test.

Table B.3: Best performing Random Forest models based on performance metrics for each pathology, listed by architectural descriptors

Test	**N**	**HF**	**LVH**	**MI**	**AR**	**AS**
EM0	E100:D16				E100:D16	E100:D16
	E150:D16		E150:D16	E150:D16		
		E200:D16	E200:D16			
			E500:D16			
M1	E100:D16			E100:D16		E100:D16
	E150:D16		E150:D16	E150:D16	E150:D16	E150:D16
	E200:D16			E200:D16	E200:D16	
	E250:D16					
	E300:D16					
		E400:D16	E400:D16			
			E450:D16		E450:D16	
			E500:D16	E500:D16	E500:D16	
M2	NONE				E100:D16	E100:D16
						E150:D16
			E200:D16	E200:D16		
		E300:D16		E300:D16		
		E350:D16		E350:D16		E350:D16
		E400:D16	E400:D16	E400:D15		
			E450:D16	E450:D16		
M3	E100:D16		E100:D16		E100:D16	
			E150:D16			E150:D16
			E200:D16	E200:D16	E200:D16	E200:D16
		E250:D16	E250:D16	E250:D16	E250:D16	
		E300:D16	E300:D16	E300:D16	E300:D16	E300:D16
			E350:D16	E350:D16	E350:D16	E350:D16

Continued on next page

Table B.3 – *Continued from previous page*

			E400:D16 E450:D16 E500:D16	 E450:D16 E500:D16		 E500:D16
M4	E150:D16 E200:D16 E250:D16 E300:D16 E350:D16 E400:D16 E450:D16	E150:D16	 E200:D16 E300:D16 E350:D16 E450:D16 E500:D16	 E450:D16	 E450:D16	 E200:D16 E350:D16 E400:D16 E450:D16 E500:D16
M5	 E400:D16	E100:D16	 E200:D16 E250:D16 	E100:D16 E300:D16	E100:D16	E100:D16
M6	 E350:D16	 E500:D16	 E350:D16 E400:D16	NONE	E100:D16	 E150:D16
M12	E100:D16	 E500:D16	 E300:D16 E350:D16 E500:D16	 E150:D16	E100:D16	 E150:D16 E250:D16
M123	E100:D16	 E200:D16 E250:D16	 E300:D16 E350:D16	 E400:D16 E450:D16 E500:D16	 E300:D16	 E250:D16
M1234	E100:D16	 E150:D16 E250:D16 E300:D16 E350:D16	 E150:D16 E250:D16 E300:D16 E350:D16 E400:D16	E100:D16	 E350:D16 E400:D16	 E150:D16
M12345	E100:D16			E100:D16	E100:D16	

Continued on next page

Table B.3 – *Continued from previous page*

	E150:D16				E150:D16	
	E200:D16					E200:D16
	E300:D16					E300:D16
	E350:D16	E350:D16				
	E400:D16		E400:D16			
	E450:D16					
	E500:D16					
M123456	E150:D16					E150:D16
	E250:D16	E250:D16		E250:D16		
		E300:D16	E300:D16			
		E350:D16				
	E400:D16					E400:D16
	E450:D16	E450:D16			E450:D16	
	E500:D16				E500:D16	

Table B.4: Averaged Performance Metrics for Best Performing Random Forest Models per Medical Test

Test	Model	Accuracy	Recall	Precision	F1-score
EM0	E100:D16	0,8333	0,2171	0,9309	0,3357
	E150:D16	0,8368	0,2264	0,9470	0,3523
	E200:D16	0,8354	0,2201	0,9493	0,3404
	E500:D16	0,8326	0,2063	0,9372	0,3243
M1	E100:D16	0,8397	0,2449	0,9374	0,3761
	E150:D16	0,8418	0,2486	0,9545	0,3834
	E200:D16	0,8397	0,2423	0,9493	0,3737
	E250:D16	0,8368	0,2334	0,9365	0,3605
	E300:D16	0,8376	0,2341	0,9514	0,3610
	E400:D16	0,8390	0,2363	0,9565	0,3644
	E450:D16	0,8383	0,2329	0,9545	0,3611
	E500:D16	0,8383	0,2331	0,9545	0,3615
M2	E100:D16	0,8376	0,2349	0,9041	0,3590
	E150:D16	0,8354	0,2242	0,8517	0,3422
	E200:D16	0,8361	0,2281	0,8613	0,3471
	E300:D16	0,8368	0,2260	0,9464	0,3433
	E350:D16	0,8383	0,2325	0,9464	0,3543
	E400:D15	0,8390	0,2394	0,9464	0,3646
	E450:D16	0,8383	0,2360	0,9454	0,3611
M3	E100:D16	0,8347	0,2178	0,9474	0,3459
	E150:D16	0,8333	0,2132	0,9444	0,3372

Continued on next page

Table B.4 – *Continued from previous page*

	E200:D16	0,8361	0,2180	0,9561	0,3457
	E250:D16	0,8354	0,2195	0,9306	0,3449
	E300:D16	0,8361	0,2172	0,9583	0,3423
	E350:D16	0,8347	0,2123	0,9524	0,3329
	E400:D16	0,8340	0,2166	0,9396	0,3394
	E450:D16	0,8333	0,2123	0,9341	0,3305
	E500:D16	0,8340	0,2123	0,9396	0,3318
M4	E150:D16	0,8291	0,1963	0,9176	0,3114
	E200:D16	0,8312	0,2015	0,9306	0,3225
	E250:D16	0,8298	0,1918	0,9358	0,3102
	E300:D16	0,8305	0,1954	0,9371	0,3150
	E350:D16	0,8312	0,1993	0,9398	0,3192
	E400:D16	0,8312	0,1991	0,9386	0,3191
	E450:D16	0,8340	0,2119	0,9398	0,3369
	E500:D16	0,8319	0,2007	0,9398	0,3206
M5	E100:D16	0,8368	0,2240	0,9434	0,3455
	E200:D16	0,8319	0,2026	0,9444	0,3193
	E250:D16	0,8298	0,1943	0,9306	0,3053
	E300:D16	0,8319	0,2019	0,9444	0,3185
	E400:D16	0,8326	0,2047	0,9524	0,3234
M6	E100:D16	0,8319	0,2066	0,9333	0,3269
	E150:D16	0,8333	0,2126	0,9288	0,3364
	E350:D16	0,8354	0,2278	0,9394	0,3561
	E400:D16	0,8326	0,2124	0,9394	0,3338
	E500:D16	0,8319	0,2103	0,9420	0,3288
M12	E100:D16	0,8425	0,2700	0,8984	0,4019
	E150:D16	0,8439	0,2694	0,8902	0,4021
	E250:D16	0,8363	0,2912	0,8967	0,4282
	E300:D16	0,8418	0,2635	0,8826	0,3926
	E350:D16	0,8411	0,2593	0,8812	0,3879
	E500:D16	0,8425	0,2630	0,8881	0,3940
M123	E100:D16	0,8432	0,2783	0,8914	0,4122
	E200:D16	0,8474	0,2774	0,9472	0,4160
	E250:D16	0,8509	0,2948	0,9479	0,4385
	E300:D16	0,8516	0,2992	0,9533	0,4434
	E350:D16	0,8502	0,2958	0,9429	0,4383
	E400:D16	0,8502	0,2860	0,9565	0,4298
	E450:D16	0,8488	0,2826	0,9327	0,4245
	E500:D16	0,8502	0,2894	0,9327	0,4325
M1234	E100:D16	0,8544	0,3130	0,9241	0,4624
	E150:D16	0,8551	0,3079	0,9330	0,4591
	E250:D16	0,8516	0,2928	0,9417	0,4409

Continued on next page

Table B.4 – *Continued from previous page*

	E300:D16	0,8530	0,2986	0,9417	0,4482
	E350:D16	0,8523	0,2944	0,9388	0,4415
	E400:D16	0,8537	0,2982	0,9373	0,4475
M12345	E100:D16	0,8706	0,3790	0,9106	0,5294
	E150:D16	0,8720	0,3842	0,9122	0,5366
	E200:D16	0,8692	0,3714	0,9127	0,5236
	E300:D16	0,8720	0,3821	0,9217	0,5356
	E350:D16	0,8727	0,3846	0,9293	0,5397
	E400:D16	0,8713	0,3855	0,9167	0,5374
	E450:D16	0,8713	0,3846	0,9167	0,5363
	E500:D16	0,8727	0,3880	0,9227	0,5412
M123456	E150:D16	0,8861	0,4570	0,9366	0,6122
	E250:D16	0,8854	0,4504	0,9403	0,6070
	E300:D16	0,8868	0,4580	0,9407	0,6132
	E350:D16	0,8840	0,4481	0,9337	0,6032
	E400:D16	0,8840	0,4503	0,9308	0,6042
	E450:D16	0,8854	0,4547	0,9375	0,6101
	E500:D16	0,8847	0,4562	0,9300	0,6094

B.3 Neural Networks Results

Each Neural Network model performance is summarised for all Medical tests. Final scores of the training and validation subsets (differentiated by $v\_$) are presented in Table B.5, with blue coloured blocks highlighting the best performing model per test.

Table B.5: Averaged Performance Metrics for Best Performing Neural Network Models per Medical Test

Test	Model	loss	accuracy	recall	precision	v_loss	v_accuracy	v_recall	v_precision
EM0	D1W16	0,4323	0,4034	0,1517	0,5768	0,4585	0,3840	0,0345	0,4545
	D1W64	0,3586	0,5333	0,3268	0,7109	0,4358	0,4135	0,1172	0,5000
	D1W256	0,2591	0,6663	0,5835	0,8344	0,4361	0,4093	0,2621	0,5891
	D2W16	0,2591	0,6663	0,5835	0,8344	0,4361	0,4093	0,2621	0,5891
	D2W64	0,3660	0,5048	0,3411	0,6972	0,4330	0,4219	0,1345	0,5909
	D2W256	0,1892	0,7635	0,7343	0,8758	0,4166	0,4388	0,3207	0,6039
	D3W16	0,4577	0,3590	0,0682	0,5468	0,4632	0,3038	0,0069	0,2222
	D3W64	0,3517	0,5533	0,3662	0,6974	0,4173	0,4304	0,2483	0,5669
	D3W256	0,1240	0,8163	0,8447	0,8945	0,3922	0,5401	0,5000	0,6808
M1	D1W16	0,4269	0,3886	0,1706	0,5919	0,4503	0,3629	0,0621	0,4865
	D1W64	0,3712	0,5227	0,2953	0,7015	0,4456	0,3882	0,1034	0,4000
	D1W256	0,2588	0,7001	0,5969	0,8200	0,4424	0,4304	0,2069	0,5660
	D2W16	0,4512	0,3263	0,1149	0,5541	0,4604	0,3249	0,0621	0,4500
	D2W64	0,3418	0,5354	0,4039	0,7132	0,4427	0,4051	0,2586	0,5000
	D2W256	0,1685	0,7941	0,7693	0,9002	0,4007	0,5232	0,4172	0,6173
	D3W16	0,4586	0,3242	0,0628	0,6195	0,4605	0,3122	0,0069	1,0000
	D3W64	0,3741	0,4921	0,3169	0,6610	0,4408	0,3797	0,2034	0,5315

Continued on next page

Table B.5 – *Continued from previous page*

	D3W256	0,1423	0,8036	0,8061	0,8918	0,4100	0,4515	0,4310	0,6098
M2	D1W16	0,4338	0,3970	0,1149	0,5333	0,4437	0,3544	0,0345	0,4348
	D1W64	0,3755	0,5037	0,2765	0,6769	0,4326	0,3755	0,1069	0,5741
	D1W256	0,2932	0,6230	0,4955	0,7677	0,4344	0,4219	0,2552	0,5441
	D2W16	0,4566	0,3168	0,0817	0,5417	0,4538	0,3671	0,0379	0,6471
	D2W64	0,3558	0,5037	0,3519	0,6701	0,4296	0,4219	0,2345	0,5075
	D2W256	0,1758	0,7835	0,7666	0,8841	0,3913	0,4979	0,4552	0,6286
	D3W16	0,4478	0,3706	0,0646	0,6316	0,4472	0,3882	0,0345	0,6250
	D3W64	0,3866	0,4530	0,2801	0,6316	0,4233	0,3671	0,1897	0,5789
	D3W256	0,1447	0,7994	0,8187	0,8872	0,3956	0,5232	0,4966	0,6316
M3	D1W16	0,4272	0,4382	0,1544	0,6014	0,4486	0,3966	0,0379	0,5238
	D1W64	0,3834	0,4699	0,2729	0,6831	0,4447	0,3629	0,1172	0,4722
	D1W256	0,2789	0,6452	0,5368	0,7973	0,4442	0,4135	0,2483	0,5180
	D2W16	0,4476	0,3559	0,0781	0,5472	0,4600	0,3713	0,0103	0,4286
	D2W64	0,3682	0,4963	0,3348	0,6782	0,4252	0,4262	0,2138	0,5536
	D2W256	0,1681	0,7761	0,7630	0,8882	0,4006	0,5021	0,4414	0,6531
	D3W16	0,4710	0,2841	0,0305	0,3400	0,4736	0,2700	0,0034	0,3333
	D3W64	0,3734	0,4963	0,3250	0,6830	0,4179	0,4430	0,1828	0,5521
	D3W256	0,1393	0,8057	0,8214	0,8858	0,3799	0,5485	0,5103	0,6820
M4	D1W16	0,4499	0,3580	0,1239	0,5349	0,4591	0,3376	0,0207	0,6667
	D1W64	0,3863	0,4794	0,2567	0,6560	0,4417	0,4262	0,0759	0,6286
	D1W256	0,2978	0,6135	0,5027	0,7898	0,4351	0,4008	0,2069	0,5882
	D2W16	0,4590	0,3453	0,0700	0,5417	0,4607	0,3797	0,0138	0,8000
	D2W64	0,3725	0,5143	0,3079	0,6874	0,4281	0,4388	0,2034	0,6941
	D2W256	0,1783	0,7761	0,7531	0,8767	0,3945	0,4895	0,4276	0,6294

Continued on next page

Table B.5 – *Continued from previous page*

	D3W16	0,4737	0,3105	0,0512	0,4286	0,4639	0,3713	0,0172	0,5556
	D3W64	0,3871	0,4773	0,2890	0,6252	0,4262	0,4219	0,2138	0,5391
	D3W256	0,1672	0,7677	0,7621	0,8654	0,4144	0,4810	0,4103	0,6263
M5	D1W16	0,4345	0,3939	0,1248	0,5697	0,4486	0,3291	0,0172	0,4545
	D1W64	0,3574	0,5396	0,3303	0,6904	0,4373	0,4135	0,1483	0,5811
	D1W256	0,2719	0,6621	0,5431	0,8045	0,4439	0,3671	0,2897	0,5563
	D2W16	0,4462	0,3749	0,0978	0,5317	0,4477	0,3586	0,0345	0,6667
	D2W64	0,3683	0,4974	0,3294	0,6746	0,4310	0,4262	0,1966	0,5758
	D2W256	0,1484	0,7899	0,7935	0,8893	0,3738	0,5021	0,4966	0,6729
	D3W16	0,4528	0,3495	0,0566	0,4846	0,4462	0,3376	0,0138	0,8000
	D3W64	0,3796	0,4889	0,3088	0,6641	0,4269	0,4051	0,2034	0,5728
	D3W256	0,1266	0,8258	0,8384	0,9042	0,3733	0,5401	0,5345	0,6828
M6	D1W16	0,4252	0,3939	0,1508	0,5695	0,4533	0,4008	0,0483	0,4242
	D1W64	0,3715	0,5037	0,3007	0,6781	0,4461	0,4008	0,1172	0,4595
	D1W256	0,2700	0,6684	0,5512	0,8079	0,4277	0,4515	0,2276	0,5280
	D2W16	0,4613	0,3390	0,0925	0,5124	0,4679	0,3460	0,0276	0,5333
	D2W64	0,3588	0,5290	0,3384	0,7167	0,4320	0,3966	0,1862	0,5934
	D2W256	0,1396	0,8163	0,8115	0,9067	0,4063	0,4937	0,4000	0,6554
	D3W16	0,4588	0,3622	0,0871	0,5673	0,4632	0,3376	0,0310	0,5000
	D3W64	0,3897	0,4731	0,2837	0,6767	0,4340	0,4346	0,1517	0,5238
	D3W256	0,1104	0,8458	0,8573	0,9200	0,3926	0,5316	0,4966	0,6344
M12	D1W16	0,4404	0,3738	0,1373	0,5484	0,4575	0,3797	0,0655	0,5429
	D1W64	0,3823	0,4784	0,2675	0,6578	0,4377	0,4008	0,0586	0,4857
	D1W256	0,2958	0,6357	0,4937	0,7757	0,4337	0,4177	0,2000	0,6304
	D2W16	0,4399	0,3738	0,1194	0,6552	0,4494	0,3924	0,0483	0,5185

Continued on next page

Table B.5 – *Continued from previous page*

	D2W64	0,3750	0,4847	0,3250	0,6654	0,4321	0,4135	0,1897	0,5556
	D2W256	0,1982	0,7476	0,7020	0,8593	0,4178	0,4641	0,3690	0,5879
	D3W16	0,4675	0,2946	0,0431	0,5275	0,4669	0,3333	0,0138	1,0000
	D3W64	0,3840	0,4773	0,2926	0,6481	0,4346	0,4008	0,1862	0,5510
	D3W256	0,1398	0,8046	0,8187	0,8854	0,3910	0,5274	0,5172	0,6173
M123	D1W16	0,4464	0,3548	0,0943	0,5097	0,4616	0,3629	0,0138	0,4444
	D1W64	0,3861	0,4900	0,2415	0,6642	0,4379	0,4051	0,1069	0,5636
	D1W256	0,3122	0,5818	0,4417	0,7822	0,4459	0,4051	0,1448	0,5316
	D2W16	0,4499	0,3400	0,0880	0,5665	0,4519	0,3418	0,0379	0,5789
	D2W64	0,3661	0,5185	0,3276	0,6887	0,4267	0,4557	0,2310	0,5234
	D2W256	0,2030	0,7497	0,7065	0,8453	0,3882	0,5021	0,4069	0,6243
	D3W16	0,4759	0,3073	0,0655	0,4867	0,4649	0,3460	0,0345	0,5556
	D3W64	0,3746	0,4921	0,3133	0,6994	0,4397	0,3966	0,2690	0,5306
	D3W256	0,1686	0,7508	0,7738	0,8663	0,4021	0,5190	0,5138	0,6183
M1234	D1W16	0,4547	0,3379	0,1041	0,5577	0,4546	0,3333	0,0276	0,7273
	D1W64	0,4038	0,4361	0,1867	0,6322	0,4426	0,3840	0,0724	0,5250
	D1W256	0,3268	0,5935	0,4174	0,7573	0,4384	0,3713	0,1931	0,5385
	D2W16	0,4567	0,3168	0,0557	0,5124	0,4521	0,3713	0,0310	0,8182
	D2W64	0,3915	0,4520	0,2496	0,6541	0,4277	0,4093	0,1414	0,5942
	D2W256	0,2329	0,6895	0,6481	0,8177	0,4115	0,4599	0,3897	0,6075
	D3W16	0,4684	0,2893	0,0557	0,5636	0,4600	0,3671	0,0034	0,2500
	D3W64	0,4071	0,4319	0,2226	0,6200	0,4268	0,4093	0,1034	0,4918
	D3W256	0,1860	0,7339	0,7280	0,8600	0,4059	0,5527	0,4759	0,6273
M12345	D1W16	0,4578	0,3485	0,0961	0,4886	0,4552	0,3966	0,0207	0,8571
	D1W64	0,4085	0,4351	0,1795	0,6173	0,4399	0,3924	0,0862	0,6250

Continued on next page

Table B.5 – *Continued from previous page*

	D1W256	0,3456	0,5544	0,3627	0,7608	0,4297	0,4008	0,1793	0,6047
	D2W16	0,4655	0,3242	0,0512	0,4254	0,4537	0,3586	0,0034	0,5000
	D2W64	0,3913	0,4382	0,2415	0,6692	0,4291	0,4346	0,1310	0,5938
	D2W256	0,2145	0,7043	0,6840	0,8467	0,3784	0,5274	0,4552	0,6600
	D3W16	0,4678	0,2957	0,0404	0,6000	0,4683	0,3122	0,0172	0,6250
	D3W64	0,3933	0,4625	0,2433	0,6362	0,4182	0,4304	0,2138	0,5905
	D3W256	0,2208	0,7064	0,6750	0,8291	0,3850	0,4979	0,4138	0,6383
M123456	D1W16	0,4560	0,3432	0,0835	0,5196	0,4474	0,3671	0,0207	0,6667
	D1W64	0,4142	0,4108	0,1768	0,6234	0,4369	0,3840	0,1069	0,6458
	D1W256	0,3540	0,5280	0,3474	0,7274	0,4313	0,4388	0,1793	0,6118
	D2W16	0,4571	0,3157	0,0727	0,5159	0,4546	0,3629	0,0276	0,8889
	D2W64	0,3822	0,4562	0,2648	0,6674	0,4140	0,4515	0,2000	0,6517
	D2W256	0,2291	0,7012	0,6499	0,8399	0,3734	0,5316	0,4138	0,6977
	D3W16	0,4696	0,3052	0,0548	0,4919	0,4509	0,4008	0,0448	0,7647
	D3W64	0,4047	0,4414	0,1966	0,6460	0,4230	0,4304	0,1483	0,5890
	D3W256	0,1657	0,7709	0,7882	0,8702	0,3781	0,5696	0,5379	0,6783

C Appendix 3: Data Tests Results

C.1 Shortlisted Models: Previous Results

Figures C.1 and C.2 show model responses to feature deletions associated with all Engineering and Medical tests. Data test feature/s selections were loosely based on these results. To read these figures, the plots for "Individual" metrics directly correlate with the x-axis labelling. For example, in Figure C.1 the "Individual Accuracy" plotted for test E2 is the accuracy for test E2 alone (where Age data was removed). However, for "Cumulative Accuracy", E2 represents the accuracy for test E12 - where all features associated with test E2 and those prior (excluding EM0) were deleted. Similarly, "Cumulative Accuracy" at E3, represents the accuracy for test E123, and so forth.

C.2 Random Forest Results

Table C.1: Performance metrics output for each Random Forests instance for each train-test ratio tested

Train-Test Split	Training Duration	Average Accuracy	Average Precision	Average Recall	Average F1-score
E100D16					
0.1-0.9	0,7922	0,8044	0,5834	0,0752	0,1196
0.2-0.8	0,9361	0,8092	0,6978	0,1065	0,1750
0.3-0.7	1,1269	0,8164	0,6994	0,1426	0,2318
0.4-0.6	1,3821	0,8207	0,7711	0,1459	0,2387
0.5-0.5	1,6767	0,8198	0,7305	0,1604	0,2552
0.6-0.4	1,7958	0,8249	0,7724	0,1898	0,2956
0.7-0.3	2,0252	0,8305	0,8192	0,2005	0,3133
0.8-0.2	2,1964	0,8347	0,8883	0,2290	0,3478
0.9-0.1	2,4073	0,8445	0,9179	0,2716	0,3989
E300D16					
0.1-0.9	2,3262	0,8046	0,5929	0,0632	0,1059
0.2-0.8	2,7747	0,8092	0,6979	0,0974	0,1611

Continued on next page

Table C.1 – *Continued from previous page*

0.3-0.7	3,3708	0,8148	0,6725	0,1301	0,2130
0.4-0.6	4,0038	0,8214	0,8029	0,1413	0,2326
0.5-0.5	4,7547	0,8229	0,7785	0,1635	0,2621
0.6-0.4	5,2733	0,8263	0,8053	0,1873	0,2930
0.7-0.3	6,1849	0,8282	0,8033	0,1904	0,3032
0.8-0.2	6,6693	0,8397	0,8958	0,2400	0,3621
0.9-0.1	7,13889	0,8445	0,9528	0,2632	0,3925
E500D16					
0.1-0.9	3,8483	0,8046	0,5855	0,0618	0,1037
0.2-0.8	4,6434	0,8078	0,6781	0,0961	0,1580
0.3-0.7	6,1680	0,8156	0,7046	0,1289	0,2116
0.4-0.6	7,6762	0,8204	0,7963	0,1411	0,2316
0.5-0.5	7,8686	0,8229	0,7899	0,1589	0,2553
0.6-0.4	8,8288	0,8267	0,7784	0,1945	0,3023
0.7-0.3	9,8492	0,8291	0,8048	0,1931	0,3051
0.8-0.2	10,9182	0,8383	0,8981	0,2347	0,3557
0.9-0.1	11,9124	0,8431	0,9253	0,2608	0,3907

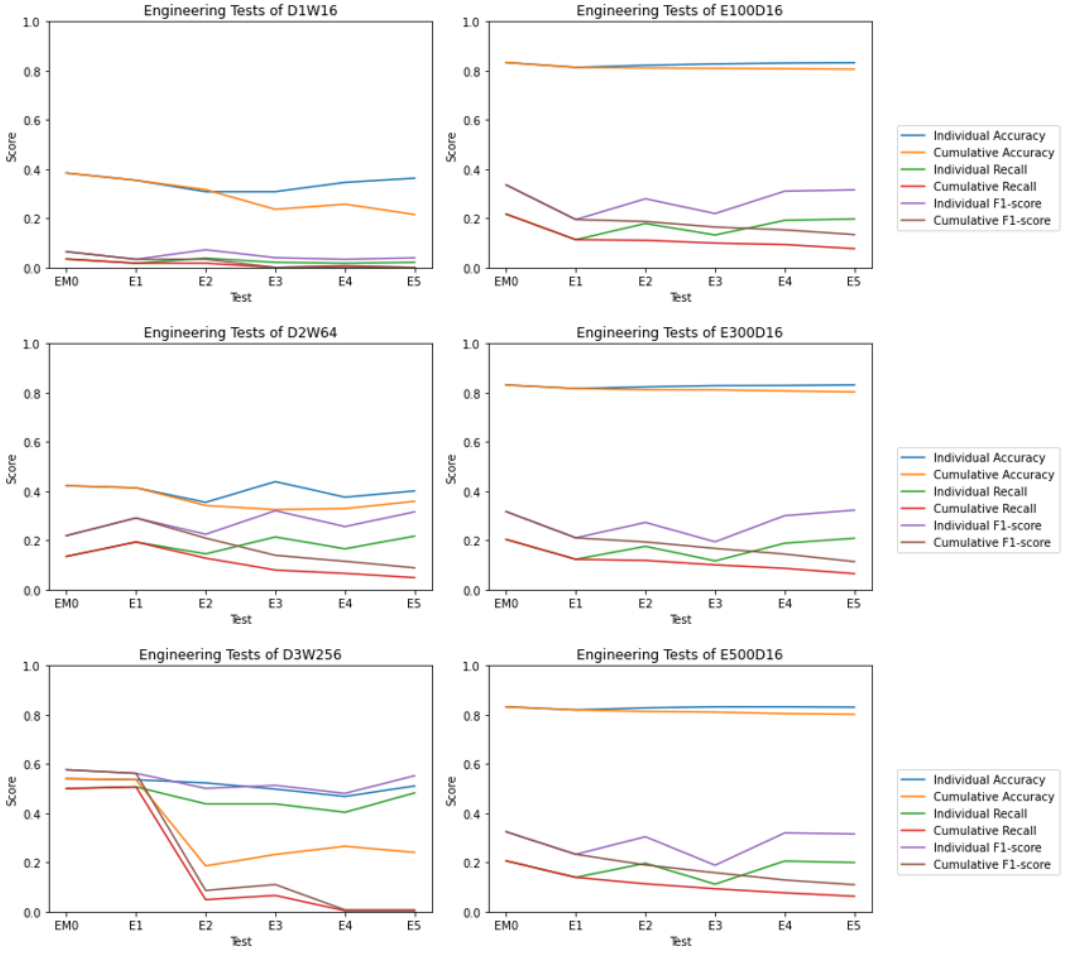

Figure C.1: Performance metric plots of shortlisted models for Engineering Tests

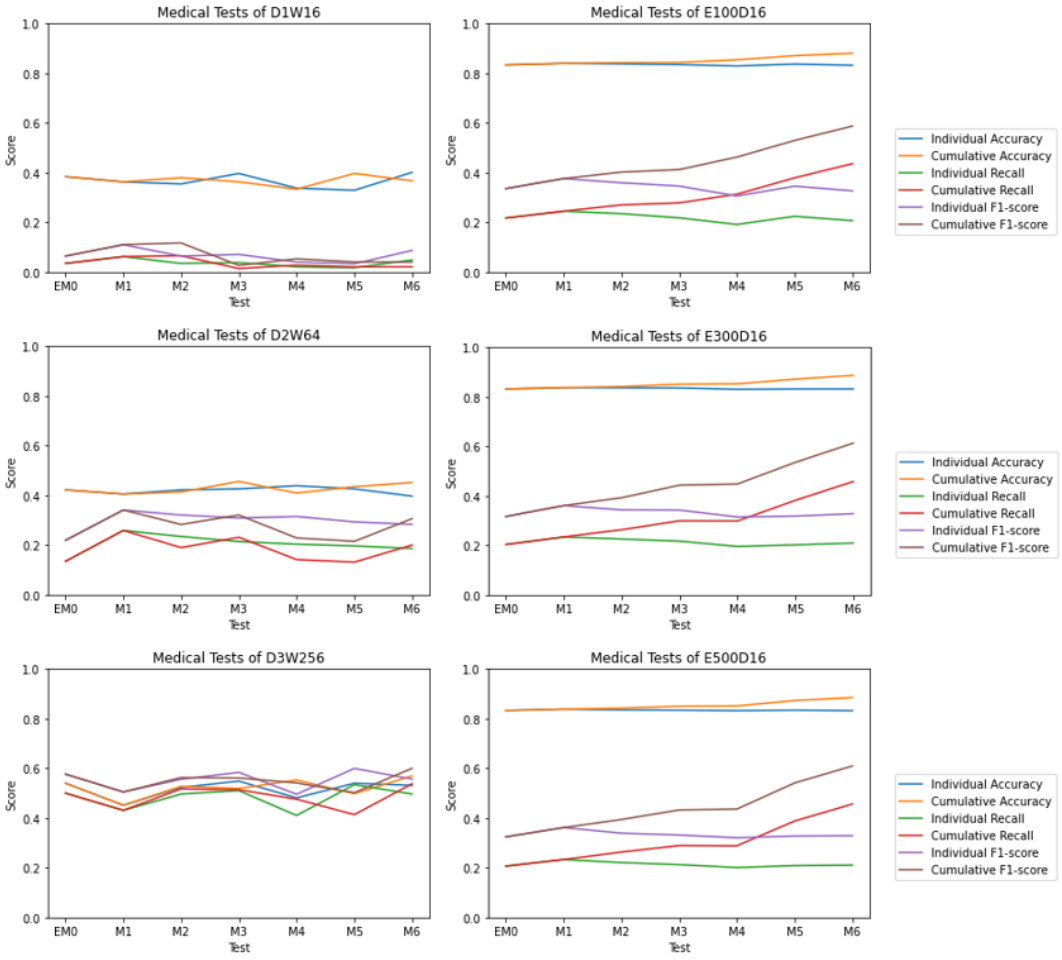

Figure C.2: Performance metric plots of shortlisted models for Medical Tests

C.3 Neural Networks Data tests Results

Table C.2: Train and Validation set performance metrics for Neural Network model instances in all Data tests

Test	Model	loss	accuracy	recall	precision	v_loss	v_accuracy	v_recall	v_precision
0,1-0,9	D1W16	0,4226	0,4576	0,1691	0,4600	0,5091	0,2298	0,0126	0,3404
	D2W64	0,2604	0,7034	0,5074	0,8625	0,4939	0,2871	0,0615	0,4216
	D3W256	0,0964	0,9068	0,8971	0,9760	0,5110	0,2054	0,0426	0,2857
0,2-0,8	D1W16	0,4488	0,3729	0,1176	0,4384	0,5030	0,2395	0,0124	0,6364
	D2W64	0,3419	0,5678	0,3493	0,7197	0,5003	0,2321	0,0760	0,4456
	D3W256	0,1825	0,7585	0,7353	0,8850	0,5178	0,1540	0,1352	0,3493
0,3-0,7	D1W16	0,4565	0,3831	0,1175	0,5213	0,4782	0,3004	0,0365	0,5538
	D2W64	0,3397	0,5324	0,3981	0,7186	0,4634	0,3486	0,1317	0,5179
	D3W256	0,1973	0,7296	0,7242	0,8555	0,4846	0,2099	0,0932	0,3882
0,4-0,6	D1W16	0,4590	0,3298	0,1227	0,5113	0,4786	0,2743	0,0518	0,5176
	D2W64	0,3678	0,5201	0,3303	0,6803	0,4502	0,3572	0,0871	0,5441
	D3W256	0,2624	0,6871	0,5758	0,7975	0,4980	0,2363	0,0376	0,4103
0,5-0,5	D1W16	0,4385	0,3699	0,1019	0,5035	0,4616	0,3226	0,0339	0,6486
	D2W64	0,3769	0,4848	0,2855	0,6723	0,4460	0,4003	0,1443	0,4880
	D3W256	0,1852	0,7720	0,7360	0,8754	0,4487	0,3936	0,3239	0,5401
0,6-0,4	D1W16	0,4573	0,3408	0,1134	0,4872	0,4689	0,2911	0,0477	0,6429
	D2W64	0,3500	0,5408	0,3711	0,7182	0,4382	0,3882	0,2138	0,5378
	D3W256	0,1821	0,7479	0,7470	0,8448	0,4492	0,4008	0,2686	0,5278
0,7-0,3	D1W16	0,4459	0,3466	0,1086	0,5889	0,4637	0,3455	0,0421	0,6000
	D2W64	0,3788	0,4771	0,2941	0,6769	0,4470	0,3624	0,1449	0,5082

Continued on next page

Table C.2 – *Continued from previous page*

	D3W256	0,2604	0,6703	0,6096	0,7933	0,4635	0,3343	0,2266	0,4619
0,8-0,2	D1W16	0,4498	0,3654	0,0844	0,5081	0,4661	0,3333	0,0483	0,5385
	D2W64	0,3746	0,5069	0,3187	0,6974	0,4449	0,3586	0,1517	0,4783
	D3W256	0,2673	0,6695	0,5646	0,7972	0,4827	0,3122	0,1655	0,4948
0,9-0,1	D1W16	0,4422	0,3709	0,0979	0,5371	0,4635	0,3277	0,0340	0,7143
	D2W64	0,4057	0,4385	0,2339	0,5731	0,4458	0,3529	0,1293	0,4419
	D3W256	0,2479	0,6723	0,6325	0,7958	0,4134	0,3613	0,4218	0,5741

D Appendix 4: Additional Information

/

D.1 Research Protocol Content

This section presents some of the details of the project, for data acquisition from the sourcing hospital, as required by their internal Ethics Committee before permissions were granted. Ethical clearance was granted by both the hospital (ref: UNIV-2020-0007) and Stellenbosch University (ref: S19/02/032) prior to data collection.

D.1.1 Randomisation, Confidentiality & Bias

Randomisation is a method of arbitrary allocation of participants when assigning (to groups) or choosing them for a study. Randomisation is advantageous in maximising the statistical power of sub-grouped analyses while minimising bias. Thus, the distribution of characteristics explored is equally present in all sub-groups. For this project, *Selective sampling* was used to identify patients for the study. This method is a type of non-probability randomised sampling based on the judgement of the primary investigator (Garg, 2016). Given the boundaries of the exclusion-inclusion criterion and the time restriction of the study, it was most suitable for the application.

Confidentiality in clinical research prioritises the protection of human participant identities and personal information. For this project, all data was extracted without any indicators of a patient's private information (anonymised). Data was exported based on internal diagnostic codes (used by the sourcing unit) associated with the abnormalities of interest. Together with randomisation, these measures *masked* subjects to the researcher, minimizing room for potential bias.

Bias is defined as the systematic deviation of a study's outcomes leading to skewed results and is often the product of a defective study design. In this particular study, the potential biases and respective corrective measures were identified and discussed below (Garg, 2016):

- **Investigator Bias**: conscious / subconscious favouring of a particular group over another. This bias is removed with the method of selection and sample scarcity. All data that satisfied the exclusion-inclusion criterion was exported without investigator bias due to lack of domain knowledge of the investigator. Furthermore, rare abnormalities were all included for the study.
- **Selection Bias**: occurs during sampling, where due to some hindrance in admission of patients for a study or their refusal to participate, the sampled data will be unrepresentative of the larger population. This bias is addressed as only existing patient cases will be exported having already undergone an echocardiograph exam.
- **Ascertainment/Information Bias**: measurement error that results from some misclassification of a patient (during diagnostic procedures). This is addressed by the structure within the Cardiac Unit. Analyses are done by experienced cardiologists, while trained staff perform echocardiograph exams. Diagnoses are further based on visually observed characteristics in conjunction with secondary measurements obtainable during echocardiograms.

D.1.2 Data Collection & Management

The amount of data acquired for the context of this project must be sufficiently large to be representative of the larger population. For the allocated time period, a sample size of 1200 (200 per pathology) was to be collected. However, given the scarcity of some of the included abnormalities, at best 50 could be extracted over the given time - hence the need for data augmentations. As per ethical requirements, data was to be viewed only the involved parties of the study at specified locations; viz. various places of work. Video data was exported (as .avi files), pre-processed as discussed in Section 3.3, making it suitable as model input (as training/testing data). The training subset will used to aid model learning, while the test set will be used to assess model performance. All programming takes place in Python (using standard packages, OpenCV, and Tensorflow) primarily, with image processing steps occurring in FIJI. All results will be represented by means of model performance metrics or graphs, accompanied by thorough discussions.

D.1.3 Project Commencement Plan

A commencement plan is usually requested to inform Ethical Committees how sensitive data will be managed once it is no longer used. These regulations are necessary to ensure appropriate measures are taken at all points to protect patient information. In this study, the data collection process concluded on 18 November 2020, with the project completion pending until submission from 15-18 November 2021. Upon completion of the project, the external hard drive used will be returned to the hospital for deletion of all data. As per requirements for research conducted in collaboration with the hospital, the final dissertation must be submitted to their internal research project manager. Additionally, any formal documents or publications must exclude the name of the sourcing hospital but can include names of staff that externally advised (as per request).

D.2 Continuous Ranked Probability Score

The Continuous Ranked Probability Score (CRPS) was the prescribed for evaluating model output in the 2015 Kaggle contest (Venugopal and Ramaswamy, 2015). It is typically used when probabilistic forecasting is being done by generalising the mean absolute error (MAE) across all probabilities. CRPS can be defined mathematically as per Equation D.2.1:

$$C = \frac{1}{600N} \sum_{m=1}^{n} \sum_{n=0}^{599} (P(y \leq n) - I[(n - V_m) \geq 0])^2 \tag{D.2.1}$$

$$C = \frac{1}{600N} \sum_{m=1}^{n} \sum_{n=0}^{599} |P(y \leq n) - I[(n - V_m) \geq 0]| \tag{D.2.2}$$

where P represents the predicted cumulative distribution and I, the Heaviside step function equal to 1 if true or 0 otherwise. In general, the smaller the score, the closer the predicted distribution is to the actual distribution (Gneiting and Raftery, 2007).

D.3 FIJI Image Processing Schematic

The flow diagram presented in Figure D.1 depicts the the individual steps and associated key parameters or settings used to generate the output used in successive steps as per Figure 3.1.

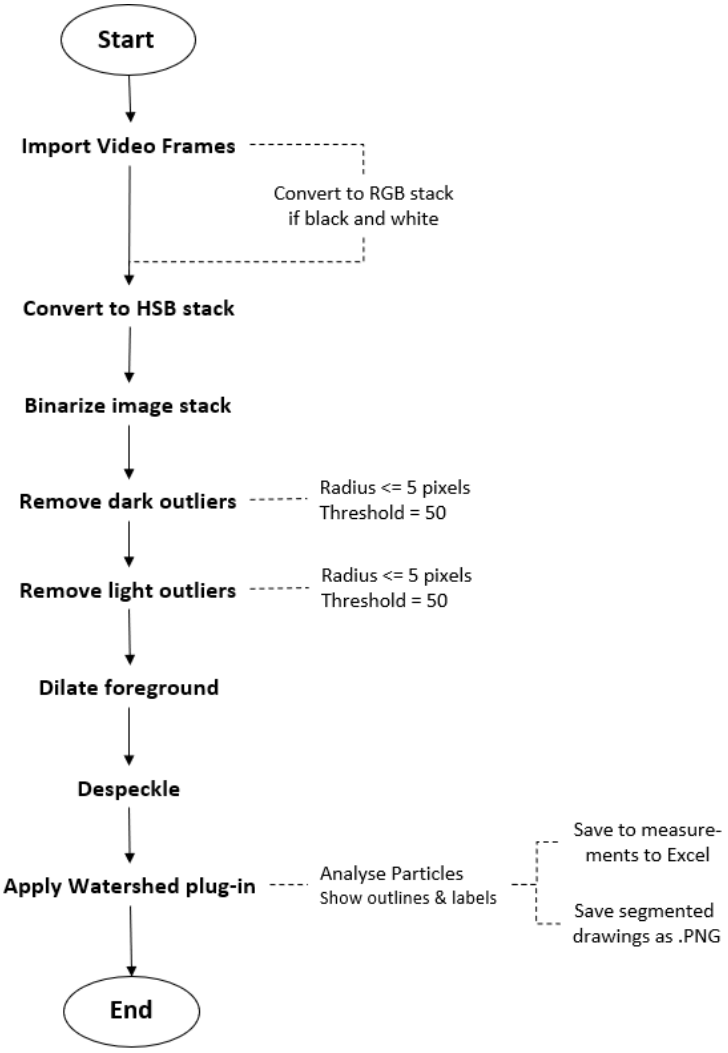

Figure D.1: Flow diagram of FIJI image processing steps implemented

List of References

Adler, P.S. and Clark, K.B. (1991). Behind the Learning Curve: A Sketch of the Learning Process. *Management Science*, vol. 37, no. 3, pp. 267–281. ISSN 0025-1909.

American Heart Association (2017). Ejection Fraction Heart Failure Measurement. Available at: https://www.heart.org/en/health-topics/heart-failure/diagnosing-heart-failure/ejection-fraction-heart-failure-measurement

Anatomy, T.M. (2017). The Heart Valves. Available at: https://teachmeanatomy.info/thorax/organs/heart/heart-valves/

Anzanello, M.J. and Fogliatto, F.S. (2011 sep). Learning curve models and applications: Literature review and research directions. *International Journal of Industrial Ergonomics*, vol. 41, no. 5, pp. 573–583. ISSN 0169-8141.

Awad, W. (2011 feb). Machine Learning Methods for Spam E-Mail Classification. *International Journal of Computer Science and Information Technology*, vol. 3, no. 1, pp. 173–184.

Babich, N. (2020 jun). What is Computer Vision and How Does it Work? An Introduction. Available at: https://xd.adobe.com/ideas/principles/emerging-technology/what-is-computer-vision-how-does-it-work/

Biga, L.M., Dawson, S., Harwell, A., Hopkins, R., Kaufmann, J., LeMaster, M., Matern, P., Morrison-Graham, K., Quick, D. and Runyeon, J. (2019 sep). 19.2 Cardiac Muscle and Electrical Activity. Available at: https://open.oregonstate.education/aandp/chapter/19-2-cardiac-muscle-and-electrical-activity/

Bizopoulos, P. and Koutsouris, D. (2019). Deep Learning in Cardiology. *IEEE Reviews in Biomedical Engineering*, vol. 12, pp. 168–193. ISSN 19411189. 1902.11122.

Bozkurt, S. (2019). Mathematical modeling of cardiac function to evaluate clinical cases in adults and children. *PLoS ONE*, vol. 14, no. 10, pp. 1–20. ISSN 19326203. Available at: http://dx.doi.org/10.1371/journal.pone.0224663

Brownlee, J. (2016). Multi-Class Classification Tutorial with the Keras Deep Learning Library.
Available at: https://machinelearningmastery.com/multi-class-classification-tutorial-keras-deep-learning-library/

Brownlee, J. (2019). How to use Learning Curves to Diagnose Machine Learning Model Performance.
Available at: https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/

Burkhoff, D.C.U. (2002). Mechanical Properties of the Heart and its Interaction with the Vascular System. *Cardiac Physiology*, pp. 1–23.

Chen, W., Yue, H., Wang, J. and Wu, X. (2014). An improved edge detection algorithm for depth map inpainting. *Optics and Lasers in Engineering*, vol. 55, pp. 69–77. ISSN 01438166.

Garg, R. (2016). Methodology for research I. *Indian Journal of Anaesthesia*, vol. 60, no. 9, pp. 640–645. ISSN 00195049.

Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., Harrington, R.A., Liang, D.H., Ashley, E.A. and Zou, J.Y. (2020). Deep learning interpretation of echocardiograms. *npj Digital Medicine*, vol. 3, no. 1, pp. 1–10. ISSN 23986352.
Available at: http://dx.doi.org/10.1038/s41746-019-0216-8

Gneiting, T. and Raftery, A.E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative Adversarial Nets.
Available at: http://www.github.com/goodfeli/adversarial

Greenlight Medical (2020 jan). 2020 Medical Technology Trends .
Available at: https://www.greenlightmedical.com/2020-medical-technology-trends/

Guhanesvar (2021). Mutual Information based Feature Selection Based for Ml | Medium.
Available at: https://guhanesvar.medium.com/feature-selection-based-on-mutual-information-gain-for-classification-and-regression-

Holbrook, R. (2021). Feature Engineering - Mutual Information.
Available at: https://www.kaggle.com/ryanholbrook/mutual-information

Jeanrenaud, X., Seiler, C., Jost, A., Kaufmann, B. and Gruner, C. (2015). What is a standard transthoracic echocardiogram performed by a cardiologist? *Cardiovascular Medicine*, vol. 18, no. 04, pp. 146–151. ISSN 1664-2031.

Jiang, J., Wang, C., Chattopadhyay, S. and Zhang, W. (2020). Road context-aware intrusion detection system for autonomous cars. In: Zhou, J., Luo, X., Shen, Q. and Xu, Z. (eds.), *Information and Communications Security*, pp. 124–142. Springer International Publishing, Cham. ISBN 978-3-030-41579-2.

Jose, G.V. (2019). Useful Plots to Diagnose your Neural Network.
Available at: `https://towardsdatascience.com/useful-plots-to-diagnose-your-neural-network-521907fa2f45`

Kaggle (2015). Second Annual Data Science Bowl.
Available at: `https://www.kaggle.com/c/second-annual-data-science-bowl`

Kaggle (2016). House Prices - Advanced Regression Techniques | Kaggle.
Available at: `https://www.kaggle.com/c/house-prices-advanced-regression-techniques`

Litjens, G., Ciompi, F., Wolterink, J.M., de Vos, B.D., Leiner, T., Teuwen, J. and Išgum, I. (2019). State-of-the-Art Deep Learning in Cardiovascular Image Analysis. *JACC: Cardiovascular Imaging*, vol. 12, no. 8P1, pp. 1549–1565. ISSN 18767591.

Lohr, J.L. and Sivanandam, S. (2015). Introduction to echocardiography. *Handbook of Cardiac Anatomy, Physiology, and Devices, Third Edition*, pp. 241–248.

Madani, A., Ong, J.R., Tibrewal, A. and Mofrad, M.R.K. (2018). Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *npj Digital Medicine*, vol. 1, no. 1, pp. 1–11. ISSN 2398-6352.
Available at: `http://dx.doi.org/10.1038/s41746-018-0065-x`

Mandes, L., Rosca, M., Ciuperca, D. and Popescu, B.A. (2020). The role of echocardiography for diagnosis and prognostic stratification in hypertrophic cardiomyopathy. *Journal of Echocardiography*, vol. 18, no. 3, pp. 137–148. ISSN 1880344X.
Available at: `https://doi.org/10.1007/s12574-020-00467-9`

Marieb, E.N. (2015). *Essentials of Human Anatomy and Physiology*. 11th edn. Pearson.

Müller, A. and Guido, S. (2016). *Introduction to Machine Learning with Python*. 1st edn. O Reilly Media, Inc., Sebastopol. ISBN 978-1-449-36941-5.

Muralidhar, K. (2021). Learning Curve to Identify Overfitting and Underfitting in Machine Learning.
Available at: `https://towardsdatascience.com/learning-curve-to-identify-overfitting-underfitting-problem-133177f38df5`

Ng, A. (2015). Machine Learning - Home | Coursera.
Available at: `https://www.coursera.org/learn/machine-learning/home/welcome`

NIH (2021). Common Data Types in Public Health Research.
Available at: https://www.nihlibrary.nih.gov/resources/subject-guides/health-data-resources/common-data-types-public-health-research

Nolen, S. (2019). GANs for Data Augmentation.
Available at: https://medium.com/abacus-ai/gans-for-data-augmentation-21a69de6c60b

OpenCV (2021). Image Segmentation with Watershed Algorithm.
Available at: https://docs.opencv.org/4.5.3/d3/db4/tutorial_py_watershed.html

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A. and Zou, J.Y. (2020). Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, vol. 580, no. 7802, pp. 252–256. ISSN 14764687.
Available at: http://dx.doi.org/10.1038/s41586-020-2145-8

Panda, A.K. (2018). DOGS VS CATS IMAGE CLASSIFIER.
Available at: https://towardsdatascience.com/fast-ai-season-1-episode-2-1-e9cc80d81a9d

Peace, I.C., Uzoma, A.O. and Abasiama Ita, S. (2015). Effect of Learning Rate on Artificial Neural Network in Machine Learning. *International Journal of Engineering Research & Technology*, vol. 4, no. 2.
Available at: www.ijert.org

Portilla, J. (2018). Python for Computer Vision with OpenCV and Deep Learning.
Available at: https://www.udemy.com/course/python-for-computer-vision-with-opencv-and-deep-learning/learn/lecture/12704179?start=0{\#}overview

Raj, B. (2018). Data Augmentation | How to use Deep Learning when you have Limited Data â Part 2.
Available at: https://medium.com/nanonets/how-to-use-deep-learning-when-you-have-limited-data-part-2-data-augmentation-c26971dc

Scikit-learn (2007). 1.12. Multiclass and multioutput algorithms.
Available at: https://scikit-learn.org/stable/modules/multiclass.html

Silva, J.F., Silva, J.M., Guerra, A., Matos, S. and Costa, C. (2018). Ejection Fraction Classification in Transthoracic Echocardiography Using a Deep Learning Approach. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, vol. 2018-June, pp. 123–128. ISSN 10637125.

TensorFlow (2021). tf.keras.losses.BinaryCrossentropy.
Available at: https://www.tensorflow.org/api_docs/python/tf/keras/losses/BinaryCrossentropy

Thapliyal, S., Pundir, Y.P., Bahuguna, A.S. and Setwal, S. (2017). Object Motion estitation using edge detection and background subtraction with block tatching algoritht. In: *Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering*, March, pp. 345–348. RICE.

The Heart Foundation (2017). Transthoracic (standard) echocardiogram.

Van der Bijl, P. (2021). Consultation on the diagnosing process and what is observed in an echocardiogram.

Vashistha, V. (2019). Computer Vision - Object Detection on Videos.
Available at: `https://www.udemy.com/course/machine-learning-on-videos-using-python/learn/lecture/20245946?start=0{\#}overview`

Venugopal, V. and Ramaswamy, S. (2015). DeepMD: Transforming How We Diagnose Heart Disease Using Convolutional Neural Networks.
Available at: `http://cs231n.stanford.edu/reports/2016/pdfs/327_Report.pdf`

Walley, K.R. (2016). Left ventricular function: Time-varying elastance and left ventricular aortic coupling. *Critical Care*, vol. 20, no. 1, pp. 1–11. ISSN 1466609X.
Available at: `http://dx.doi.org/10.1186/s13054-016-1439-6`

World Heart Federation (2017 may). Cardiovascular diseases in South Africa Factsheet.
Available at: `https://world-heart-federation.org/wp-content/uploads/2017/05/Cardiovascular_diseases_in_South_Africa.pdf`

Yui, T. (2019). Understanding Random Forest. How the Algorithm Works and Why it Is So Effective.
Available at: `https://towardsdatascience.com/understanding-random-forest-58381e0602d2`

Zhang, J., Gajjala, S., Agrawal, P., Tison, G.H., Hallock, L.A., Beussink-Nelson, L., Lassen, M.H., Fan, E., Aras, M.A., Jordan, C.R., Fleischmann, K.E., Melisko, M., Qasim, A., Shah, S.J., Bajcsy, R. and Deo, R.C. (2018). Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy. *Circulation*, vol. 138, no. 16, pp. 1623–1635. ISSN 15244539.

Zulkifli, H. (2018). Understanding Learning Rates and How It Improves Performance in Deep Learning.
Available at: `https://towardsdatascience.com/understanding-learning-rates-and-how-it-improves-performance-in-deep-learning-d0d4059`