Alkan, Muhammet (2025) *Multimodal machine learning framework for outcome prediction in congenital heart disease.* PhD thesis.

https://theses.gla.ac.uk/85428/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses
https://theses.gla.ac.uk/
research-enlighten@glasgow.ac.uk

Multimodal Machine Learning Framework for Outcome Prediction in Congenital Heart Disease

Muhammet Alkan

Advisor: Dr Fani Deligianni

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow

April 2025

Abstract

Congenital Heart Disease (CHD) affects approximately 1.2 million newborns annually worldwide, with around 4,600 cases occurring in the UK each year. CHD encompasses a complex set of structural heart defects that pose challenges in early diagnosis, risk stratification, and treatment planning. Traditional methods employed for predicting clinical outcomes constrained by the pronounced anatomical and functional heterogeneity, limited number of datasets, and single-modal clinical markers, which often hinders the development of generalisable models in congenital heart diseases. Recent advancements in the field of Machine Learning (ML) and Deep Learning (DL) offer opportunities to integrate multi-modal data sources, thereby enabling a more comprehensive understanding of patient health. This thesis explores a multi-modal machine learning framework designed to improve CHD classification and outcome prediction, by integrating multi-modal data and geometric learning.

A significant challenge encountered during the course of this research is the heterogeneity characteristic of clinical data sources. Patient records contain Electrocardiogram (ECG) signals, cardiopulmonary exercise testing metrics and unstructured clinical documentation, each with different formats and level of completeness. Furthermore, the inherent anatomical and physiological heterogeneity of CHD increases the complexity of predictive performance. It is important to note that a model trained on one subtype may exhibit suboptimal performance when applied to a different CHD presentation, making generalisation across diverse patient populations a challenge. This thesis attempts to bridge these gaps by leveraging Riemannian geometry for the purpose of feature extraction, employing covariance augmentations to generate more data, and utilising multi-modal data integration to maximise predictive potential.

Risk prediction models are statistical or machine learning-based frameworks designed to estimate the likelihood of future adverse events for a given patient or population. In the domain of cardiology, these models facilitate predictions about a variety of outcomes, including the risk of mortality and the progression of the disease. This, in turn, serves to inform the development of early intervention and treatment strategies. They often rely on features extracted from clinical data, including ECGs, laboratory results, imaging data, and patient demographics to generate meaningful insights. However, developing accurate risk prediction models with small sample sizes presents several challenges such as limited generalisation, high variance, reduced reliability, and an insufficient representation of rare cases, particularly due to the low prevalence of related events and the inherent imbalances in datasets. Furthermore, models constructed solely on mortality data often suffer from significant imbalances, which can compromise their predictive performance. To address these challenges, this thesis explores the use of Cardiopulmonary Exercise Testing (CPET) as a surrogate for mortality, providing a novel approach to enhance model accuracy even with limited data. This key contribution not only aims to improve the reliability of risk predictions but also demonstrates the potential for developing robust predictive models that can better inform clinical decisions and improve patient outcomes in the CHD population.

Geometric deep learning can be defined as a subfield of machine learning that involves the utilisation of manifold-based, or topology-aware methodologies, for the extraction of features from structured data. Unlike conventional deep learning models, which assume inputs are organised in a regular format, such as image or text, geometric deep learning preserves spatio-temporal relationships and dependencies inherent in medical signals like ECGs. In this thesis, the covariance structure of ECG signals plays a fundamental role in enhancing risk prediction models, given that ECG readings exhibit correlated variations across different leads. The utilisation of covariance matrices to represent signals in Riemannian space ensures the preservation of higher-order relationships and can generate more stable and generalisable features, thereby reducing the impact of small sample sizes.

Machine learning applications in CHD research have traditionally focused on heartbeat classification, arrhythmia detection, and patient risk stratification based primarily on ECGs interpretation. While deep learning architectures have demonstrated promising results, challenges remain in model generalisability, dataset diversity, and clinical utility. This thesis explores the development of a multi-modal machine learning framework designed to incorporate a variety of clinical indices. The framework utilises multiple data modalities including medical health records and ECGs, with the objective of enhancing the precision and reliability of outcome prediction models. Furthermore, regression models are employed to assess cardiopulmonary exercise test results, providing insights into cardiac function of the patients. Text-mining techniques are also applied to extract meaningful clinical information from physician notes, enabling richer data-driven assessments of patient conditions.

By leveraging multiple data modalities, including medical health records and ECGs, this research aims to enhance the precision and reliability of outcome prediction models by providing a more comprehensive understanding of patient health. The scope encompasses the identification and digitisation of multiple data sources, the design and implementation of relevant machine learning models, and the evaluation of the framework's performance in clinical settings. The integration of multi-modal data enhances the ability to capture complex cardiac abnormalities, thus offering a more comprehensive approach to diagnosis. The findings from this thesis contribute to the growing research on machine learning and congenital heart disease outcomes. We present a data-driven pathway for improving classification and outcome prediction, addressing key challenges such as imbalanced datasets, model generalisability and multi-modal data integration. By expanding dataset accessibility, future research can enhance the application of machine learning models in CHD, thus supporting clinical decision-making and improving patient care.

Contents

Abstract **i**

List of Abbreviations **xiii**

Acknowledgements **xvi**

Declaration **xvii**

1 Introduction **1**

1.1 Motivations 2

1.2 Scope of the Thesis 3

1.3 Research Questions 3

1.4 Contributions 4

1.4.1 List of publications 6

1.4.2 Code and data availability 8

1.5 Thesis Outline 8

2 Background on EHR & ML **10**

2.1 Electronic Health Records 10

2.2 From Descriptive Analytics to Predictive Analytics 12

2.3 Model Development and Validation Strategies of Machine Learning Models . . 13

2.3.1 Clinical decision support systems 17

2.4 From Machine Learning and Statistical Models to Clinical Decision Support Systems 19

2.5 Performance Validation in Clinical Decision Support Systems 21
  2.5.1 Calibration of clinical prediction models 22
  2.5.2 Calibration in deep learning models for clinical decision support 23
  2.5.3 Assessing bias in clinical predictive models 24
2.6 Explainability as a central component of Human-Centered CDSS 26
  2.6.1 Interpretability vs explainability 26
  2.6.2 Explainability in healthcare applications 27
  2.6.3 Evaluating explainability in clinical decision support systems 28
2.7 Machine Learning Models 29
  2.7.1 Logistic regression 29
  2.7.2 Support vector machine 30
  2.7.3 Convolutional neural network 31
  2.7.4 Riemannian geometry 31
2.8 Data Augmentation Techniques 33
  2.8.1 Resampling (oversampling or undersampling) 33
  2.8.2 Mixup 34
  2.8.3 Domain-knowledge guided augmentation 34
  2.8.4 Deep generative models 36
2.9 Projection of ECG data 36
2.10 Visualisation of high-dimensional spaces 37
2.11 Conclusions 38

3 Background on Congenital Heart Disease 39
3.1 Congenital Heart Disease 39
  3.1.1 Prevalence in the population 41
  3.1.2 Adults congenital heart disease 41
  3.1.3 Predictive models 41
3.2 Electrocardiograms 43
3.3 Clinical Models 47

3.4 Research gaps & questions . . . 49
3.5 Conclusions . . . 52

4 ECG Extraction and Representation in Congenital Heart Disease 54
4.1 Abstract . . . 54
4.2 Introduction . . . 55
4.3 Methods . . . 57
  4.3.1 Linkage of CHD EHR . . . 57
  4.3.2 ECG data preprocessing . . . 58
  4.3.3 Digitisation of 12-lead ECGs . . . 59
  4.3.4 Validation of the digitisation algorithm . . . 60
4.4 Results . . . 62
4.5 Conclusions . . . 64

5 ECG Classification with Riemannian Geometry 67
5.1 Abstract . . . 67
5.2 Introduction . . . 68
5.3 Methods . . . 70
  5.3.1 Data . . . 70
  5.3.2 From common to multiple tangent spaces . . . 70
  5.3.3 Augmentation of covariance matrices . . . 72
  5.3.4 Stratified patient leave-out . . . 74
  5.3.5 Pre-trained models . . . 74
  5.3.6 Transfer learning . . . 77
5.4 Results . . . 79
  5.4.1 Classification results with the pre-trained model . . . 83
5.5 Conclusions . . . 87
5.6 Discussion . . . 87

6 Predicting Cardiopulmonary Exercise Testing Outcomes in Congenital Heart Disease 89
6.1 Abstract . . . 89
6.2 Introduction . . . 90
6.3 Related Work . . . 91
6.3.1 CPET as outcome variables . . . 91
6.4 Methods . . . 93
6.4.1 Building a regression model . . . 93
6.4.2 Enhancing feature extraction with Riemannian geometry . . . 95
6.4.3 Regressing patient specific data out . . . 96
6.5 Results . . . 96
6.5.1 From regression to classification . . . 102
6.6 Discussion . . . 105

7 Integrating Information from Clinical Letters with ECGs 107
7.1 Abstract . . . 107
7.2 Introduction . . . 108
7.3 Methods . . . 110
7.3.1 Data extraction from clinical letters . . . 110
7.3.2 Data processing and populating the database . . . 110
7.3.3 Dimensionality reduction for visualization . . . 113
7.3.4 Integrating the information with ECGs . . . 114
7.4 Results . . . 115
7.4.1 Regression with combined information . . . 118
7.4.2 Classification with combined information . . . 123
7.5 Discussion . . . 127

8 Conclusions and Future Work 130
8.1 Summary . . . 130

8.1.1 Chapter 4:
ECG Extraction and Representation in Congenital Heart Disease 130
8.1.2 Chapter 5:
ECG Classification with Riemannian Geometry 132
8.1.3 Chapter 6:
Predicting Cardiopulmonary Exercise Testing Outcomes in CHD 134
8.1.4 Chapter 7:
Integrating Information from Clinical Letters with ECGs 136
8.2 Future Work 137

A Appendix 139
A.1 Beta Distribution 139
A.2 Vectorcardiogram 140
A.3 Dower Transformation 141

Bibliography 143

List of Tables

3.1 Summary of ECG based deep learning approaches for CHD classification and prognosis. 50

5.1 Ablation study: AUC scores for different methods. 79

5.2 Performance metrics of the machine learning models. (With **bold** we highlight the best six models). 80

5.3 Effect of covariance augmentations on label-wise distances to class centroids. . 82

5.4 Classification results on ECGs, with pre-trained model. 84

5.5 Classification results on ECGs, with fine-tuned model. 86

6.1 Regression results on tabular data (using **vendor** list), with baseline LR. 97

6.2 Regression results on tabular data (using **vendor2** list), with baseline LR. . . . 98

6.3 Regression results on tabular data, with SVM model. 99

6.4 Regression results on ECGs, with SVM model. 101

6.5 Classification results using regression groups with SVM model. 103

7.1 List of extracted variables from the clinical letters. 111

7.2 Completeness of data extraction for all variables. 116

7.3 Regression results on ECGs + clinical letters, with SVM model. 120

7.4 Regression results on ECGs + clinical letters, with SVM model using covariance augmentations. 121

7.5 Classification results using regression groups with SVM model. 124

List of Figures

2.1 K-fold cross validation method (k=3). 14
2.2 Performance evaluation of ML models. 16
2.3 Interpretable vs explainable models [117]. 27
2.4 Model architecture. 31
2.5 Visualisation of oversampling technique. 33
2.6 Example application of mixup on signals. 34
2.7 Augmentations on ECG signals using VCG space. 35
2.8 Example vectorcardiograms. 37
2.9 Different perplexity values (from 2 to 100). 38

3.1 Overview of anatomical variations of different congenital heart diseases. 40
3.2 Example beats for each class (Lead I). 44
3.3 Placement of leads in a 12-lead ECG, 10 physical electrode positions. 45
3.4 Key components of an ECG trace. 46

4.1 Top 15 anatomic diagnoses for 1409 patients. 57
4.2 Example ECG PDF document. 59
4.3 Digitisation and linkage of ECG PDF documents that use vectorised graphic format to store ECG waveforms. 60
4.4 ECG alignment using first 50 ECGs on lead I of patients with a Mustard procedure. 60
4.5 P, Q, R, S, and T points on a single ECG wave. 62
4.6 Comparison of our digitisation results with the Paper-ECG[204]. 63
4.7 Bland-Altman plots between vendor and extracted values. 64

5.1 Average VCGs across anatomical defects in CHD. 68
5.2 Mapping ECG signals to tangent space using Riemannian manifold. 71
5.3 Multiple tangent space concept. 72
5.4 t-SNE visualisations on the tangent space. 73
5.5 Stratified patient leave-out setup. 74
5.6 Architecture of the pretrained model [207]. 76
5.7 AUC scores for the best six approaches. 81
5.8 Confusion matrices of DEF, TS_COV and MTS_COV approaches with an MLP classifier. 81
5.9 AUC for the best model (MTS_COV). 82
5.10 Confusion matrix comparison of the pre-trained model in different settings. . . 83
5.11 Confusion matrices of the fine-tuned model using last five, seven and eight layers. 85
5.12 Confusion matrices of the fine-tuned model using last twelve and all the layers. 85
5.13 Confusion matrix comparison of the best pre-trained model with the single tangent space (TS_COV) and multiple tangent space (MTS_COV) approaches. . . 86

6.1 Some important variables from a CPET PDF document. 94
6.2 Vendor features extracted from the ECG letters. 95
6.3 Histogram of CPET documents. 95
6.4 Overall process of using ECG signals. 96
6.5 Kernel Density Estimations (KDEs) of the baseline model LR (using **vendor** list). 97
6.6 Regression line plots of the baseline model LR (using **vendor** list). 98
6.7 Kernel Density Estimations (KDEs) of the baseline model LR (using **vendor2** list). 98
6.8 Regression line plots of the baseline model LR (using **vendor2** list). 99
6.9 Kernel Density Estimations (KDEs) of the SVM model (using **vendor** list). . . 100
6.10 Regression line plots of the SVM model (using **vendor** list). 100
6.11 Kernel Density Estimations (KDEs) of the SVM model (on ECGs). 101
6.12 Regression line plots of the SVM model (on ECGs). 101

6.13 Data distribution of each class, with CPET groups. 103
6.14 Comparison of t-SNE visualizations on the tangent space. 104
6.15 Comparison of learning curves for the classification model. 104
6.16 Learning curve projection for the classification model on ratio of ventilation to carbon dioxide production (VE/VCO_2). 105

7.1 Summary of the data extraction steps. 112
7.2 Text to matrix transformation using CountVectorizer. 114
7.3 t-SNE dimensionality reduction of diagnosis, intervention, medication, and ESC classification. 116
7.4 Selection of t-SNE initialisation parameters. 117
7.5 Clustering results. 119
7.6 Kernel Density Estimations (KDEs) of the SVM model. 120
7.7 Regression line plots of the SVM model. 121
7.8 Kernel Density Estimations (KDEs) of the SVM model using covariance augmentations. 122
7.9 Regression line plots of the SVM model using covariance augmentations. . . . 122
7.10 Predicting cardiopulmonary exercise biomarkers: comparison of predicted vs. real values with fitted regression lines 122
7.11 Comparison of t-SNE visualizations on the tangent space. 125
7.12 Calibration plot comparing predicted probabilities with observed outcomes. . . 125
7.13 Comparison of learning curves for the classification model. 127
7.14 Learning curve projection for the classification model on VE/VCO_2. 127

A.1 Average VCGs across anatomical defects in CHD. 140
A.2 Frank's orthogonal lead system [238]. 141
A.3 Augmentations on ECG signals using VCG space. 142

List of Abbreviations

ACHD Adult Congenital Heart Disease 5, 41, 57

ASD Atrial Septal Defect 39, 40, 43, 58, 61, 70, 84–86

BELCODAC BELgian COngenital heart disease Database combining Administrative and Clinical data 108

BMI Body Mass Index 26, 96

BNF British National Formulary 110

CHD Congenital Heart Disease i–iii, 1–5, 8, 11, 19, 38–43, 47–49, 51, 52, 54–56, 64, 66–68, 70, 87, 89–93, 105–110, 115, 118, 127–129, 132–136, 140

CHF Chronic Heart Failure 92

CHI Congenital Heart Initiative 108

CNN Convolutional Neural Network 29, 31, 48

CONCOR CONgenital CORvitia 108

CPET Cardiopulmonary Exercise Testing ii, 2, 4, 5, 89–93, 95, 96, 105–107, 109, 111, 112, 127–129, 135, 136

DL Deep Learning i, 18, 29, 43, 47, 49, 88

DNN Deep Neural Network 47, 75, 76

ECE Expected Calibration Error 23, 24

ECG Electrocardiogram i–iii, 2–5, 9, 19, 31, 34–36, 43–49, 51, 52, 54–70, 74–76, 79, 80, 83, 87, 89–91, 93–97, 99, 100, 102, 103, 105–109, 112, 114, 115, 118, 122–124, 127–132, 135–138, 140–142

EHR Electronic Health Record 10, 19, 20, 24, 25

ESC European Society of Cardiology 110, 113, 115, 118

JSI Jaccard Similarity Index 114, 115

KDE Kernel Density Estimation 97–100, 120, 121

LOOCV Leave-One-Out Cross Validation 15

LR Logistic Regression 29, 30, 47, 93, 95, 96, 99, 118, 120

LSTM Long Short-Term Memory 48

MDM Minimum Distance to Means 31, 32, 79

ML Machine Learning i, 18–20, 54–56, 60, 67, 95

MLP Multi-Layer Perceptron 79–81, 86

MNN Multimodal Neural Network 48

NICOR National Institute for Cardiovascular Outcomes Research 108

NLP Natural Language Processing 20, 109, 128

NYHA New York Heart Association 49

OCR Optical Character Recognition 89, 90, 111

OUE Oxygen Uptake Efficiency 92

OvR One-vs-Rest 81

PA Pulmonary Atresia 57, 58, 61, 70, 83–86

ResNet Residual Network 24, 48

RMSE Root Mean Square Error 97

RNN Recurrent Neural Network 48

SACCS Scottish Adult Congenital Cardiac Service 108, 109

SPD Symmetric Positive Definite 32, 70, 72, 96

SPLO Stratified Patient Leave-Out 74, 76, 93

SVM Support Vector Machine 30, 54, 57, 74, 79–81, 93, 96, 99, 102, 118, 120, 123

SWEDCON SWEDish registry of CONgenital heart disease 108

t-SNE t-distributed Stochastic Neighbor Embedding xii, 37, 38, 72, 82, 104, 113–115

ToF Tetralogy of Fallot 40, 43, 47, 57, 58, 61, 70, 83–85, 92

VCG Vectorcardiogram 35, 36, 47, 68, 79, 80, 140–142

VSD Ventricular Septal Defect 39, 40

VCO_2 carbon dioxide production 89, 92, 107

VE pulmonary ventilation 89, 92, 107

VE/VCO_2 ratio of ventilation to carbon dioxide production xii, 89, 92, 93, 95, 97–107, 112, 120–124, 126, 127

VO_2 oxygen consumption 89, 91–93, 95, 97–104, 106, 107, 112, 120–124, 127, 128

Acknowledgements

I would like to express my sincere gratitude to all those who have supported me throughout my PhD journey. First and foremost, I would like to thank my supervisor, Dr. Fani Deligianni, for her invaluable guidance and expertise throughout this research. Your insights and constructive feedback have been instrumental in shaping my research and helping me grow as a scholar. Additionally, this research would not have been possible without the clinical support of Dr. Gruschen Veldtman, whose help was invaluable in shaping my research. Thank you for your contributions and for making this journey a rewarding experience. I would also like to extend my appreciation to the members of my viva committee, Dr. Mark McGill, Dr. Edmond S. L Ho and Dr. Stathis Hadjidemetriou, for their thoughtful suggestions and encouragement. Your diverse perspectives have enriched my work and inspired me to think critically about my research.

I am thankful to my colleagues and friends in Glasgow for creating a positive and supportive environment that has enriched my experience. The discussions, bi-weekly sessions, occasional picnics and camaraderie we shared made this journey not only productive but also enjoyable.

I would like to acknowledge the sponsorship and resources provided by the Turkish Ministry of National Education through the YLSY scholarship, which made this research possible.

On a personal note, I would like to express my heartfelt gratitude to my family for their unwavering love and support. I would like to dedicate this thesis to my wife, Sena, whose unconditional love, encouragement and support have been a guiding light throughout my academic studies. Her understanding and encouragement during the inevitable ups and downs of this journey have been invaluable, and her unwavering support has been a constant source of strength. I am profoundly grateful for her presence in my life, which has made this achievement possible. In closing, I humbly thank Allah for giving us patience, strength and our little joy Mihrimah.

Declaration

Name: Muhammet Alkan

Registration Number: xxxxxxx

I certify that the thesis presented here for examination for a PhD degree of the University of Glasgow is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it) and that the thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice.

The copyright of this thesis rests with the author. No quotation from it is permitted without full acknowledgement.

I declare that the thesis does not include work forming part of a thesis presented successfully for another degree.

I declare that this thesis has been produced in accordance with the University of Glasgow's Code of Good Practice in Research.

I acknowledge that if any issues are raised regarding good research practice based on review of the thesis, the examination may be postponed pending the outcome of any investigation of the issues.

Chapter 1

Introduction

It is estimated that around 1.2 million babies worldwide are born with Congenital Heart Disease (CHD) each year [1]. CHD refers to a complex group of structural defects of the heart, which present challenges for medical diagnosis and treatment. In the UK, it affects approximately 1 in every 100 babies born, which is estimated to result in approximately 4,600 cases annually [2]. The prevalence of CHD emphasises the importance of early detection and treatment in order to improve the outcomes for those affected.

Clinical care for CHD has significantly improved over the years, allowing most children with the condition to reach adulthood. However, managing the symptoms of adult CHD and ensuring a good quality of life requires careful monitoring of disease progression and timely clinical interventions. This includes adjusting medications and providing appropriate treatments as needed. Despite the advancements in the domain of surgical interventions and long-term disease management, the ability to predict risk remains critical for the optimisation of treatment strategies and the enhancement of long-term prognoses for affected individuals.

Traditional methods employed for outcome prediction are constrained by the limited datasets and markers, which often hinders the comprehensive nature of the diseases. A significant limitation inherent in many traditional models is their incapacity to effectively integrate heterogeneous data sources, resulting in suboptimal predictive performance and a partial understanding. However, recent developments in the field of machine learning have facilitated the creation of predictive models that possess the capacity to integrate diverse data sources, thereby promoting a more comprehensive understanding of patient health.

This thesis presents a multi-modal machine learning framework designed for predicting clinical outcomes in patients with congenital heart disease, addressing the limitations of traditional models. By leveraging multiple data modalities including medical health records and Electrocardiograms (ECGs), and geometry-driven data transformations such as Riemannian feature extraction, this research aims to enhance the precision and reliability of outcome prediction models. Additionally, this thesis explores transfer learning techniques for adapting state-of-the-art pre-trained models to resource-constrained hospital environments, ensuring that advanced predictive methodologies remain applicable and scalable in real-world healthcare settings.

1.1 Motivations

The rising prevalence of CHD underscores the need for advanced predictive models that can provide accurate outcome predictions for the patients. While the overall survival rates for individuals with CHD have improved significantly, the low prevalence of related events and the inherent imbalances within this population presents a unique challenge for predictive models. Even with large datasets, the infrequency of adverse outcomes makes it difficult for traditional models to succeed in accurately predicting risks and complications. This limitation has motivated us to utilise Cardiopulmonary Exercise Testing (CPET) as a proxy for mortality in our machine learning model. This approach not only serves as a key contribution to the field but also answers a critical question: can we successfully develop predictive models with limited data by relying on this surrogate? By incorporating CPET data, we aim to enhance the model's ability to assess the risk of complications and guide long-term management for patients with CHD.

Predictive models can also help tailor treatment plans to individual needs by predicting how different patients might respond to various interventions [3]. Additionally, these models can also assess the risk of complications or adverse outcomes in patients with CHD, helping healthcare providers prioritise high-risk patients for more intensive monitoring and care [4]. For adults with CHD, predictive models can monitor disease progression and predict potential complications, guiding long-term management and follow-up care [5]. Given the complexity and variability of

CHD presentations, traditional prediction methods often fall short due to their reliance on limited and singular data modalities. By developing a multi-modal machine learning framework, we aim to improve the precision and reliability of outcome predictions, ultimately enhancing patient care and supporting clinical decisions. This approach represents a key contribution to the field, as it not only addresses the challenges posed by the low prevalence of related events, inherent imbalances within CHD population and the limitations of traditional prediction methods, but it also highlights the potential for the development of more sophisticated predictive models even in the context of limited data.

1.2 Scope of the Thesis

This thesis focuses on the development of a multi-modal machine learning framework for outcome prediction in congenital heart disease. The scope encompasses the identification and digitisation of relevant data sources, the design and implementation of relevant machine learning models, and the evaluation of the framework's performance in clinical settings. The research will leverage data from medical records and ECGs to create a comprehensive predictive model. By integrating these data sources, the thesis aims to provide a more comprehensive understanding of patient health and improve the accuracy of outcome predictions.

1.3 Research Questions

1. **How can we develop effective diagnostic classification and risk prediction models for CHD using small and extremely heterogeneous populations?**

   The fundamental question guiding this study is concerned with the exploration of methodologies and techniques that can be applied in the context of limited data availability and heterogeneous patient characteristics.

2. **Can we exploit the ECG waveforms to develop risk prediction models in CHD?**

   Another objective of this study is to explore the potential of ECG waveforms in the development of risk prediction models for CHD. Additionally, the study will investigate the potential of surrogate outcomes, such as CPET variables, with a view to determine whether they can be utilised for prediction purposes and how incorporating CPET variables as surrogate outcomes can improve the accuracy of these predictions.

3. **How to enhance the performance of prognostic/risk prediction models under extremely small populations?**

   Explore methods for combining data from multiple sources, such as ECGs and clinical letters, to create a more comprehensive dataset. And implement robust cross-validation techniques tailored for small datasets to ensure reliable model evaluation and prevent overfitting.

4. **How to improve feature extraction applying geometric learning on ECGs?**

   Investigate the potential of Riemannian geometry to capture the intrinsic geometric properties of ECG signals and their relevance to CHD classification. Explore the use of geometric augmentations on the same space to generate more data that are clinically relevant and to enhance the diversity of training samples.

1.4 Contributions

This thesis presents a multi-modal machine learning framework designed for predicting clinical outcomes in patients with congenital heart disease. Through a structured approach, it integrates multi-modal data processing, and geometric learning principles to address the complexities of CHD diagnostics in resource-constrained clinical environments. The contributions of this thesis span multiple chapters, as outlined below:

Establishing the role of AI in CHD diagnostics (Chapter 2 & 3): The initial chapters establish the foundation for the investigation of machine learning applications in cardiology. These chapters discuss the current challenges, limited data availability, and computational constraints in hospital settings. With a literature review highlighting the strengths and weaknesses of the related works, thereby establishing the foundation for the proposed methodologies.

Digitisation of ECG signals for computational analysis (Chapter 4): This chapter develops a pipeline for transforming ECG signals from PDF documents into structured digital datasets, overcoming barriers associated with non-digitised clinical formats. This step is instrumental in facilitating the transition to automated data processing, thereby ensuring compatibility between ECG signals and machine learning models employed for diagnostic and predictive applications.

Application of Riemannian geometry to ECG feature extraction (Chapter 5): This chapter demonstrates the novel application of Riemannian embeddings for enhancing ECG feature extraction. By applying covariance matrix transformations and tangent space projections, this chapter significantly enhances classification accuracy, offering a mathematically structured approach for interpreting ECG signals in clinical research.

Pre-trained models in resource-constrained environments (Chapter 5): This chapter also examines the pre-trained state-of-the-art model approaches, highlighting their potential for efficient adaptation in data-limited hospital settings. Findings suggest that fine-tuning pre-trained models does not always yield optimal results due to specialised CHD patient demographics, reinforcing the necessity for tailored solutions.

Predicting CPET Outcomes in CHD Patients (Chapter 6): This chapter introduces CPET as a surrogate of mortality and demonstrates the effectiveness of Riemannian geometry and the covariance augmentation technique in predicting outcomes from cardiopulmonary exercise testing, by applying both regression and classification models. To our knowledge, this is the first time that a Riemannian framework has been successfully exploited for the prediction of patient outcomes in Adult Congenital Heart Disease (ACHD). The chapter assesses the benefits of the geometric learning approach with ablation studies, demonstrating that this approach yields superior performance for risk stratification in patients with CHD.

Integration of multi-modal patient data (Chapter 7): A key innovation in this thesis is the integration of ECG data with clinical information, which facilitate the structuring of unstructured patient records. This approach, when combined with the proposed geometric learning augmentations, enhances the robustness of classification and regression models. And demonstrates that it enables machine learning models to better capture anatomical and physiological variations associated with CHD.

Discussion, limitations, and future directions (Chapter 8): This chapter provides a reflection on the findings, outlining practical challenges such as dataset size limitations, generalisability concerns, and computational constraints. It proposes future research opportunities, advocating for collaborative efforts to access larger datasets, lightweight models for hospital deployment, and advanced multi-modal feature extraction strategies.

1.4.1 List of publications

Chapter 7

- Alkan M, Veldtman G, and Deligianni F. "Predicting Cardiopulmonary Exercise Testing Outcomes in Congenital Heart Disease Through Multi-modal Data Integration and Geometric Learning". arXiv:2503.14239. 2025. (preprint for Scientific Reports journal)
- Alkan M, Huijsdens H, Jones Y, and Deligianni F. "Electronic Health Records: Towards Digital Twins in Healthcare". Smart and Connected Healthcare. Springer 2025. (Book chapter).
- Alkan M, Zakariyya I, Leighton S, Sivangi K, Anagnostopoulos C, and Deligianni F. "Artificial Intelligence-Driven Clinical Decision Support Systems". Smart and Connected Healthcare. Springer 2025. (Book chapter).

Chapter 6

- Alkan M, Veldtman G and Deligianni F. "Machine Learning Prediction Models of CPET as a Surrogate of Mortality in CHD Patients". The Living Lab: Healthcare Innovation Symposium. 2024. (Abstract).

Chapter 5

- Alkan M, Veldtman G and Deligianni F. "Riemannian Prediction of Anatomical Diagnoses in Congenital Heart Disease based on 12-lead ECGs". IEEE International Symposium on Biomedical Imaging. ISBI 2024.

- Alkan M, Deligianni F, Anagnostopoulos C and Veldtman G. “Prediction of Anatomical Diagnoses in Congenital Heart Disease”. Medical Image Understanding and Analysis. MIUA 2023. (Abstract).

- Alkan M, Deligianni F, Anagnostopoulos C and Veldtman G. “Prediction of Anatomical Diagnoses in Congenital Heart Disease”. AI-UofG-2023: AI in Pervasive Well-Being and Healthy Ageing Event. 2023 (Abstract/Poster Presentation).

Chapter 4

- Alkan M, Deligianni F, Anagnostopoulos C, Zakariyya I and Veldtman G. “Digitization and Linkage of PDF Formatted 12-lead ECGs in Adult Congenital Heart Disease”. CJC Pediatric and Congenital Heart Disease. 2025.

- Alkan M, Leighton S, Krishnadas R, Cavanagh J, Mallikarjun PK, Gkoutos GV, Everard L, Singh SP, Freemantle N, Fowler D, Jones PB, Sharma V, Murray R, Wykes T, Drake RJ, Buchan I, Lewis SW and Birchwood M, Deligianni F. “Challenges in Developing Risk Prediction Models based on EHR”. AI-UofG-2023: AI in Pervasive Well-Being and Healthy Ageing Event. 2023 (Abstract/Poster Presentation).

- Verma S, Alkan M, Deligianni F, Anagnostopoulos C, Diller G, Walker L, Johnston FC, Danton M, Walker H, Swan L, Hunter A, McGuire A, Dawes M, Stott S, Lyndsey M, Walker N, Veldt- man G. “Development of a Semiautomated Database for Patients With Adult Congenital Heart Disease”. Canadian Journal of Cardiology. 2022.

- Alkan M, Leighton S, Krishnadas R, Cavanagh J, Mallikarjun PK, Gkoutos GV, Everard L, Singh SP, Freemantle N, Fowler D, Jones PB, Sharma V, Murray R, Wykes T, Drake RJ, Buchan I, Lewis SW and Birchwood M, Deligianni F. “Challenges in Developing Risk Prediction Models based on EHR”. 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2022.

- Alkan M, Leighton S, Krishnadas R, Cavanagh J, Mallikarjun PK, Gkoutos GV, Everard L, Singh SP, Freemantle N, Fowler D, Jones PB, Sharma V, Murray R, Wykes T, Drake

RJ, Buchan I, Lewis SW and Birchwood M, Deligianni F. “Challenges in Developing Risk Prediction Models based on EHR”. The 13th Annual Scottish Informatics & Computer Science Alliance (SICSA) PhD Conference, 2022 (Abstract/Poster Presentation).

1.4.2 Code and data availability

The study was conducted in accordance with the ethical guidelines established by the Institutional Governance Division of NHS Golden Jubilee National Hospital and the Health Research Authority. Approval was granted under the designated project ID (23/SC/0436, IRAS project ID: 335717), thus ensuring adherence to the regulatory standards for medical research and the protection of patient data. Ethical considerations included the safeguarding of patient confidentiality, adherence to informed consent protocols, and maintenance of transparency in data collection and analysis. These ethical safeguards ensure that the research contributes in a manner that upholds the highest standards of integrity and responsibility in healthcare research.

To ensure transparency and reproducibility, some of the figures and code related to this study are publicly accessible on GitHub at `https://github.com/alkanmuhammet/ISBI`. Additionally, all remaining figures and code will also be publicly available soon, thus facilitating further validation and exploration of the methodologies presented.

1.5 Thesis Outline

The thesis is organised into eight chapters as follows:

- Chapter 2: Introduces the background on electronic health records and the potential of machine learning in healthcare. It provides an overview of clinical decision supports systems, their evolution, and the role of machine learning models in enhancing healthcare delivery and decision-making processes.

- Chapter 3: Discusses the background and significance of congenital heart disease, and ECGs which is a key tool in CHD assessment. This chapter explores the clinical presentation and existing challenges in diagnosing and treating CHD, emphasising the need for the clinical predictive models.

- Chapter 4: Outlines the data extraction methodology, including data collection, preprocessing, and digitisation. It details the various sources of data used in the study, the techniques employed for cleaning and transforming raw data, and the strategies for ensuring data quality and validation.

- Chapter 5: Presents the results and analysis of the ECG classification models with Riemannian geometry. This chapter explains the theoretical foundations of Riemannian geometry, its application in ECG classification, and the performance metrics used to evaluate the model's effectiveness. This chapter also discusses pre-trained state-of-the-art model approaches highlighting their limitation in different patient demographics and examines transfer learning approaches reinforcing the necessity for tailored solutions.

- Chapter 6: Presents the results and analysis of the classification and regression models on cardiopulmonary exercise test results. It covers the development of models, the variables considered, and the statistical techniques used to assess model accuracy.

- Chapter 7: Examines the results and analysis of integrating information from clinical letters with ECGs. This chapter discusses the data extraction techniques employed to extract relevant information from clinical letters, the integration process with ECG data, and the impact on predictive model performance.

- Chapter 8: Discusses the implications of the findings, potential limitations, and future research directions. It provides a comprehensive summary of the study's contributions to the field, areas for improvement, and suggestions for further investigation to enhance the understanding and management of congenital heart disease.

Chapter 2

Background on EHR & ML

2.1 Electronic Health Records

Electronic Health Records (EHRs) are digital representations of a patient's health-related data. They comprise a heterogeneous collection of data types, including both structured and unstructured data to support clinical decision-making. Structured data includes demographic details, laboratory test results, medication prescriptions and diagnostic codes, which are systematically organised and easily searchable [6]. Conversely, unstructured data includes clinical notes, radiology reports and patient narratives which provides rich contextual insights but poses challenges for automated processing [6, 7].

In contrast to the systematic monitoring of patients in clinical trials (at intervals such as 6 months, 1 year, or 5 years), the data in electronic health records is not collected at regular intervals but is instead recorded when an event occurs. For example, an individual may be admitted to hospital for a period of three days, undergo extensive medical follow-up during their time in hospital, and then not have any further contact with the medical system for the following year. This irregular data collection pattern reflects the real-world healthcare dynamics but also introduces complexities in patient analysis [8].

The transition from paper-based to electronic health records has had a profound impact on the management of healthcare data, enhancing the quality of clinical decision-making and opening new avenues for medical research [9]. It has also led to a shift from independent records to integrated databases. These transitions, which take place over a long period of time, present challenges and requirements for successful integration, even the different health systems are organised very differently [10, 11].

Clinical registries are of great importance in modern healthcare systems, serving as comprehensive repositories of patient data focused on specific conditions or diseases. These systems are designed to collect, store, and analyse information over extended periods, offering invaluable insights into patient care and treatment outcomes. The primary functions of clinical registries include monitoring and enhancing the quality of care, evaluating the efficacy of various treatments, and providing a solid foundation for research initiatives in targeted medical areas. Research findings indicate that clinical registries play a significant role in enabling healthcare providers to assess the impact of interventions and refine best practices [12]. Several notable examples of clinical registries have been established across Europe, each contributing significantly to their respective fields. The Dutch CONCOR Registry, initiated in 2005, stands as the national registry and DNA bank for patients with congenital heart disease in the Netherlands [13]. This registry has played an important role in advancing understanding and treatment of congenital heart conditions in the country, facilitating the monitoring of Congenital Heart Disease (CHD) prevalence, identification of long-term complications and optimisation of treatment strategies [14].

Clinical registries focus on specific conditions or diseases, whereas clinical databases encompass a broader range of health information from a variety of sources. These databases are not limited to particular conditions but instead offer a comprehensive overview of patient health data. A prime example of this broader approach is the hospital database, which typically contains a wealth of information on various medical conditions, treatments, and patient outcomes within a single healthcare institution or network. In the context of evolving healthcare systems, clinical registries and databases have become invaluable tools for supporting day-to-day patient care, driving medical research, informing policy decisions and contributing to the advancement of medical knowledge and practice [14, 15]. As healthcare continues to evolve, these data collection and analysis systems will undoubtedly contribute to the shaping of the future of patient care and medical research.

2.2 From Descriptive Analytics to Predictive Analytics

The processes of healthcare information retrieval may be conceptualised as a pathway that progresses from descriptive analytics through to diagnostic analytics, predictive analytics, and finally prescriptive analytics [16, 17]. Descriptive analytics employs techniques such as data aggregation, data mining, and intuitive visualisation to facilitate comprehension of historical data [18]. Descriptive analytics provides answers to questions such as the number of patients admitted to a hospital in a given year, the number of patients who died within 30 days, or the number of patients who contracted an infection [19]. In other words, descriptive analytics offers intuitive ways to summarise the data via histograms and graphs and show the data distribution properties. However, one limitation of descriptive analytics is that it has limited ability to guide decision-making because it is based on a snapshot of the past [20]. This limitation can be addressed by progressing through diagnostic, predictive and prescriptive analytics, each offering a more in-depth understanding of patient care and operational efficiency.

Diagnostic analytics is a specific type of analytics that employs the examination of data in order to ascertain the underlying cause of a given phenomenon. Diagnostic analytics may entail the utilisation of correlation techniques with the objective of discerning interrelationships between clinical variables, treatments, and drugs [21]. The application of predictive analytics enables the estimation of the probable outcome and likelihood of an event. For instance, one might wish to predict the mortality risk of a patient, the length of hospitalisation, or the risk for infection [22]. The objective of predictive analytics is to utilise historical data in order to provide valuable insights into potential future scenarios. The demand for predictive analytics is driven by the need for evidence-based approaches to predict and avoid adverse events. Crucially, predictive analytics facilitate early intervention, which can save lives and enhance quality of life.

Prescriptive analytics, on the other hand, aim to make decisions that yield optimal outcomes. Prescriptive analytics seek to quantify the impact of prospective decisions by offering guidance on potential outcomes before the decision is made [23]. Consequently, they provide recommendations regarding actions that take into account the results of predictive analytics [24]. In other words, prescriptive analytics are instrumental in transforming a prediction model into a clinical decision support model.

2.3 Model Development and Validation Strategies of Machine Learning Models

Model evaluation and selection is a critical step in the machine learning development process. In an ideal scenario, we would have access to data that perfectly represents the entire target population. In this case, we could train and test the machine learning model using the same data, and the error rate obtained would closely reflect the true error rate when the number of samples is very large. However, in reality, the error rate obtained when training and testing on the same dataset is positively biased [25, 26]. This is because the model has been exposed to the same data during both the training and testing phases, which can lead to an overly optimistic estimate of its performance. To address this issue, in real-world applications, it is common to split the available data into two separate sets: a training set and a testing set. Typically, around 70% of the data is used for training the model, while the remaining 30% is reserved for testing its performance on unseen examples [27]. By separating the training and testing data, we can better evaluate the model's ability to make accurate predictions on new, unseen data, which is crucial for deploying the model in real-world applications [28].

In real-world applications, we estimate the empirical risk based on a limited number of testing samples. This involves measuring the loss function with respect to our trained classifier, as discussed in [29]. Empirical risk estimation is achieved through the computation of the average loss over the data points (m is the number of samples) according to a loss function L, which penalises the differences between the predicted values $f(x)$ and the actual targets y.

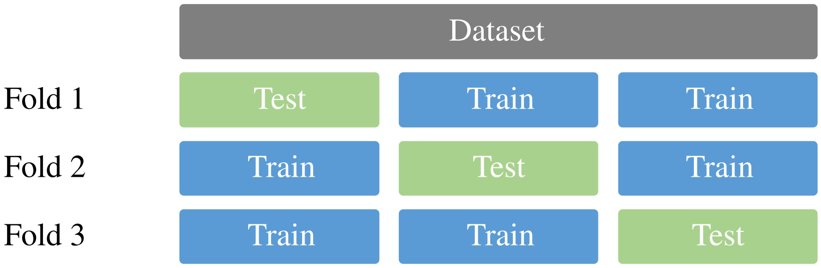
$$R_S(f) = \frac{1}{m}\sum_{i=1}^{m} L(y_i, f(x_i)) \tag{2.1}$$

The error modelled based on the Bernoulli distribution is employed to calculate the bound on the error, thereby providing an indication of the potential deviation between the empirical risk estimation and the true risk with a high probability $(1-\delta)$ of accuracy [30]. In Equation 2.2,

m represents the number of samples and δ represents the confidence parameter which quantifies the probability that our estimate is accurate.

$$E = \sqrt{\frac{1}{2m'} \ln(\frac{2}{\delta})} \tag{2.2}$$

Variations in the empirical risk estimation can arise from several factors. These include random variations in the testing set, the training set, the learning algorithm itself, and even the noise inherent in the data classes being considered [25–27, 31, 32]. One key advantage of the hold-out method (where the training and testing sets are independent) is that it provides some guarantees about the model's performance on data it has not been trained on before [33]. However, we must also consider the confidence intervals around the empirical risk estimation. For example, when evaluating a machine learning algorithm, we should not assume a Gaussian distribution of the loss error, as the errors may be clustered near zero [34].



	Dataset		
Fold 1	Test	Train	Train
Fold 2	Train	Test	Train
Fold 3	Train	Train	Test

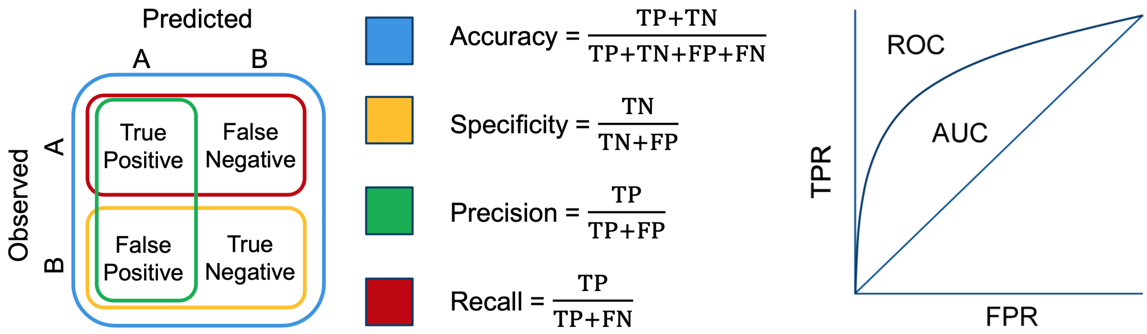
Figure 2.1: K-fold cross validation method (k=3).

The k-fold cross validation is one of the most popular error estimation approach in machine learning model training and evaluation [26, 35]. In this method, the dataset is divided into k distinct parts or "folds" like in Figure 2.1. During each iteration, one of these folds is reserved for testing, while the remaining $(k-1)$ folds are used for training the model. This process is repeated k times, with a different fold serving as the test set each time. By doing so, we obtain k separate estimates of the classifier's error rate. These estimates can then be averaged to give the mean performance of the algorithm across the different folds. Examining the variability of the error estimates across the k iterations can also provide valuable insights into the algorithm's stability and robustness [35]. A key advantage of k-fold cross-validation is that the test samples are independent between the different folds, as there is no overlap. This helps to ensure a more reliable and unbiased assessment of the model's generalisation capabilities [35, 36].

A potential issue that can arise with standard k-fold cross validation is that the data may not be evenly distributed across classes. This problem becomes worse when dealing with imbalanced class data, which is common in healthcare applications [37, 38]. To address this, we can employ a technique called stratified k- fold cross validation. In this approach, the folds are created in a way that ensures the class distribution within each fold closely matches the original class distribution in the overall dataset [27]. A special case of k-fold cross validation is Leave-One-Out Cross Validation (LOOCV). In this case, the value k is set to the number of samples in the dataset, meaning that each sample is used as the test set once while the remaining $(k-1)$ samples are used for training. LOOCV has the advantage of utilising most of the available data, which can result in relatively unbiased classifier. However, this comes at the cost of significant computational expense, especially as the dataset size grows [27]. While LOOCV may provide better performance estimates in datasets with extreme values, it is important to note that this is not a guarantee of an unbiased classifier, especially when dealing with small datasets [26, 35]. The underlying assumption of LOOCV is that the training set is representative of the true data distribution, which may not always hold [35].

One key aspect of validation techniques such as k-fold cross-validation and LOOCV is that the estimate is not based on a single, fixed classifier. Instead, the model is retrained each time, producing a new classifier with each iteration. This approach has both advantages and disadvantages. The primary advantage is that it allows us to assess the stability of machine learning models across different data partitions [26, 34, 35]. However, the disadvantage is that when comparing the performance of different algorithms, we must remember that we are comparing the average performance estimates of various classifiers, rather than a single, fixed classifier as in the holdout method [26, 34, 35].

One way to describe the performance of classification algorithm is through a confusion matrix as in Figure 2.2b. This square matrix has rows and columns equal to the number of classes. The diagonal elements represent the true positives and true negatives, assuming "positive" refers to one class and "negative" to another. The off-diagonal elements indicate false positives and false negatives. From the confusion matrix, we can derive several performance metrics. For instance, accuracy is the ratio of correctly predicted observations to the total observations. Specificity, or the true negative rate, shows how well the classifier identifies negative cases, while recall also called sensitivity, or the true positive rate, indicates how well it identifies positive cases [39]. Precision reflects the positive predictive value for a class. The F1 score combines recall and precision into a single metric, weighting them evenly [40].


Predicted
A
B
Observed
A
B
True Positive
False Negative
False Positive
True Negative
Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
Specificity = $\frac{TN}{TN+FP}$
Precision = $\frac{TP}{TP+FP}$
Recall = $\frac{TP}{TP+FN}$
ROC
AUC
TPR
FPR

(a) Confusion matrix (b) Performance metrics (c) The ROC curve

Figure 2.2: Performance evaluation of ML models.

An alternative method to assess the performance of a machine learning algorithm is by using a Receiver Operating Characteristic (ROC) curve, as shown in Figure 2.2c. The ROC curve plots the false positive rate on the horizontal axis and the true positive rate on the vertical axis. The true positive rate, or sensitivity, indicates how well the classifier identifies positive cases, while the false positive rate can be expressed as $(1 - \mathit{specificity})$ [41]. Thus, the ROC curve illustrates the trade-off between sensitivity and specificity for the classifier. It is generated by varying the threshold for classifying positive and negative cases. This makes the ROC curve a comprehensive measure of performance, as it considers different threshold settings [41, 42]. It is used not only to analyse the behaviour of machine learning algorithms but also for model selection by identifying the optimal threshold region [42]. For a random classifier, the ROC curve would be a straight diagonal line. The area under the curve (AUC) summarises the classifier's performance, with higher values indicating better performance [43]. The AUC is often used to compare different classifiers. Additionally, there are various extensions of the ROC and AUC for multi-class scenarios [44–47].

When comparing the performance of one algorithm against another, or multiple algorithms across one or more datasets, it is common to use null hypothesis statistical testing. Several statistical tests can validate the performance of two algorithms, but it is crucial to consider the assumptions underlying these tests [31]. For instance, the paired t-test assumes a normal distribution, independence of measurements, and an adequate sample size. If the normality assumption is violated, non-parametric tests are often used [48]. One such test is the Wilcoxon signed-rank test, an alternative to the paired t-test. This test is based on the ranks of the absolute differences, making it more robust to outliers [31, 48]. It is important to note that both parametric and non-parametric tests can be manipulated by increasing the number of samples, potentially affecting the results of the null hypothesis statistical testing approach [31, 48].

2.3.1 Clinical decision support systems

The domain of clinical research requires a sophisticated methodology for the assessment of predictive models, which must extend beyond the conventional data science approach [49, 50]. As we explore this intricate domain, it is essential to address a series of pivotal questions that influence the advancement and verification of machine learning models, ultimately determining their efficacy as decision-support systems in healthcare. In prediction modelling, our primary focus is on estimating the risk of adverse events based on a combination of factors. We seek to understand not only the predictive power of these factors but also their individual contributions to the model's decision-making process. This understanding is crucial, as it allows us to incorporate subject matter knowledge into the modelling pipeline, thereby bridging the gap between data-driven insights and clinical expertise [51–53].

The selection of patient data for model development is of critical importance. It is not uncommon for data to be collected for purposes other than the study at hand, which raises questions about their representativeness [54, 55]. It is therefore necessary to examine whether the patient records truly reflect the population for which the study is intended [56, 57]. Furthermore, the treatment of prognostic factors and their effects presents a unique challenge. While traditional studies often consider treatment effects to be negligible compared to prognostic factors, there are instances where these effects warrant specific attention [58, 59]. Adjusting for baseline prognostic factors can offer significant advantages in estimating treatment effects applicable to individual patients [60].

The reliability and completeness of predictor measurements present another challenge in model development. Incomplete datasets are a common occurrence, with missing values for potential predictors [61]. The approach to handling these missing data can have a significant impact on the model's performance and validity [62]. While complete case analysis, which excludes patients with any missing values, is a straightforward solution, it often results in a loss of valuable information [63]. More sophisticated methods, such as imputation techniques that leverage correlations between variables, offer a more nuanced approach to preserving data integrity [62].

The selection of the prediction outcome is of paramount importance in clinical research. Outcomes such as 30-day mortality rates are frequently relevant to a variety of research questions [55]. However, it is not merely the nature of the outcome that plays a role, but also its frequency within the dataset. This frequency effectively determines the sample size, which in turn influences the statistical power and reliability of the model [60]. As we proceed through these considerations, we appreciate that the development of machine learning models for clinical decision support systems is a multifaceted process. It necessitates a delicate equilibrium between statistical precision, clinical relevance, and practical applicability [49]. By addressing these crucial questions and challenges, we establish a foundation for more robust and reliable predictive models that can genuinely enhance clinical decision-making and, ultimately, patient care.

The development of Machine Learning (ML) models for clinical decision support systems requires careful consideration of various factors. The handling of predictor variables is a crucial aspect, with categorical variables often requiring expert knowledge for appropriate concatenation or fusion [49]. Continuous variables may be modelled linearly, but this approach can affect interpretation and performance, especially when relationships are non-linear [64]. Model specification and variable selection are crucial elements of the development process. Traditional stepwise selection techniques in regression models can be unreliable when applied to small sample sizes or rare events [65]. Deep Learning (DL) models rely on global explainability methods to identify influential input parameters [66]. The principle of Occam's razor is frequently employed, favouring models with fewer parameters to prevent overfitting and enhance generalisation [67].

Digital twins can be used by researchers to conduct experiments and analyse data in more detail, in a controllable and repeatable environment [68]. This can be achieved through the collection of patients' existing health data and the translation of the growing amount of such data into knowledge that is relevant for decision making [69]. Such patient data can be used to monitor various health indicators and generate important insights. Although this approach has recently been adopted in the healthcare, it is widely used in engineering [70]. Simulation of an object or a system is one of the essential parts in engineering for predictive analysis, such as determining when maintenance is required or observing how the system responds to different inputs [71]. Adopting this engineering practice, it is possible to create a digital twin that is paired with the patient's data. This enables healthcare professionals to process large amounts of patient data, leading to more personalised and effective care [72]. In our work [73], we have identified the opportunity to develop digital twins of CHD patients by extracting information over a decade from clinical letters, which involved diagnoses, diagnostic complexity, interventions, arrhythmia, medications, and demographic data. We believe this information can also be linked to the patient's 12-lead Electrocardiogram (ECG) dataset when in digitised format. This provides the opportunity to develop robust ML models for diagnoses and risk prediction and explore their capabilities.

2.4 From Machine Learning and Statistical Models to Clinical Decision Support Systems

In the early stages, statistical models were employed to analyse clinical data and predict outcomes, relying on predefined algorithms with the assumption that the data was structured. The initial versions of clinical decision support systems were rule-based, utilising if-then logic derived from clinical guides and expert knowledge to provide decision support [74, 75]. The widespread adoption of EHRs provided a rich source of patient data, which enabled more sophisticated analysis and real-time decision support [76]. Integration with EHRs enabled clinical decision support systems to provide more comprehensive support and diagnostic assistance [77].

The advent of machine learning models enabled the analysis of vast quantities of clinical data, facilitating the identification of patterns and the prediction of outcomes. This approach has led to improvements in diagnostic accuracy without relying on hand-curated features [78–81]. In the other hand, the application of Natural Language Processing (NLP) techniques has enabled the extraction and interpretation of information from unstructured data sources within EHRs, such as clinical notes [82–84].

It has been demonstrated that clinical decision support systems based on machine learning can enhance patient safety, reduce errors and improve the overall quality of care [85]. These systems are capable of processing large volumes of data rapidly, thereby providing clinicians with timely and accurate information to inform their decision-making processes [86]. However, integrating clinical decision support systems with existing healthcare infrastructure can be complex and costly, and there can be concerns about the reliability of ML models [87–89]. The effectiveness of ML models also depends on the quality and completeness of the data used to train them [85]. As we explore this challenging domain, it is essential to address a series of key questions that form the development and validation of machine learning models, ultimately determining their efficacy as decision support systems in healthcare.

The foundations of the validation process of clinical prediction models have been presented in Steyerberg [90]. The fundamental research question or hypothesis represents the core of this process. The selection of the prediction outcome is of paramount importance in clinical research. Outcomes such as 30-day mortality rates are frequently relevant to a variety of research questions. However, it is not merely the nature of the outcome that carries significance, its frequency within the dataset is also a crucial factor [91]. The frequency in question has an impact on the sample size, which in turn affects the statistical power and reliability of the model [60, 91, 92].

In prediction modelling, our primary focus is on estimating the risk of adverse events based on a combination of factors. We seek to understand not only the predictive power of these factors but also their individual contributions to the model's decision-making process [93]. This understanding is crucial, as it allows us to incorporate subject matter knowledge into the modelling pipeline, bridging the gap between data-driven insights and clinical expertise [94–96].

The selection of patient data for model development is of critical importance. It is not uncommon for data to be collected for purposes other than the study at hand, which raises questions about their representativeness [60]. It is therefore necessary to examine whether the patient records truly reflect the population for which the study is intended. Furthermore, the treatment of prognostic factors and their effects presents a unique challenge. While traditional studies often consider treatment effects to be negligible compared to prognostic factors, there are instances where these effects warrant specific attention [97]. Adjusting for baseline prognostic factors can offer significant advantages in estimating treatment effects applicable to individual patients [98, 99].

The reliability and completeness of predictor measurements pose another hurdle in model development. Incomplete datasets are common, with missing values for potential predictors [100]. The approach to handling these missing data can significantly impact the model's performance and validity [101]. While complete case analysis – excluding patients with any missing values – is a straightforward solution, it often results in a loss of valuable information [102]. More sophisticated methods, such as imputation techniques that leverage correlations between variables, offer a more nuanced approach to preserving data integrity [61, 63, 103].

In considering these issues, it becomes evident that the development of machine learning models for clinical decision support systems is a complex and multifaceted process. There must be a careful balance between maintaining statistical rigour, ensuring clinical relevance, and practical applicability. By addressing these core questions and challenges, we can facilitate the creation of more robust and reliable predictive models which can, ultimately, enhance clinical decision-making and improve patient care.

2.5 Performance Validation in Clinical Decision Support Systems

Validation of prediction models tailored for clinical decision support systems should occur through both internal and external methods [90, 104]. Internal validation, using techniques like split-sample validation, cross-validation, or bootstrapping, assesses reproducibility within the development population [105]. External validation, involving patients from different populations, tests the model's generalisability across various settings and demographics [106].

While internal validation techniques provide valuable insights into a model's performance, external validation serves as a crucial complement in assessing predictive models for clinical use [3, 107, 108]. This process involves testing the model on data that is entirely separate from the development dataset, often collected from different institutions or time periods [109]. Despite the growing number of publications on prediction models, studies employing both internal and external validation remain relatively scarce [110]. This highlights the challenges in establishing predictive models as reliable decision support systems [111].

Successful external validation strengthens confidence in a model's clinical utility. It demonstrates that the model's predictions remain accurate across different patient populations and healthcare settings [49]. This robustness is essential for establishing the model as a trustworthy component of clinical decision support systems [112]. In other words, external validation offers a more rigorous test of a model's generalisability, revealing how well it performs in diverse real-world scenarios. It helps identify potential overfitting issues that may not be apparent through internal validation alone [108]. By subjecting the model to new, unseen data, researchers can gauge its true predictive power and assess its potential as a reliable decision support tool [3, 110].

In essence, external validation acts as a bridge between theoretical model development and practical clinical application. It provides the evidence needed to justify the integration of predictive models into healthcare decision-making processes, ultimately contributing to improved patient care and outcomes.

2.5.1 Calibration of clinical prediction models

In clinical practice, the validation of machine learning models extends beyond traditional prediction performance metrics. While measures like AUC, precision, and F1 scores are crucial, they do not fully capture a model's clinical utility. Effective clinical decision support systems require assessment of underlying risk estimates and clinical usefulness, which can be subjective and application-dependent.

Model calibration is a critical aspect of validation, referring to the agreement between observed outcomes and predictions [50]. For instance, if a model predicts a 15% risk of 30-day mortality, approximately 15 out of 100 patients with such a prediction should experience the outcome. Calibration is typically assessed using flexible calibration curves, which plot estimated risks against observed proportions of events.

Two key measures of calibration are the calibration-in-the-large (alpha) and calibration slope (beta). A well-calibrated model should have an alpha close to zero and a beta close to one. However, these measures alone do not guarantee perfect calibration across all risk levels. Visualisation through calibration plots helps identify areas of over or underestimation.

Poor calibration can arise from various factors, including differences in patient characteristics or disease prevalence between development and validation populations, changes in healthcare practices over time, model over-fitting, and measurement errors in medical data. Strategies to improve calibration include model refitting, continuous updating, and addressing population shifts dynamically. Sample size significantly impacts calibration assessment, with at least 200 events and non-events recommended for precise evaluation. In smaller datasets, evaluating moderate calibration through intercept and slope calculations may suffice.

The importance of calibration in clinical settings cannot be overstated. A poorly calibrated model, even with high discrimination, can lead to misleading or potentially harmful clinical decisions. For example, in predicting in vitro fertilization (IVF) success rates, overestimation could give false hope and expose patients to unnecessary risks. In conclusion, comprehensive validation of clinical decision support systems must encompass discrimination, calibration and clinical usefulness.

2.5.2 Calibration in deep learning models for clinical decision support

Calibration is also a critical aspect of deep learning models in clinical decision support systems, particularly for establishing trustworthiness with users. A well-calibrated model provides confidence estimates that accurately reflect the probability of correct predictions. For instance, a model with 90% confidence should be correct 90 out of 100 times. In practice, perfect calibration is unattainable, but we aim to approximate it. The Expected Calibration Error (ECE) is a common metric used to assess calibration, measuring the difference between confidence and accuracy across prediction bins.

Recent studies have shown that deeper and more complex neural networks tend to be poorly calibrated, despite high accuracy [113, 114]. Interestingly, increasing model depth or the number of convolutional filters per layer tends to worsen calibration error while improving predictive performance. For example, a 110-layer Residual Network (ResNet) model demonstrated high accuracy but poor calibration, potentially limiting its reliability as a decision support tool. The causes of miscalibration in deep networks are not fully understood, but they appear to correlate with model complexity and capacity. Conversely, techniques like weight decay (L2 regularisation) can help reduce calibration error. These findings suggest that in complex networks, over-fitting may manifest in probability estimates rather than classification errors.

For healthcare applications, reliable confidence measures are crucial. Users of clinical decision support systems need to be aware of the confidence level in disease diagnostics. Efforts to improve calibration in deep neural networks include architectural modifications and adjustments to training and optimisation strategies. While the ECE is widely used, it has limitations. The choice of bin number involves a bias-variance trade-off, and the metric may not fully capture calibration in multi-class problems. Additionally, it can be affected by cancellation effects between over- and under-confident predictions. Adaptive binning schemes have also been proposed to enhance the stability of calibration measurements.

In conclusion, while deep learning models can achieve high accuracy, their calibration remains a significant challenge. Addressing this issue is essential for developing trustworthy and effective clinical decision support systems. Future research should focus on refining calibration metrics and developing techniques to improve the reliability of confidence estimates in complex neural networks.

2.5.3 Assessing bias in clinical predictive models

EHRs present unique challenges and opportunities in predictive modelling. The concept of "informative presence" in EHRs refers to the potential information carried by the presence or absence of patient data at any given time point. This phenomenon differs from missing data, as there is no intention to collect data from healthy individuals. For example, EHRs can be inherently biased because sicker individuals are monitored more frequently. This type of informative presence, indicates that the frequency of health records can reflect a patient's health status. "Informative observation" extends this concept to the timing, frequency, and patterns of longitudinal observations in EHRs, which can provide insights into a patient's evolving health state [115]. While these phenomena can complicate causal or association studies, they also offer potential sources of implicit information that can enhance predictive models.

Therefore, the evaluation of predictive models should extend beyond discrimination and calibration to encompass potential biases that may introduce systematic errors. A framework for assessing bias in clinical predictive models has been proposed, focusing on four key domains: participant selection, variable and predictor selection, outcome assessment, and analysis [116]. This framework emphasises the importance of appropriate inclusion and exclusion criteria, consistent predictor definition and assessment across participants, and the use of standardised outcome definitions. Biases can stem from flaws in study design, execution, or data analysis. To identify such biases, a comprehensive approach is necessary, considering the model's intended use, target population, predictors, and predicted outcomes. The timing between predictor assessment and outcome determination is crucial for meaningful results. Additionally, the analysis should consider sample size adequacy, handling of continuous and categorical predictors, and appropriate performance evaluation methods. Researchers are advised to avoid selecting predictors based solely on univariate analysis. Models can be categorised as having low, medium, or high risk of bias based on the assessment of these domains. Notably, prediction models developed without external validation should generally be considered high risk, except when based on very large datasets.

In conclusion, while identifying and addressing bias is crucial for developing robust clinical predictive models, the inherent characteristics of EHRs, such as informative presence and observation, present both challenges and opportunities. Researchers must carefully interpret results while also exploring innovative ways to leverage these implicit data patterns to improve prediction accuracy.

2.6 Explainability as a central component of Human-Centered CDSS

2.6.1 Interpretability vs explainability

Interpretability and explainability, while often used interchangeably, have distinct meanings in machine learning. Interpretability is an inherent model property, whereas explainability involves post-hoc methods to elucidate non-interpretable models. Consider a network assessing patient risk based on factors like Body Mass Index (BMI), age, smoking habits, alcohol consumption, and blood pressure. For an 85-year-old female patient with a BMI of 32 (class 1 obesity or low-risk obesity), high blood pressure, and no smoking or alcohol use, the system labels her 'at risk' recommending medication. However, this output alone may not suffice for a doctor to trust and act upon the model's decision. Understanding the reasoning behind the model's output becomes crucial, highlighting the importance of explainability in complex models, especially in critical domains like healthcare where comprehending the decision-making process is essential for informed and ethical treatment decisions.

Explainability in machine learning models addresses crucial questions about a model's performance, success conditions, and decision factors. Consider that in heart failure prediction age is a significant factor, with individuals over 60 having a 60% probability. Furthermore, a BMI exceeding 25 increases risk by 20%, as does smoking for over a decade. High blood pressure is also correlated with heart failure. An explainable model should identify these key input factors, quantify their impact on the decision, and elucidate why they are significant. This insight into the model's underlying function aids result interpretation, clarifies decision-making processes, and helps understand model failures in noisy conditions. Essentially, explainability provides transparency into the model's inner workings, enabling users to comprehend not just what the model predicts, but why it makes those predictions.

Figure 2.3 provides a simplified overview of interpretable and explainable models. On the left, we see inherently interpretable models such as decision trees, linear regression, and logistic regression. These models have been widely used in clinical practice and decision-making due to their simple construction and easily understandable results. The straightforward nature of these models allows practitioners to readily interpret their outputs and understand the reasoning behind decisions. Consequently, these models do not require additional methods to explain their results, as their decision-making process is transparent by design. This intrinsic interpretability distinguishes them from more complex models that may require additional explanation techniques to elucidate their outputs.

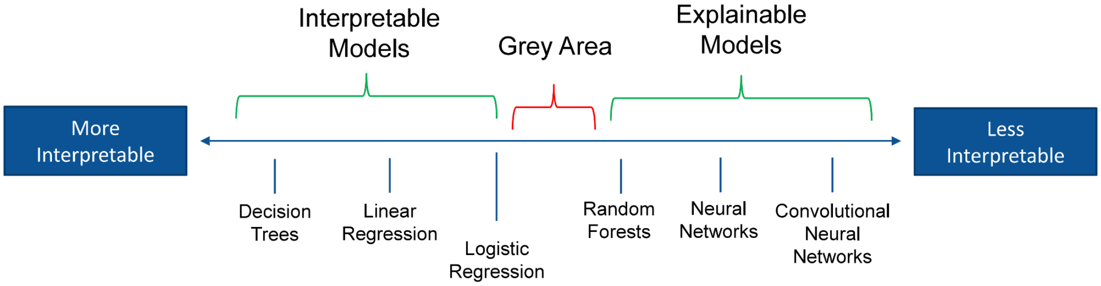


Figure 2.3: Interpretable vs explainable models [117].

While linear regression and decision trees offer high interpretability, they often fall short in performance, especially given the complexity of modern datasets and available computational resources. Sacrificing predictive accuracy for inherent interpretability is increasingly seen as suboptimal in many applications. Instead, the focus has shifted towards developing methods to explain high-performing, complex models. This approach aims to harness the superior predictive power of sophisticated algorithms while still providing insights into their decision-making processes.

2.6.2 Explainability in healthcare applications

In healthcare applications, explainability is particularly critical for assessing model stability, visualising relationships affecting outcomes, and enabling ethical analysis, especially concerning minority groups. It also facilitates patient involvement in decision-making processes and allows for the evaluation of privacy risks associated with complex model representations. This transparency is essential for maintaining the confidence of healthcare professionals, patients, and end-users. Moreover, explainability aids in monitoring model performance over time, as data distributions may shift and affect outcomes.

The significance of explainability extends to various stakeholders. Clinicians require trustworthy models that contribute to scientific knowledge. Patients need assurance of fair treatment and absence of hidden biases. Data scientists and developers utilise explainable models for debugging and improving product efficiency. Management must ensure regulatory compliance, while regulatory bodies need to certify model adherence to legislation.

Explainability is associated with several key objectives, including trustworthiness, causality inference, transferability, informativeness, confidence, fairness, accessibility, interactivity, and privacy awareness. While an explainable model may not guarantee absolute trust or prove causality, it can provide valuable insights into potential causal relationships and help validate results from other inference techniques.

2.6.3 Evaluating explainability in clinical decision support systems

The evaluation of explainability methods in clinical decision support systems is a crucial aspect of their development and implementation. This process considers not only the technical aspects of the explanations but also their effectiveness for end-users, whether they are healthcare professionals or lay individuals.

Evaluation frameworks can be categorised into application-grounded experiments with end-users and human-grounded experiments with lay individuals. These experiments may employ both qualitative and quantitative approaches. Quantitative evaluations often utilise metrics based on questionnaires assessing the usefulness, satisfaction, and interest provided by the system's explanations. They may also measure human-machine task performance in terms of accuracy, response time, error likelihood, and error detection ability.

While mathematical methods provide strong evidence for the efficacy of explainability techniques, evaluating machine learning models with human users is essential to understand both the model's performance and its integration into human-in-the-loop systems.

Explanations typically provide three types of information: the importance of features or attributes to the model, including their interactions; the reasoning behind specific predictions; and an approximation of the complex model using a simpler, interpretable surrogate model such as rule-based systems, decision trees, or linear models.

The evaluation of explanations involves assessing both the model's intrinsic interpretability and the quality of its approximation by interpretable explanations. Key aspects of this assessment include clarity (consistency of rationale for similar instances), parsimony (complexity and compactness of the explanation), fidelity (accuracy in describing the task model), and soundness (truthfulness to the task model).

Attribution-based explanations, which identify the input features most relevant to the model's decision, are common in post hoc explanation methods. While these explanations may not fully satisfy the sufficiency property, they often meet the parsimony criterion if the features are understandable to humans. For clinicians, such explanations are valuable in comparing the model's decision-making process with their own clinical knowledge.

User-based evaluation, both quantitative and qualitative, is crucial in understanding how trust in AI models affects overall system performance in human-in-the-loop scenarios. This approach bridges the gap between technical performance and practical utility in clinical settings, ensuring that explainable AI systems not only perform well mathematically but also integrate effectively into clinical workflows and decision-making processes.

2.7 Machine Learning Models

We evaluated a range of learning algorithms ranging from classical statistical models such as Logistic Regression (LR) to DL architectures like Convolutional Neural Network (CNN). Also, we tried to improve the deep learning results by using different geometric learning implementations like Riemannian.

2.7.1 Logistic regression

Logistic Regression (LR) is a method used to model binary classification tasks as a linear function of the relationship between the input variables and the target variable [118]. It predicts a dependent variable by using the relationship between existing independent variables. LR can be expressed as in the Equation 2.3 where $p(x)$ indicates the probability of an event, β_i indicates the coefficients and the x_i indicates the variables.

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ ... \ + \beta_m x_m \tag{2.3}$$

The interpretable nature of LR coefficients facilitates the evaluation of the individual predictor contributions and the integration of clinical knowledge [119, 120]. Furthermore, it serves as a robust baseline for binary clinical prediction, exhibiting fast training times and optimal performance when linear relationships are apparent [49, 121, 122].

2.7.2 Support vector machine

Support Vector Machines (SVMs) are a widely used machine learning algorithm that is employed for a variety of classification tasks. The fundamental objective of an SVM model is to identify the optimal decision boundary, the point at which the separation between different classes is maximised [123]. It constructs a hyperplane in a high-dimensional space, thereby ensuring the largest possible margin between data points from different classes [124]. Given a set of labelled training data, SVM finds the support vectors which are the critical data points closest to the hyperplane determining the classification boundary. The SVM decision function $f(x)$ can be formulated as in Equation 2.4, where w represents the weight vector, x represents the feature vector, and b is the bias term. In the context of classification, the SVM predicts the class label based on the *sign* of the decision function, as in Equation 2.4:

$$f(x) = sign\left(w^T x + b\right) \tag{2.4}$$

SVM is particularly effective in high-dimensional spaces and is commonly used for applications such as image recognition, medical diagnosis, and bioinformatics [124]. The ability to handle both linear and non-linear classification through the utilisation of kernel functions, makes it a versatile tool in machine learning [25, 124].

2.7.3 Convolutional neural network

Convolutional Neural Network (CNN) is a type of feed forward neural network mostly used on computer vision problems, especially image classification. Due to its success in feature extraction, it is often applied to other supervised learning problems to perform classification [125]. Each layer is fully connected to the following layers. By training it on many subjects with similar properties, CNN can generalise the features learned from the examples and then can recognise similar patterns from a new subject. The model shared in the Figure 2.4 was utilised in the CNN predictions.

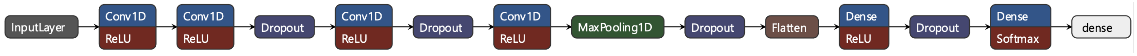


Figure 2.4: Model architecture.

Typical CNN building blocks include convolutional layers, non-linear activation functions (e.g., ReLU), pooling and batch normalization. Stacking these blocks enables hierarchical feature extraction from low-level edges to high-level parts. Convolutions share weights across spatial positions, allowing CNNs to efficiently learn local patterns and generalise well across image translations and modest deformations when trained on sufficiently diverse datasets [126]. Transfer learning with pre-trained CNNs (fine-tuning or using networks as fixed feature extractors) extends their applicability to related supervised tasks with limited data, including medical imaging, remote sensing, and time-series classification after suitable input preprocessing [127, 128]. Regularisation techniques (e.g. dropout, augmentation, weight decay) along with architecture architecture choices (e.g. ResNet, DenseNet, EfficientNet) have been shown to help mitigate over-fitting and improve optimisation for very deep models [28, 129].

2.7.4 Riemannian geometry

Riemannian geometry is a branch of differential geometry that studies smooth manifolds and curved spaces with unique geometries [130]. Using this geometry on electroencephalography (EEG) data has proven to be very useful in effectively mapping data points and transforming time series into covariance matrices [131]. We can perform classification in the tangent space of the covariance matrices estimated from the ECGs or we can use a Minimum Distance to Means

(MDM) classifier that functions directly on the covariance matrices. MDM takes covariance matrices as input and then tries to perform classification by the nearest centroid. A centroid is estimated for each of the classes and then, for each new point, the class is estimated according to the nearest centroid.

The covariance $\sigma(x,y)$ of two random variables x and y with n samples is calculated as in Equation 2.5.

$$\sigma(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{2.5}$$

And then, with the covariance $\sigma(x,y)$, we can calculate entries of the covariance matrix, where our dataset is expressed by the matrix X as in Equation 2.6.

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T \tag{2.6}$$

Mean of covariance matrices according to the Riemannian metric is estimated as in Equation 2.7, where weight is expressed by w and distance is expressed by d.

$$\arg\min_{\mathbf{C}} \sum_i w_i d_R(\mathbf{C}, \mathbf{C}_i)^2 \tag{2.7}$$

MDM operates on covariance matrices by representing each class with a mean (centroid) on the appropriate manifold and assigning new samples to the class whose centroid is nearest under a chosen metric [132]. For covariance-based MDM the affine-invariant distances are commonly used because covariance matrices live on the manifold of Symmetric Positive Definite (SPD) matrices. The utilisation of such metrics has been demonstrated to preserve geometry and improve classification accuracy over Euclidean distance, as Euclidean distance ignores curvature which can result in an underestimation or overestimation of the similarity between SPD matrices [133]. Affine-invariant metrics preserve important properties like positive-definiteness and meaningful geodesic structure, which often improves classification accuracy in applications like brain–computer interfaces and other biomedical signal classification tasks where covariance captures spatial/temporal structure [134].

2.8 Data Augmentation Techniques

Despite the ever-increasing capabilities of deep learning models, the success is primarily based on statistical correlations between input data and predefined labels. This fully supervised learning approach heavily depends on the variety and volume of curated datasets, which are assumed to be well-balanced and to have identical distributions between training and testing data. However, these assumptions are not always valid and handling imbalanced data is a common challenge in machine learning, especially in fields like healthcare. Various techniques have been developed to create diverse training samples, thereby improving generalisation and robustness.

2.8.1 Resampling (oversampling or undersampling)

Oversampling can be applied by increasing the number of instances in the minority class by duplicating or generating synthetic samples (e.g., SMOTE - Synthetic Minority Over-sampling Technique). On the other hand, undersampling is applied to reduce the number of instances in the majority class to balance the dataset.

In a machine learning problem, we need to make sure that we are oversampling or undersampling the dataset only after we split dataset into training, testing and validation. If we perform oversampling or undersampling before splitting the dataset into training and testing, there is a high probability that our model will suffer from data leakage. Visualisation of oversampling technique can be seen in Figure 2.5.

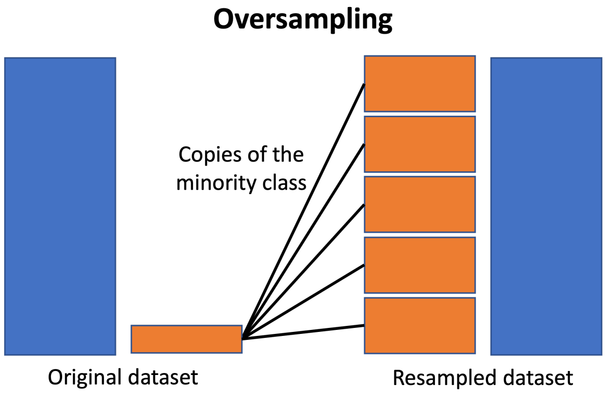


Figure 2.5: Visualisation of oversampling technique.

2.8.2 Mixup

This robust augmentation strategy involves linearly interpolating between training samples in both input and output spaces. Mixup has been widely adopted across various domains, including ECG and EEG, as it introduces a useful inductive bias that helps control model complexity and improves generalisation by creating virtual training samples that are not identical to the originals.

In this technique, samples are mixed-up to create a more balanced distribution that a model can effectively learn. This helps to improve accuracy and calibration in poorly represented (minority) classes. An example of its application can be seen in Figure 2.6.

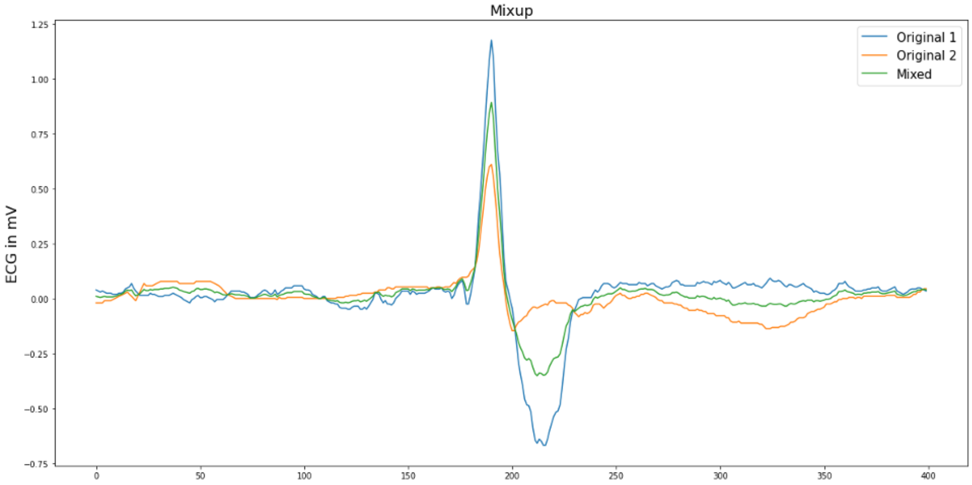


Figure 2.6: Example application of mixup on signals.

This procedure can be defined as follows (λ is a float between 0 and 1):

$$signal_{new} = \lambda * signal_1 + (1 - \lambda) * signal_2 \tag{2.8}$$

$$target_{new} = \lambda * target_1 + (1 - \lambda) * target_2 \tag{2.9}$$

In Figure 2.6, λ value is selected as 0.5 to apply mixup procedure. The value of λ should be a float between 0 and 1. Choosing a higher λ can create virtual samples that are more distant from the training samples.

2.8.3 Domain-knowledge guided augmentation

Domain knowledge guided augmentation techniques are frequently used by machine learning researchers to improve the performance of models by incorporating specific domain knowledge into the data augmentation process. This approach leverages expert knowledge to create more relevant and diverse training data, which can improve the model's ability to generalise and perform well on real-world tasks.

Data augmentation in the spatial domain is one of the most widely used domain knowledge guided augmentation techniques to improve the performance of machine learning models, especially in areas such as computer vision and medical imaging. The following is a short summary of the most commonly applied geometric transformation strategies.

- **Flipping:** Horizontal and vertical flipping to create mirror images.
- **Rotation:** Rotating images by various degrees to introduce different perspectives.
- **Scaling:** Zooming in or out on images to simulate different distances.
- **Translation:** Shifting images along the x-axis or y-axis to create variations in position.

Physiologically inspired augmentation methods provide another domain-knowledge guided augmentation technique, which is tailored to the specific physiological characteristics of the data. For instance, Gopal et al. proposed a method for 12-lead ECG signals that involves mapping these signals to Vectorcardiogram (VCG) space, applying 3D spatial augmentations, and then back-projecting to the original ECG space. This approach is motivated by the need to account for application-specific data variations. An example workflow can be seen in Figure 2.7.

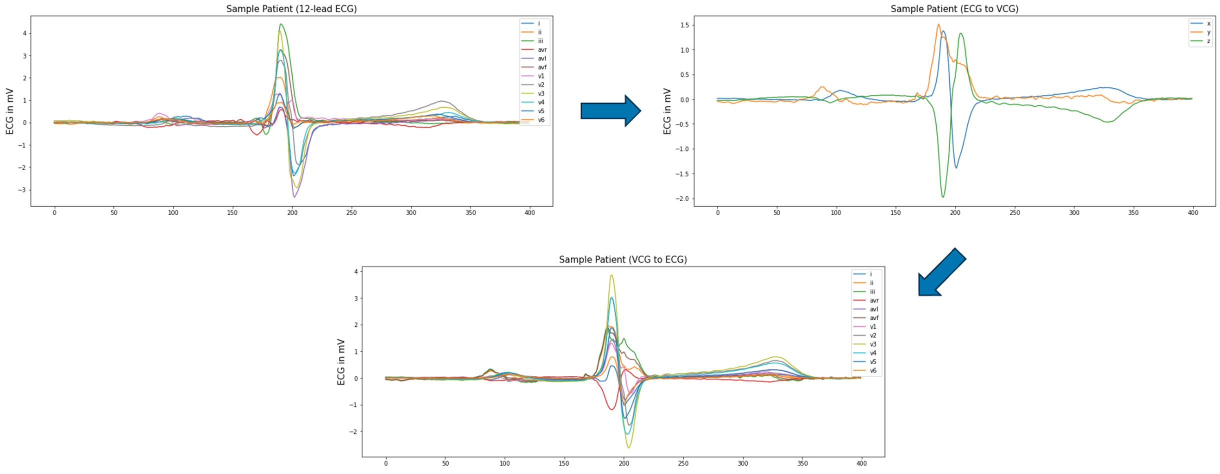


Figure 2.7: Augmentations on ECG signals using VCG space.

2.8.4 Deep generative models

Deep generative models are a powerful tool for data augmentation, with applications including the generation of synthetic data that mimics real-world samples, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) [135, 136]. They can generate high-quality synthetic data samples that enhance the diversity and size of training datasets, which is especially useful in scenarios with limited data. GANs comprise two neural networks (generative and discriminative) that are trained concurrently. In contrast, VAEs represent another type of generative model that learns to encode data into a hidden space and subsequently decode it back to the original space.

In summary, effective data augmentation techniques are essential for improving the performance of machine learning models, especially in healthcare applications. By leveraging a combination of physiological insights, domain knowledge, and advanced augmentation methods like GANs, researchers can create robust models capable of generalising well across diverse datasets.

2.9 Projection of ECG data

Projection of ECGs rather than the ECG signals itself may be more useful for the classification. 3D VCGs is a technique that tries to represent electrical changes in human heart, its magnitude and orientation with time [137]. Those changes can then be plotted in 3D space as a snapshot of electrical activity of the heart. The three leads in 3D space are represented by right-left axis (X), head-to-feet axis (Y) and front-back axis (Z), with the following equations[138].

$$X = -(-0.172\ V1 - 0.074\ V2 + 0.122\ V3 + 0.231\ V4 + 0.239\ V5 + 0.194\ V6 + 0.156\ DI - 0.010\ DII) \tag{2.10}$$

$$Y = (0.057\ V1 - 0.019\ V2 - 0.106\ V3 - 0.022\ V4 + 0.041\ V5 + 0.048\ V6 - 0.227\ DI + 0.887\ DII) \tag{2.11}$$

$$Z = -(-0.229\ V1 - 0.310\ V2 - 0.246\ V3 - 0.063\ V4 + 0.055\ V5 + 0.108\ V6 + 0.022\ DI + 0.102\ DII) \tag{2.12}$$

We tried to use those 3D representations on training rather than signal itself to see if it makes any difference on the results. In Figure 2.8, we plotted all classes by using raw signal data and aligned signal data of a random patient.

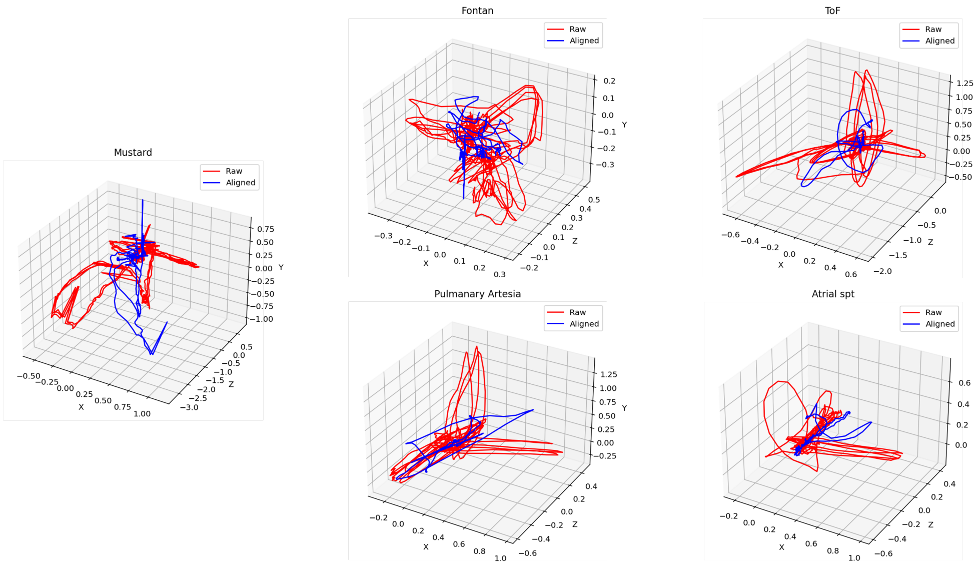


Figure 2.8: Example vectorcardiograms.

2.10 Visualisation of high-dimensional spaces

t-distributed Stochastic Neighbor Embedding (t-SNE) simplifies more complex non-linear or multidimensional data into 2 or 3 dimensions that are easier to visualise and interpret [139]. Such techniques can be considered as projections of the data, similar to how photos encode information about a three-dimensional world into 2 dimensional visual displays. t-SNE transforms the similarity, assigned as a distance, as a probability distribution over all the potential patient pairs. Thus, if two patients' clinical characteristics are close in high dimensional space, i.e. in the initial complex dataset, then they are going to be assigned with a high probability of being neighbour points (i.e. located in close proximity). The aim of the low dimensional embedding/projections is to approximate the data distributions as closely as possible with the original, high-dimensional dataset. Mathematically, the similarity between patients x_i and x_j is the conditional probability $p_{i/j}$ defined in Equation 2.13 as follows

$$p_{i/j} = \frac{\exp\left(-\left|x_i - x_j\right|^2 / 2\sigma_i^2\right)}{\Sigma_{k\neq i} \exp\left(-\left|x_i - x_k\right|^2 / 2\sigma_i^2\right)} \tag{2.13}$$

An important parameter in t-SNE is the perplexity, that relates to the variance of the Gaussian distribution, and it can be interpreted as a smoothness factor of the number of neighbour points. Evidently, these probabilities are generated by quantifying and transforming norms (in this case we adopt the Euclidean norm/distance) between data points reflecting the probabilistic similarity between them. It is evident that the changes in perplexity will result in changes in the t-SNE visualisation. Since these visualisations are interactive the user will adjust them manually. Figure 2.9 shows t-SNE plots for five different perplexity values.

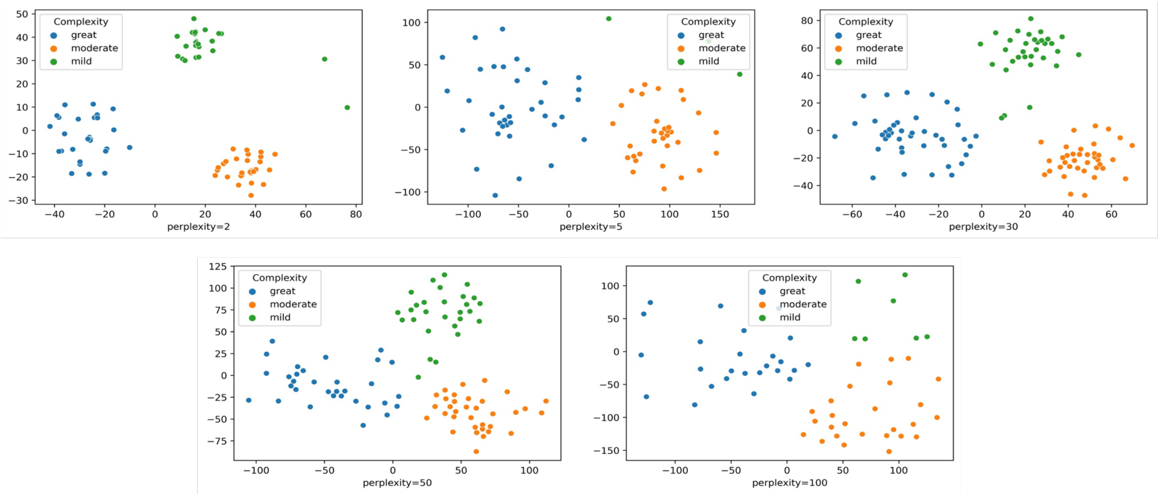


Figure 2.9: Different perplexity values (from 2 to 100).

2.11 Conclusions

In this chapter, the foundational concepts and methodologies that underpin our research are explored, with a focus on Riemannian geometry, data augmentation techniques and the calibration of clinical models. Riemannian geometry provides a powerful mathematical framework for analysing complex data structures, particularly in the context of medical signal analysis. It facilitates more accurate representation of the underlying anatomical and physiological variations in patients with CHD. Furthermore, this thesis places particular emphasis on the significance of data augmentation techniques, which have been demonstrated to enhance the diversity and robustness of training datasets, consequently leading to improvements in model performance and generalisability. Finally, the critical need for the calibration of clinical models is addressed, with the objective being to ensure that their predictions align with real-world outcomes, thereby enhancing their utility in clinical decisions. The integration of these elements is instrumental in the development of more effective predictive models that have the potential to significantly advance the diagnosis and management of CHD.

Chapter 3

Background on Congenital Heart Disease

3.1 Congenital Heart Disease

Cardiovascular disease is a term for any disease that affects the heart or blood vessels. Heart disease is a type of cardiovascular disease and is a general term for various conditions that affect the structure and function of the heart. All heart diseases can be classified as cardiovascular diseases, but not all cardiovascular diseases can be classified as heart disease. Congenital Heart Disease (CHD) includes various heart defects present from birth, affecting the heart's structure and function [140, 141]. Symptoms can range from rapid heartbeat and breathing to difficulty feeding in babies. Causes are often unknown but can include genetic conditions, maternal infections, certain medications or poorly controlled diabetes during pregnancy [142]. Diagnosis is typically made before or shortly after birth, with treatments varying from monitoring to surgery [143]. The following represent some of the most common types of CHD, each affecting cardiovascular function in distinct ways:

- **Septal defects:** Conditions such as Atrial Septal Defect (ASD) and Ventricular Septal Defect (VSD) are included in this category. These conditions are characterised by the presence of abnormal openings between the heart's chambers, which can potentially result in inefficient blood circulation and oxygenation [140].
- **Coarctation of the aorta:** This condition is characterised by the narrowing of the body's main artery, which has the potential to restrict blood flow and thereby cause hypertension

and heart strain [144].

- **Pulmonary valve stenosis:** This condition is characterised by the narrowing of the pulmonary valve, which consequently results in impaired blood flow from the heart to the lungs, resulting in increased cardiac workload [142].

- **Tetralogy of Fallot (ToF):** This complex condition comprises four abnormalities: a ventricular septal defect, pulmonary stenosis, an overriding aorta, and right ventricular hypertrophy. The consequence of these abnormalities is cyanosis and reduced oxygenation of blood [143].

Figure 3.1 illustrates the anatomical variations associated with different types of CHD, including ASD, VSD and ToF. The top image represents a normal heart, while the lower images illustrate the structural abnormalities characteristic of each defect, highlighting the impact of CHD on cardiac anatomy and function.

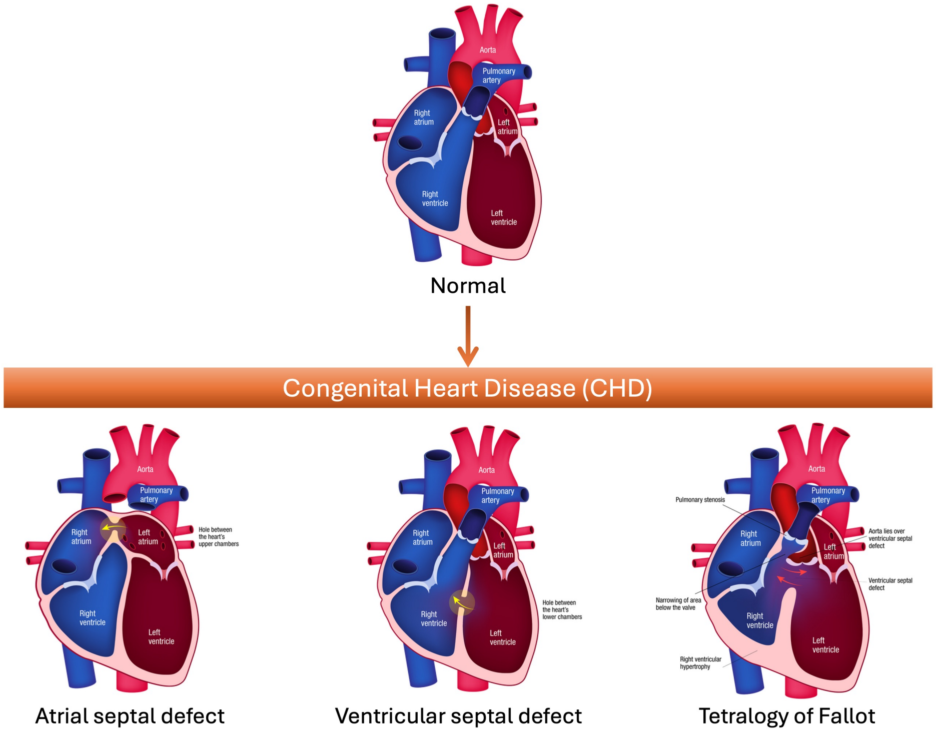


Figure 3.1: Overview of anatomical variations of different congenital heart diseases.

3.1.1 Prevalence in the population

CHD is one of the most common types of birth defects. Globally, it is diagnosed in about 1 in 110 births, which translates to approximately 1.2 million babies each year [1]. In the UK, it affects approximately 1 in every 100 babies born, which is estimated to result in approximately 4,600 cases annually [2]. The prevalence of CHD emphasises the importance of early detection and treatment in order to improve the outcomes for those affected.

3.1.2 Adults congenital heart disease

CHD in adults, also known as Adult Congenital Heart Disease (ACHD), refers to the presence of heart defects from birth that persist into adulthood. Many people with CHD are now living well into adulthood thanks to advances in medical and surgical treatments [145]. Some adults may have been diagnosed and treated as children, while others might not discover their condition until later in life. The monitoring and management of ACHD is essential for ensuring that these individuals maintain a good quality of life. Regular follow-up with specialists is an effective method for preventing and managing issues at an early stage [146]. For example, in the case of women with ACHD, proper management and counselling are vital for ensuring safe pregnancies and reducing risks for both the mother and the child [147].

3.1.3 Predictive models

Predictive models are essential tools in many fields because they allow us to anticipate future outcomes and make informed decisions based on data. These models can predict how different patients might respond to various treatments, enabling doctors to tailor treatment plans to individual needs [148]. Predictive models can also assess the risk of complications or adverse outcomes in patients with CHD, helping healthcare providers prioritise high-risk patients for more intensive monitoring and care [149, 150]. For adults with CHD, predictive models can monitor disease progression and predict potential complications, guiding long-term management and follow-up care.

Predictive models have been widely used in clinical settings to improve patient care and outcomes, such as:

- **Risk Stratification:** Predictive models assist in the stratification of patients based on their risk of developing specific conditions or experiencing adverse events. Such complications may include heart failure, arrhythmias, stroke, or pulmonary hypertension. In traditional practice, physicians relied on standard clinical risk scores. However, recent technological advances have led to the development of machine learning models that utilise large-scale datasets, including electronic health records, imaging data and laboratory test results, with the aim of identifying patterns that may indicate elevated risk [148]. These models enable early detection of potential health issues, thereby enabling healthcare professionals to implement preventative measures that improve patient outcomes. Advanced computational techniques, such as deep learning, have demonstrated superior accuracy in risk assessment when compared with traditional methods, helping prioritise patients who may require closer monitoring or early interventions [4, 148].

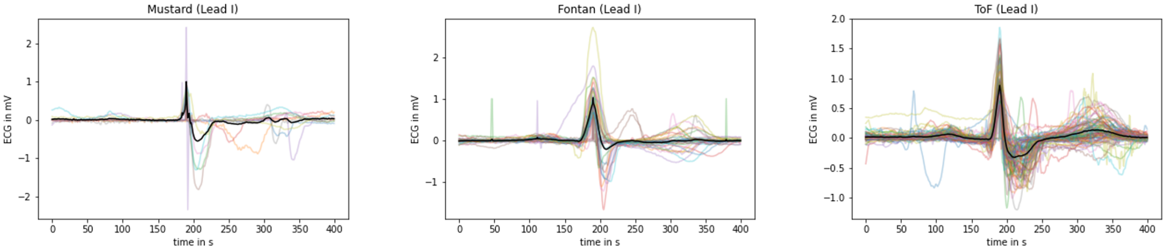
- **Diagnosis and Prognosis:** Predictive models are also employed to estimate the risk of existing disease (diagnostic) and future clinical outcomes (prognostic). These models utilise a combination of demographic factors, biomarkers and imaging data to provide a clearer picture of disease severity [5, 151]. For example, for prenatal CHD detection, the utilisation of AI-driven diagnostic models have improved early identification of cardiac anomalies through the use of fetal echocardiography [152, 153]. Additionally, prognostic models facilitate the prediction of long-term survival, likelihood of post-surgical complications, and potential deterioration in cardiac function [154–157]. This information enables healthcare providers to plan interventions more effectively, offering timely treatments that result in a reduction in morbidity and mortality associated with CHD.

- **Personalised Treatment Plans:** Predictive models facilitate the creation of personalised treatment plans by predicting how patients may respond to different treatments. By analysing patient-specific variables, including cardiac anatomy, comorbidities, and prior responses to treatment, predictive models can assist clinicians in selecting the most effective approach, thus enabling them to personalise treatment plans based on individual patient characteristics. For example, machine learning-based models can predict the success rates of different types of surgical procedures or medications, and post-operative care allowing physicians to tailor their interventions accordingly [3, 158]. In complex cases, such as single-ventricle defects, predictive models help determine the optimal timing for surgeries, improving survival rates and overall patient well-being [159].

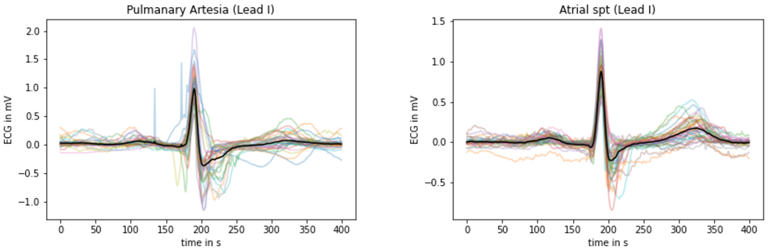
- **Resource Allocation:** In healthcare systems, predictive models facilitate the optimal allocation of resources by identifying patients who are most likely to benefit from specific interventions or require more intensive care [160]. Hospitals can utilise AI-driven models to predict patient admission rates, surgical demands, and intensive care unit requirements. By analysing historical patient data, seasonal trends, and demographic shifts, hospitals can proactively allocate resources in order to prevent shortages and minimise operational bottlenecks [161]. The enhancement of workflow management, the prevention of resource shortages, and the reduction of healthcare costs are all key benefits of this approach. In CHD treatment, the utilisation of predictive models has the potential to assist in scheduling of follow-ups, and optimising diagnostic procedures, ultimately minimising delays in care delivery and improving overall patient satisfaction [148].

3.2 Electrocardiograms

Electrocardiogram (ECG) is a valuable tool in the analysis of congenital heart defects. In particular, it can reveal heart blocks and congenital heart defects that may be missed clinically by providing diagnostic and prognostic information. The more serious the congenital heart defect, the more likely it is to find abnormal changes on the ECG [162]. ASD represents 30% of congenital heart diseases detected in adults, while ToF represents approximately 10% [163]. It is important to take into account the main anatomical aspects of various CHD conditions like ASD and ToF, which may reveal their characteristic ECG patterns individually (see Figure 3.2). Studies in the literature applying Deep Learning (DL) techniques on ECG signals are generally focused on heartbeat detection, identification (annotation) and arrhythmia (irregular heartbeat) classification [164].



(a) Mustard (b) Fontan (c) Tetralogy of Fallot



(d) Pulmonary Atresia (e) Atrial Septal Defect

Figure 3.2: Example beats for each class (Lead I).

A 12-lead ECG is a standard diagnostic tool used to assess cardiac electrical activity. It consists of 10 electrodes that are placed on the body with the intention of generating 12 different views of the heart, as seen in Figure 3.3. The positioning of these electrodes is outlined as follows:

Limb Leads:

- **RA (Right Arm)** – Anywhere between the right shoulder and right elbow.
- **LA (Left Arm)** – Anywhere between the left shoulder and left elbow.
- **RL (Right Leg)** – Anywhere below the right torso and above the right ankle.
- **LL (Left Leg)** – Anywhere below the left torso and above the left ankle.

Chest (Precordial) Leads:

- **V1** – 4th intercostal space, right sternal border.
- **V2** – 4th intercostal space, left sternal border.
- **V3** – Midway between V2 and V4.

- **V4** – 5th intercostal space, mid-clavicular line.
- **V5** – 5th intercostal space, anterior axillary line (same level as V4).
- **V6** – 5th intercostal space, mid-axillary line (same level as V4).

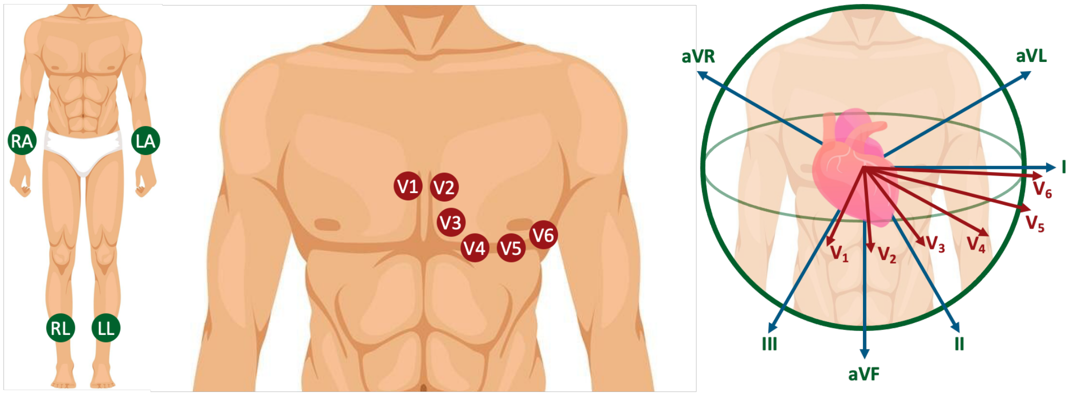


Figure 3.3: Placement of leads in a 12-lead ECG, 10 physical electrode positions.

ECG is a diagnostic procedure that provides a graphical representation of the electrical activity of the heart corresponding to different phases of the cardiac cycle. There are a number of characteristic wave patterns within the ECG for a single heart cycle, as seen in Figure 3.4:

- **P Wave**: Represents atrial depolarisation, initiating atrial contraction. It is small and rounded, reflecting the movement of electrical impulses through the atria.
- **Q Wave**: Represents the initial phase of ventricular depolarisation, corresponding to early ventricular depolarisation. It is the first downward deflection in the QRS complex.
- **R Wave**: Represents the primary phase of ventricular depolarisation as the electrical signal moves through the ventricles. It is the largest upward spike in the ECG.
- **S Wave**: Represents the completion of ventricular depolarisation process. It is a downward deflection following the R wave.
- **T Wave**: Represents ventricular repolarization, which marks the recovery phase before the next cycle.

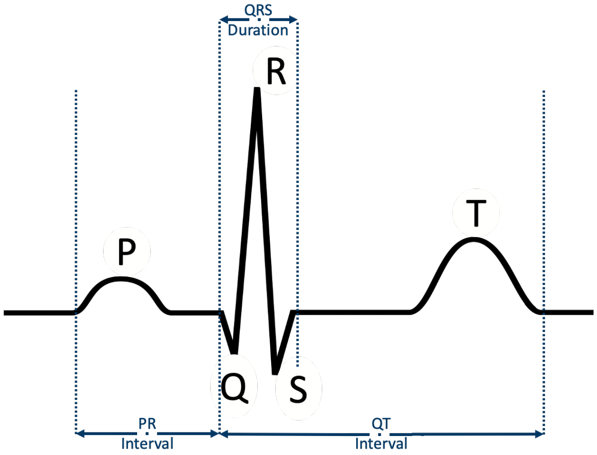


Figure 3.4: Key components of an ECG trace.

Each ECG waveform corresponds to a particular mechanical event within the heart, thereby assisting clinicians in the diagnosis of abnormalities, including but not limited to heart blocks, arrhythmias, and congenital heart defects. ECG trace also consists of several key components as seen in Figure 3.4, which reflect the heart's electrical activity and provide crucial insights into cardiac function of the patient.

- **PR Interval**: Measures the time from the start of atrial depolarisation (P wave) to the onset of ventricular depolarisation (QRS complex). It is widely acknowledged that the PR interval is of greater clinical significance, as it quantifies the period commencing from the onset of atrial depolarisation (the P wave) and culminating in the initiation of the QRS complex.

- **QT Interval**: Measures the total time taken for ventricular depolarisation and repolarization, extending from the start of the QRS complex to the end of the T wave. Prolongation of this interval may serve as a potential indicator of an increased risk of developing arrhythmias.

- **QRS Complex**: Corresponds to ventricular depolarisation, which in turn marks ventricular contraction (systole). The QRS duration and its morphology is critical for the detection of abnormal electrical activity.

3.3 Clinical Models

Since, in-hospital mortality and length of stay are key patient outcomes that reflect quality of healthcare, several risk-prediction models have been developed. Most common clinical models use Logistic Regression (LR), due to its interpretability and effectiveness in binary classification tasks [165, 166]. Recently, DL has been proposed to exploit large population data and develop more robust and personalised clinical decision support systems [167]. There are only a few studies on CHD (for adults or children) that focus on ECG analysis and classification using machine learning techniques due to the lack of diverse and large labelled data [149, 150, 168, 169]. Most previous works were implemented as a binary problem to detect whether a patient has the condition or not with aim to reduce mortality and morbidity [169, 170].

Vullings et al. [171] proposed a Deep Neural Network (DNN) to classify 3-dimensional fetal Vectorcardiogram (VCG) as either healthy or CHD (binary classification). Their architecture incorporated multiple convolutional layers optimised for feature extraction and classification, each with residual connections and leaky rectified linear units as activation functions. ECG measurements were performed at six different medical centres in the Netherlands. The data collected consisted of various types of CHD, including septal defects, ToF. However, a binary classification was utilised in this study: either CHD or not. By utilising 386 measurements (266 from the healthy, 120 from the CHD) of 30 mins ECG, their model improved CHD detection rates at an accuracy rate of 0.76 compared with traditional ultrasonic examination. It is important to note that the study did not provide any additional metrics beyond accuracy, such as sensitivity, specificity, or F1 score which are important for evaluating model performance comprehensively.

Du et al. [172] proposed a Residual Network to classify child ECGs as either CHD or not (binary classification). The model utilised in this study comprised 51 layers, aiding the effective extraction of temporal and spatial characteristics from the ECG signal with skip connections. These skip connections were implemented to help mitigate the vanishing gradient problem and improve feature propagation across layers. By utilising 68,969 child ECGs (58,624 from the healthy, 10,345 from the CHD) of 10 seconds records on Guangzhou Women and Children's Medical Center, their model improved CHD detection rates. In the independent test set, the accuracy of the proposed model was 0.92, the sensitivity was 0.74, and the specificity was 0.94. The performance exceeds other individual CHD screening indicators, which shows that ECG can be considered as a great value for CHD identification and can be included in the screening process.

Kim et al. [173] proposed a Long Short-Term Memory (LSTM) to classify heart disease with ECGs. The dataset was obtained from UCR Time Series Archive and was pre-processed in two steps: extraction of heartbeats and normalisation of heartbeat length. Then 5,000 heartbeats were randomly selected for training. The dataset utilised in the study encompassed a range of cardiac conditions, with the primary focus on the classification of general heart disease rather than CHD. There was no info about the labels, but the accuracy of classification was 0.97. Also, they did not share any metric results other than accuracy.

Liang et al. [174] proposed a Residual Network (ResNet) to classify child ECGs as either CHD or not (binary classification). Since most of the ECGs in the dataset consist of only 9 channels (I, II, III, AVR, AVL, AVF, V1, V3, V5), they reduced the small number of 12 channels ECGs to 9 channels, leaving out 3 channels (V2, V4, V6). They implemented two different training approaches with using whole ECG and cardiac cycle segment to assess which one might have better performance. By utilising 72,626 child ECGs (59,042 from the healthy, 13,584 from the CHD), they used the same model structure and the same splits for training. The performance of the model was greatly improved when cardiac cycle segmentation was used, in particular the sensitivity increased from 0.74 to 0.80. The accuracy of the proposed model was 0.94, the sensitivity was 0.80, and the specificity was 0.95.

Yuan et al. [175] proposed three different models to compare the classification effect and accuracy rate of Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Multimodal Neural Network (MNN) in ECG classification of CHD patients. The dataset was obtained from MIT-BIH (Massachusetts Institute of Technology-Beth Israel Hospital) as single-lead ECGs of 127 patients with CHD. ECGs were classified as normal or bundle branch block beat (NB), supraventricular abnormal rhythm (SA), abnormal ventricular beat (AV), fusion beat (FB), and uncategorised beats (UC). The results showed that the classification results of the MNN (0.98 accuracy) were better than the single CNN and the RNN. Also, they did not share any metric results other than accuracy.

Diller et al. [176] proposed a DL architecture to categorise diagnostic group, disease complexity, and New York Heart Association (NYHA) class in a large cohort of patients with adult CHD. Rather than using ECG signals, they only included some ECG parameters like resting heart rate (b.p.m.), QRS duration (ms), and QTc duration (ms) along with laboratory and exercise parameters. By utilising 13,649 ECGs and 44,421 medical reports of 10,019 adult patients (age 36.3 ± 17.3 years), their model achieved an accuracy of 0.91, 0.97, and 0.90 on categorised diagnosis, disease complexity, and NYHA class.

So far only a binary classification using ECG data has been performed on CHD patients, child or adult. Making this distinction (either CHD or not) is an important step, but there is not yet a deep learning study on the classification of disease types. In addition, most studies have been done entirely on ECG data, only one study has combined different data types as seen in Table 3.1. Their work was based on derived ECG parameters instead of raw ECG signals. Beyond the classification type, other potential limitations include the reliance on a single data modality (e.g. ECG alone) which may restrict predictive performance in comparison to multimodal approaches. Furthermore, the size and diversity of the dataset are of crucial importance in ensuring the generalisability of the model, since imbalanced datasets may lead to biased predictions that favour more common CHD presentations.

3.4 Research gaps & questions

There are a number of research gaps and questions within the field of predictive models for CHD, including issues related to data quality and imbalanced datasets, integration with existing clinical systems, and the limited utilisation of multi-modal data. The following will address some of the key issues in this regard.

- **Data Quality and Availability:** Predictive models rely considerably on high-quality and diverse datasets for the purpose of accurate risk stratification and outcome prediction. However, many of these datasets exhibit inconsistencies, missing values, and class imbalances for rare conditions as they may suffer from limited sample sizes. This is particularly

Table 3.1: Summary of ECG based deep learning approaches for CHD classification and prognosis.

Author	Cohort	Dataset	Methods	Results	Limitations
Vullings	Child	386 ECGs: - 266 healthy - 120 CHD 30 mins ECGs - 287 for training - 99 for testing	Deep neural network 3-dimensional fetal vectorcardiogram binary classification (CHD or not)	76% accuracy no any other metrics	* 1 ECG per patient * Data augmentation was used (probably mixed ??)
Due et al.	Child	68,969 ECGs - 58,624 healthy - 10,345 CHD 10 seconds ECGs - 58,969 for training - 10,000 for testing	Residual Network with 51 layers binary classification (CHD or not)	92% accuracy 74% sensitivity	* 1 ECG per patient * Nothing about data augmentation
Yuan et al.	Child	MIT-BIH database single-lead ECGs - 82,600 for training - 21,000 for testing	compared 3 models: -Convolutional Neural Network (CNN) -Recurrent Neural Network (RNN) -Multi-modal Neural Network (MNN) classification classes: -normal or bundle branch block beat (NB) -supraventricular abnormal rhythm (SA) -abnormal ventricular beat (AV) -fusion beat (FB) -uncategorised beats (UC)	MNN performs well -98,5% accuracy CNN RNN -98% accuracy	* 127 patients with CHD * Nothing about multiple entries or data augmentation
Du et al.	Child	72,626 ECGs - 59,042 healthy - 13,584 CHD 72,626 patients - 62,626 for training - 10,000 for testing	9 ECG channels (I, II, III, AVR, AVL, AVF, V1, V3, V5) Residual of Residual (RoR) network (RNN with two layers of shortcut) train with: -whole ECG -cardiac cycle segment	better performance using cardiac cycle segments 94% accuracy 80% sensitivity	* 1 ECG per patient * Nothing about data augmentation
Kim et al.	Adult	UCR Time Series Archive 5,000 ECGs	LSTM transformed ECG into the sequences (equally divided 20 bins) no info about labels	97% accuracy no any other metrics	* Nothing about data split * Nothing about patient numbers * Nothing about multiple entries or data augmentation
Diller et al.	Adult	13,649 ECGs 44,421 medical reports 10,019 patients (age 36.3 ± 17.3 years)	DL-models to categorise: - diagnostic group - disease complexity - New York Heart Association (NYHA) class combined data from: - demographic - laboratory - ECG parameters - exercise data	accuracy of -91.1% -97.0% -90.6%,	* 1.36 ECG per patient * Single patients could have multiple entries (multiple visits to the clinic)
Mayourian et al.	Child Adult	Boston Children's Hospital database 12-lead ECGs - 112,804 for training - 112,575 for testing	Convolutional Neural Network (CNN) 5-year mortality prediction	0.79% AUROC 0.17% AUPRC	* 2.8 ECG per patient * group stratified data split * Single patients could have multiple entries (multiple visits to the clinic)

evident in the case of CHD conditions, which are frequently under-represented and lead to biased models that favour more prevalent conditions [148]. Addressing these issues requires more robust data pre-processing frameworks and improved data augmentation approaches, with the objective of achieving a balanced representation of rare cases. As the development of accurate predictive models hinges on the availability of high-quality and diverse datasets, there is also a need for more standardised and comprehensive datasets incorporating a range of different data types.

- **Limited Utilisation of Multi-modal Data:** Predictive models often rely on a single data source, such as electronic health records or clinical findings. However, incorporating different data modalities, including ECGs, electronic health records, laboratory test results, demographic factors and imaging data, has the potential to enhance the performance of predictive models and facilitate more precise patient risk assessments [177]. Therefore, it is worth focusing on the development of architectures that incorporate different data modalities while preserving clinical interpretability and relevance.

- **Model Generalisability:** It is evident that a considerable number of predictive models perform well on specific datasets but fail to generalise to other datasets/populations [178]. In the context of CHD, ensuring generalisability is crucial for clinical applicability and real-world effectiveness [179]. Such predictive models should demonstrate the ability to perform well on unseen data beyond the training dataset. Consequently, there is a clear need that aim to develop models which are both robust and applicable across a range of demographic groups and healthcare settings. In order to address these challenges, the employment of standardised cross-validation techniques is essential, with models undergoing testing on independent datasets prior to their clinical deployment.

- **Clinical Utility:** Although predictive models demonstrate potential and good performance, their integration into routine clinical workflows continues to present a significant challenge [3]. Many healthcare institutions continue to use legacy systems that present significant challenges in terms of data retrieval and model development, which consequently creates barriers to implementation of such models [180]. So, future research should explore strategies for seamless data integration, enhanced data preprocessing frameworks and clinician-friendly interfaces to ensure predictive models enhance decision-making without compromising standard care protocols [3, 181].

- **Ethical and Privacy Concerns:** As predictive models become more integrated into healthcare systems, the utilisation of these models raises a number of ethical and privacy concerns particularly in relation to the handling of sensitive health data [182, 183]. One major challenge is the security of patient data, as predictive models often rely on large-scale electronic health records, imaging data, and demographic information [183]. So, there is a need for further research to develop frameworks for the ethical use of such models and to guarantee the privacy and security of patient data.

Addressing these questions has the potential to yield more precise, generalisable and ethical predictive models, which can significantly enhance the diagnosis, treatment, and management of patients with CHD.

3.5 Conclusions

In summary, the existing literature and clinical practice surrounding predictive models for CHD face several significant challenges that this thesis aims to address. One of the primary issues is the quality and availability of data, as many datasets suffer from inconsistencies, missing values and class imbalances, particularly for rare CHD conditions. This limitation not only prevents accurate risk stratification and outcome prediction but also leads to biased models that favour more prevalent conditions. In order to overcome these challenges, there is a need for robust data pre-processing frameworks and improved data augmentation strategies that ensure a balanced representation of rare cases, alongside the development of standardised datasets.

Additionally, the limited utilisation of multi-modal data presents another critical gap. Current predictive models often rely on single data sources, which restricts their performance and the precision of patient risk assessments. By incorporating diverse data modalities such as ECGs and clinical letters this thesis aims to enhance model performance while maintaining its clinical relevance.

By addressing the research gaps and questions identified, this thesis aims to contribute to the development of more precise, generalisable, and well-calibrated predictive models, ultimately enhancing the diagnosis, treatment and management of patients with congenital heart disease.

Chapter 4

ECG Extraction and Representation in Congenital Heart Disease

4.1 Abstract

12-lead Electrocardiograms (ECGs) form an essential part of the late follow-up of adults with Congenital Heart Disease (CHD). Such ECGs are most frequently reviewed by clinicians in paper or PDF formats. These visual representations of the original vector data do not easily lend themselves to be directly analysed with the increasingly powerful Machine Learning (ML) algorithms that hold promise in risk prediction and early prevention of adverse events. In this work, we set out to recreate the original digital signals from ECG PDF documents by a series of data processing steps, validate accuracy of the process, and demonstrate its potential utility in research. Using 4153 ECG PDF documents from 436 CHD patients, we created a "pipeline" to successfully digitise the visually represented ECG vector datasets. We then proceed with the validation of the digitised ECG dataset using several features that are also calculated by the vendor, such as QRS duration, PR interval and ventricular rate. We confirmed a strong correlation with the vendor measured ECG parameters including PR interval ($R = 0.952, p < 0.001$), QRS duration ($R = 0.969, p < 0.001$) and ventricular rate ($R = 0.980, p < 0.001$). Further, using Support Vector Machine (SVM), a well-established ML model we demonstrate the ability of the digitised ECG dataset to accurately predict anatomic diagnosis in CHD. Digitisation of PDF

formatted ECG signal data can be accomplished with good accuracy and can be used in clinical research in CHD.

4.2 Introduction

With advances in medical sciences, more than 90% of patients with even the most complex CHD survive into adult life [184]. With this transformational success of medical and surgical advances, there has emerged a growing burden of morbidity and mortality as individuals enter later into adult life. Increasing attention is being drawn to the ability to predict which individuals are likely to deteriorate, and those at risk of mortality [185–189]. ECGs are a fundamental aspect of such assessments [176, 190–192]. They have been demonstrated to carry important diagnostic and prognostic information that correlate with the outcomes [176, 193–196]. For example, in CHD, clinicians use derived ECG measures such as QRS duration, QRS fragmentation, QRS axis and R-wave height along with PR intervals as indicators of underlying anatomy and prior surgery, disease state and for mortality prognosis in some specific settings.

More recently ML algorithms promise to extract relevant ECG features and exceed the performance of more traditional and more manual approaches in disease diagnosis and prognostication [172, 197]. In a recent review, Helman et al. [198] highlighted that the development of ML algorithms to process 12-lead ECGs for CHD is a promising direction, which currently remains unexplored due to the lack of diverse and large datasets.

We have identified the opportunity to develop digital twins of CHD patients by extracting information over a decade from clinical letters, which involved diagnoses, diagnostic complexity, interventions, arrhythmia, medications, and demographic data [73]. We believe this information can also be linked to the patient's 12-lead ECG dataset when in digitised format. This provides the opportunity to develop robust ML models for diagnoses and risk prediction and explore their capabilities.

In clinical practice, 12-lead ECGs are often interpreted from scanned in or printed PDF documents. Valuable information can be derived from such visual manual assessment, such as cardiac rhythm and rate, hemodynamic characteristics like pulmonary hypertension, and prognostic information relating to sudden cardiac death risk. ML on the actual raw signal data enhance further diagnostic and prognostic capacity. Due to proprietary software and analysis of raw ECG signal data, such "raw' data is not readily available for de-novo or empiric data analysis or use in further research. Being able to accurately reproduce the digitised ECG signals from PDF documents can potentially enable signal data to be used in developing ML models for risk stratification and diagnosis.

For ECG digitisation, all current approaches focus on image processing (i.e. on scanned images or PDF documents converted into a desired image format) rather than the conversion to the vector data represented by the PDF document [199, 200], and this in particular has not previously been evaluated in CHD. By focusing on PDF documents, we recreate ECG PDF documents created by a specific ECG machine via reverse digitisation, instead of converting them into images and applying image processing techniques.

We propose a stepwise approach as a pipeline that assemble several steps that are combined to extract 12-lead ECG digital signals from ECG PDF documents. We adopt open-source generic Python libraries to extract information about the text and the geometric objects present in ECG PDF documents, e.g., PyMuPDF [16]. We then validate the digitised dataset using derived ECG features such as QRS duration, PR interval and heart rate, and further demonstrate its potential research utility by using it to predict anatomic diagnosis associated with CHD. This is an initial step towards mortality prediction, which requires significantly larger numbers of subjects to achieve. Summarising, the contributions of this chapter are as follows:

- **Development of an ECG Digitisation Pipeline:** A data extraction pipeline was established for the systematic conversion of PDF-formatted 12-lead ECG data into digital vector format, thereby facilitating computational analysis. This approach has been shown to overcome the limitations of conventional image-based approaches and to enable computational analysis.

- **Validation of Digitised ECG Data:** The accuracy and reliability of the digitisation process were evaluated through a comparison with vendor-measured parameters. Strong correlations ($R > 0.94$, $P < 0.05$) were observed for key features such as PR interval, QRS duration, and ventricular rate, demonstrating the robustness of the proposed pipeline.

- **Facilitation of Risk Prediction Models:** By enabling the transformation of ECG data into analysable digital signals without any artefacts, this approach opens new opportunities for the development of risk prediction models. Utilising an SVM model, this chapter demonstrated that digitised ECG data can effectively predict anatomic diagnosis in Adult Congenital Heart Disease (ACHD) patients, thus highlighting its potential in clinical research and risk prediction.

4.3 Methods

4.3.1 Linkage of CHD EHR

Study approval was obtained by the Institutional Governance Division of the NHS Golden Jubilee National Hospital. Demographic information including age, sex, anatomic diagnoses, and prior surgical intervention were extracted from clinical letters as previously described in [73]. The most common condition was Tetralogy of Fallot (ToF), in 197 patients, followed by Pulmonary Atresia (PA), in 96 patients. The top 15 anatomic diagnoses are summarised in Figure 4.1.

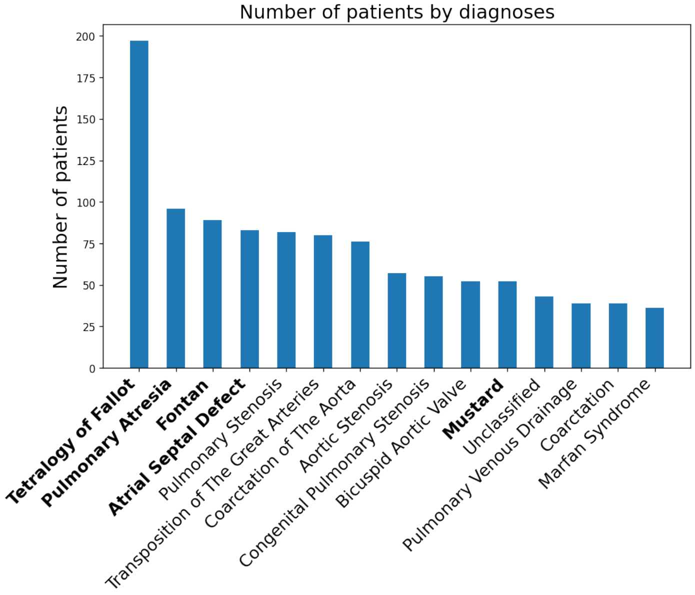


Figure 4.1: Top 15 anatomic diagnoses for 1409 patients.

Among these conditions, we have selected the 5 most common conditions (shown in bold in Figure 4.1) for diagnosis prediction including:

1. Tetralogy of Fallot (ToF) (Diagnosis List)
2. Pulmonary Atresia (PA) (Diagnosis List)
3. Fontan (Intervention & Diagnosis List)
4. Atrial Septal Defect (ASD) (Diagnosis List)
5. Mustard (Intervention List) including other atrial switch procedures.

Information from clinical letters was linked to ECG PDFs based on the patient chart number. Patients were excluded if they did not have available ECGs or were in a permanent dysrhythmia, such as atrial fibrillation or flutter, or were atrioventricular paced. This led to 436 patients being included in our investigation. Of these 436 patients, 173 patients had ToF, 77 patients had Atrial Septal Defect (ASD), 73 patients had PA, 66 patients had Fontan and 47 patients had Mustard. Overall, 4153 ECG PDF documents were extracted from these 436 patients with diagnoses as outlined.

4.3.2 ECG data preprocessing

12-lead ECG data were extracted from ECG PDF documents obtained via the Marquette™ 12SL by GE Healthcare analysis program [201]. For resting ECGs, the analog voltage potential is digitised into 4.88-μV units at a rate of 4kHZ. The software down-samples the signal to 500 samples per second and represents a value every 0.05mm on a chart. The ECG signal has been pre-processed to remove noise and QRS template matching was employed to extract ECG features and export ECG waveforms to PDF documents in a vectorised format. Each PDF document contains 12 leads in a specific order (Lead I, II, III, aVF, aVR, aVL, V1, V2, V3, V4, V5, V6), and provides only 2.5-second strip for each lead as in Figure 4.2.

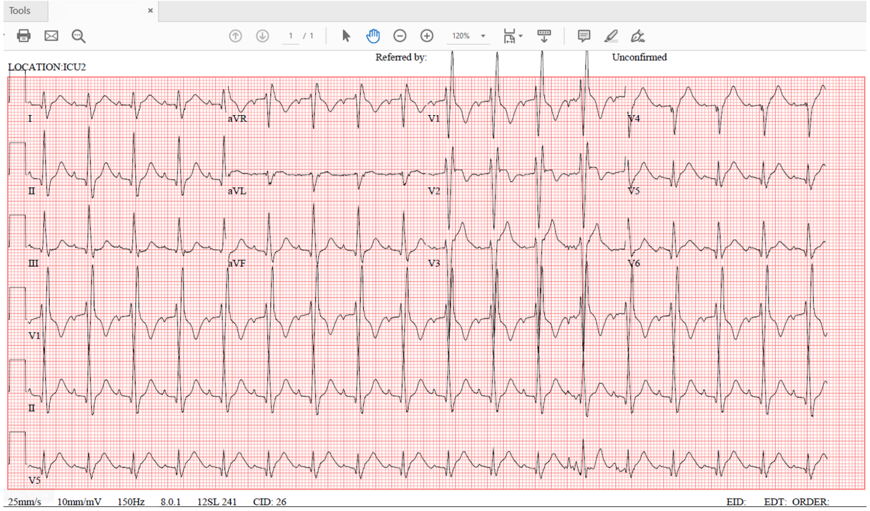


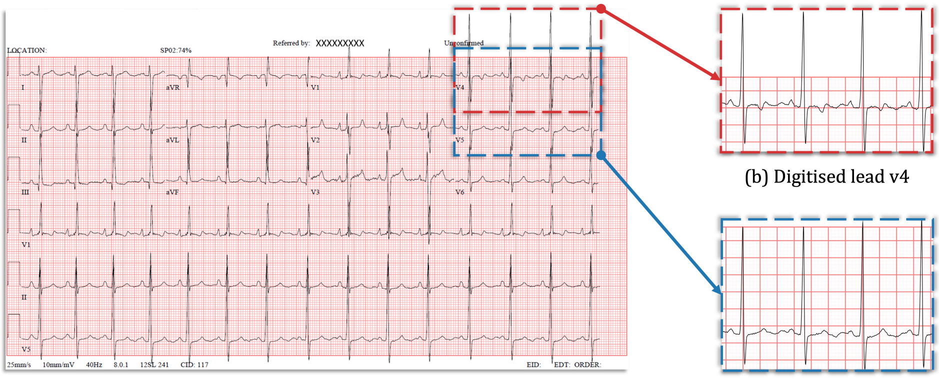
Figure 4.2: Example ECG PDF document.

4.3.3 Digitisation of 12-lead ECGs

We developed a new algorithm using vector drawings on the ECG PDF documents to digitise ECGs without user intervention. PyMuPDF [202] library was used to extract information about the graphics points present in the PDF documents. The algorithm follows the steps below and it is described in detailed at Algorithm 1.

1. Locates the coordinates of each lead (position on the page).
2. Obtains the scale indicator (to indicate size and timing intervals) for later use in the digitisation step.
3. Locates the lead specific graphics points to extract each lead signal separately.
4. Using the graphics points on the PDF document, our algorithm extracts line points for each lead and digitises them individually.
5. After the digitisation step, it scales the extracted signal in mV using the scale indicator obtained in the second step.
6. As the final step, it matches extracted ECG signals with the determined output column of the original data of the 436 patients.

An example of the ECG document with two of the extracted leads is shown in Figure 4.3.



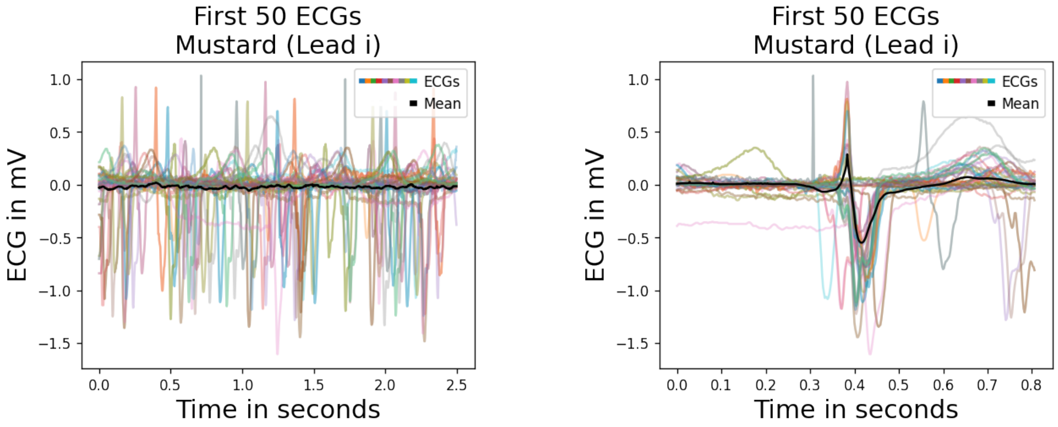
(b) Digitised lead v4

(a) Example ECG document with 12-leads

(c) Digitised lead v5

Figure 4.3: Digitisation and linkage of ECG PDF documents that use vectorised graphic format to store ECG waveforms.

To facilitate analysis of ECG data with ML algorithms, we standardised them by aligning the ECGs so that the peaks of the R-waves intersected and QRS complexes were digitally synchronised across all leads. An example of the average signal of the aligned ECGs for Mustard is shown in Figure 4.4 and its characteristic of the underline abnormality.



(a) ECGs (Mustard, Lead I)

(b) Aligned ECGs (Mustard, Lead I)

Figure 4.4: ECG alignment using first 50 ECGs on lead I of patients with a Mustard procedure.

4.3.4 Validation of the digitisation algorithm

For validation of our Algorithm 1, we compared vendor derived ECG intervals with derived intervals from the digitally aligned extracted ECG signals. All the patients were checked to assess the accuracy of the digitisation algorithm. The onsets and offsets for the points P, Q, R, S, and T in all 12 leads should be determined first in order to calculate the metrics like QRS

Algorithm 1: ECG Digitisation Pseudocode that shows each algorithmic step.

```
Data: 12-lead ECG PDF documents
Result: 12-lead ECG signals
initialise a folder with all ECG PDF documents;
while there is a PDF document do
    Locate the coordinates of each lead (position on the page);
    Obtain the scale indicator (to indicate size and timing intervals);
    if locate is successful then
        for each lead do
            Extract line points;
            Digitise the lead;
    if digitisation is successful then
        for each lead do
            Scale the digitised lead in mV;
            Save the lead;
        if set label then
            Get the pre-defined output column;
            Match extracted ECG signals with the output column;
        Save all the extracted data;
```

duration, PR interval and ventricular rate. An example illustration can be found in Figure 4.5, on a single beat. In the vendor algorithm [201], onsets are defined as the earliest deflection in any 12 leads, and offsets are defined as the latest deflection in any 12 leads. The QRS duration is measured in milliseconds from the earliest Q onset in any lead to the latest S offset in any lead. Similarly, the PR interval is measured in milliseconds from the earliest P onset in any lead to the QRS onset (or earliest Q onset) in any lead. For the ventricular rate (beats per minute), the number of beats is counted and divided by the time difference in minutes between the first and last beat. For all our corresponding calculations of QRS duration, PR interval and ventricular rate, the NeuroKit2 [203] library was used. We followed a processing pipeline similar to the vendors guidelines [201] to enable meaningful comparison of the final measurements.

Of the 436 patients, 194 patients were female (44.4%). The most common condition was ToF (39.9%), followed by ASD (17.6%) and PA with VSD (16.7%). Mean ECG age was 33 years (SD 11.7 and 75-25% IQR (40,23)). In a number of ECGs there were overlaps between the lead signals on the original PDF documents, as can be seen in Figure 4.3. This included lead labels, as well as large QRS complexes impinging on the lead displayed below it, in particularly

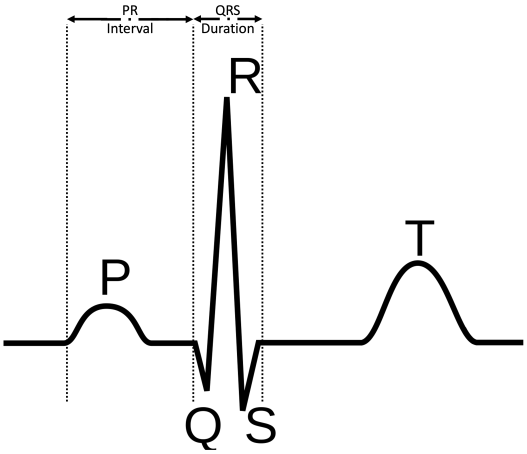


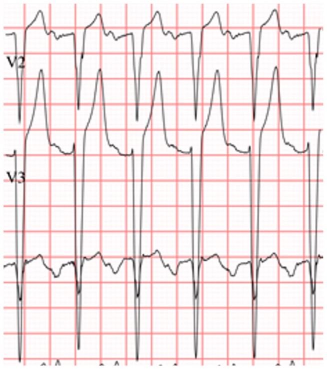
Figure 4.5: P, Q, R, S, and T points on a single ECG wave.

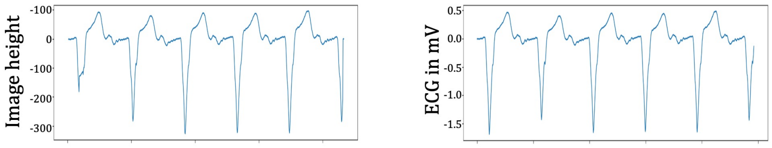
the precordial leads (V2-V6). We developed an algorithm to successfully overcome this issue using vector drawings on the original ECG PDF documents. Sample extraction results for the leads V4 and V5 can be seen in Figure 4.3.

4.4 Results

All the artefacts were overcome without further user intervention, as shown in Figure 4.6 on another patient. We compared our work with prominent open-source approaches found in the literature to digitise ECGs like Paper-ECG[204]. Figure 4.6 demonstrates the advantages of our methods by comparing ECG segment extraction of the two algorithms side-by-side, respectively. Original signals from the corresponding ECG document are shown in Figure 4.6a, for leads v2 and v3. All the original signals have some difficulties like text overlaps of lead names and letter notes, signal overlaps between different leads. Those difficulties may lead the digitisation fail, as detailed in the following examples. While initial QRS complex in Figure 4.6a for v2 is distorted and abbreviated in a vertical direction in Figure 4.6b because of the lead name overlapping with the signal, our algorithm can correct this in Figure 4.6c. Figure 4.6d shows overlapping QRS complexes from the above lead impinging on the QRS complexes of the present lead. Our algorithm also corrects this in Figure 4.6e. Similarly, the baseline shift artefact in Figure 4.6d is corrected as shown in Figure 4.6e.

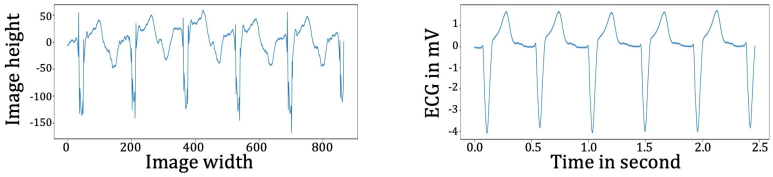
Bland-Altman plots are plotted to compare the measurement of the three variables using two





(b) Paper-ECG on lead v2

(c) Our algorithm on lead v2



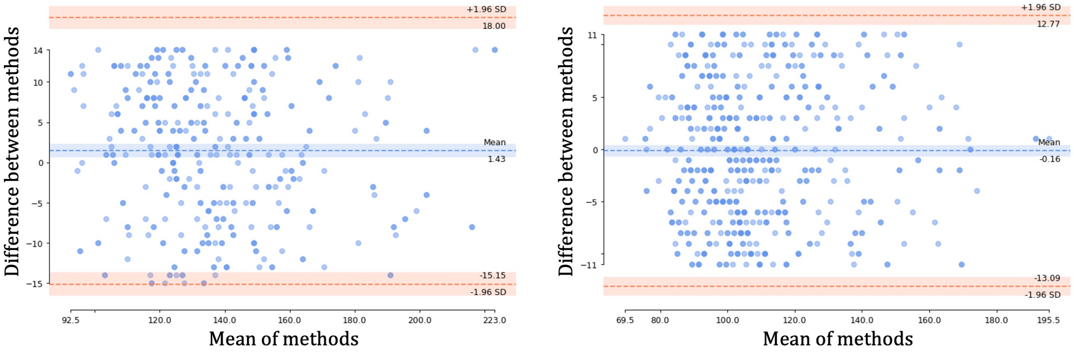
(a) Original leads (v2 & v3)

(d) Paper-ECG on lead v3

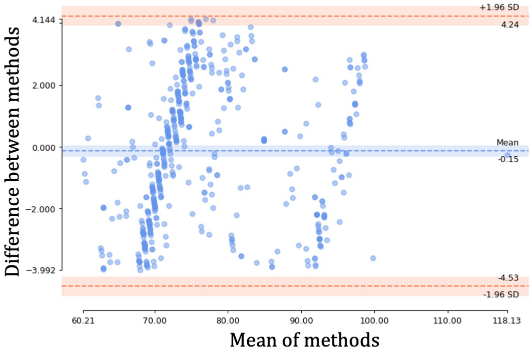
(e) Our algorithm on lead v3

Figure 4.6: Comparison of our digitisation results with the Paper-ECG[204].

different algorithms, vendor and our implementation. The mean of the two measurements is plotted on the x-coordinate while y-axis is the difference between the two algorithms, to show the agreement or disagreement between the two results. PR interval, QRS duration and ventricular rate are calculated and plotted in Figure 4.7a, 4.7b and 4.7c, respectively. For all the calculations, similar techniques as in the vendor manual were implemented, as closely as possible. Ventricular rates derived from the digitised ECG correlated well with the original vendor rendered ECG, as can be seen in Figure 4.7a. Similarly, in Figure 4.7b and 4.7c, the correlation between vendor calculated and digitised ECG calculated values for both QRS duration and PR interval is displayed respectively. Plots depict the agreement between the vendor calculated values on original signals and user calculated values on extracted signals. It also shows that there is no bias as the mean difference between two measurements is not consistently positive or negative. Pearson correlation coefficients are also reported along with the two-sided p-value, to show that there is strong correlation between the results. We also performed a null hypothesis test to calculate the significance of the correlation coefficient and to decide whether the relationship between the results is strong enough to be used to model the relationship. Null hypothesis assumes that the correlation coefficient is not significantly different from zero, and hence there is not a significant relationship between the variables. As the p-value is less than the chosen significance level (α = 0.05), we reject the null hypothesis. It indicates that there is sufficient evidence to conclude that there is a statistically significant correlation between the two results.



(a) PR interval (R: 0.94, $P < 0.05$) (b) QRS duration (R: 0.94, $P < 0.05$)



(c) Vent. rate (R: 0.97, $P < 0.05$)

Figure 4.7: Bland-Altman plots between vendor and extracted values.

4.5 Conclusions

ECGs are powerful tools when used at a population scale to identify poor cardiovascular outcomes. ECGs can be used to provide a sense of the physiological and structural condition of the heart, while also providing valuable diagnostic clues. Currently analysis of large-scale data requires access to the original ECG datasets, which is embedded in a codified fashion within the vendor analysis software and is not readily accessible for analysis. Instead, clinicians tend to use manual evaluation of printed ECGs or review of ECG PDF documents for interpretation. Digitisation of such 12 lead ECGs as we have been able to demonstrate in this manuscript, is potentially very helpful in facilitating larger scale research including risk stratification, cross institutional research, and in registry data recording.

In this work, we suggest linking ECG data with clinical letters and subsequently extracting accurately labelled digitised ECG waveforms for further analysis with machine learning algorithms. To our knowledge, this is the first time that ECG data from CHD patients are extracted from PDF documents and labelled automatically. Our proposed framework does not require additional user manipulation and we have extensively validated it by estimating ECG features, such as PR interval, QRS duration and ventricular rate and comparing the values with the vendor corresponding values.

There are several digitisation algorithms implemented by different researchers and vendors [199, 200, 204], but to our knowledge all of them work on pixelated images captured from the ECG PDF documents, or from the scanned ECG papers. This usually results in quality loss, text along with the ECG waveforms overlap and the digitisation process fails to accurately depict the original signal. Here we used open-source tools to extract ECG vectorised graphical information from the PDF documents. In this way, we were able to reconstruct the signal accurately.

The digitised ECG overcomes artefacts present in the Paper-ECG [204] digitised ECG traces in Figure 4.6, and accurately corrects these without any user intervention. For the first time, we demonstrate the ability for a relatively simple machine learning model to accurately predict diagnosis in the selected five conditions including Atrial Septal Defect, Mustard, Single Ventricle physiology with a Fontan circulation, Pulmonary Atresia and Tetralogy of Fallot. This preliminary research application and utility demonstrates encouraging potential for the technique. We believe that digitisation of the ECG will facilitate further research used in large datasets that can use PDF formats of the ECG.

The ability of the algorithm we developed to correct artefacts on the original ECG document was particularly gratifying. We demonstrate for the first time, the ability to correct for overlapping QRS complexes, baseline drift in the ECG, and text overlapping QRS or other parts of the ECG traces in patients with CHD. Though there are open-source works in the literature to digitise ECGs like Paper-ECG [204], all such software currently needs user intervention to locate all the 12-leads with a corresponding bounding box. With our automated approach, we can not only digitise but also export the corrected ECG waveforms to PDF documents to avoid overlapping and allow clearer interpretation. So, it allows us to reconfigure the 12 ECG leads from the PDF raw data including individual lead vector data rather than having just an image for each lead as we are able to capture all the vectors on the PDF document.

We validated the ECG digitisation process by comparing vendor rendered ECG data like QRS duration, ventricular rate and PR interval between the original ECG and the digitised version. We were able to demonstrate a strong correlation between the measurements. Inherent variability in the intervals is expected as the techniques used by the vendor to measure particular intervals vary considerably. For example, some vendors use only simple band-pass filtering while others use template matching to detect QRS complexes. In future studies, we plan to create a ground-truth dataset of PDFs to further validate our digitisation algorithm.

In our case, raw ECG data was not readily available even for data analysis purposes. It is a common practice in the UK as access to raw data requires additional fees. Many hospitals use standardised forms/reports that are generated as PDFs. These documents can be easily shared across different hospitals and platforms, making it easier for healthcare providers to access and digitise information without needing specialised hardware. This is the main reason why we implemented such a digitisation pipeline, to enable further research on the raw data. We are also aware that this is not always the case, depending on the agreement with the provider. As they do not support for such research, not guarantee to access raw data for us. Our digitisation pipeline can be easily updated for different PDF structures/layouts of ECGs. We have tried several open-source works and non-commercial toolboxes to digitise ECGs (like Paper-ECG [204]), but the results were poor on the patients with CHD.

Chapter 5

ECG Classification with Riemannian Geometry

5.1 Abstract

Congenital Heart Disease (CHD) is a relatively rare disease that affects patients at birth and results in extremely heterogeneous anatomical and functional defects. 12-lead Electrocardiogram (ECG) signal is routinely collected in CHD patients because it provides significant biomarkers for disease prognosis. However, developing accurate Machine Learning (ML) models is challenging due to the lack of large available datasets. Here, we suggest exploiting the Riemannian geometry of the spatial covariance structure of the ECG signal to improve classification. Firstly, we use covariance augmentation to mix samples across the Riemannian geodesic between corresponding classes. Secondly, we suggest to project the covariance matrices to their respective class Riemannian mean to enhance the quality of feature extraction via tangent space projection. We perform several ablation experiments and demonstrate significant improvement compared to traditional machine learning models and deep learning on ECG time series data.

5.2 Introduction

12-lead ECG is a very common diagnostic and prognostic tool in cardiac diseases. The reason behind this is that it is easy to acquire and it reflects cardiac anatomy and function in high spatio-temporal resolution. In fact, recent work showed that 12-lead ECGs can be converted into 3D representations of the direct cardiac activity during a heart beat based on Vectorcardiograms (VCGs) [205, 206]. An example of these representations is shown in Figure 5.1 and it reveals signatures of anatomical defects in CHD based on our data. Several deep learning techniques have been proposed to classify cardiac rhythms and estimate the risk of adverse effects [207]. These methods showed impressive results with large datasets that include millions of patients and ECG recordings. However, it is not clear how they can extend in relatively rare and extremely heterogeneous cases.

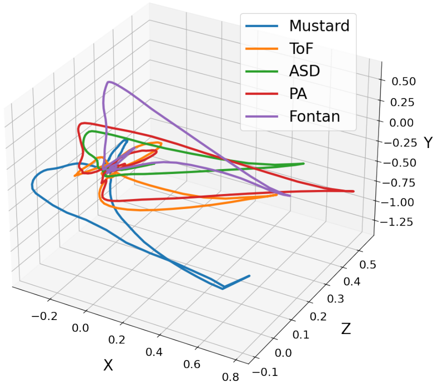


Figure 5.1: Average VCGs across anatomical defects in CHD.

In CHD, abnormalities in structure and function are present at birth and affect around 1% of babies. In other words, patients are born with genetic defects that differ significantly from the cardiac abnormalities emerging later in life. Therefore, the efficiency of deep learning methods developed on a broader population is limited due to lack of large scale representative data and extreme physiological variations in both anatomy and function. Inspired by successful work on Riemannian classification [208], we proposed to use the covariance structure of the 12-lead ECGs to predict anatomic diagnosis associated with CHD as an initial step toward mortality prediction. Summarising, the contributions of this chapter are as follows:

- **Application of Riemannian Geometry to ECG-Based Diagnosis:** This chapter proposes the integration of Riemannian geometry techniques for the purpose of analysing 12-lead ECG data, with the objective of improving the classification accuracy in the context of congenital heart disease. The utilisation of spatial covariance structure of the ECG signals enables more robust feature extraction and data representation, improving the classification metrics of the models in clinical research.

- **Covariance Augmentation for Improved Data Representation:** Covariance augmentation is utilised to allow ECG samples to be mixed across the Riemannian geodesic, thereby generating more samples in a controlled manner and avoiding data imbalance issues by improving feature representation and model robustness. This approach has been shown to enrich the quality of the dataset, thereby facilitating enhanced model generalisation and reduced risk of overfitting.
- **Multiple Tangent Space Projections:** This chapter introduces a multiple tangent space projection approach, improving feature extraction by mapping covariance matrices onto the Riemannian mean of each class. This approach facilitates the extraction of a data-driven representation of ECG signals, not only by utilising the tangent space in which the data resides, but also the other tangent spaces. It has been demonstrated that this approach provides relevant information regarding the location of a data points in relation to other tangent spaces, which refines feature mapping for improved classification.
- **Ablation Experiments Validating Model Performance:** The study conducts a series of ablation experiments, demonstrating that the proposed methods significantly outperform conventional classification models, thereby demonstrating their utility in clinical applications. These techniques enhance the performance of machine learning models by leveraging the subtle variations captured in ECG signals, which may otherwise be overlooked by conventional approaches.

5.3 Methods

5.3.1 Data

The digitised 12-lead ECGs from patients with CHD under regular follow-up at the Scottish Adult Congenital Cardiac Service based at the Golden Jubilee National Hospital in Scotland were utilised as the dataset. ECGs in atrial flutter or atrial fibrillation were excluded, as were atrioventricular paced rhythms, as one of our primary aims was to use ECGs in sinus rhythm to predict diagnosis. The most common condition was Tetralogy of Fallot (ToF) (39.9%), followed by Atrial Septal Defect (ASD) (17.6%) and Pulmonary Atresia (PA) (16.7%). Mean ECG age was 33 years (SD 11.7 and 75-25% IQR(40,23)). Patients with no documented ECGs or those in atrial flutter or fibrillation or other heart rhythm abnormalities including being paced, at the time of the ECG were excluded. For patients with more than one anatomic diagnosis, the dominant diagnosis was considered the primary diagnosis. Extracted 436 patients are summarised as follows: 173 patients with ToF, 77 patients with ASD, 73 patients with PA, 66 patients with Fontan and 47 patients with Mustard.

5.3.2 From common to multiple tangent spaces

ECG signal Y is described as a $12 \times n$ time series data. For Riemannian manifold classifier, the covariance structure of the multichannel (12-lead) ECG signal is estimated and projected into a flat space. This process allows more accurate estimation of linear operations. We applied spatial filtering F to enhance the signal-to-noise ratio and remove the artifacts [209]. Then, the covariance matrix is estimated using the Equation 5.1, which reflects the correlations between each pair of the leads.

$$\mathbf{C} = F(Y) \cdot F(Y)^{\mathsf{T}} \tag{5.1}$$

Since covariance matrices are Symmetric Positive Definite (SPD) matrices, they must be analysed in a Riemannian manifold rather than Euclidean space. Riemannian metric rather than Euclidean metric is used to project covariance matrices onto tangent space while respecting their geometry [210]. To achieve this, typically, the covariance matrices are projected onto a common tangent space based on Equation 5.2 [208].

$$\mathbf{V}_{\mathbf{i}}^{\mathbf{C}} = \text{upper}\left(\mathbf{C}^{-\frac{1}{2}} \text{Log}_{\mathbf{C}}\left(\mathbf{C}_i\right) \mathbf{C}^{-\frac{1}{2}}\right) \tag{5.2}$$

Where $\mathbf{C_i}$ is the covariance matrix to be projected onto the tangent space at point $\mathbf{C}$, which represents the Riemannian mean of all the covariance matrices. This projection enhances the performance of classifiers that depend on distance metrics between the sample covariance matrices and it has been successful in processing high-dimensional neurophysiological data [133, 210, 211]. However, it assumes that $\mathbf{C}$ and $\mathbf{C_i}$ are relatively close (see Figure 5.2).

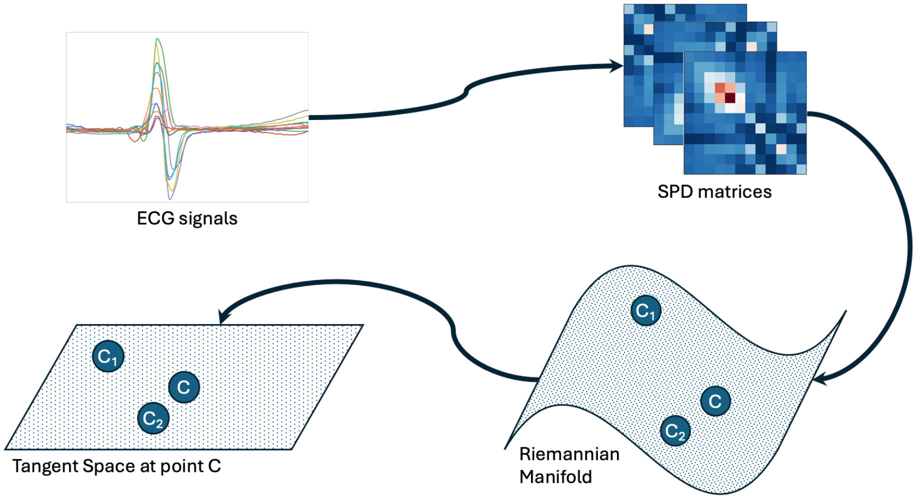


Figure 5.2: Mapping ECG signals to tangent space using Riemannian manifold.

We hypothesise that projecting each covariance matrix to its corresponding class mean will improve the quality of the mapping, since the distance will be smaller compared to the corresponding distance with the global Riemannian mean. Thus, we fitted a different tangent space for each class and then combined each of the outputs into a new feature vector. After tangent space projection, each covariance matrices is represented as a vector $\mathbf{V}$ of size $n \times (n+1)/2$, where n is the dimension of the covariance matrices. Each covariance matrix is mapped into tangent space by keeping the upper triangular part of the resulting symmetric matrix as denoted in Equation 5.2. Also, an illustration of the multiple tangent space concept can be seen in Figure 5.3. Each output of the tangent spaces can be combined into a single enriched feature vector that is fed as input to the classification model. It also allows us to balance potential data issues by using covariance augmentation, described below, on feature vectors of the underrepresented classes. All the algorithmic steps of our proposed approach are demonstrated in Algorithm 2.

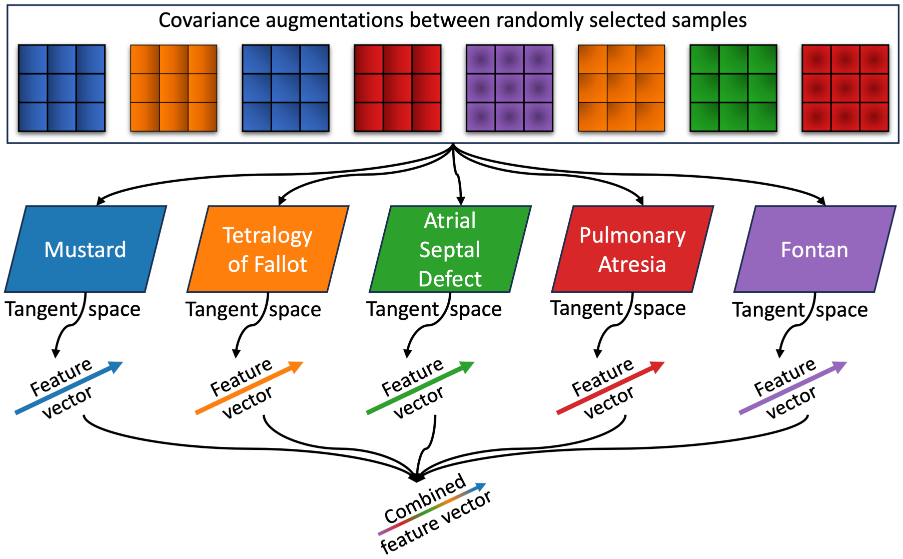


Figure 5.3: Multiple tangent space concept.

5.3.3 Augmentation of covariance matrices

Since we have a limited dataset, we use a covariance mixing technique to generate more samples in a controlled way similar to [212]. To apply mixing, we sample α from a beta distribution on the interval [0, 1], and compute the weighted Riemannian mean according to the Riemannian distance metric between the randomly selected covariance matrices. We also tried to control mixing by restricting the range of sampled values for α, but the best results were obtained with no restrictions. Riemannian mean that minimises the sum of squared Riemannian distances to the given two SPD matrices was calculated to find the weighted Riemannian mean as in Equation 5.3, where w_i represents a weight matrix generated using the α value, d_R represents the Riemannian distances to the SPD matrices.

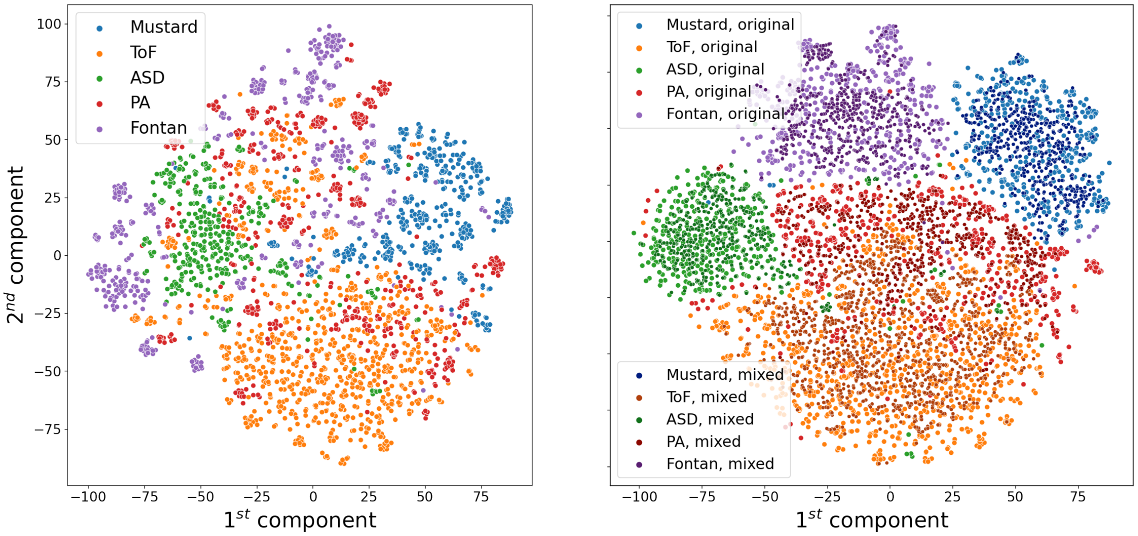
$$\mathbf{C}_{\mathbf{aug}} = \arg\min_{\mathbf{C}} \sum_i w_i \, d_R(\mathbf{C}, \mathbf{C}_i)^2 \tag{5.3}$$

Instead of mixing all the data we have, we tried to focus only on the classes that are not easily distinguishable by the model. t-distributed Stochastic Neighbor Embedding (t-SNE) vi-

Algorithm 2: Prediction of anatomical diagnoses.

```
Data: 12-lead ECG signals
Result: Anatomic diagnosis associated with CHD
initialise SPLO splits and align R Peaks;
while there is a 12-lead ECG do
    calculate covariance matrices based on the
    spatial filtering $F$;
    if mixup then
        mixup on covariance matrices using Eq. 5.3;
    if projection then
        if multiple tangent space then
            fit a tangent space for each class;
            project covariance matrices to those tangent spaces using Eq. 5.2;
            combine each output to get a feature vector to train on the model
        else
            fit only a single tangent space;
            project covariance matrices to the
            tangent space using Eq. 5.2;
            get feature vector to train on the model
    make classification;
```

sualisations [213] on the tangent space using only original data, and mixed data combined with the original are shown in Figures 5.4a and 5.4b, respectively.



(a) t-SNE without covariance augmentations on tangent space

(b) t-SNE with covariance augmentations on tangent space

Figure 5.4: t-SNE visualisations on the tangent space.

5.3.4 Stratified patient leave-out

Training-testing split of data was repeated 100 times based on pseudo-randomised, Stratified Patient Leave-Out (SPLO) evaluation to ensure that the testing set of patients was representative of all the classes. One patient for each class is randomly selected for testing. All of testing patients' ECG data are removed from training and only data from the rest of patients are used in the training. This training procedure is repeated 100 times in a pseudo-randomised manner and the average performance results (accuracy, AUC and F1 macro) along with standard deviation are reported. On average, each patient has 10 ECG recordings (for a very small number of patients this number can vary from 2 to 40). Firstly, the data are split into different sets and subsequently further processing, such as covariance augmentations, is performed only on the training data. Using such an approach, summarised in Figure 5.5, SPLO enhances the capability of the Support Vector Machine (SVM) model to predict unseen data by reducing bias [32]. This approach ensures that each class is proportionally represented in both the training and test sets, thereby preventing overfitting and underfitting.

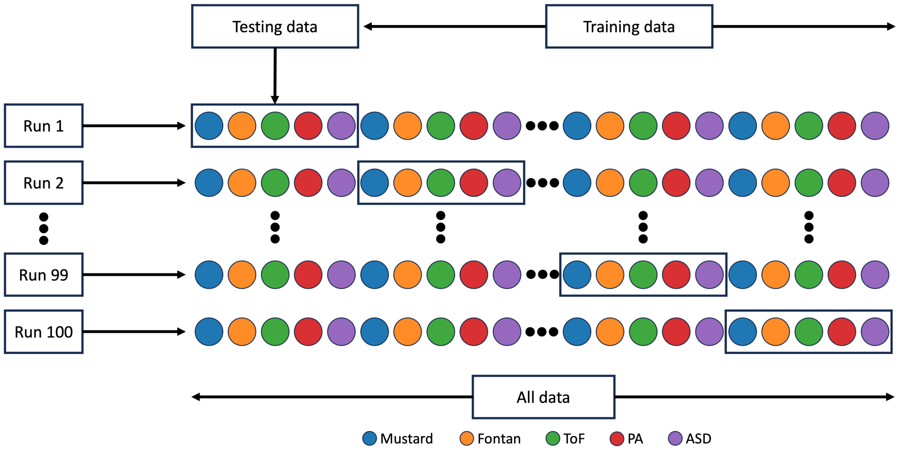


Figure 5.5: Stratified patient leave-out setup.

5.3.5 Pre-trained models

In machine learning, a pre-trained model refers to a type of model that has undergone initial training on a large and diverse dataset. This initial training phase is designed to capture broad features, patterns and relationships across the full range of data in the larger set. A variety of organisations, researchers and communities are involved in the development of pre-trained models. One such example is Google's Bidirectional Encoder Representations from Transformers (BERT), a language model originally implemented in the English language with a base model comprising over 110 million parameters [214]. Following training on the Toronto BookCorpus (800 million words) and English Wikipedia (2.5 billion words), BERT demonstrates the substantial computational resources required for such models [215]. Such entities employ large-scale data and machine learning methodologies to develop pre-trained models that can be fine-tuned for specific tasks in both practical and cost-effective manner [216].

Although pre-trained models offer a convenient means of reducing training time and computational power, their use without fine-tuning may not yield the desired results [217]. It may therefore be necessary to perform fine-tuning on the intended data set in order to ensure suitability for the task at hand. A further training phase utilising supervised learning techniques on a task-specific dataset enables the pre-trained model to be optimally fine-tuned for the target task, while simultaneously leveraging the general knowledge acquired during the pre-training phase. This approach employs transfer learning, whereby knowledge acquired during the pre-training phase is transferred to the target task, thereby enhancing performance and reducing the necessity for extensive task-specific data [216].

The use of pre-trained models helps to overcome the issue of data scarcity by taking advantage of the vast and diverse datasets available during the pre-training phase. This enables the model to generalise more effectively from a smaller, task-specific dataset during the fine-tuning process. Furthermore, this approach improves the model's performance on the target task, as it is capable of effectively recognising patterns and features that are common across different datasets. As a result of the model having already acquired general features during the pre-training phase, the fine-tuning phase requires less computational power and time, thereby enhancing the overall process's efficiency.

Riberio et al. developed a state-of-the-art Deep Neural Networks (DNNs) model for the automatic diagnosis of 12-lead ECGs [207]. The model was trained on a large dataset comprising over 2 million labelled ECGs , using supervised learning techniques where the model learned to map ECG inputs to corresponding diagnostic labels. The 6 diagnostic labels were as follows: 1st degree AV block (1dAVb), right bundle branch block (RBBB), left bundle branch block (LBBB), sinus bradycardia (SB), atrial fibrillation (AF) and sinus tachycardia (ST). The DNNs model demonstrated superior performance in identifying 6 types of abnormalities in 12-lead ECGs , outperforming cardiology resident medical doctors with F1 scores above 80% and specificity exceeding 99%.

Their ECG data utilised in the paper and our own ECG data extracted from the PDF documents exhibit similar characteristics. The study employed ECGs obtained from patients aged 16 years and above. And, 12 different leads of the ECGs , that are sampled at 400 Hz, populated in the following order: $DI, DII, DIII, AVR, AVL, AVF, V1, V2, V3, V4, V5, V6$. As not all signals have the same duration, both dimensions were padded with zeros to ensure that they were of an identical size. We also employed the same strategy while fine-tuning the model with our data. And, same patient split strategy (SPLO) was used to ensure consistency both with the previous results and between the different results.

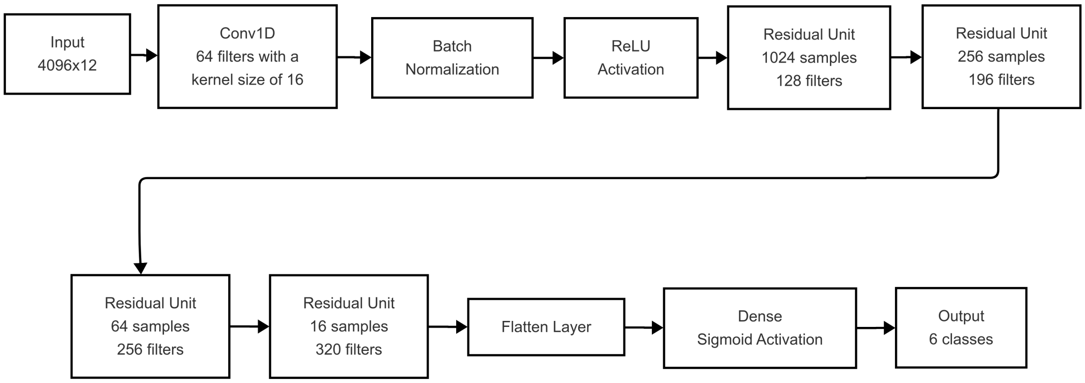


Figure 5.6: Architecture of the pretrained model [207].

The model receives an input tensor with dimensions (N, 4096, 12), where N is the batch size, 4096 is the number of data points, and 12 represents the 12 leads of the ECG signal. The structure of the model comprises multiple residual blocks, where each block incorporates convolutional layers, batch normalization, and ReLU activation functions. The residual connections enhance the training of deeper networks by enabling the effective propagation of gradients through the network while addressing the issue of vanishing gradients. The final layer of the network comprises a dense layer with 6 units, which correspond to the 6 types of ECG abnormality. The output is then passed through a softmax activation function, which generates probabilities for each class. Architecture of the pretrained model can be seen in Figure 5.6.

5.3.6 Transfer learning

Transfer learning represents a strategy in machine learning, whereby a model developed for a task is repurposed as the starting point for another model addressing another related task. This approach is effective in scenarios where there is limited data or when the new task has similarities with the original task. Especially in the domain of image classification, researchers tend to adopt a preliminary approach, utilising a pre-trained model that has been previously trained to categorise visual elements and has acquired general attributes such as edges and shapes.

Fine-tuning is employed to provide additional training to the model whose parameters have been previously adjusted through an initial learning phase. It involves utilising the knowledge of the trained model as a starting point and subsequently training the model on a smaller, task-specific dataset, thereby enabling the model's adaptation to the particular task at hand. One of the techniques employed to adapt a pre-trained model to a specific task is fine-tuning by freezing all but the last layer. The freezing of a layer prevents the weights associated with that layer from being updated during the training process. This approach is employed to ensure the preservation of the knowledge that has already been acquired by the pre-trained model.

In order to achieve the desired outcome, it is also necessary to replace the final layer with a new layer that is appropriate for the specific task at hand. In our case, as it was a classification task, the final layer may be a dense layer with a softmax activation function. Subsequently, the model should be fine-tuned by training it on the specific dataset. As only the weights of the final layer are updated, the model can rapidly adapt to the new task without losing the features learned during the pre-training phase.

The freezing of the majority of the model's layers results in a reduction in the time required for training, as only a limited number of parameters need to be updated. Another advantage of this approach is better performance, as the pre-trained layers have already captured useful features, resulting in an improvement in performance with less data.

Feature extraction

This strategy involves the utilisation of a pre-trained model to extract relevant features from the intended dataset. The extracted features are subsequently fed into another model for the designated task. The entire process can be summarised in the following sequence of steps.

1. Load the pre-trained model and pass the data through it.
2. Extract the features using the output of intermediate layers.
3. Train a new model using these features as inputs.

Fine-tuning

Fine-tuning involves the initialisation of the training process with a pre-trained model and then progresses to the training of a new and relatively smaller dataset. This is usually done by selecting a model that has been trained on similar and larger amounts of data. The entire process can be summarised in the following sequence of steps.

1. Load the pre-trained model.
2. Replace or add new layers suitable for the new task and outputs.
3. Continue training the model on the new dataset.

Freezing layers

A similar approach to fine-tuning is freezing the layers of a pre-trained model, which prevents any updates to their weights during the training process. Instead, only the newly added layers undergo training process. The entire process can be summarised in the following sequence of steps.

1. Load the pre-trained model.
2. Set the trainable attribute to "False" for the frozen layers.
3. Replace or add new layers (after the frozen layers) suitable for the new task and outputs.
4. Continue training the model on the new dataset, updating only the non-frozen layers.

5.4 Results

Three different classifiers were used to compare our feature extraction framework: SVM, Minimum Distance to Means (MDM), and Multi-Layer Perceptron (MLP). We have done ablation studies with MLP model as well with baseline models SVM and MDM. MDM performs classification by the nearest centroid. A centroid of the covariance matrices is estimated for each of the classes and then, for each new covariance sample, the class is estimated according to the nearest centroid. MLP model was trained with the different feature extraction frameworks as mentioned in Table 5.1. Table 5.1 shows the different strategies applied for augmentation and covariance projection. (DEF) denotes application of the classification directly on the R peak aligned 12-lead ECG data as the default setting. (VCG) represents augmentations based on ECG VCG space. Dower transformation was used to project 12-lead ECG data into a 3-dimensional VCG space. Subsequently, we applied rotations along all three orthogonal axes from 5 to 45 degrees and projected the augmented VCG back into 12-lead ECG space [206]. (VCG) augmentation performed poorly in classifying covariance matrices (VCG_COV). In the next ablation steps, (COV) represents that covariance matrices were used. The final step was the projection of these matrices into tangent space (single: TS or multiple: MTS). The best results were obtained using multiple tangent spaces with covariance augmentations (MTS_COV).

Table 5.1: Ablation study: AUC scores for different methods.

Name	**Feature Extraction Framework**				**AUC (MLP)**	**AUC (SVM)**	**AUC (MDM)**
	VCG Augmentations	Covariance Matrix	Covariance Augmentations	Tangent Space			
DEF	✗				0.73 ± 0.05	0.80 ± 0.08	NA
VCG	✓				0.76 ± 0.08	0.59 ± 0.04	NA
COV	✗	✓			0.77 ± 0.08	0.50 ± 0.09	0.76 ± 0.09
VCG_COV	✓	✓			0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00
TS	✗	✓	✗	Single	0.82 ± 0.08	0.72 ± 0.07	0.80 ± 0.07
TS_COV	✗	✓	✓	Single	0.83 ± 0.07	0.77 ± 0.06	0.82 ± 0.06
MTS	✗	✓	✗	Multiple	0.81 ± 0.09	0.78 ± 0.08	NA
$\boldsymbol{MTS\_COV}$	✗	✓	✓	Multiple	**0.84 ± 0.06**	**0.82 ± 0.08**	NA

Table 5.2 provides a comparison on the performance metrics (Accuracy, AUC, F1 macro) of the approaches described in ablation Table 5.1 as well as a baseline model with SVM applied on the aligned ECG time-series data DEF_{2D}. Using tangent space projection on MLP resulted in a

9% increase in the AUC score and an 11% increase in the accuracy. Using covariance augmentations on the single tangent space also yielded a further increase of 1% in the AUC score and a 3% increase in the accuracy. Taking a step further and using multiple tangent space projections with covariance augmentations resulted in an 11% increase in the AUC score and a 16% increase in the accuracy. Better results were obtained by projecting the augmented covariance matrices into multiple tangent space while respecting their underlying geometry.

Table 5.2: Performance metrics of the machine learning models.
(With **bold** we highlight the best six models).

	Accuracy	**AUC**	**F1 macro**
MLP (MTS_COV)	**0.71 ± 0.10**	**0.84 ± 0.06**	**0.69 ± 0.13**
MLP (MTS)	**0.64 ± 0.17**	**0.81 ± 0.09**	**0.63 ± 0.18**
MLP (TS_COV)	**0.69 ± 0.12**	**0.83 ± 0.07**	**0.65 ± 0.13**
MLP (TS)	**0.66 ± 0.15**	**0.82 ± 0.08**	**0.63 ± 0.17**
MLP (VCG_COV)	0.17 ± 0.06	0.50 ± 0.00	0.05 ± 0.01
MLP (VCG)	0.62 ± 0.13	0.76 ± 0.08	0.54 ± 0.15
MLP (DEF)	0.55 ± 0.12	0.73 ± 0.05	0.48 ± 0.09
SVM (MTS_COV)	**0.66 ± 0.17**	**0.82 ± 0.08**	**0.65 ± 0.17**
SVM (MTS)	0.61 ± 0.13	0.78 ± 0.08	0.56 ± 0.16
SVM (TS_COV)	0.59 ± 0.10	0.77 ± 0.06	0.55 ± 0.11
SVM (TS)	0.56 ± 0.12	0.72 ± 0.07	0.53 ± 0.12
SVM (VCG_COV)	0.17 ± 0.06	0.50 ± 0.00	0.05 ± 0.01
SVM (DEF_{2D})	0.65 ± 0.14	0.80 ± 0.08	0.62 ± 0.15
MDM (TS_COV)	**0.67 ± 0.14**	**0.82 ± 0.06**	**0.66 ± 0.11**
MDM (TS)	0.65 ± 0.15	0.80 ± 0.07	0.63 ± 0.13
MDM (VCG_COV)	0.17 ± 0.06	0.50 ± 0.00	0.05 ± 0.01
MDM (COV)	0.63 ± 0.15	0.76 ± 0.09	0.56 ± 0.17

Both SVM and MLP models were trained on R peak aligned ECG data (DEF), but for the SVM model 3D ECG data were reshaped to 2D (DEF_{2D}). SVM model achieved 7% better results at this stage for the AUC score. Also, some 3D rotations on the ECG data were tried by projecting it to the VCG space with the help of Dower transformations. Only MLP model was trained on these 3D augmented ECG data (VCG) and achieved 3% better results for the AUC score. But, they caused very poor results with the next steps that include covariance matrices (VCG_COV). VCG means without any augmentations for each class can be seen in Figure 5.1.

Figure 5.7 shows how the top six models compare statistically where statistical significance is based on corrected paired t-test [218]. All the models except SVM achieved better results on

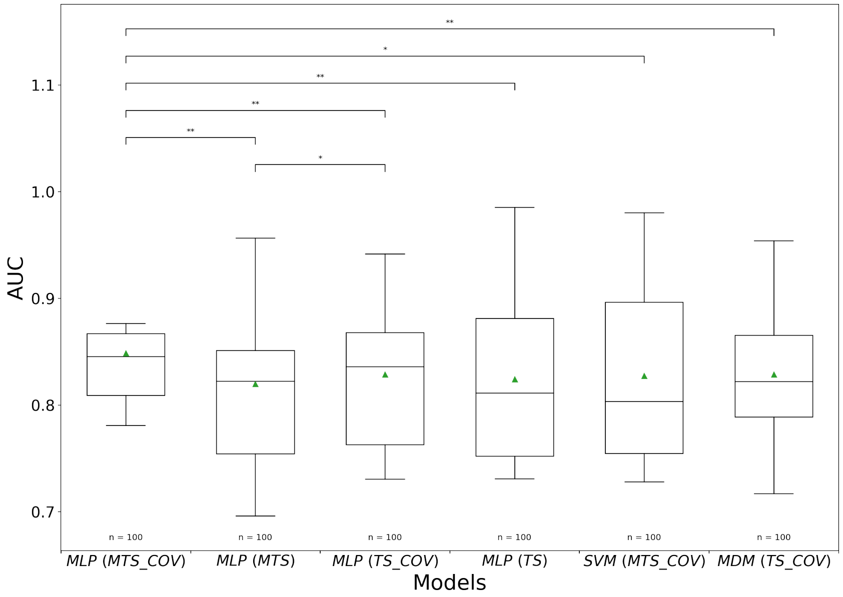
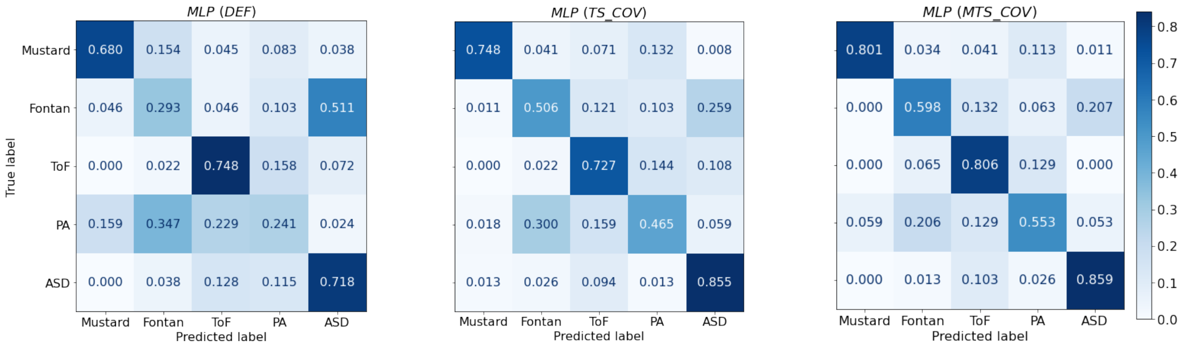


Figure 5.7: AUC scores for the best six approaches.

the tangent space and peaked at an AUC score of 82%. There was a statistically significant difference between the AUC results of the $MLP(MTS\_COV)$ model compared to the other models in Figure 5.7. As shown in Table 5.2, using multiple tangent spaces provided an improvement of 11% for the MLP model and 2% for the SVM model in the AUC score.

There was a statistically significant difference between the AUC results of the $MLP(MTS\_COV)$ model compared to the other models in Figure 5.7. The confusion matrices in Figure 5.8a, 5.8b and 5.8c, reflect the improvement in performance with the application of the multiple tangent space projection and covariance augmentation. Furthermore, Figure 5.9 shows AUC score for each class on the best model MTS_COV, evaluated using One-vs-Rest (OvR) strategy.



(a) MLP

(b) MLP on tangent space

(c) MLP on multiple tangent space

Figure 5.8: Confusion matrices of DEF, TS_COV and MTS_COV approaches with an MLP classifier.

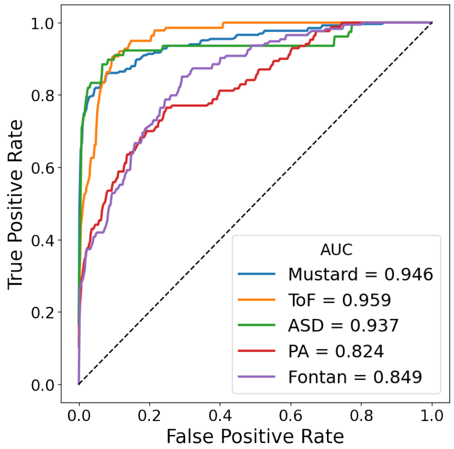


Figure 5.9: AUC for the best model (MTS_COV).

In order to quantify the effect of covariance augmentations in Riemannian space on t-SNE embeddings (see Figure 5.4), the label-wise distances to each class centroid were computed. For each class we report mean, minimum and maximum Euclidean distances from points to their centroid in the t-SNE embedding. As demonstrated in Table 5.3, these augmentations not only reduce intra-class dispersion (points of the same class move closer to their centroid) but also tighten cluster structure as evidenced by increased separation, which reflected in both the minimum and maximum distances to each centroid.

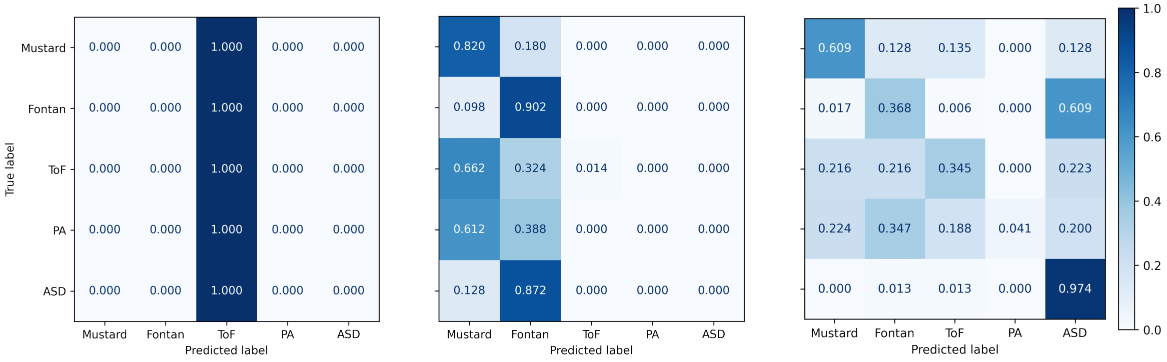
Table 5.3: Effect of covariance augmentations on label-wise distances to class centroids.

	mean	**std**	**min**	**max**
ASD, without augmentations	24.395	20.331	1.454	134.273
ASD, with augmentations	20.466	13.561	0.756	84.403
PA, without augmentations	44.360	17.438	6.694	112.768
PA, with augmentations	28.743	15.315	1.852	95.420
ToF, without augmentations	38.371	16.433	2.074	102.396
ToF, with augmentations	31.023	13.429	0.258	78.450
Fontan, without augmentations	29.144	13.165	5.783	117.708
Fontan, with augmentations	20.046	10.566	0.469	84.345
Mustard, without augmentations	26.318	12.901	0.873	128.842
Mustard, with augmentations	19.070	9.468	0.447	119.332

5.4.1 Classification results with the pre-trained model

Firstly, the pre-trained model was tested on our ECG dataset in order to evaluate how the process of fine-tuning affects the model's capacity to achieve better results. Without any fine-tuning and data augmentations, the classification outcomes were predominantly focused on the more dominant class (ToF). As can be seen in Figure 5.10a, using pre-trained model without any fine-tuning is not achieving good results.

The pre-trained model then fine-tuned by training on our ECG dataset and only the weights of the final layer are updated, while the weights of the remaining layers were maintained in a frozen state. Additionally, the dataset was augmented with randomly selected duplicates, without compromising the integrity of the training and testing processes. As can be seen in Figure 5.10b, it achieved better results in terms of accuracy and AUC, but poor confusion matrix results. An accuracy value of 0.456 was obtained, but the fine-tuned model only classifies the results as belonging to one of two classes: mustard and fontan. This may be indicative of potential overfitting, as our dataset contained an even distribution of data across all classes.



(a) without any fine-tuning and data augmentations

(b) with augmentations (copying)

(c) with augmentations (resample)

Figure 5.10: Confusion matrix comparison of the pre-trained model in different settings.

A further attempt was made with the use of augmentations via resampling, employing the same methodology of fine-tuning. As can be seen in Figure 5.10c, it achieved better results in terms of accuracy and AUC, as well as a better confusion matrix representation. An accuracy value of 0.445 was obtained, but the fine-tuned model categorises the outcomes as almost no outcomes in one category, PA. In our particular case, the utilisation of a state-of-the-art pre-trained model was insufficient to yield the desired level of accuracy and reliability in the results obtained. Even with the incorporation of data augmentation techniques, the desired level of reliable results was not achieved.

All the experiment results are shared in Table 5.4, while incorporating data augmentations techniques demonstrated improved results, it did not achieve the desired level of accuracy as reflected in the F1 score. So, it was not enough fine-tuning only the last layer of the pre-trained model to achieve optimal results, as the deeper layers of the model are capable of capturing more complex and task-specific features.

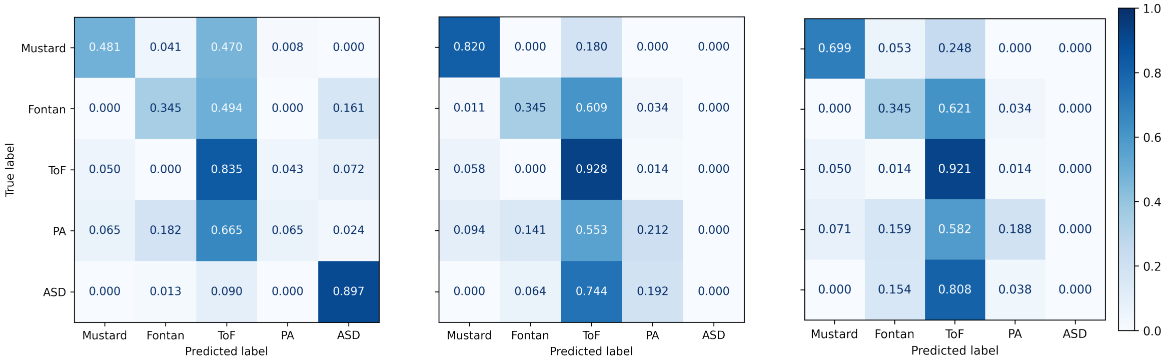
Table 5.4: Classification results on ECGs, with pre-trained model.

	Accuracy	**AUC**	**F1 score**
Without fine-tuning	0.172	0.500	0.058
Fine-tuning with augmentations (copying)	0.456	0.599	0.242
Fine-tuning with augmentations (resample)	0.445	0.680	0.378

In order to enable the model to accommodate these more profound and complex characteristics at the deeper layers, the number of layers undergoing training was increased. For the first experiment, all but the last five layers were kept frozen, resulting in a total of 26,245 trainable parameters. The same data augmentation technique utilising resampling was employed for all subsequent experiments. As illustrated in Figure 5.11a, the pre-trained model demonstrated enhanced performance in terms of accuracy (0.466), AUC (0.698), and F1 score (0.461), along with a more optimal confusion matrix representation. However, the challenge of classifying outcomes into a single category for PA class remained.

The same procedure was followed this time with the last seven layers unfrozen, resulting in a total of 108,165 trainable parameters. As illustrated in Figure 5.11b, the pre-trained model underperformed in terms of AUC (0.673), and F1 score (0.415). Furthermore, the confusion matrix representation exhibited more issues than those observed in previous models. While the challenge of classifying the results into a single category for the PA class persisted, a similar problem was observed for the ASD class. As the total number of trainable layers increased, the classification process became progressively more centralised, and it was focused solely on the ToF class for those problematic classes.

The same procedure was repeated this time with the last eight layers unfrozen, resulting in a total of 1,746,565 trainable parameters. As illustrated in Figure 5.11c, the pre-trained model underperformed in terms of accuracy (0.491), AUC (0.653), and F1 score (0.386). A similar confusion matrix representation with the same issues was observed for the ASD and PA classes, with a particular focus on the ToF class.



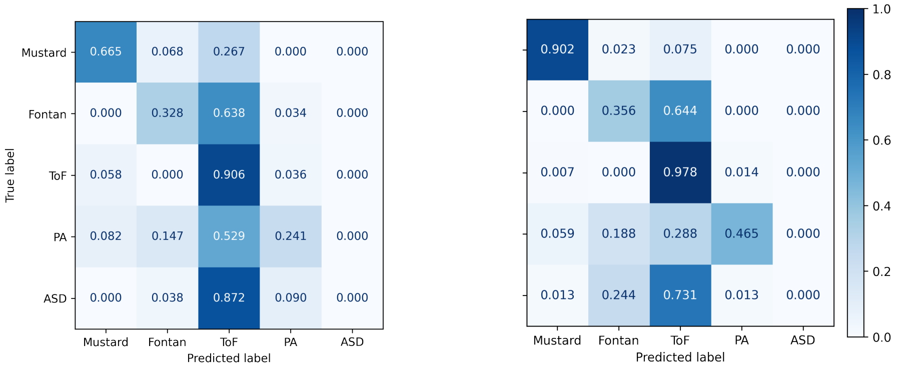
(a) Fine-tuning last five layers

(b) Fine-tuning last seven layers

(c) Fine-tuning last eight layers

Figure 5.11: Confusion matrices of the fine-tuned model using last five, seven and eight layers.

Before unfreezing all the layers, the same procedure was repeated this time with the last twelve layers unfrozen, resulting in a total of 1,747,205 trainable parameters. As illustrated in Figure 5.12a, the pre-trained model underperformed in terms of accuracy (0.485), AUC (0.651), and F1 score (0.389). A similar confusion matrix representation with the same issues was observed for the ASD and PA classes, with a particular focus on the tetralogy of fallot class.



(a) Fine-tuning last twelve layers

(b) Fine-tuning all layers

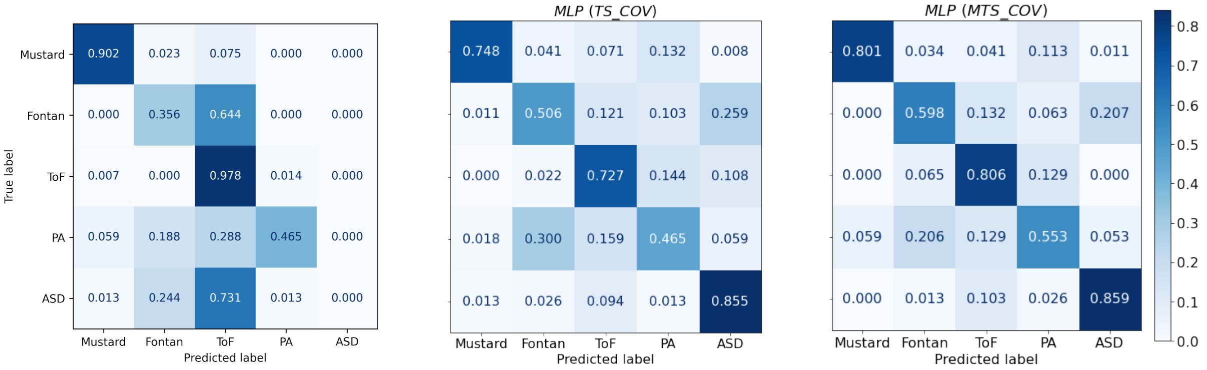
Figure 5.12: Confusion matrices of the fine-tuned model using last twelve and all the layers.

Given the absence of improvement over the last steps, a further round of fine-tuning was conducted utilising all layers. The same procedure was repeated this time with all the layers unfrozen, resulting in a total of 6,416,789 trainable parameters. As illustrated in Figure 5.12b, the pre-trained model demonstrated best performance so far in terms of accuracy (0.625), AUC (0.724), and F1 score (0.501), along with a more optimal confusion matrix representation. However, although the problems in PA class improved, the challenge of classifying outcomes into a single category for the ASD class remained. All the experimental results shared in Table 5.5 indicate that this approach was insufficient to achieve optimal results.

Table 5.5: Classification results on ECGs, with fine-tuned model.

	Accuracy	**AUC**	**F1 score**
All layers	0.625	0.724	0.501
Last layer	0.445	0.680	0.378
Last 5 layers	0.466	0.698	0.461
Last 7 layers	0.536	0.673	0.415
Last 8 layers	0.491	0.653	0.386
Last 12 layers	0.485	0.651	0.389

The confusion matrices in Figure 5.13b and 5.13c reflect the improvement in performance with the application of the single and then multiple tangent space projections, in comparison with the pre-trained model in Figure 5.13a. Notably, the MLP model on a single tangent space projection outperformed the best configuration of the pre-trained model, highlighting the effectiveness of the proposed approach.



(a) Fine-tuning all layers

(b) MLP on tangent space

(c) MLP on multiple tangent space

Figure 5.13: Confusion matrix comparison of the best pre-trained model with the single tangent space (TS_COV) and multiple tangent space (MTS_COV) approaches.

5.5 Conclusions

We demonstrated promising results on 12-lead ECG classification of anatomical diagnosis in CHD. Our proposed projection of the augmented covariance matrices to multiple Riemannian spaces yields significantly better results in improving classification performance with small and extremely imbalanced 12-lead ECG data. The proposed approach, based on Riemannian geometry, demonstrates significant improvements in diagnostic accuracy, providing a mathematically rigorous framework that aligns with improved results. While the study establishes the efficacy of Riemannian geometry, it also highlights areas for further investigation. The limitations in sample size and data imbalance in this study offer opportunities for future research to refine and expand upon this approach. By addressing these limitations, the integration of Riemannian methods into clinical practice has the potential to enhance the accuracy of diagnostic tools in healthcare.

5.6 Discussion

This chapter explores the application of Riemannian geometry to improve CHD diagnosis accuracy using 12-lead ECG data. By employing covariance matrices and tangent space projection, the study addresses challenges associated with high-dimensional and heterogeneous physiological data. The proposed approach presents a compelling feature extraction framework for analysing complex signal structures, demonstrating improved classification performance compared to traditional machine learning and deep learning models, as reflected in ablation studies.

The utilisation of Riemannian geometry provides a mathematically rigorous approach to analysing the non-Euclidean nature of covariance matrices derived from ECG signals. Unlike traditional methods that may overlook such geometric properties, the proposed framework ensures that the underlying signal structure is preserved during analysis. The results with higher diagnostic accuracy validate the efficacy of this approach. These findings underscores the importance of tailored methodologies for ECG analysis.

However, several limitations must be considered. The relatively limited diversity of the dataset suggests the need for validation across broader populations and some rare CHD presentations. Expanding data sources and leveraging Riemannian geometry techniques may further enhance model generalisability.

In this chapter, we also acknowledge the significant advancements achieved through existing Deep Learning (DL) models in the field of medical diagnostics. Numerous state-of-the-art architectures have demonstrated remarkable performance across various domains, including regression and prediction tasks. However, the application of these models in our context has been constrained by the limited resources available at the hospital. These constraints include hardware limitations, restricted computational power, and the unavailability of large annotated datasets typically required for training such models to their full potential.

Given these limitations, we consider to adopt a transfer learning approach by fine-tuning a pre-trained model, which is widely regarded as a resource-efficient alternative. Fine-tuning enables leveraging features learned from large-scale datasets and adapting them to specific tasks with smaller datasets. Despite this strategy, our experiments revealed that even comprehensive fine-tuning, adjusting all layers of the pre-trained models, did not yield optimal results. These findings suggest that the pre-trained models, while effective in their original domains, may not fully capture the complex nuances of our dataset, which is inherently characterised by unique patient demographics with congenital heart disease.

This observation highlights the challenges of applying generalised DL models to specific clinical data, especially when data availability is limited. It underscores the need for tailored approaches that account for specific clinical contexts. Future work could involve exploring hybrid methods, such as combining traditional feature engineering with deep learning techniques, or developing lighter, domain-specific models that are more adaptable to resource-constrained environments. Additionally, expanding collaborations to access larger and more diverse datasets may alleviate some of the limitations observed in this study.

Chapter 6

Predicting Cardiopulmonary Exercise Testing Outcomes in Congenital Heart Disease

6.1 Abstract

Cardiopulmonary Exercise Testing (CPET) provides a comprehensive assessment of functional capacity by measuring key physiological variables including oxygen consumption (VO_2), carbon dioxide production (VCO_2), and pulmonary ventilation (VE) during exercise. Previous research has established that parameters such as peak VO_2 and ratio of ventilation to carbon dioxide production (VE/VCO_2) serve as robust predictors of mortality risk in chronic heart failure patients. In this study, we leverage CPET variables as surrogate mortality endpoints for patients with Congenital Heart Disease (CHD). Our methodology began with digitising Electrocardiograms (ECGs) to obtain quantifiable waveforms and the core innovation of our approach lies in exploiting the Riemannian geometric properties of covariance matrices derived from 12-lead ECGs to develop robust regression and classification models. For this purpose, we also digitised a total of 595 CPET documents (all the test results are stored in image format) using Optical Character Recognition (OCR) techniques and linked all the data together. Through extensive ablation studies, we demonstrated how Riemannian embeddings enhanced by covariance augmentation

techniques in Riemannian space consistently produced superior predictive performance compared to conventional approaches.

6.2 Introduction

CPET is a specialised type of exercise to assess a patient's functional capacity and exercise ability. The information collected about the heart and lungs during the test is used to understand whether the physical response to exercise is normal or abnormal, as this is clinically important. CPET is also a valuable tool in patient assessment and management, as it provides an objective measure of exercise capacity which can be used to identify patients at risk. The findings are presented as a set of variables (like oxygen consumption) that can be compared with reference values obtained from a healthy population. According to a recent survey [219], CPET is employed in 68% of the hospital departments in the UK to help in the assessment of patients undergoing major procedures or surgeries.

Using the same set of patients (436 patients with CHD), we digitised all the available CPET PDF documents to obtain the required exercise variables. Since the CPET results were stored as images in a PDF document by the vendor, as in Figure 6.1, they were first converted to text using OCR packages in Python and then all exercise results were saved in a folder. Combining these important exercise variables with the ECG signals, we aim to develop robust regression and classification models for predicting the important outcomes on CPET, which are reliable indicators of mortality and morbidity in CHD patients. Summarising, the contributions of this chapter are as follows:

- **Application of Riemannian Geometry to ECG-Based Diagnosis:** An important contribution of this study is the utilisation of Riemannian geometric properties of covariance matrices derived from 12-lead ECGs, thereby enabling more structured and robust feature extraction. The chapter demonstrates how Riemannian embeddings improve classification accuracy, allowing for enhanced diagnostic insights in CHD.
- **Reiterating the Efficacy of Covariance Augmentation Methods:** The chapter employs the previously introduced covariance augmentation technique, which was presented in

earlier chapters, for the purpose of refining the feature representation and enhancing the generalisation of predictive models. By leveraging Riemannian space transformations, the proposed methodology enhances the diagnostic capabilities of machine learning models applied to ECG signals.

- **Experimental Validation of Tangent Space Representation for ECGs:** The chapter conducts extensive ablation experiments to demonstrate that projecting ECG signals onto tangent space significantly improves predictive performance for CHD patients. The chapter also establishes baseline models using vendor-provided ECG features, such as PR interval and QRS duration, to compare traditional feature-based models with advanced geometric learning techniques.

6.3 Related Work

6.3.1 CPET as outcome variables

CPET provides invaluable insights into a patient's cardiorespiratory fitness by measuring key variables like oxygen consumption, carbon dioxide production, heart rate and ventilatory parameters [220]. VO_2 is a crucial cardiopulmonary exercise variable that serves as a reliable predictor of mortality and morbidity in patients with CHD. This concept is rooted in the work of Hill et al., who introduced the idea that there is an individual exercise intensity at which VO_2 no longer increases [221]. Consequently, VO_2 peak represents the limit of cardiorespiratory capacity and is a strong indicator of whether individuals reach maximal conditions at the end of a CPET. The normal range for VO_2 peak is typically between 25-35 ml/kg/min. Studies have shown that lower peak oxygen consumption (VO_2 peak) values during CPET are associated with higher risks of mortality, especially in patients undergoing major surgeries [222].

On the other hand, VO_2 %pred is the percentage of the predicted maximum oxygen consumption based on factors such as age, sex, and height. This measure is useful for assessing ventilatory efficiency, and may indicate potential respiratory and/or cardiac limitations. The normal range for VO_2 %pred is typically within the range of 60 to 85 percent. Another important variable is the VE/VCO_2 ratio, which measures the relationship between pulmonary ventilation (VE) and carbon dioxide production (VCO_2). VCO_2 represents the maximum amount of carbon dioxide that can be produced during the exercise. Thus, VE/VCO_2 ratio can be summarised as the required ventilation to eliminate the CO_2 produced during the test. The normal range for VE/VCO_2 is typically between 20 and 30. Previous works have demonstrated that these variables are independent predictors of a high mortality risk in patients with Chronic Heart Failure (CHF) [223, 224]. Nanas et al. confirmed that the VE/VCO_2 slope is a strong, independent predictor of high mortality risk in CHF patients [223]. Another study demonstrated that the VE/VCO_2 slope is a significant predictor of cardiac-related hospitalisations in CHF patients [225]. VO_2 and VE/VCO_2 slope values have been correlated with long-term mortality risk in adults with CHD, with increased risk observed in cases of low VO_2, low heart rate reserve, and high VE/VCO_2 in non-cyanotic heart diseases [224].

Efficiency of VO_2 is also a good indicator in predicting mortality and morbidity in CHD patients and is calculated by dividing the VO_2 by the logarithm of the minute ventilation (VE), yielding a value of an Oxygen Uptake Efficiency (OUE). OUE measures how efficiently the body uses oxygen during the CPET and can be expressed as a slope (OUES) or a plateau (OUEP) depending on the relationship between VO_2 and VE [226]. Several studies have reported that OUE is lower in patients with CHD compared to healthy controls, and lower OUE is associated with increased risk of morbidity and mortality [220, 226–228]. Bongers et al. [229] conducted a study that is relevant to our work as it examines the same set of diagnoses, including Fontan, Tetralogy of Fallot (ToF), and Mustard. The study compared OUE in children with CHD and the results showed that OUE was significantly lower in children with CHD than in healthy children, with Fontan patients having the lowest OUE among the CHD groups.

Given the challenges of directly modelling mortality risk due to low prevalence, CPET offers more than just a pragmatic statistical solution. It provides a dynamic approach to patient monitoring that extends beyond mortality, which is an extreme outcome. By tracking continuous variables such as VO_2 peak and VE/VCO_2 ratio, clinicians can capture subtle changes in a patient's physiological functioning, allowing for early detection of declining health, personalised intervention strategies, and more proactive medical management. This approach enables healthcare providers to assess and predict cardiovascular risk with greater sensitivity, tracking the patient's functional capacity and potential health trajectories long before critical events might occur. Recent advancements also suggest the potential of predicting CPET outcomes from ECG data using machine learning algorithms, which could further streamline the assessment process [230].

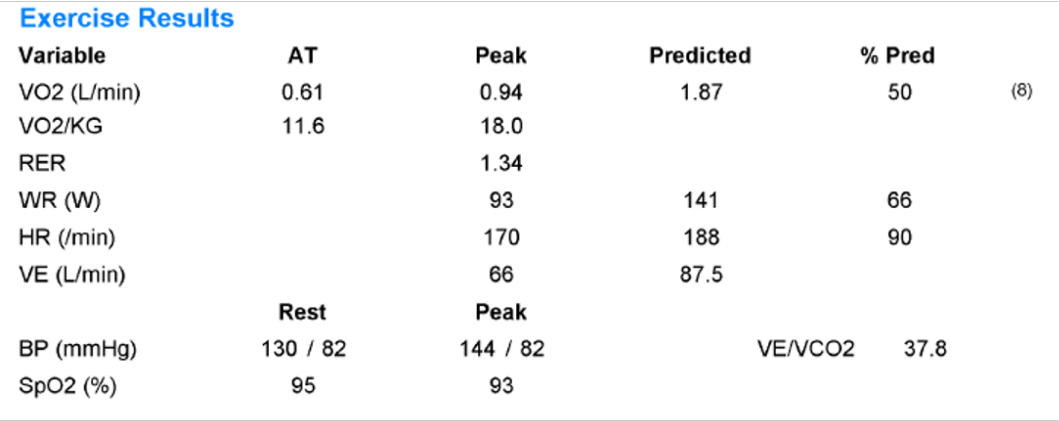
6.4 Methods

This study aims to utilise VO_2 and VE/VCO_2 during CPET as a means of predicting mortality in patients with CHD. These surrogate outcomes have demonstrated their efficacy as independent indicators of a high mortality risk in given demographic [231]. A covariance mixing regularisation technique for augmentation was employed for ECGs. Similar to mixup approach [232], this approach performs the interpolation on a Riemannian manifold with respect to the underlying covariance matrices. Building upon our previous work [233], efficiency of the augmentation technique was validated with ablation studies on a dataset of patients with CHD that was extremely small and imbalanced. This technique led to enhancements in both classification and regression problems, which were the best results.

6.4.1 Building a regression model

In our regression problem, the dependent variables are the selected CPET outcomes like VO_2 and VE/VCO_2, while the independent variables (or predictor variables) are the ECG intervals as well as ECG mappings into tangent space. These independent variables are fused to generate predictions and account for the variation in the dependent variable, objective. Two different regression models were used to compare the effect of fusing different input types: Support Vector Machine (SVM) and Logistic Regression (LR). We have done ablation studies with SVM model as well with baseline model LR. Same patient split strategy (Stratified Patient Leave-Out (SPLO)) was used to ensure consistency between the different results. VO_2 (peak and %pred) and VE/VCO_2, which are the most important outputs of CPET, are used as label.

During CPET, a breathing test is carried out in order to assess the flow of air in and out of the lungs. A mouthpiece is fitted in to measure the respiratory rate and levels of certain gases. The patient is asked to ride an exercise bike for the duration of the test, and the body's response to exercise is assessed. During the exercise, the heart rate is also monitored by small electrodes positioned on the chest. As can be seen in the Figure 6.1, all these values are reported after the exercise. For the blood pressure (BP) and Oxygen saturation (SPO2), both resting and peak values during the exercise were reported. For the other values, Oxygen uptake (VO2), Workload (WR) and Heart Rate (HR), peak, predicted and percentage of the predicted relative to the healthy reference values were reported.



Exercise Results

Variable	AT	Peak	Predicted	% Pred	
VO2 (L/min)	0.61	0.94	1.87	50	(8)
VO2/KG	11.6	18.0			
RER		1.34			
WR (W)		93	141	66	
HR (/min)		170	188	90	
VE (L/min)		66	87.5		
	Rest	**Peak**			
BP (mmHg)	130 / 82	144 / 82	VE/VCO2	37.8	
SpO2 (%)	95	93			

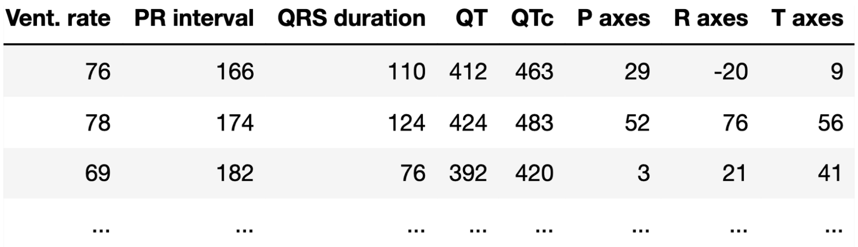
Figure 6.1: Some important variables from a CPET PDF document.

Different types of input data were used for the implemented models, such as ECG signals and tabular data containing calculated ECG parameters. Some of the calculated ECG parameters like QRS duration and PR interval can be found on the ECG PDF documents alongside with the ECG signals. Also, using the vendor manual, similar calculations performed on the extracted ECG signals to derive those vendor features.

ECG Data Extraction - Vendor Features

ECG measurements (vendor features) extracted from the ECG PDF documents alongside with the ECG signals. Similar calculations performed on the extracted ECG signals to derive the vendor features; ventricular rate, PR interval and QRS duration. Extracted vendor features for a group of patient ECG PDF documents can be seen in Figure 6.2.

Most patients have at least 1 exercise result (see Figure 6.1), although this is not the case for



Vent. rate	PR interval	QRS duration	QT	QTc	P axes	R axes	T axes
76	166	110	412	463	29	-20	9
78	174	124	424	483	52	76	56
69	182	76	392	420	3	21	41
...	...	...	...	...	...	...	...

Figure 6.2: Vendor features extracted from the ECG letters.

every patient. Currently, we have a total of 595 exercise documents (CPET), categorised as in Figure 6.3.

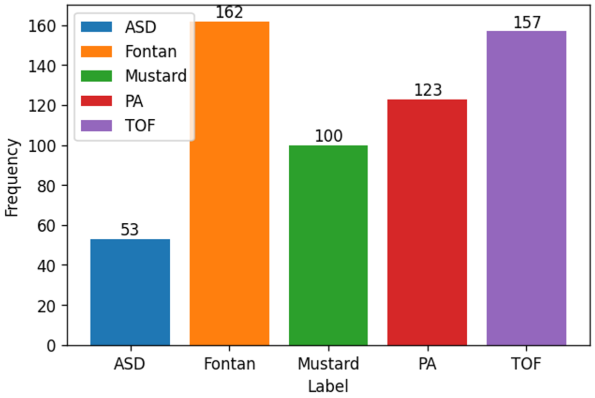


Figure 6.3: Histogram of CPET documents.

As we have a limited amount of data, we started to experiment with a rather simple Machine Learning (ML) model to get a baseline result. LR was trained on both the derived ECG measurements (vendor features) and the calculated ECG measurements separately to predict the CPET variables, VO_2 and VE/VCO_2. After the training, the models were tested on the corresponding testing set that does not include any data from the patients in the training set. Vendor list separated into two sets to match first set (*vendor*) with the calculated ECG measurements, as follows:

- **vendor:** PR interval, QRS duration, Vent. rate
- **vendor2:** QT, QTc, P axes, R axes, T axes

6.4.2 Enhancing feature extraction with Riemannian geometry

We also experimented on extracted ECG signals by mapping each patient to a tangent space. Overall process can be seen in Figure 6.4. We exploit tangent spaces to project covariance matrices and extract more coherent features. SVM was trained on the projected tangent space of the ECG aligned time-series data. ECG data were also aligned by using R peak points. After the training, the models were tested on the corresponding testing set that does not include any data from the patients in the training set.

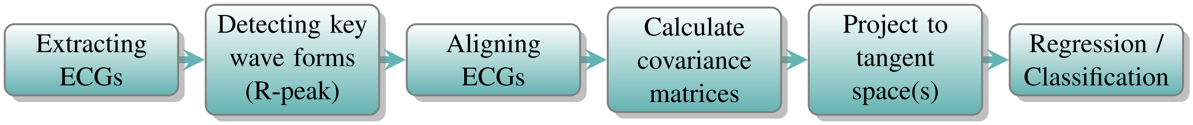

Extracting ECGs
Detecting key wave forms (R-peak)
Aligning ECGs
Calculate covariance matrices
Project to tangent space(s)
Regression / Classification

Figure 6.4: Overall process of using ECG signals.

As covariance matrices are Symmetric Positive Definite (SPD) matrices, they must be analysed in a Riemannian manifold. Riemannian distance metric is used to project covariance matrices onto tangent space while respecting their geometry. To achieve this, typically, the covariance matrices are projected onto a common tangent space based on Equation 5.2. This projection enhances the performance of models that depend on distance metrics between the sample covariance matrices, and it has been successful in processing high-dimensional neurophysiological data.

6.4.3 Regressing patient specific data out

As CPET assesses the capacity of an individual's cardiovascular system, a number of reference values have been proposed based on sex, age and Body Mass Index (BMI). Since these values are based on population and body characteristics, they were regressed out in order to avoid any potential bias in the model's predictive capabilities. The objective is to generate an output that is dependent completely on the desired input (such as ECG signals), irrespective of the characteristics of the patient.

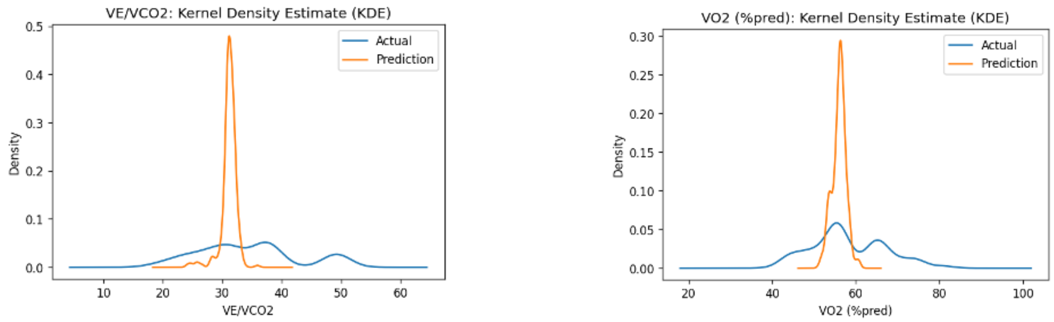
6.5 Results

First experiments, to establish a baseline, were carried out with ECG features using tabular data, such as PR interval, QRS duration and ventricular rate. Baseline model LR were trained using both vendor and calculated ECG features. For all the regression tasks, following evaluation metrics are reported; R2, adjusted R2, Root Mean Square Error (RMSE) and correlation coefficient.

Table 6.1: Regression results on tabular data (using **vendor** list), with baseline LR.

	R^2	**Adjusted** R^2	**RMSE**	**Correlation Coefficient**	**Predicted Label**	**Input Data**
Vendor ECG features	-0.080	-0.086	8.695	0.196	VE/VCO_2	PR interval, QRS duration, Vent. rate
Vendor ECG features	-0.093	-0.099	9.221	-0.001	VO_2 (%pred)	PR interval, QRS duration, Vent. rate
Vendor ECG features	-0.054	-0.060	0.444	0.051	VO_2 (peak)	PR interval, QRS duration, Vent. rate
Calculated ECG features	-0.171	-0.178	9.056	-0.025	VE/VCO_2	PR interval, QRS duration, Vent. rate
Calculated ECG features	-0.106	-0.112	9.274	-0.021	VO_2 (%pred)	PR interval, QRS duration, Vent. rate
Calculated ECG features	-0.058	-0.064	0.445	0.030	VO_2 (peak)	PR interval, QRS duration, Vent. rate

Table 6.1 summarises all the evaluation metrics for each prediction label, VE/VCO_2, VO_2 (%pred) and VO_2 (peak). There was no correlation between the prediction of the model and the actual exercise value. Therefore, using ECG parameters alone was not sufficient to create a good regression model. These ECG parameters are often reported in ECG PDF documents but are not sufficient to train a regression model. Kernel Density Estimations (KDEs) are also plotted in Figure 6.5 to show the distribution of the actual and predicted values. These plots show that all predictions fall within a narrow range, while actual values vary over a wide range.

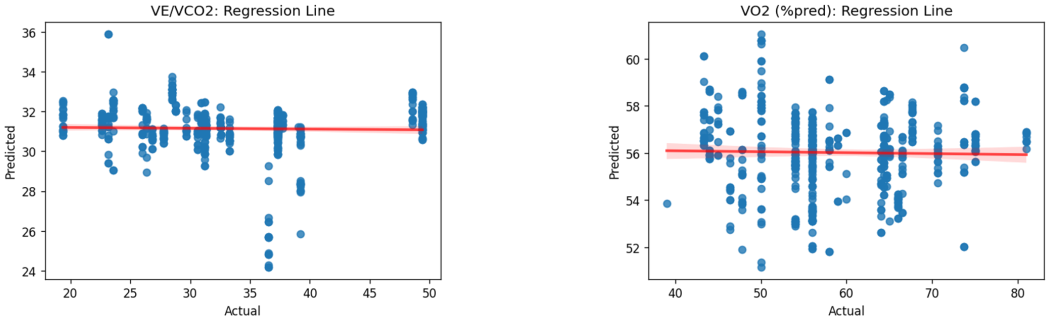


(a) KDE for VE/VCO_2 (b) KDE for VO_2 (%pred)

Figure 6.5: Kernel Density Estimations (KDEs) of the baseline model LR (using **vendor** list).

Regression line plots are also shared in Figure 6.6 to show the relationship between the actual and the predicted values. These plots show that all predictions fall within a narrow range, while actual values vary over a wide range. And hence, a flatter regression line is observed which is also reflected poorly in the results.

After these experiments on the vendor and the calculated ECG measurements, other list of vendor features (*vendor2*) are used in similar manner to get the regression results on. This time,



(a) Regression line for VE/VCO_2 (b) Regression line for VO_2 (%pred)

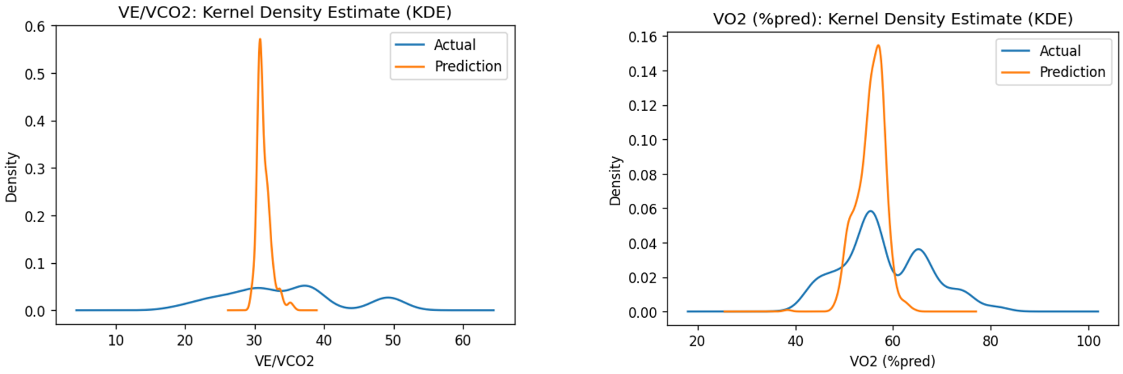
Figure 6.6: Regression line plots of the baseline model LR (using **vendor** list).

only the vendor features QT, QTc, P, R and T axes were used. Table 6.2 summarises all the evaluation metrics for each prediction label, VE/VCO_2, VO_2 (%pred) and VO_2 (peak). Some of the results achieved an improvement over those of the first set of vendor features (*vendor*), VO_2 (%pred) and VO_2 (peak). However, there was no evidence of improvement for VE/VCO_2, indicating a lack of consistency between the results.

Table 6.2: Regression results on tabular data (using **vendor2** list), with baseline LR.

	R^2	**Adjusted R^2**	**RMSE**	**Correlation Coefficient**	**Predicted Label**	**Input Data**
Vendor ECG features	-0.166	-0.176	9.034	-0.117	VE/VCO_2	QT, QTc, P axes, R axes, T axes
Vendor ECG features	0.009	0.001	8.774	0.328	VO_2 (%pred)	QT, QTc, P axes, R axes, T axes
Vendor ECG features	0.069	0.060	0.417	0.354	VO_2 (peak)	QT, QTc, P axes, R axes, T axes

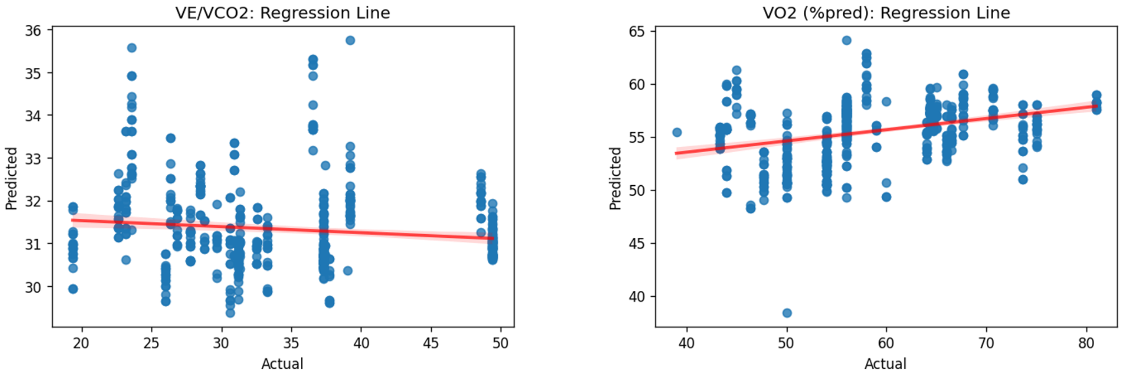
KDE are also plotted in Figure 6.7 to show the distribution of the actual and predicted values. Again, as in the first set of vendor features (*vendor*), these plots show that all predictions fall within a narrow range, while actual values vary over a wide range.



(a) KDE for VE/VCO_2 (b) KDE for VO_2 (%pred)

Figure 6.7: Kernel Density Estimations (KDEs) of the baseline model LR (using **vendor2** list).

In addition to KDE plots, regression line plots are also shared in Figure 6.8 as a way to visually depict the relationship between actual and predicted values. Again, as in the first set of vendor features (*vendor*), these plots show that all predictions fall within a narrow range, while actual values vary over a wide range. It also shows that a flatter regression line is observed, which is also reflected in the results.



(a) Regression line for VE/VCO_2 (b) Regression line for VO_2 (%pred)

Figure 6.8: Regression line plots of the baseline model LR (using **vendor2** list).

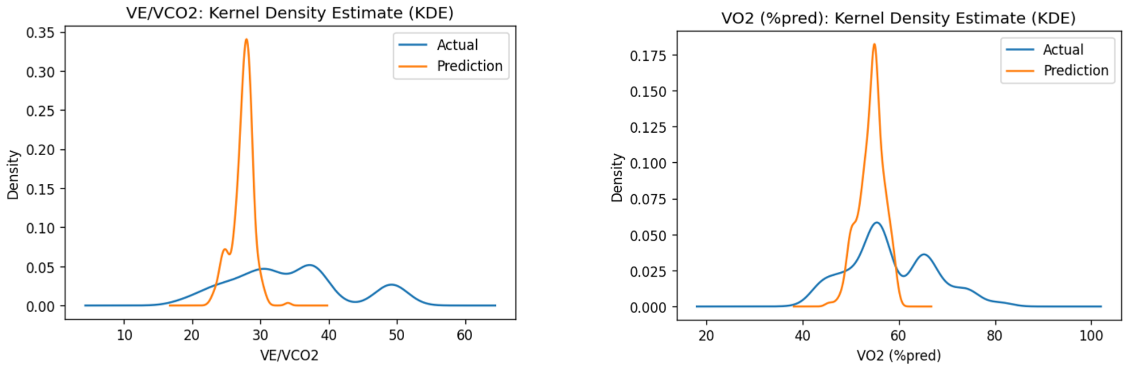
In addition to the baseline model (LR), SVM was also trained using vendor features to demonstrate that ECG parameters alone are insufficient for creating a reliable regression model. As illustrated in Table 6.3, a comprehensive overview of the evaluation metrics for each prediction label is provided, including VE/VCO_2, VO_2 (%pred), and VO_2 (peak). A small improvement in performance was observed across all metrics when compared to the initial experiment results presented in Table 6.1, with the exception of VO_2 (%pred) and VO_2 (peak) in Table 6.2.

Table 6.3: Regression results on tabular data, with SVM model.

	R^2	**Adjusted R^2**	**RMSE**	**Correlation Coefficient**	**Predicted Label**	**Input Data**
Vendor ECG features	-0.109	-0.118	10.613	0.247	VE/VCO_2	PR interval, QRS duration, Vent. rate
Vendor ECG features	-0.177	-0.183	9.567	0.188	VO_2 (%pred)	PR interval, QRS duration, Vent. rate
Vendor ECG features	0.024	0.018	0.427	0.243	VO_2 (peak)	PR interval, QRS duration, Vent. rate

KDE are also plotted in Figure 6.9 to show the distribution of the actual and predicted values. Again, as in the first set of vendor features (*vendor*), these plots show that all predictions fall within a narrow range, while actual values vary over a wide range.

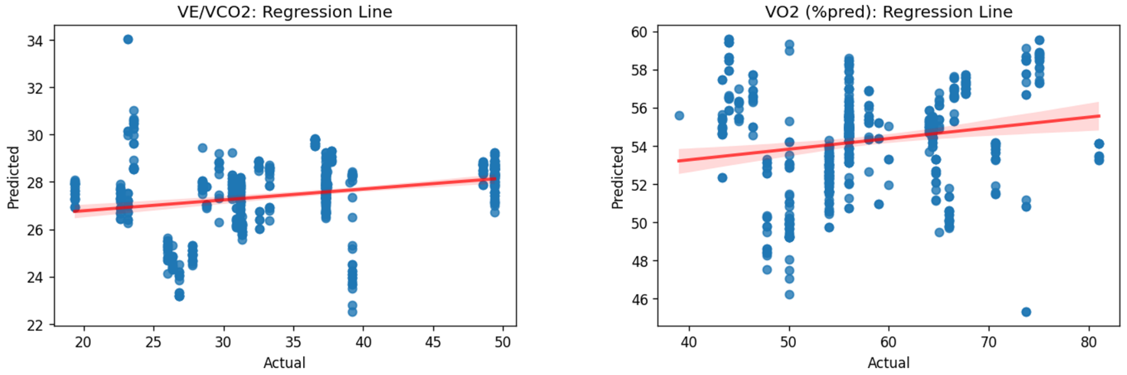
In addition to KDE plots, regression line plots are also shared in Figure 6.10 as a way to visually depict the relationship between actual and predicted values. Again, as in the first set of



(a) KDE for VE/VCO_2 (b) KDE for VO_2 (%pred)

Figure 6.9: Kernel Density Estimations (KDEs) of the SVM model (using **vendor** list).

vendor features (*vendor*), these plots show that all predictions fall within a narrow range, while actual values vary over a wide range. It also shows that a flatter regression line is observed, which is also reflected in the results.



(a) Regression line for VE/VCO_2 (b) Regression line for VO_2 (%pred)

Figure 6.10: Regression line plots of the SVM model (using **vendor** list).

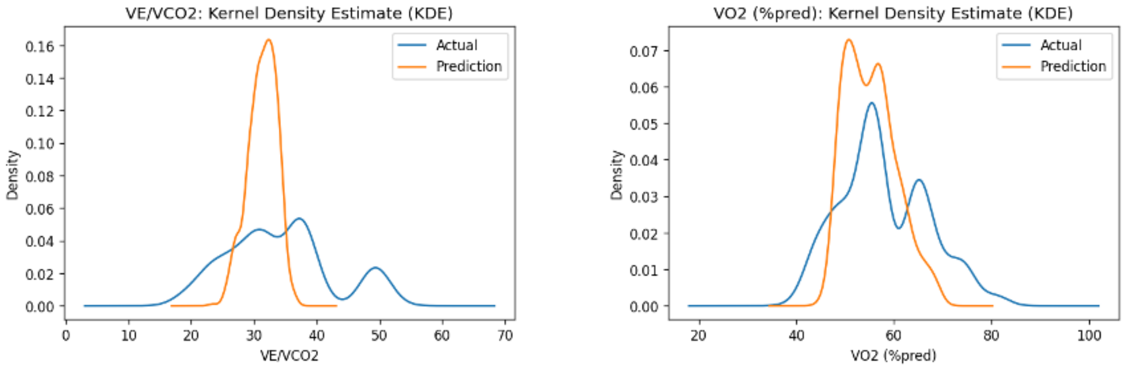
As there was no correlation between the prediction and the actual value, and all predictions fall within a narrow range, we explored tangent space mappings. Similar to our previous works, we first calculated covariance matrices of the ECG signals and then map all to the tangent space using Riemannian distance. It yielded better results and also a better density plot of the predictions, as can be seen in Figure 6.11.

Table 6.4 summarises all the evaluation metrics for each prediction label, VE/VCO_2, VO_2 (%pred) and VO_2 (peak). The correlation between the model's prediction and the actual exercise value was better, reaching 0.57 for VO_2 (peak). It was way better than the parameters that are often reported in ECG PDF documents, such as PR interval and QRS duration. KDEs plotted in

Table 6.4: Regression results on ECGs, with SVM model.

	R^2	**Adjusted** R^2	**RMSE**	**Correlation Coefficient**	**Predicted Label**	**Input Data**
Tangent Space	0.011	0.082	8.178	0.349	VE/VCO_2	ECGs
Tangent Space	0.218	0.046	7.892	0.550	VO_2 (%pred)	ECGs
Tangent Space	0.303	0.149	0.358	0.573	VO_2 (peak)	ECGs

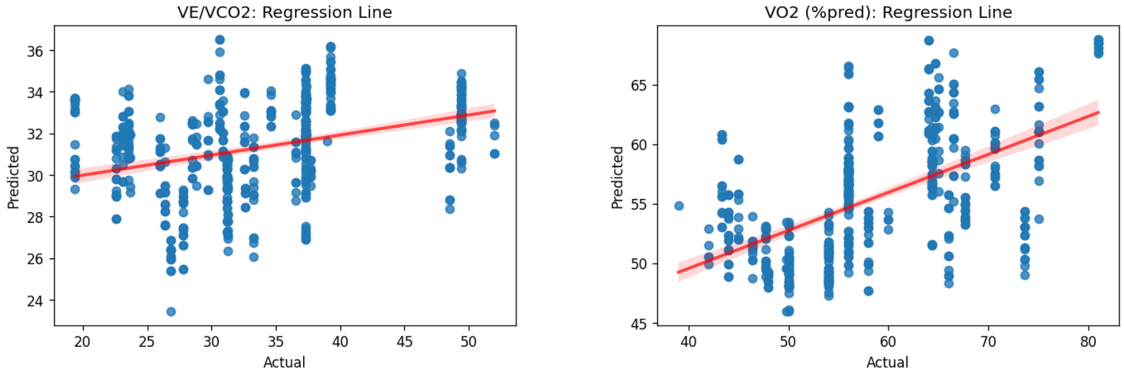
Figure 6.11 to show the distribution of the actual and predicted values. These plots show that all predictions are much closer to the distribution of actual exercise values.



(a) KDE for VE/VCO_2 (b) KDE for VO_2 (%pred)

Figure 6.11: Kernel Density Estimations (KDEs) of the SVM model (on ECGs).

Regression line plots are also shared in Figure 6.12 as a way to show the relationship between actual and predicted values. As there was an improvement over the previous results, this was also reflected in the regression lines. These plots show that all predictions are close to the range of actual values. It also shows that a more sloped regression line is observed, which is also reflected in the results.



(a) Regression line for VE/VCO2 (b) Regression line for VO_2 (%pred)

Figure 6.12: Regression line plots of the SVM model (on ECGs).

6.5.1 From regression to classification

In order to transform the regression problem into a relatively simple classification problem, clinical groups were determined for each label of VE/VCO_2, VO_2 (%pred) and VO_2 (peak) with data distribution of each class as shown in Figure 6.13, with the grouping boundaries are indicated by red lines. The VE/VCO_2 and VO_2 (peak) variables were grouped into two categories, whereas VO_2 (%pred) was grouped into four categories. We have done ablation studies with a Support Vector Machine (SVM) model and evaluated the model performance using multiple metrics that involve the accuracy, AUC and F1 score.

- VE/VCO_2
    - Less than 35
    - Above 35
- VO_2 (peak)
    - Less than 1.2
    - Above 1.2
- VO_2 (%pred)
    - Less than 50
    - 50 to 65
    - 65 to 75
    - Above 75

Initial experiments were carried out without any augmentation in order to establish a baseline. However, it was observed that utilising vendor features derived from the ECG letters as inputs to the SVM model did not yield optimal results, as previously reported in Table 6.3 on the regression section. So, without utilising vendor features, the baseline SVM model was employed with two separate inputs: ECGs alone, and ECGs with covariance augmentations. A notable improvement (up to 5% on accuracy and AUC, 4% on F1 score) was observed when the

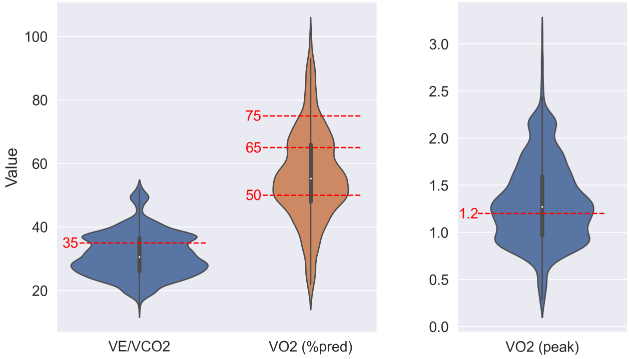


Figure 6.13: Data distribution of each class, with CPET groups.

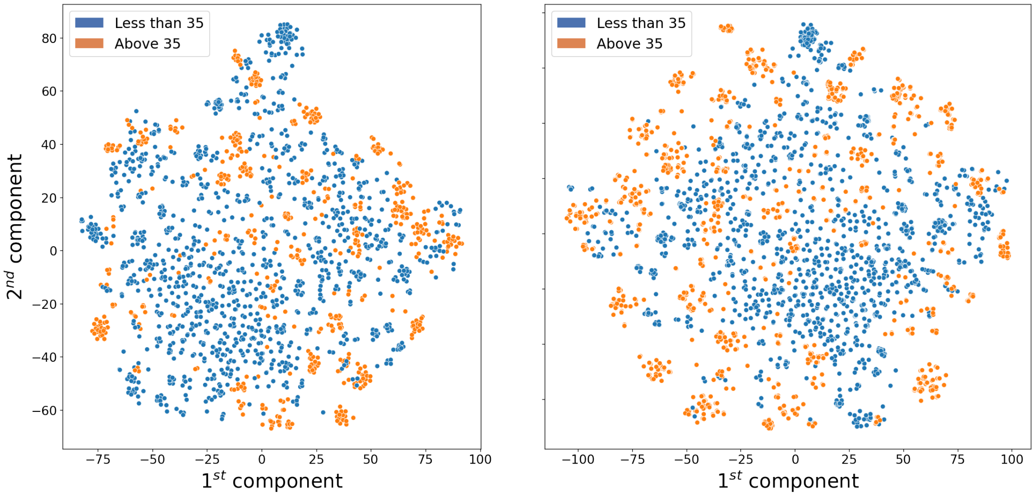
ECG data combined with the covariance augmentations, which proved to be advantageous in terms of the classification metrics.

Table 6.5: Classification results using regression groups with SVM model.

Predicted Label	Input	Augmentation	Accuracy	AUC	F1 macro
VE/VCO_2	ECGs		0.672 ± 0.212	0.577 ± 0.141	0.534 ± 0.181
VE/VCO_2	ECGs	covariance	**0.704 ± 0.195**	**0.622 ± 0.183**	**0.572 ± 0.197**
VO_2 (%pred)	ECGs		0.302 ± 0.163	0.506 ± 0.225	0.229 ± 0.127
VO_2 (%pred)	ECGs	covariance	**0.324 ± 0.132**	**0.520 ± 0.225**	**0.236 ± 0.106**
VO_2 (peak)	ECGs		0.614 ± 0.137	0.585 ± 0.162	0.535 ± 0.167
VO_2 (peak)	ECGs	covariance	**0.657 ± 0.150**	**0.620 ± 0.126**	**0.553 ± 0.152**

Building upon our previous work [233], we employed a sophisticated augmentation technique using Riemannian geometry on covariance matrices. The weighted Riemannian mean is computed by minimising the sum of squared Riemannian distances to the given matrices, as formalised in Equation 5.3. A comprehensive summary of all results is provided in Table 6.5, with the most optimal results indicated in bold. It was observed that applying covariance augmentations only to ECGs was able to increase accuracy, AUC and the F1 score for all classes by up to 5%. Our approach to use a sophisticated augmentation technique yielded the best results. The utilisation of Riemannian geometry augmentations on the covariance matrices of ECGs has been shown to produce features that are more coherent, as evidenced by the regression problems previously discussed.

Furthermore, Figure 6.15 illustrates the learning curves for the classification model, demonstrating how increasing the number of patient samples improves performance. It demonstrates a clear trend: as the sample size (number of patients) increases, the performance of the model

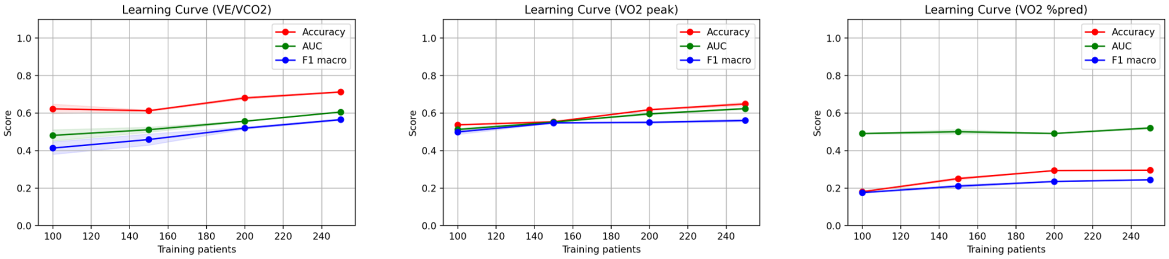


(a) ECG without augmentations

(b) ECG with augmentations

Figure 6.14: Comparison of t-SNE visualizations on the tangent space.

improves significantly. Initially, with a smaller set of patients, the model exhibited high variance which resulted in inconsistent predictions. However, as more patient data was incorporated, accuracy, AUC, and F1 score steadily improved. Figure 6.15a shows the impact of sample size on model performance for VE/VCO_2, illustrating that as the number of patient samples increased from 100 to 250, accuracy rose from 0.62 to 0.71, AUC improved from 0.48 to 0.62, and F1 macro score increased from 0.41 to 0.57. These observations suggest that an expanded dataset improves the model's ability to generalise effectively, leading to a reduction in the probability of overfitting and an enhancement in its predictive robustness.



(a) VE/VCO_2

(b) VO_2 (peak)

(c) VO_2 (%pred)

Figure 6.15: Comparison of learning curves for the classification model.

Moreover, Figure 6.16 presents the projected model scores for VE/VCO_2, providing accuracy, AUC and the F1 score, across a range of sample sizes up to 400 patients. This rough analysis reveals that with 300 patient samples, accuracy reaches at 0.81, AUC reaches 0.61, and

F1 score reaches at 0.62. As the dataset expands to 400 patient samples, accuracy improves to 0.89, AUC rises to 0.68, and F1 score reaches 0.73. This trend underscores the advantage of incorporating more patient data, as it enables the model to learn more complex patterns and enhance prediction performance. These findings underscore the importance of continuous data collection and model refinement to optimise predictive accuracy in healthcare applications.

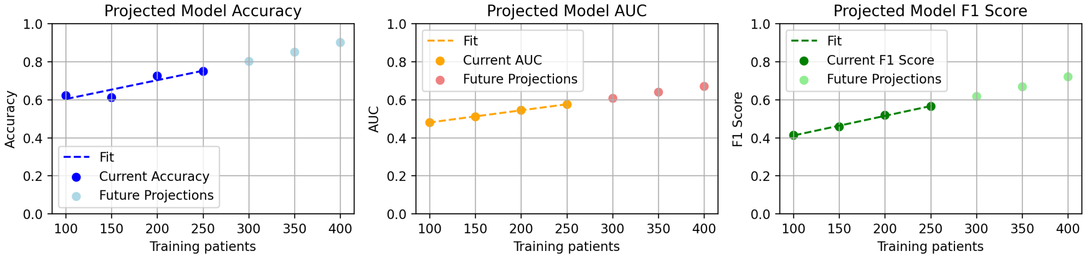


Figure 6.16: Learning curve projection for the classification model on VE/VCO_2.

6.6 Discussion

Medical data analysis, particularly concerning CHD, poses challenges due to the inherent complexity and imbalanced nature of the data. Large machine learning models often struggle in such contexts, primarily due to their sensitivity to data distribution and inefficiency in learning from datasets with limited examples. Our approach leverages the geometric properties of Riemannian spaces, offering a more robust and discriminative feature space for machine learning models. By utilising the non-Euclidean nature of the data, our method captures intrinsic geometrical structures that are frequently overlooked by conventional methods. Our study demonstrated promising results in predicting surrogate mortality for patients with congenital heart disease. The proposed projection of augmented covariance matrices to Riemannian spaces significantly improved performance with small and extremely imbalanced 12-lead ECG data. The experimental results support the hypothesis that the proposed solution is effective for both regression and classification problems.

This chapter explores the importance of geometric learning and augmentation techniques in predicting cardiopulmonary exercise testing (CPET) outcomes in patients with congenital heart disease (CHD). Our results demonstrate that traditional ECG parameters alone are insufficient for accurately predicting critical outcomes such as oxygen consumption (VO_2) and ventilatory efficiency (VE/VCO_2). This results align with the previous literature indicating that a multifactorial approach is necessary to capture the complexity of cardiovascular responses in CHD patients[146].

The application of Riemannian geometry for feature extraction from ECG signals represents a significant advancement in our methodology. By projecting covariance matrices onto tangent spaces, we were able to derive features that capture the underlying geometric structure of the data better. This approach improved the correlation between predicted and actual CPET outcomes. The enhanced performance of our models when using tangent space mappings suggests that traditional linear methods may overlook critical relationships in high-dimensional data. This finding is particularly relevant in the context of CHD, where the physiological responses to exercise can demonstrate complexity.

While the results of this study are promising, certain limitations should be acknowledged, including the relatively small sample size which may affect the generalisability of the findings. Future research should prioritise validating the proposed models in larger and more diverse cohorts to ensure their applicability across different patient populations. Furthermore, although this study focused primarily on some of the CPET outcomes, examining additional CPET variables and their interactions could provide further insights in CHD patients. Additionally, reliance on historical clinical data may introduce potential biases related to changes in treatment protocols or advancements in medical technology over time. To maintain the relevance and accuracy of the models in clinical practice, continuous updates to both the dataset and the framework may be required.

In conclusion, this chapter highlights the potential of leveraging geometrical techniques to enhance the prediction of CPET outcomes in congenital heart disease. By employing Riemannian based feature extraction methods, we can capture the intrinsic geometry of the data and thus improve our understanding of exercise capacity and its implications for patient health.

Chapter 7

Integrating Information from Clinical Letters with ECGs

7.1 Abstract

Cardiopulmonary Exercise Testing (CPET) provides a comprehensive assessment of functional capacity by measuring key physiological variables including oxygen consumption (VO_2), carbon dioxide production (VCO_2), and pulmonary ventilation (VE) during exercise. Previous research has established that parameters such as peak VO_2 and ratio of ventilation to carbon dioxide production (VE/VCO_2) ratio serve as robust predictors of mortality risk in chronic heart failure patients. In this study, we leverage CPET variables as surrogate mortality endpoints for patients with Congenital Heart Disease (CHD). To our knowledge, this represents the first successful implementation of an advanced machine learning approach that predicts CPET outcomes by integrating Electrocardiograms (ECGs) with information derived from clinical letters. Our methodology began with extracting unstructured patient information—including intervention history, diagnoses, and medication regimens—from clinical letters using natural language processing techniques, organising this data into a structured database. We then digitised ECGs to obtain quantifiable waveforms and established comprehensive data linkages. The core innovation of our approach lies in exploiting the Riemannian geometric properties of covariance matrices derived from both 12-lead ECGs and clinical text data to develop robust regression and

classification models. Through extensive ablation studies, we demonstrated that the integration of ECG signals with clinical documentation, enhanced by covariance augmentation techniques in Riemannian space, consistently produced superior predictive performance compared to conventional approaches.

7.2 Introduction

CHD represents the most frequent form of congenital anomaly, with an incidence rate of 1 in 180 births in the United Kingdom. With the advancements in surgical techniques, the overall survival rate has increased to over 94%, providing more individuals with the opportunity to reach adult life. Similar improvements have been observed in other developed countries, with the prevalence of adult CHD now documented at approximately 4 per 1000 adults. This highlights the need for the development of clinical databases capable of identifying, characterising and monitoring local and regional CHD patient populations.

A number of developed countries have already invested a significant period of time in establishing national CHD databases, which have subsequently become an invaluable resource for a range of purposes including the planning of healthcare services, facilitating research, and the monitoring of trends in outcomes. Notable databases include the CONgenital CORvitia (CONCOR) registries, the BELgian COngenital heart disease Database combining Administrative and Clinical data (BELCODAC), the SWEDish registry of CONgenital heart disease (SWEDCON), and the National Institute for Cardiovascular Outcomes Research (NICOR) in the United Kingdom [234, 235]. More recently, the Congenital Heart Initiative (CHI) in the United States has also been established. Despite their efficacy, these databases are labour-intensive, requiring substantial time and resources, and may be prone to inaccuracies.

In Scotland, the care of adults diagnosed with CHD is centralised and commissioned by the National Service Division of NHS Scotland, which is hosted by the Golden Jubilee National Hospital in Glasgow. The Scottish Adult Congenital Cardiac Service (SACCS) is currently responsible for the care of an estimated 3,000 patients diagnosed with moderate or complex forms of CHD, in addition to a further 7,000 to 8,000 patients with less severe cases of CHD, who are seen less frequently. The latter group with less severe cases of CHD is primarily cared for by cardiologists at local cardiac centres. The objective was to create a clinical database with a minimal dataset of approximately 1400 patients. This would facilitate the accurate identification and analysis of all SACCS CHD patients' diagnoses, interventions, medications, and demographic data. Summarising, the contributions of this chapter are as follows:

- **Integration of Multi-Modal Data:** This chapter presents a novel approach for predicting Cardiopulmonary Exercise Testing (CPET) outcomes in Congenital Heart Disease (CHD), by integrating both structured and unstructured patient data. The study incorporates ECG signals with clinical letters using Natural Language Processing (NLP) techniques, transforming unstructured text into meaningful data representations. This integration has been shown to enhance predictive performance by converting heterogeneous clinical records into structured formats, thereby enabling more effective analysis.

- **Application of Riemannian Geometry:** Another contribution of this study is the utilisation of Riemannian geometric properties of covariance matrices derived from both 12-lead ECGs and clinical information, thereby enabling more structured and robust feature extraction. The chapter demonstrates the advantages of non-Euclidean feature spaces and how Riemannian embeddings improve classification accuracy, allowing for enhanced diagnostic insights in CHD.

- **Reiterating the Efficacy of Covariance Augmentation Methods:** The chapter employs the previously introduced covariance augmentation technique, which was presented in earlier chapters, for the purpose of refining the feature representation and enhancing the generalisation of predictive models. By leveraging Riemannian space transformations, the proposed methodology enhances the diagnostic capabilities of machine learning models.

- **Experimental Validation of Tangent Space Representation:** The chapter conducts extensive ablation experiments to demonstrate that projecting both ECG signals and clinical information onto tangent space significantly improves predictive performance for CHD patients. The integration of multi-modal patient data with advanced geometric learning methods has been demonstrated to produce superior results in terms of model performance, both for regression and classification.

7.3 Methods

7.3.1 Data extraction from clinical letters

First draft of the data extraction algorithm is experimented on a small batch of 60 randomly selected letters, to capture a generic image of different writing styles of the cardiology consultants. Total of 60 letters, 12 letters per cardiology consultant, were randomly selected to develop an algorithm that is sensitive to different writing styles. The algorithm first converts all the PDF letters into text files and stores them into a secondary folder, to make the letters easy to process and work on. Next, using key words appearing in the text and regular expression techniques, the algorithm extracts demographic and clinical information including hospital number, date of birth, name, sex, diagnosis, intervention, medication, clinic date, postal code, health board, height, weight, and arrhythmia class of each patient. The extracted data is then preprocessed to remove all the unnecessary tokens such as unwanted characters, bullet points, double spaces. The letters that do not contain the diagnosis, intervention and medication information are marked as incomplete letters and then excluded from the data extraction. At the final step, after extracting all the information, European Society of Cardiology (ESC) adult CHD lesion complexity classification is assigned by matching keywords in the diagnosis section of the letter. These keywords are populated using a permutation of different cardiac condition names for each classification. ESC classification of mild, moderate, and great complexity of CHD was used for the complexity assignment. Medication names were obtained from the British National Formulary (BNF) published list of medication and matched for all medications listed in the patient letter. Intervention names were taken from published CHD treatment lists.

7.3.2 Data processing and populating the database

In total, 17 clinical variables were extracted from the letters into columns of a structured table as summarised in Table 7.1. These 17 columns of data are stored as text file and then populated to a database (NoSQL/MongoDB). All the required checks were made to ensure that the extracted information stored correctly in the populated database without any problems. For any possible changes on the database after the first initialization, it reads the text output created by the data extraction algorithm and then crosschecks the current database to identify whether the patient is new or not. It identifies new patients if the hospital number is not already stored on the database and updates information for patients already present in the database. With such control mechanism, it allows only the single latest version of the patient information to be present on the database.

Table 7.1: List of extracted variables from the clinical letters.

Name	Type	Explanation
CHI Number	String	The unique identifier of the patient (Community Health Index)
Name	String	The full name of the patient
DOB	String	The date of birth of the patient
Diagnosis	String	All the text in the diagnosis section of the patient’s clinical letter
Diagnosis List	String	List of extracted diagnoses
Intervention	String	All the text in the intervention section of the patient’s clinical letter
Intervention List	String	List of extracted interventions
Medication	String	All the text in the medication section of the patient’s clinical letter
Medication List	String	List of extracted medications
ESC Classification	String	ESC classification as mild, moderate and great complexity of CHD
Arrhythmia	String	Arrhythmia class of the patient
Clinic Date	Date	The date of visit of the patient.
Health Board	String	Health board where the postcode is located
Postcode	String	Postcode of the patient’s usual place of residence
Sex	String	Sex at birth
Height	Integer	Height of the patient
Weight	Integer	Weight of the patient

After all the steps, the populated database is full of updated information and ready to be inspected by the clinicians and to be used for various data analysis and data visualization tasks. Total of 1409 clinical letters were uploaded to the database, from different cardiologists over a span of 2-years’ time. All the steps, data extraction and populating a database, are performed automatically so that any user intervention is not necessary and clinicians do not have to go through each patients’ clinical letter manually.

In addition to clinical data, 595 cardiopulmonary exercise test (CPET) documents from the same patient set were digitised using Optical Character Recognition (OCR) techniques, with all data linked through unique patient identifier numbers. Typically, each patient may have 1 to 3 CPET documents corresponding to different testing sessions conducted over time, depending on the patient's follow-up schedule and clinical needs. However, it is important to note that the availability of CPET documents depends on clinical circumstances, as well as the assessment of the necessity and feasibility of conducting such testing for each individual patient. Although the cohort consists of 436 patients, the available CPET documents are restricted to a subset of 258 patients who have undergone the testing.

For the purpose of this study, we used the CPET document that was closest in date to the ECG data for each patient. This approach ensures that the data reflects the most relevant physiological status corresponding to the ECG measurements, providing a more accurate and aligned analysis. Utilising the CPET document closest to the ECG date, we aim to minimise potential discrepancies that could arise from changes in the patient's condition over time. Key variables such as VO_2 and VE/VCO_2 were selected as surrogate outcomes for the assessment of mortality risk. Furthermore, a similar selection approach was also employed for the clinical letters to gather the closest letter in date to the ECG data for each patient, to ensure that the information was as relevant as possible. In total, 17 clinical variables were extracted from the letters into columns of a structured table. With the exclusion of demographic data, list of extracted diagnoses (Diagnosis List), interventions (Intervention List) and medications (Medication List) were combined to gather important information about history of the patients. As illustrated in Figure 7.1, a comprehensive overview of the entire data extraction process is provided.

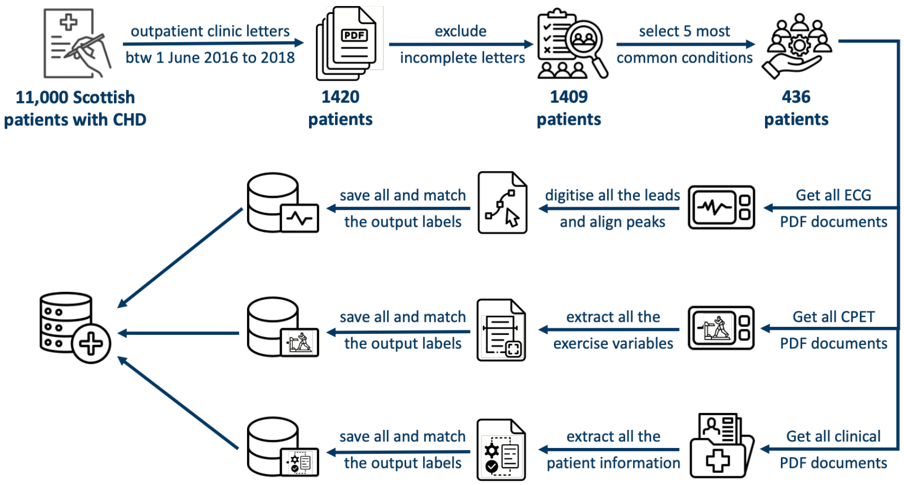


Figure 7.1: Summary of the data extraction steps.

7.3.3 Dimensionality reduction for visualization

t-distributed Stochastic Neighbor Embedding (t-SNE) is a data visualization technique that simplifies more complex data into 2 or 3 dimensions that are easier to visualise and interpret. As t-SNE projects the data into a low-dimensional embedding space, it uses the distance between each data point as a way of calculation of the similarity. Therefore, if 2 patients' clinical characteristics are close in high dimensional space, they are going to be assigned with a high probability of being neighbour points and located in close proximity of each other. The aim of these embeddings is to approximate the data distributions as closely as possible with the original high-dimensional dataset. As a pilot exploration of the database, we applied unsupervised learning to the following domains including diagnosis, intervention, medication, and ESC classification lists. In this way, we aim to qualitatively assess the data-driven low-dimensional embeddings that form notional clusters of patients with relation to our knowledge about the data.

The aim of the low dimensional embedding is to approximate the distributions as close as possible with the original, high-dimensional data. Mathematically, the similarity between patients $\mathbf{x}_i$ and $\mathbf{x}_j$ is the conditional probability $\mathbf{p}_{i/j}$ with

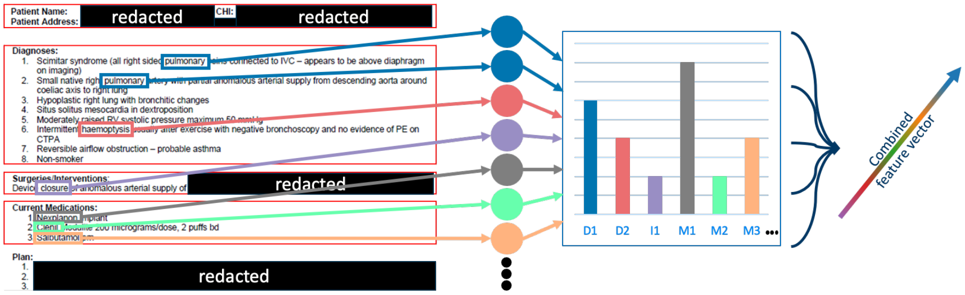
$$p_{i/j} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)} \tag{7.1}$$

An important parameter in t-SNE is the perplexity, that relates to the variance of the Gaussian distribution, and it can be interpreted as a smoothness factor of the number of neighbour points. It is reasonable to conclude that changes in perplexity value will consequently lead to modifications in the t-SNE visualisation. Thus, if the perplexity values is small, there will be a minimal number of pairs that exhibit any degree of attraction, resulting in an embedding that is relatively a bubble-like, rounded shape. This phenomenon can be observed in the embeddings depicted in the first two columns of the Figure 7.4. Conversely, if the perplexity is high, clusters will tend to contract into more compact structures. In Figure 7.4, we show how perplexity affects the t-SNE results and also highlight with a red rectangle the results we choose to demonstrate in Figure 7.3.

Moreover, k-means clustering was employed to facilitate a comparison between the data-driven clusters and the default grouping based on disease complexity. The optimal number of clusters was determined through a process of 100 bootstrap repetitions of t-SNE, followed by k-means clustering. The degree of agreement between the k-means clusters and the disease complexity grouping was estimated based on the Jaccard Similarity Index (JSI), which is the ratio of the intersection of two clusters divided by their union. JSI ranges from 0 to 1, with higher values indicating greater similarity between two clusters.

7.3.4 Integrating the information with ECGs

From the initially extracted 17 clinical variables in Table 7.1, we focused on three key categories of patient history: diagnoses, interventions, and medications, excluding demographic data. To integrate this clinical history information with the ECG signals in tangent space, we developed a comprehensive text preprocessing pipeline. The pipeline transformed clinical letters into structured numerical representations through the following steps. First, we converted the textual content into a sparse matrix representation, where each row corresponds to a unique clinical letter and each column represents a distinct term found across all letters. This matrix captured term frequencies, quantifying the occurrence patterns of medical terminology within each document. We then concatenated these word-frequency matrices with the tangent space mapping matrices derived from ECG signals to create unified input features for our model. This fusion approach enabled us to simultaneously leverage both the semantic patterns present in clinical documentation and the geometric properties inherent in ECG signals, providing a more comprehensive representation of each patient's cardiac condition. The combined feature space preserved both key features of clinical assessments and the characteristics of electrical cardiac activity.



Patient Name:
Patient Address: redacted **CHI:** redacted

Diagnoses:
1. Scimitar syndrome (all right sided pulmonary veins connected to IVC – appears to be above diaphragm on imaging)
2. Small native right pulmonary artery with partial anomalous arterial supply from descending aorta around coeliac axis to right lung
3. Hypoplastic right lung with bronchitic changes
4. Situs solitus mesocardia in dextroposition
5. Moderately raised RV systolic pressure maximum 50 mmHg
6. Intermittent haemoptysis usually after exercise with negative bronchoscopy and no evidence of PE on CTPA
7. Reversible airflow obstruction – probable asthma
8. Non-smoker

Surgeries/Interventions:
Device closure of anomalous arterial supply of redacted

Current Medications:
1. Nexplanon implant
2. Clenil Modulite 200 micrograms/dose, 2 puffs bd
3. Salbutamol prn

Plan:
1.
2.
3.
redacted

Figure 7.2: Text to matrix transformation using CountVectorizer.

For the augmentation of clinical data, a similar weighted mean approach in Equation 5.3 is applied to the vectors that contain clinical letter information, without any distance function. The two vectors were averaged according to the specified weight α. All the augmentation is applied in the same method in order to maintain consistency in the value of α for both ECG and clinical data augmentations of the same data. To qualitatively validate the effectiveness of this approach, we employed t-SNE visualizations [213] on the tangent space, comparing representations using only original data and those combined with the mixed data, as illustrated in Figure 7.11.

7.4 Results

A total of 1409 patient letters were analysed from a cohort of approximately 11,000 Scottish patients with CHD, of whom 3000 were under regular follow-up. The mean age of the cohort was 35.4 years (standard deviation [SD] 14.6; interquartile range [IQR] 45–24), with 737 patients (52.3%) identified as male and 672 patients (47.6%) as female. 284 patients (20.1%) exhibited mild complexity, 698 patients (49.5%) exhibited moderate complexity, and 369 patients (26.1%) exhibited severe complexity according to the ESC classification. 43 patients had no classifiable CHD (3.1%), such as anomalous right subclavian artery. And 8 patients (0.5%) did not have CHD, such as carcinoid disease. The completeness of data extraction was tested for each variable and is summarised in Table 7.2.

Figure 7.3 illustrates the t-SNE embeddings within the domains of diagnosis, intervention, medication, and ESC classification lists. The mean JSI across 100 bootstrap repetitions was 0.28, 0.2, 0.14, and 0.97 for diagnosis, intervention, medication, and ESC classification, respectively. These values align with the qualitative results presented in Figure 7.5.

The selection of optimal t-SNE parameters and their impact on the embeddings are illustrated in Figure 7.4. In Figure 7.4, we show how perplexity affects the t-SNE results and also highlight with a red rectangle the results we choose to demonstrate in Figure 7.3. In Figure 7.4, each column corresponds to a perplexity parameter of 3, 5, 30, 50 and 100, respectively. Each row corresponds to diagnosis, intervention, medication and ESC classification embeddings, respectively.

Table 7.2: Completeness of data extraction for all variables.

Attribute	**Completeness (%)**
CHI Number	100
Name	100
DOB	100
Diagnosis	100
Diagnosis List	96.95
Intervention	90.84
Intervention List	81.62
Medication	95.39
Medication List	91.55
ESC Classification	96.95
Arrhythmia	13.77
Clinic Date	100
Health Board	100
Postcode	100
Sex	100
Height	53.16
Weight	47.13

Also, in Figure 7.5, the low-dimensional embeddings are coloured according to disease complexity and are compared with data-driven clustering (k-means). First row shows the stability score of k-means clustering based on 100 bootstrap repetitions. The middle row shows the embeddings colored based on the data-driven (k-means) clustering and the bottom row shows the same low-dimensional embeddings coloured according to the disease complexity. The low-dimensional embeddings have been labelled according to the a priori data by human design, and

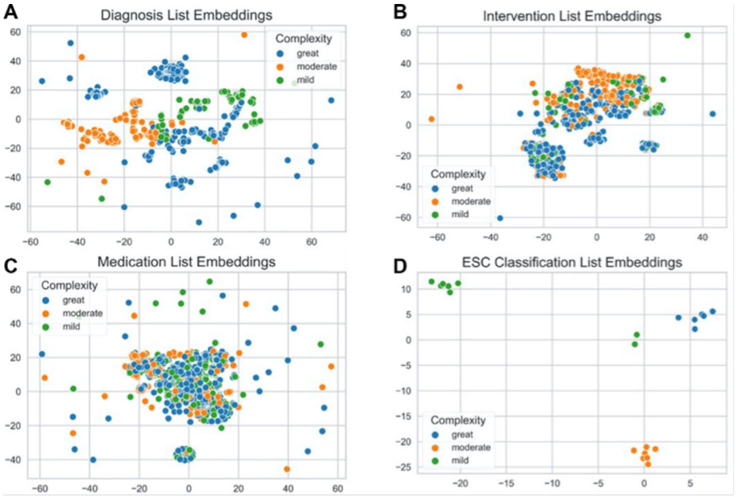


Figure 7.3: t-SNE dimensionality reduction of diagnosis, intervention, medication, and ESC classification.

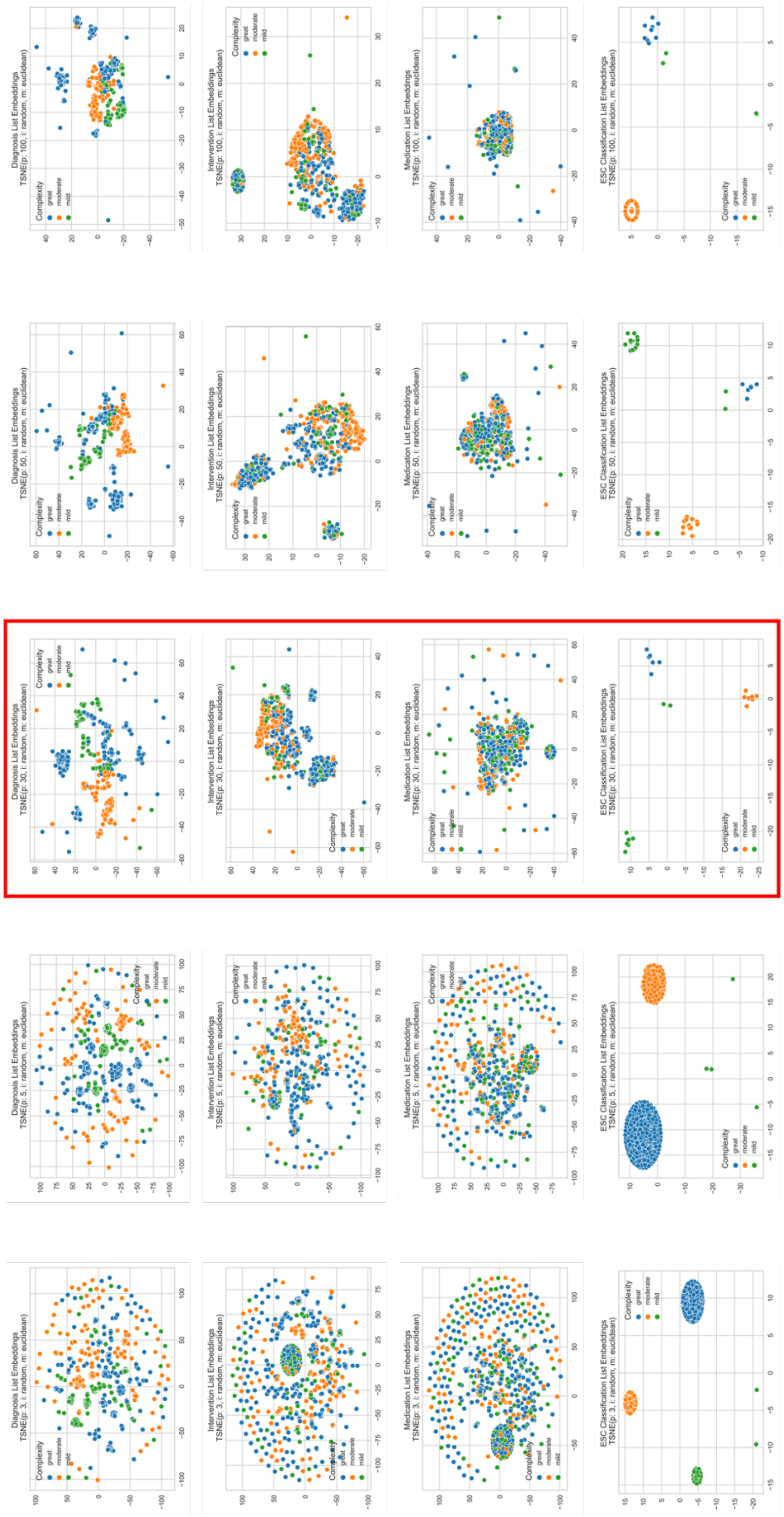


Figure 7.4: Selection of t-SNE initialisation parameters.

thus represent the diagnostic components of the data set, classified as either mild, moderate or severe complexity. The corresponding ESC classification lists are indicated by the colours blue, orange, and green, respectively. The results indicate a direct link between disease complexity and ESC-based embedding, which is also reflected in the almost perfect agreement between the embeddings and the classification. However, there is no evidence that the diagnostic embeddings overlap with the diagnostic classification. Furthermore, the diagnostic embeddings resemble disease complexity more closely than the interventions and the medication embeddings, as demonstrated quantitatively in Figure 7.3.

A random sample comprising 5% of patient letters was subjected to manual examination to evaluate the accuracy of the data extraction algorithm as applied to diagnosis, intervention and medications. The dataset was evaluated manually by an independent skilled observer (F.C.J.), a final-year medical student working in the CHD department, with regard to the accuracy in extracting diagnostic, interventional and medication-related data. The data extraction algorithm was able to accurately identify diagnoses, intervention, and medications in 94%, 93%, and 93% accurately. Inaccuracies were caused by the following factors: a genetic diagnosis not listed (DiGeorge syndrome) in 1 case; in another letter, diagnoses were listed with social history in the original letter, and this was extracted into the diagnostic list; in another 3 letters, clinical diagnoses and interventions were listed in the same section under diagnosis, and this was inappropriately listed as diagnoses only; in another 2 letters that did not list either diagnoses or interventions, the algorithm incorrectly retrieved diagnostic and intervention history.

7.4.1 Regression with combined information

Another series of experiments were performed combining information from clinical letters with ECG signals projected onto tangent space. All the information about history of the patients regarding to interventions, diagnoses and medications is used. More specifically, frequency count of each unique word is computed and it transformed into a numerical matrix of word counts. These matrices are then concatenated with the tangent space mapping matrices to feed the model. To compare the effect of fusing different input types, we employed two regression models: Support Vector Machine (SVM) and Logistic Regression (LR). Ablation studies were

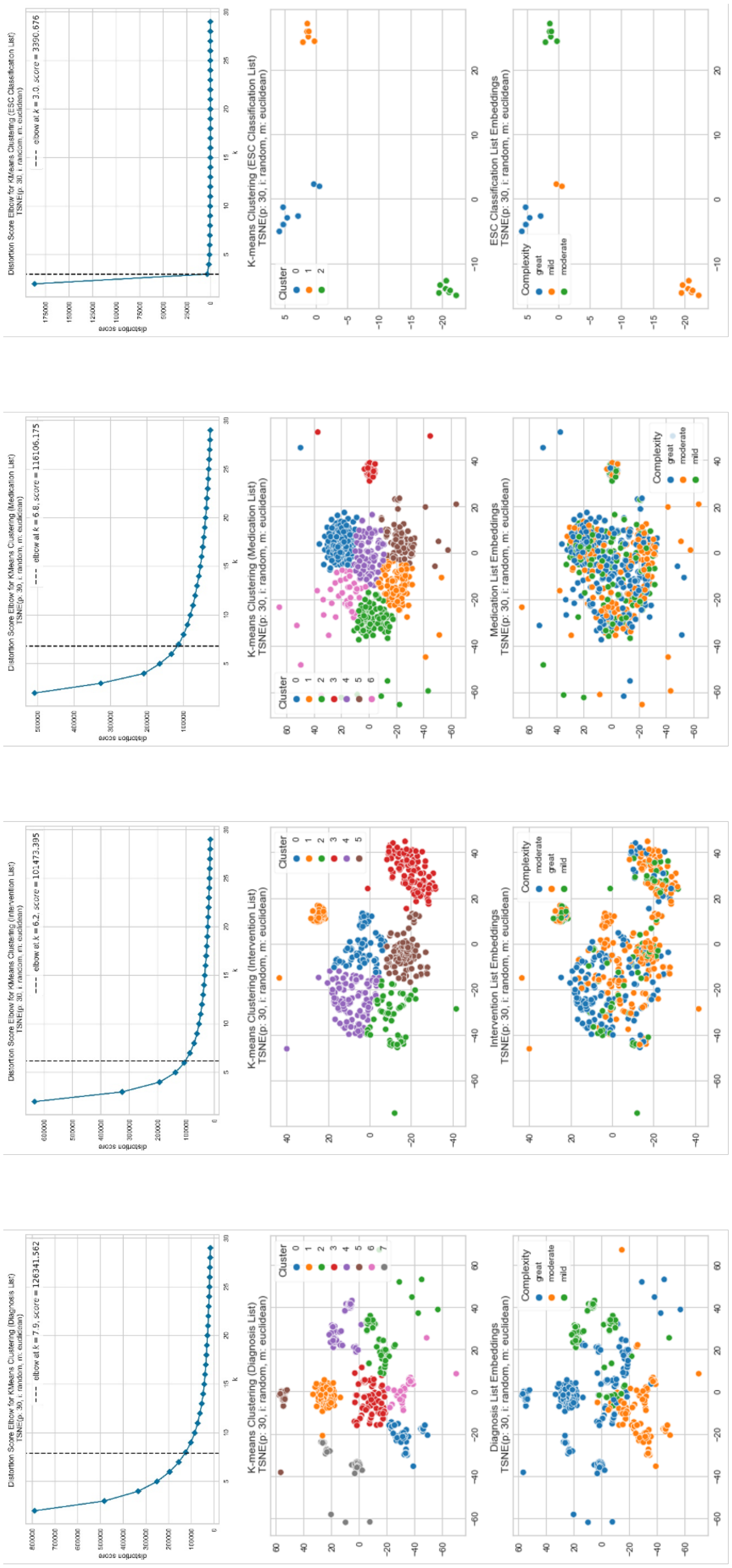


Figure 7.5: Clustering results.

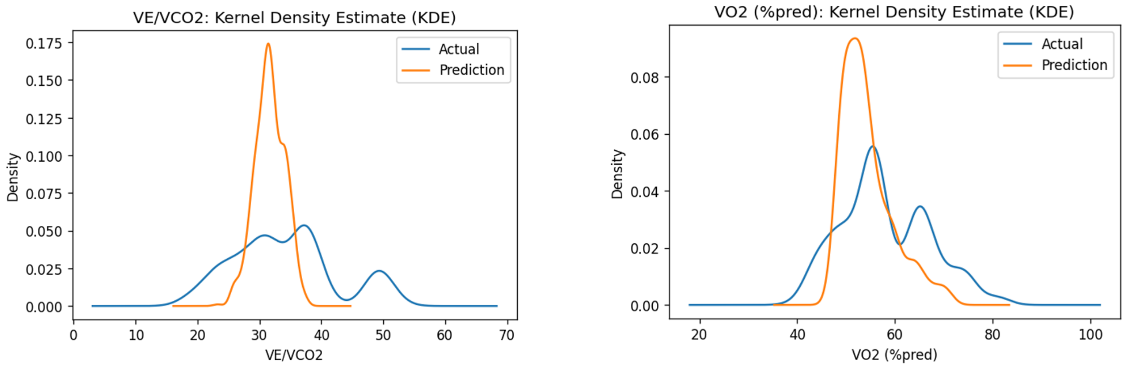
conducted using the SVM model, with the LR model serving as the baseline from the previous chapter.

The data was split into training and testing sets 100 times using a pseudo-randomised, stratified patient leave-out evaluation. This method ensured that the testing set was representative of all classes by randomly selecting one patient from each class for the testing set, while the remaining patients populated the training set. For each run, the SVM model was trained on the training set and then tested on the corresponding testing set, which did not include any data from the patients in the training set. This stratified patient leave-out method enhances the model's ability to predict unseen data by reducing bias [32]. The same patient split strategy was used to ensure consistency across different results. Table 7.3 summarises all the results, it achieved better results in terms of all the evaluation metrics. The correlation between the model's prediction and the actual exercise value was much better, reaching 0.66 for VO_2 (peak).

Table 7.3: Regression results on ECGs + clinical letters, with SVM model.

	R^2	**Adjusted R^2**	**RMSE**	**Correlation Coefficient**	**Predicted Label**	**Input Data**
Tangent Space	0.111	0.084	7.754	0.478	VE/VCO_2	ECGs + clinical letters
Tangent Space	0.263	0.101	7.662	0.658	VO_2 (%pred)	ECGs + clinical letters
Tangent Space	0.395	0.261	0.333	0.660	VO_2 (peak)	ECGs + clinical letters

Kernel Density Estimations (KDEs) plotted in Figure 7.6 to show the distribution of the actual and predicted values. These plots show that all predictions are much closer to the distribution of actual exercise values. These plots also demonstrated a positive alignment with the better results.

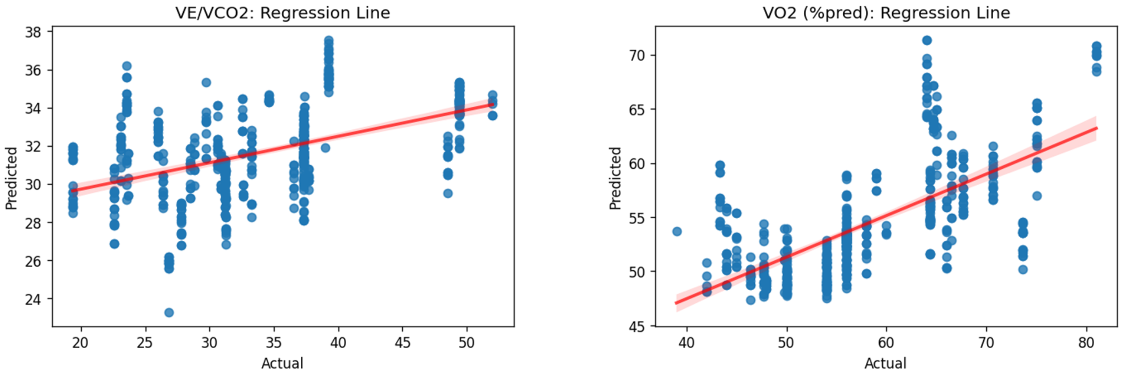


(a) KDE for VE/VCO_2

(b) KDE for VO_2 (%pred)

Figure 7.6: Kernel Density Estimations (KDEs) of the SVM model.

Regression line plots are also shared in Figure 7.7 to show the relationship between actual and predicted values. Improvements in the results are reflected here as well as in the KDE plots. These plots show that all predictions are close to the range of actual values. It also shows that a more sloped regression line is observed, which is also reflected in the results.



(a) Regression line for VE/VCO_2 (b) Regression line for VO_2 (%pred)

Figure 7.7: Regression line plots of the SVM model.

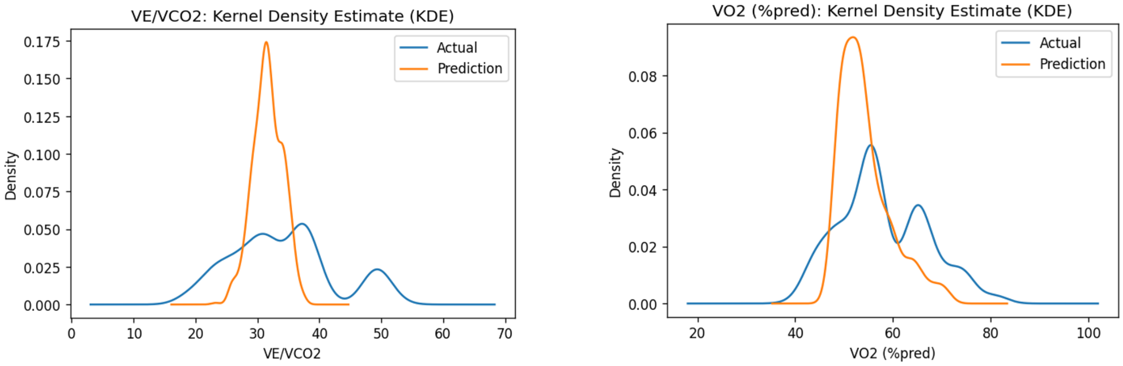
Covariance augmentation technique, which has been successfully applied before to further optimise the results, was also applied here in order to improve the results. Table 7.4 summarises all the results, which achieved better results in terms of all the evaluation metrics.

Table 7.4: Regression results on ECGs + clinical letters, with SVM model using covariance augmentations.

	R^2	**Adjusted R^2**	**RMSE**	**Correlation Coefficient**	**Predicted Label**	**Input Data**
Tangent Space	0.111	-0.084	7.754	0.478	VE/VCO_2	ECGs + clinical letters
Tangent Space (aug)	0.130	-0.060	7.670	0.491	VE/VCO_2	ECGs + clinical letters
Tangent Space	0.263	0.101	7.662	0.658	VO_2 (%pred)	ECGs + clinical letters
Tangent Space (aug)	0.243	0.076	7.769	0.662	VO_2 (%pred)	ECGs + clinical letters
Tangent Space	0.395	0.261	0.333	0.660	VO_2 (peak)	ECGs + clinical letters
Tangent Space (aug)	0.395	0.261	0.333	0.666	VO_2 (peak)	ECGs + clinical letters

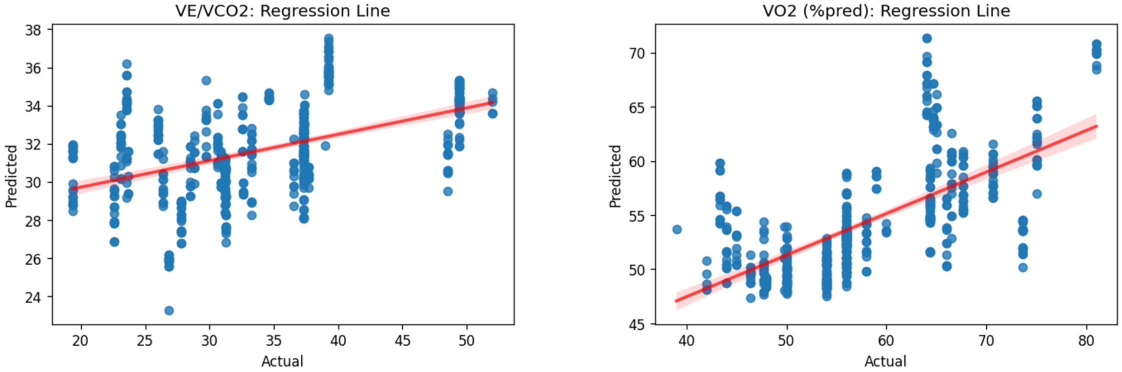
KDEs plotted in Figure 7.8 to show the distribution of the actual and predicted values. These plots show that all predictions are much more closer to the distribution of actual exercise values. These plots also demonstrated a positive alignment with the better results.

Regression line plots are also shared in Figure 7.9 to show the relationship between actual and predicted values. Improvements in the results are reflected here as well as in the KDE plots. These plots show that all predictions are close to the range of actual values. It also shows that a more sloped regression line is observed, which is also reflected in the results.



(a) KDE for VE/VCO_2 (b) KDE for VO_2 (%pred)

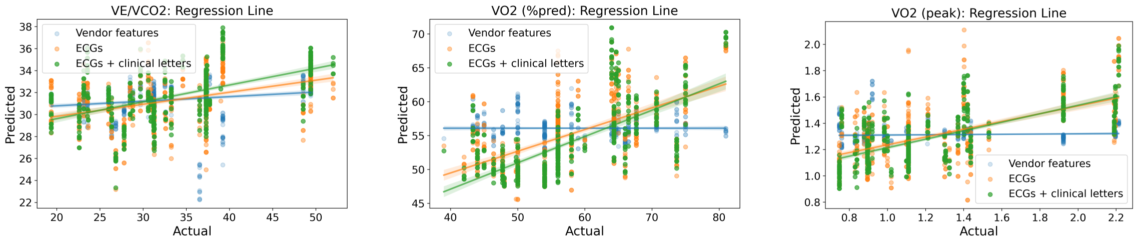
Figure 7.8: Kernel Density Estimations (KDEs) of the SVM model using covariance augmentations.



(a) Regression line for VE/VCO_2 (b) Regression line for VO_2 (%pred)

Figure 7.9: Regression line plots of the SVM model using covariance augmentations.

In order to provide a more concise summary of the each step, a regression plot for each label was also provided in Figure 7.10. The plots illustrate three regression lines for each step, starting from the vendor features and proceeding to the combination of ECGs and clinical letters. These plots not only illustrate the capability of the model but also provide an overview of the improvements achieved.



(a) VE/VCO_2 (b) VO_2 (%pred) (c) VO_2 (peak)

Figure 7.10: Predicting cardiopulmonary exercise biomarkers: comparison of predicted vs. real values with fitted regression lines

7.4.2 Classification with combined information

Preliminary experiments were conducted without the utilisation of any augmentation in order to establish a baseline as outlined in the preceding chapter. Another series of experiments were performed combining information from clinical letters with ECG signals projected onto tangent space. All the information about history of the patients regarding to interventions, diagnoses and medications is used. These matrices are then concatenated with the tangent space mapping matrices to feed the model.

The regression problem was transformed into a classification problem by assigning clinical groups for each label of VE/VCO_2, VO_2 (%pred) and VO_2 (peak). Specifically, VE/VCO_2 and VO_2 (peak) variables were categorised into two groups, while VO_2 (%pred) variable was divided into four groups. As demonstrated in Figure 6.13, the separation of the groups is indicated by dashed red lines. A series of ablation studies were conducted employing an SVM model to examine the impact of various factors, including the fusion of different input types and the application of different augmentation techniques. To ensure the reliability and reproducibility of the results, the same patient split strategy and training setup were consistently used across all experiments. The performance of the classification model was evaluated using multiple metrics, including accuracy, Area Under the Curve (AUC), and F1 macro score. These metrics were reported alongside their corresponding mean and standard deviation values to provide a comprehensive assessment.

Using tangent mappings derived from the ECGs as inputs to the SVM model did not yield the best results. However, a notable improvement (up to 8% on accuracy and AUC, 5% on F1 score) was observed when the ECG data combined with the clinical letter data, which proved to be advantageous in terms of the classification metrics. In order to enhance the result and the balance of the class distribution, the experiments were repeated with the augmentations, both on ECGs and clinical letters.

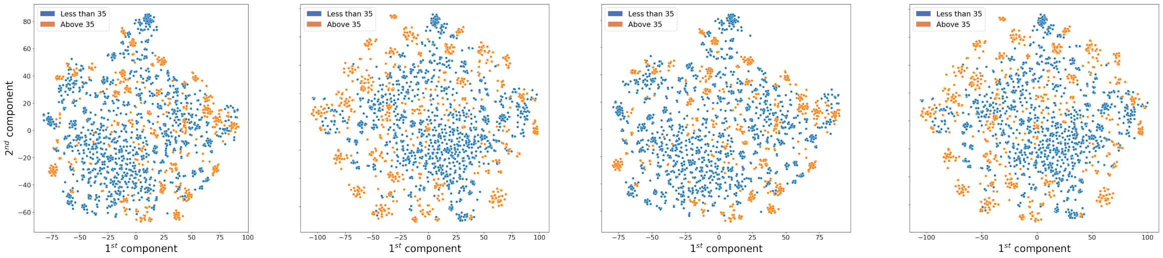
We employed the same sophisticated augmentation technique using Riemannian geometry on covariance matrices, building on our previous work [233]. The weighted Riemannian mean is computed by minimising the sum of the squared Riemannian distances to the given matrices, as formalised in the equation 5.3. In this work, we also experiment with the clinical letters to demonstrate the employed augmentation technique's prominence on tabular data. And compared it with another widely-used and rather simple augmentation technique on tabular data, known as resampling.

Table 7.5: Classification results using regression groups with SVM model.

Predicted Label	**Input**		**Augmentations**		**Accuracy**	**AUC**	**F1 macro**
	ECGs	Clinical letters	ECGs	Clinical letters			
VE/VCO_2	✓				0.672 ± 0.212	0.577 ± 0.141	0.534 ± 0.181
VE/VCO_2	✓		covariance		0.704 ± 0.195	0.622 ± 0.183	0.572 ± 0.197
VE/VCO_2	✓	✓			0.696 ± 0.227	0.655 ± 0.179	0.587 ± 0.236
VE/VCO_2	✓	✓	covariance	simple	0.675 ± 0.240	0.620 ± 0.179	0.541 ± 0.238
VE/VCO_2	✓	✓	covariance	covariance	**0.738 ± 0.192**	**0.663 ± 0.180**	**0.617 ± 0.206**
VO_2 (%pred)	✓				0.302 ± 0.163	0.506 ± 0.225	0.229 ± 0.127
VO_2 (%pred)	✓		covariance		0.324 ± 0.132	0.520 ± 0.225	0.236 ± 0.106
VO_2 (%pred)	✓	✓			0.316 ± 0.124	0.529 ± 0.229	0.239 ± 0.093
VO_2 (%pred)	✓	✓	covariance	simple	0.312 ± 0.116	0.537 ± 0.255	0.239 ± 0.095
VO_2 (%pred)	✓	✓	covariance	covariance	**0.331 ± 0.124**	**0.540 ± 0.245**	**0.257 ± 0.093**
VO_2 (peak)	✓				0.614 ± 0.137	0.585 ± 0.162	0.535 ± 0.167
VO_2 (peak)	✓		covariance		0.657 ± 0.150	0.620 ± 0.126	0.553 ± 0.152
VO_2 (peak)	✓	✓			0.693 ± 0.163	0.601 ± 0.136	0.525 ± 0.169
VO_2 (peak)	✓	✓	covariance	simple	0.642 ± 0.111	0.607 ± 0.093	0.544 ± 0.120
VO_2 (peak)	✓	✓	covariance	covariance	**0.708 ± 0.091**	**0.658 ± 0.134**	**0.617 ± 0.142**

A comprehensive summary of all results is provided in Table 7.5, with the most optimal results indicated in bold. It was observed that applying covariance augmentations only to ECGs was able to increase accuracy, AUC and the F1 score for all classes by up to 5%. When clinical letters were incorporated, a similar outcome was observed with the covariance augmentations applied both to ECGs and clinical letters. It was able to increase accuracy, AUC and the F1 score up to 9% for all classes. However, using simple augmentation techniques as opposed to the covariance augmentations on clinical letters resulted in poor performances, as presented in Table 7.5. It also demonstrates the importance of the augmentation technique that is tailored to the particular problem.

Our approach to fuse information derived from ECGs and clinical letters, and to use a sophisticated augmentation technique yielded the best results. The utilisation of Riemannian geometry augmentations on the covariance matrices of ECGs and clinical letters has been shown to produce features that are more coherent in comparison to those obtained through simple augmentation techniques. The efficacy of our methodology was evaluated with ablation studies, which demonstrated that the integration of ECGs and clinical data yielded the best results. Furthermore, these studies underscored the significance of the augmentation technique that is tailored to address specific problems.



(a) ECG without augmentations

(b) ECG with augmentations

(c) ECG + clinical letters without augmentations

(d) ECG + clinical letters with augmentations

Figure 7.11: Comparison of t-SNE visualizations on the tangent space.

In order to assess the reliability of the models, calibration plots are another valuable tool in the context of medical applications. The purpose of these plots is to facilitate a comparison between the predicted probabilities of an event occurring and the actual observed outcomes [50, 236]. This process provides insights into the calibration of the model, thereby determining whether the model is well-calibrated or whether it over- or underestimates risks. A perfectly calibrated model produces predictions where the estimated probabilities closely match observed frequencies. A well-calibrated model ensures that risk scores align with real-world probabilities, thus enabling clinicians to make accurate, data-driven decisions.

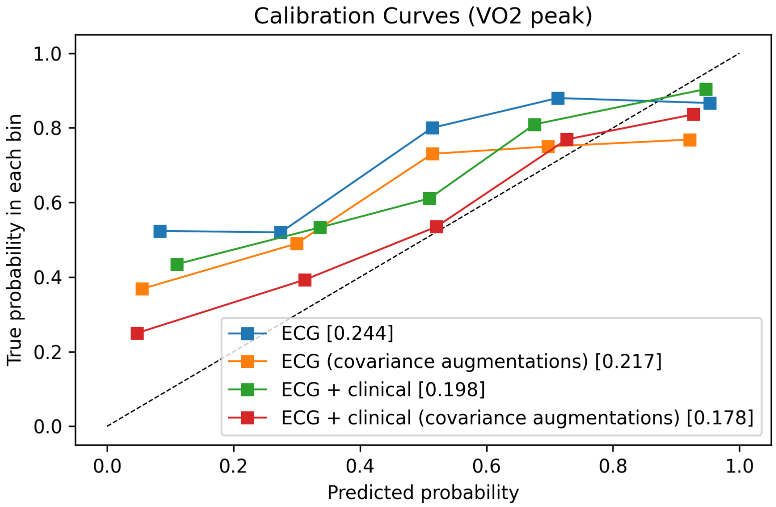


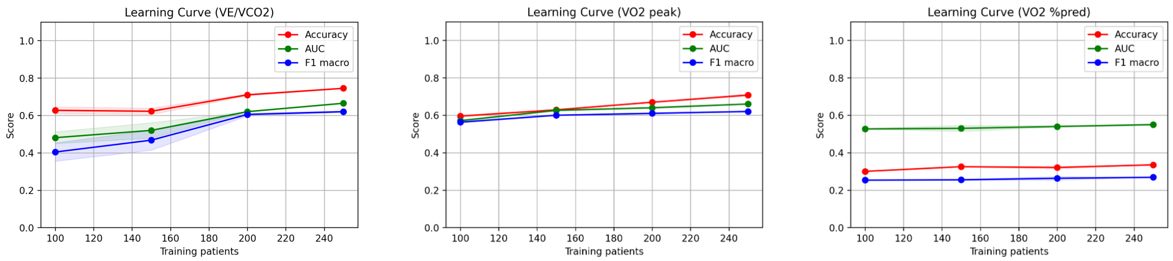
Figure 7.12: Calibration plot comparing predicted probabilities with observed outcomes.

As demonstrated in Figure 7.12, leveraging covariance augmentation and tangent space projections improve the robustness of the estimates, consequently leading to better calibration plots. Additionally, multi-modal data integration that provides complementary patient information contributes to reducing calibration errors, as demonstrated in Figure 7.12. These calibration plots reveal that the models incorporating covariance augmentations and multi-modal data demonstrated higher reliability, with predicted probabilities aligning closely with actual observed outcomes. The reliability of these results is further supported by the utilisation of Brier loss scores, which are presented in square brackets in Figure 7.12. The Brier loss is a metric used to evaluate the performance of a probabilistic prediction (p) in terms of its alignment with actual outcomes (y), and can be used to assess how well a classifier is calibrated. It is measured on a scale ranging from 0 to 1, with lower values indicating greater accuracy in prediction, and is defined as follows for a dataset with n samples:

$$BrierLoss = \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2 \tag{7.2}$$

Furthermore, Figure 7.13 illustrates the learning curves for the classification model, demonstrating how increasing the number of patient samples improves performance. It again demonstrates a clear trend: as the sample size increases, the performance of the model improves significantly. Initially, with a smaller set of patients, the model exhibited high variance which resulted in inconsistent predictions. However, as more patient data was incorporated, accuracy, AUC, and F1 score steadily improved as we observed in the previous chapter. Figure 7.13a shows the impact of sample size on model performance for VE/VCO_2, illustrating that as the number of patient samples increased from 100 to 250, accuracy rose from 0.63 to 0.74, AUC improved from 0.48 to 0.66, and F1 macro score increased from 0.41 to 0.62. These observations suggest that an expanded dataset improves the model's ability to generalise effectively, leading to a reduction in the probability of overfitting and an enhancement in its predictive robustness.

Moreover, Figure 7.14 presents the projected model scores for VE/VCO_2, providing accuracy, AUC and the F1 score, across a range of sample sizes up to 400 patients. This rough analysis reveals that with 300 patient samples, accuracy reaches at 0.77, AUC reaches 0.72, and F1 score reaches at 0.71. As the dataset expands to 400 patient samples, accuracy improves



(a) VE/VCO_2 (b) VO_2 (peak) (c) VO_2 (%pred)

Figure 7.13: Comparison of learning curves for the classification model.

to 0.88, AUC rises to 0.85, and F1 score reaches 0.86. This trend underscores the advantage of incorporating more patient data, as it enables the model to learn more complex patterns and enhance prediction performance. These findings underscore the importance of continuous data collection and model refinement to optimise predictive accuracy in healthcare applications.

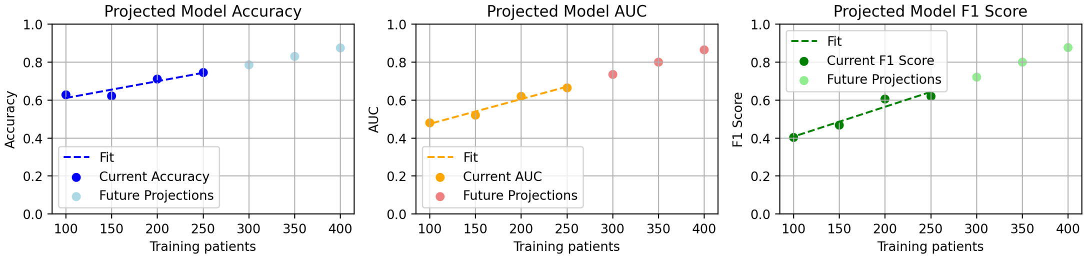


Figure 7.14: Learning curve projection for the classification model on VE/VCO_2.

7.5 Discussion

This chapter explores the potential of multi-modal data integration to predict CPET outcomes in CHD patients, combining ECG data and clinical text information. The study proposes a framework that integrates Riemannian geometry (on ECGs) and natural language processing (on clinical letters), thus addressing the challenges associated with handling diverse data types and optimising predictive performance. The proposed approach demonstrates significant improvements in predictive accuracy, emphasising the value of synthesising diverse data sources to capture the complexity of CHD. The integration of clinical letters with ECGs and the utilization of Riemannian geometry augmentations yielded the best results. Clinical letters often contain vital information about a patient's medical history, symptoms, and other relevant details not captured by ECG signals alone. Integrating this textual information with ECG data provides a more comprehensive and informative feature set. The experimental results demonstrate the efficiency of the proposed augmentation technique in generating more coherent features.

Previous studies exploring clinical models for predicting mortality and disease complexity in CHD patients have highlighted the difficulty in stratifying risk for CHD patients due to the low prevalence of adverse events. To address this challenge, we innovatively adopted CPET outcomes to provide a more dynamic evaluation of patient condition and risk. To our knowledge, this represents the first attempt to build a sophisticated machine learning model predicting CPET results as outcome variables. The improved correlation between our model's predictions and actual exercise values, particularly for VO_2 (peak), demonstrates the potential of our approach in clinical settings even with limited data.

One of the key contributions of this chapter is the demonstration that multi-modal data fusion enhances predictive accuracy for CPET outcomes, which are crucial indicators of functional capacity in CHD patients. This aligns with existing research on multi-modal learning in healthcare, which emphasises the importance of integrating diverse data types to capture the multifaceted nature of medical conditions. The proposed approach leverages the geometric structure of covariance matrices derived from ECG data while extracting meaningful features from clinical text using NLP methods. This multi-modal data approach underscores the importance of synthesising diverse sources of information to capture the multidimensional nature of CHD. Compared to traditional single-modality machine learning models (on vendor features), this framework provides a more comprehensive representation of patient data.

The AUC values for the classification problem range from 0.54 to 0.66 (see Table 7.5), indicating limited discriminative ability of the model. These values suggest that the clinical classification groups may require further refinement or the addition of more patients to improve predictive accuracy. In contrast, the regression model exhibits a strong positive correlation, with r-values ranging from 0.49 to 0.67 (see Table 7.4), indicating a meaningful relationship between the predictor variables and the outcomes. While the classification framework requires optimization, the regression results are encouraging and point to clinically relevant associations. Furthermore, our learning curve analysis (see Figure 7.14) suggests that with an additional 300 patient samples, the model may achieve performance levels suitable for clinical application.

Future research directions include exploring advanced natural language processing techniques to extract richer information from clinical letters and investigating the integration of multi-modal data sources, such as imaging, to provide more comprehensive patient profiles. In conclusion, our study demonstrates the significant potential of combining clinical letters with ECG data and leveraging Riemannian geometry to enhance predictive performance in CHD patients. The use of CPET as an outcome variable provides a dynamic and physiologically relevant endpoint that better reflects functional capacity and cardiovascular reserve compared to static measurements alone. This approach effectively addresses challenges posed by small, imbalanced datasets while providing more accurate patient risk assessments. However, it is important to acknowledge the limitations of the model, including the requirement for validation using larger and more diverse datasets to ensure generalisability. Also, incorporation of the additional data modalities has the potential to further enrich the predictive framework.

Chapter 8

Conclusions and Future Work

8.1 Summary

This thesis explores the development of a multi-modal machine learning framework designed for predicting clinical outcomes in patients with congenital heart disease. We have discussed how multiple data modalities, including medical health records and Electrocardiograms (ECGs), can be applied in different tasks to enhance the precision and reliability of outcome prediction models. The scope encompasses the identification and digitisation of multiple data sources, the design and implementation of relevant machine learning models, and the evaluation of the framework's performance in clinical settings.

In this section, we summarise the contributions and conclusions of each chapter.

8.1.1 Chapter 4: ECG Extraction and Representation in Congenital Heart Disease

The data extraction chapter demonstrates the contribution of digitising 12-lead ECG PDF documents, particularly for patients with adult congenital heart disease. By recreating digital signals from vector data in PDFs, the study addresses a challenge in leveraging machine learning models for better diagnostic and prognostic capabilities. The findings demonstrate a strong correlation between digitised values and vendor-calculated ECG metrics, thereby validating the efficacy of the developed algorithm. Furthermore, the capacity to predict diagnoses through the utilisation of a machine learning model underscores the promising potential of such digitised datasets in clinical research.

The automated pipeline for digitisation is particularly important as it addresses limitations in manual analysis, overlapping signals, and text artefacts. Linking these ECG data with clinical letters further adds value by enabling structured, labelled data for the model training.

It has been demonstrated that there are some limitations, including the automation of document processing, the incorporation of larger and more diverse datasets, and the adaptation of the algorithm to function with various formats. These limitations highlight areas where further development and refinement are essential for broader applicability and improved outcomes:

1. **File Collection:** The current process for the collection and organisation of ECG documentation involves manual handling, requiring staff to deposit these files into a designated analysis folders. This manual process not only adds significant time and effort to the workflow but also increases the potential for human error, which has the capacity to compromise data integrity. The automation of this process or the enhancement of digital solutions could improve efficiency, reduce the workload on staff, and ensure more accurate and timely access to ECG data for analysis.

2. **Signal Duration:** It is important to acknowledge that the study is limited by the duration of ECGs extracted that spans less than one second, in contrast to the ECGs utilised in the other studies referenced in Table 3.1. When combined with the low prevalence of the outcomes, the limited signal duration may restrict the ability of the framework extracting useful information. In order to enhance the functionality of the framework developed, future research should aim to incorporate larger and more diverse datasets which would provide longer signal durations and richer features.

3. **Dataset Size:** While the digitisation of 12-lead ECG PDF documents has shown promise, it is important to acknowledge that the study is limited by the relatively small size of the dataset used for data extraction and analysis. The limited number of cases may restrict the generalisability of the findings. In order to enhance the applicability of the methodology developed, future research should aim to incorporate larger and more diverse datasets

which would provide a more comprehensive evaluation of the methodology and its potential for clinical implementation.

4. **File Format:** The current methodology is specifically designed for ECGs stored as vectorised PDF documents, which allows for precise data extraction and analysis. However, it should be noted that this approach is not applicable to pixelated or scanned images, as such formats may lead to inaccuracies. It may be beneficial to explore alternative techniques or preprocessing steps that can effectively handle various file formats.

8.1.2 Chapter 5: ECG Classification with Riemannian Geometry

The chapter introduces a novel application of Riemannian geometry to improve the classification of congenital heart disease using 12-lead ECG data. By leveraging the spatial covariance structure of ECG signals, the approach addresses the challenges posed by the heterogeneity of Congenital Heart Disease (CHD) cases and the limited availability of large datasets. The use of covariance augmentation and tangent space projection enhances feature extraction, enabling the model to better capture the underlying anatomical and functional variations in CHD patients. The results demonstrate significant improvements in classification performance compared to traditional machine learning and deep learning models, highlighting the potential of Riemannian geometry in biomedical signal analysis.

The chapter also emphasises the importance of tailored methodologies for rare and complex conditions like CHD, where standard deep learning approaches may alter due to data scarcity and extreme variability. The integration of Riemannian geometry provides a robust framework for analysing ECG data, paving the way for more accurate diagnostic tools and personalised treatment strategies. The chapter addresses RQ1 and RQ4 (see section 1.3) by investigating the potential of Riemannian geometry to capture the intrinsic geometric properties of ECG signals and their relevance to CHD classification in the context of limited data availability and heterogeneous patient characteristics.

The chapter also examines the advantages of pre-trained state-of-the-art models in ECG applications, emphasising their ability to leverage large-scale datasets for more efficient feature extraction and reduced computational costs. Whilst pre-trained models and transfer learning strategies provide a resource-efficient alternative to training models from scratch, they are not always effective for clinical applications involving unique patient demographics, such as those with CHD. The findings serve to emphasise the limitations of pre-trained models, particularly in scenarios involving datasets that are characterised by unique patient demographics.

The chapter also addresses the challenges of implementing such models in hospital environments with limited resources. Despite the effectiveness of transfer learning, our findings indicate that even extensive fine-tuning across all layers of the pre-trained model does not always yield optimal results. This highlights the limitations of general-purpose feature representations when applied to datasets with distinct characteristics. Consequently, it underscores the necessity for tailored deep learning approaches that take into account domain-specific nuances and computational manageability.

It has been demonstrated that there are some limitations, including the incorporation of larger and more diverse datasets, adaptation of the proposed method to function with various CHD subtypes and restricted computational power for model development. These limitations highlight areas where further development and refinement are essential for broader applicability and improved outcomes:

1. **Dataset Size:** The study acknowledges the limitations posed by the relatively small size of the dataset, which may impact the generalisability of the models including the pre-trained model. While the intention behind using such pre-trained models is to provide a resource-efficient alternative to training from scratch, the findings indicate that these models may not always be successful for downstream tasks. However, it is important to emphasise that the size of the dataset is not the only factor, the prevalence of the outcome being predicted also plays a critical role. Moreover, it is highly impractical to expect large sample sizes in CHD, unless datasets are merged across countries and continents, where deep learning approaches might outperform the current methodology. Nevertheless, the findings of this study demonstrate that combining geometric features with deep learning offers a promising future direction. This hybrid approach could enhance the robustness and applicability of the model across diverse populations and clinical settings, ultimately

improving diagnostic and prognostic capabilities.

2. **Complexity of CHD:** The extreme heterogeneity in CHD anatomy and physiology poses challenges for model training and evaluation. While the proposed method shows promise, its performance may vary across different CHD subtypes.

3. **Generalisability:** The methodology is specifically tailored for CHD and may require some adaptation for other cardiac or non-cardiac conditions. While the underlying principles of the approach could potentially be relevant to a broader range of diseases, it should be noted that each condition may present unique anatomical, physiological and clinical characteristics which may require relevant modifications to the methodology.

4. **Computational Power:** A significant constraint of this study is the computational limitations imposed by the necessity to conduct the research on a hospital computer, which was required to facilitate access to the patient data. This reliance on limited hardware capabilities may have constrained the implementation of more advanced methods, potentially limiting the model's performance. In order to address this challenge, it is crucial to explore alternative approaches that prioritise resource efficiency and clinical adaptability, ensuring that the models are optimised for real-world healthcare settings.

8.1.3 Chapter 6: Predicting Cardiopulmonary Exercise Testing Outcomes in CHD

The chapter demonstrates the effectiveness of incorporating Riemannian geometry into the feature extraction process with small and imbalanced CHD data. It enhances classification accuracy through structured covariance matrix transformations, and facilitates more structured and robust feature extraction for CHD diagnostics. Additionally, as previously discussed in earlier chapters, the chapter reemphasises the efficacy of the covariance augmentation technique in enhancing feature representation and improving generalisation in predictive models.

Ablation studies have confirmed the enhanced performance of tailored augmentation techniques in comparison to their more simplistic counterparts. These studies have demonstrated the benefits of Riemannian geometry in generating coherent, discriminative features, which is a crucial aspect in machine learning models. This approach enables a richer understanding of the underlying relationships within the data, leading to enhancements in diagnostic classification and regression models. The chapter addresses RQ1, RQ2 and RQ4 (see section 1.3) by investigating the potential of ECG waveforms enhanced with Riemannian geometry in the development of risk prediction models for CHD and the potential of surrogate outcomes such as Cardiopulmonary Exercise Testing (CPET) variables.

It has been demonstrated that there are some limitations, including the incorporation of larger and more diverse datasets, and the adaptation of the proposed method to function with various CHD subtypes. These limitations highlight areas where further development and refinement are essential for broader applicability and improved outcomes:

1. **Dataset Size:** The chapter acknowledges the limited size of the dataset, which consists of ECGs as the primary input and CPETs as a surrogate for mortality. The limited number of patients with both ECG and CPET data may restrict the ability to precisely capture the relationship between ECG and CPET outcomes. Future research should focus on incorporating larger datasets, as well as efforts to increase the availability of CPET data.

2. **Generalisability:** The chapter focused on a specific set of anatomical diagnoses, which may restrict the generalisability of the findings. Expanding the approach to include other complex cardiac conditions would enhance its generalisability, ultimately contributing to improved diagnostic and prognostic capabilities.

3. **Additional Data:** Integration of additional patient data, such as imaging or additional clinical records, could further improve diagnostic accuracy and predictive capabilities. The integration of additional data sources has the potential to facilitate the identification of critical factors that influence patient outcomes, thereby enhancing the model's performance. This comprehensive data integration could ultimately lead to better-informed clinical decisions and improved patient care.

8.1.4 Chapter 7:

Integrating Information from Clinical Letters with ECGs

The chapter discusses the data extraction techniques employed to extract relevant information from clinical letters, the integration process with ECG data, and the impact on predictive model performance. And demonstrates the effectiveness of combining ECGs and clinical letters in CHD diagnostics using advanced augmentation techniques based on Riemannian geometry. The results highlight the importance multi-modal data integration (ECGs and clinical records) for enhancing the robustness of classification and regression models.

The chapter underscores the potential of integrating information from clinical letters with ECG data to enhance diagnostic precision and model performance in patients with CHD. The integration of clinical letters, which contain crucial patient information such as diagnoses, interventions, and medications, with digitised ECG signals facilitates a more comprehensive analysis of patient health. Furthermore, it enables machine learning models to better capture anatomical and physiological variations associated with CHD. The chapter addresses all the research questions (in section 1.3) by investigating the potential of combining data from multiple sources (ECGs and clinical letters) enhanced with Riemannian geometry in the development of risk prediction models for CHD and the potential of surrogate outcomes (CPET variables).

The incorporation of CPET outcomes into risk prediction models provides clinicians with several important advantages. The primary and most important outcome of the risk prediction model is that it can function as both a diagnostic and a functional assessment metric for patients with CHD. Secondly, model outcomes can also function as a predictive instrument, directly informing clinical decision-making.

It has been demonstrated that there are some limitations, including the incorporation of larger and more diverse datasets, and exploring alternative methods for extracting features from clinical letters. These limitations highlight areas where further development and refinement are essential for broader applicability and improved outcomes:

1. **Dataset Size:** Although promising, the study had a relatively small cohort of 436 patients which may impact the internal validation of the model. We conducted a thorough internal

validation, which suggests that the model should generalise to new patients within the same setting. However, the generalisability of our results to other centres remains less certain, highlighting the need for external validation with larger datasets. Notably, the learning curves presented in the chapter indicate that increasing the sample size to double the current number of subjects could yield better results that are clinically relevant.

2. **Additional Data:** While the integration of clinical letters has demonstrated improvements, the integration of additional data modalities, such as imaging or laboratory results, has the potential to further refine the model's performance. By incorporating a broader range of clinical data, the methodology could capture a more comprehensive view on patient health, leading to improved diagnostic accuracy and predictive capabilities.

3. **Feature Extraction:** While the existing methods provided valuable information, exploring alternative methods for extracting features from clinical letters could further improve the outcomes. By investigating and implementing alternative feature extraction strategies, future research could enhance the data utilised in the model, ultimately leading to improved diagnostic accuracy and predictive capabilities.

8.2 Future Work

While this study demonstrates the potential of geometric deep learning for predictive analysis, there are several points for future research that could further enhance model performance and applicability. One key direction involves the incorporation of additional multi-modal data sources, particularly imaging and genomic data, to enhance prediction accuracy. The incorporation of modalities such as echocardiography or cardiac imaging could facilitate the utilisation of structural and functional information beyond ECG signals, thereby potentially providing a more comprehensive representation of cardiovascular health and strengthening predictive capabilities. The integration of genomic data has the potential to provide heritable risk information that can complement other features. This may result in enhanced individualised risk stratification and clinical understanding. The practical deployment of such models can be facilitated by the use of distillation techniques to enable the utilisation of more extensive training on centralised resources while simultaneously permitting lightweight inference on limited hospital hardware.

Additionally, an important consideration is the robustness of our method under varying electrode configurations. The covariance-based feature extraction approach is dependent on the spatial relationships between different ECG leads, meaning that a reduction in electrode count could impact feature quality and overall model performance. Future studies could investigate methods for adapting these scenarios with fewer ECG electrodes, such as optimising lead selection or incorporating signal reconstruction techniques. Evaluating the impact of reduced spatial resolution on classification accuracy would provide valuable insights into the adaptability of our approach in settings where resources are limited.

Beyond these considerations, further efforts could focus on refining model interpretability and explainability, ensuring that predictions are clinically meaningful and actionable. The extension of the dataset to include a larger and more diverse cohort would also enhance generalisability of deep learning models, thereby addressing potential biases introduced by limited sample representation. It is acknowledged that continued advancements in multi-modal learning and feature extraction methods will be instrumental in improving the accuracy and clinical utility of the models for cardiovascular health assessment. So, combining geometric and deep learning may offer a promising future direction and could enhance the robustness and applicability of the model across diverse populations and clinical settings, ultimately improving diagnostic and prognostic capabilities.

Appendix A

A.1 Beta Distribution

The beta distribution is a continuous probability distribution defined on the interval $[0,1]$, and characterised by two positive shape parameters α and β [237]. It is often used for modelling random variables bounded between 0 and 1, such as proportions or probabilities. The probability density function (PDF) of the beta distribution is defined as:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{A.1}$$

where $x \in [0,1]$, $\alpha > 0$, $\beta > 0$, and $B(\alpha,\beta)$ denotes the beta function which serves as a normalisation constant ensuring the total probability equals one. Its flexibility arises from the shape parameters: varying α and β can produce distributions that are uniform, U-shaped, skewed, or symmetric. The beta function $B(\alpha,\beta)$ is defined as:

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt \tag{A.2}$$

The mean and variance of the beta distribution are given by:

$$\text{Mean} = \frac{\alpha}{\alpha+\beta} \tag{A.3}$$

$$\text{Variance} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{A.4}$$

This makes the beta distribution a natural choice for modelling proportions, probabilities and uncertainties in constrained domains. In our work, the beta distribution is used in the context of covariance matrix augmentation to control mixing between two random covariance matrices. To apply augmentation, we sample α from a beta distribution on the interval [0, 1], and compute the weighted Riemannian mean according to the Riemannian distance metric between the randomly selected covariance matrices. Using such technique we were able to generate more samples in a controlled way similar to [212].

A.2 Vectorcardiogram

A Vectorcardiogram (VCG) represents the spatial trajectory of the heart's electrical activity as a vector in three-dimensional space, typically projected onto orthogonal leads (X,Y,Z) [137]. While the standard 12-lead ECG records voltage differences over time across multiple electrodes, the VCG captures the instantaneous cardiac electrical vector and is useful for characterising spatial aspects of depolarization and repolarization [205, 206]. An example of these representations is shown in Figure A.1 and it reveals signatures of anatomical defects in CHD based on our data.

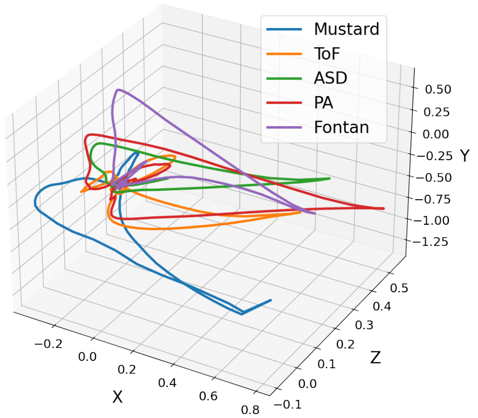


Figure A.1: Average VCGs across anatomical defects in CHD.

The primary advantage of VCG lies in its ability to preserve spatial information about depolarization and repolarization, which can enhance diagnostic capabilities in conditions such as ventricular hypertrophy and conduction abnormalities. VCGs are obtained either directly through orthogonal lead systems (such as Frank lead system) or derived mathematically from standard ECG leads via linear transformation matrices (such as Dower transformation). The Frank's orthogonal lead system to obtain VCG is the most widely used system for its simplicity. The bipolar circuit representing Frank's orthogonal lead system, depicted in Figure A.2, comprises seven electrodes whose position is denoted by capital letters: I, E, C, A, M, F and H.

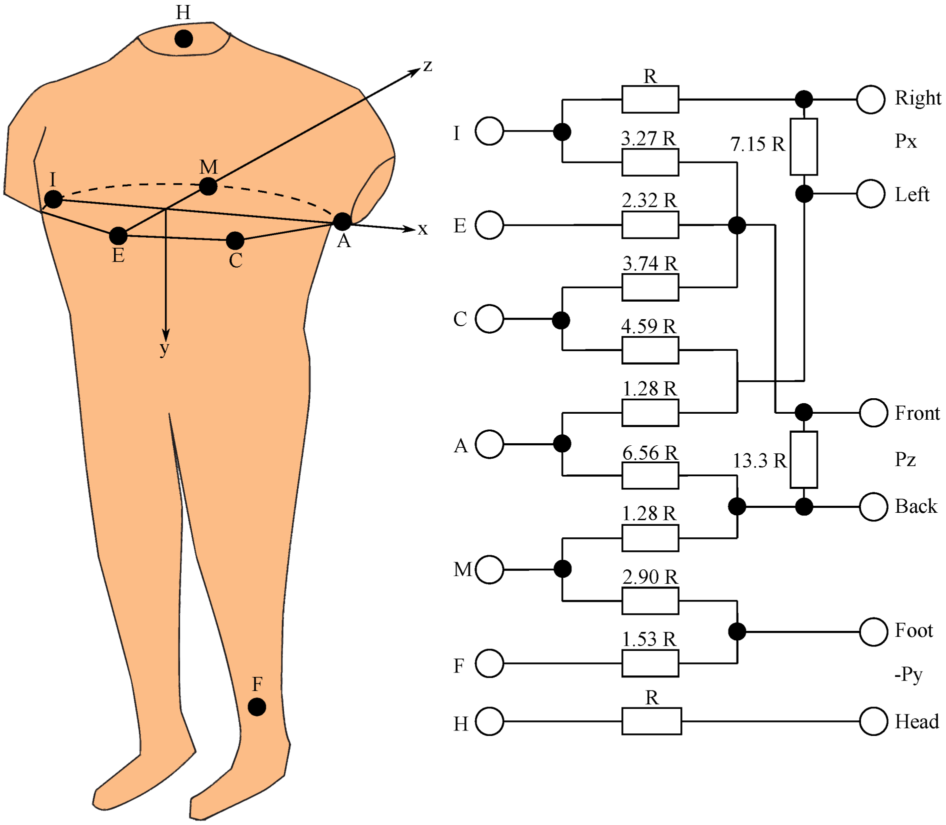


Figure A.2: Frank's orthogonal lead system [238].

A.3 Dower Transformation

The Dower transformation provides a linear mapping between the Frank VCG leads and the standard 12-lead ECG. There are two variants that are commonly referenced: the Dower (forward) matrix which maps Frank leads to 12-lead ECG approximations, and the inverse Dower (or inverse transformation) which estimates Frank leads from the 12-lead ECG. Some transformations use 8 independent leads (leads I, II, V1–V6) to avoid extra linear combinations (III, aVR, aVL, aVF). The linear transformation can be defined as:

$$\mathbf{V} = \mathbf{M} \cdot \mathbf{E} \tag{A.5}$$

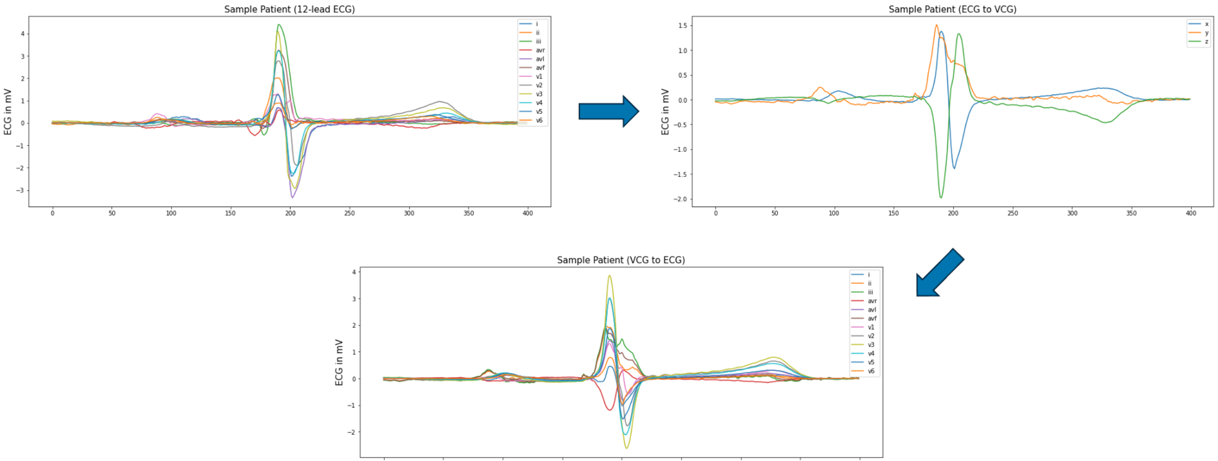
where $\mathbf{E}$ contains the ECG lead signals and $\mathbf{M}$ is the Dower coefficient matrix determined empirically from population studies [239]. This approach facilitates retrospective VCG analysis without requiring direct VCG acquisition hardware, thus enabling spatial vector analysis from routine ECG recordings. The vectors in Figure A.2, right-left axis P_X, head-to-feet axis P_Y and front-back axis P_Z are defined with the following equations [138].

$$P_X = -(-0.172V1 - 0.074V2 + 0.122V3 + 0.231V4 + 0.239V5 + 0.194V6 + 0.156DI - 0.010DII) \tag{A.6}$$

$$P_Y = (0.057V1 - 0.019V2 - 0.106V3 - 0.022V4 + 0.041V5 + 0.048V6 - 0.227DI + 0.887DII) \tag{A.7}$$

$$P_Z = -(-0.229V1 - 0.310V2 - 0.246V3 - 0.063V4 + 0.055V5 + 0.108V6 + 0.022DI + 0.102DII) \tag{A.8}$$

An example application of Dower transformations (both forward and inverse) for 12-lead ECG signals involving mapping these signals to VCG space, applying 3D spatial augmentations and then back-projecting to the original ECG space can be seen in Figure A.3.


Sample Patient (12-lead ECG)

Sample Patient (ECG to VCG)

Sample Patient (VCG to ECG)

Figure A.3: Augmentations on ECG signals using VCG space.

Bibliography

[1] British Heart Foundation, *Circulatory diseases factsheet*, British Heart Foundation, pp. 1–12, 2023.

[2] S. Petersen, V. Peto, and M. Rayner, "Congenital heart disease statistics," British Heart Foundation, Tech. Rep., 2003.

[3] O. Efthimiou, M. Seo, K. Chalkou, T. Debray, M. Egger, and G. Salanti, "Developing clinical prediction models: A step-by-step guide," *BMJ*, vol. 386, 2024.

[4] N. Hassan *et al.*, "Road map for clinicians to develop and evaluate ai predictive models to inform clinical decision-making," *BMJ Health & Care Informatics*, vol. 30, e100784, 2023.

[5] L. Bonnett, K. Snell, G. Collins, and R. Riley, "Guide to presenting clinical prediction models for use in clinical settings," *BMJ*, vol. 365, 2019.

[6] J. Oliveira, C. Costa, and R. Antunes, "Data structuring of electronic health records: A systematic review," *Health and Technology*, pp. 1–17, 2021.

[7] C. Tam *et al.*, "Combining structured and unstructured data in emrs to create clinically-defined emr-derived cohorts," *BMC Medical Informatics and Decision Making*, vol. 21, pp. 1–10, 2021.

[8] D. Zhang, C. Yin, J. Zeng, X. Yuan, and P. Zhang, "Combining structured and unstructured data for predictive models: A deep learning approach," *BMC Medical Informatics and Decision Making*, vol. 20, pp. 1–11, 2020.

[9] G. Baniulyte, N. Rogerson, and J. Bowden, "Evolution–removing paper and digitising the hospital," *Health and Technology*, vol. 13, pp. 263–271, 2023.

[10] S. Zandieh, K. Yoon-Flannery, G. Kuperman, D. Langsam, D. Hyman, and R. Kaushal, “Challenges to ehr implementation in electronic-versus paper-based office practices,” *Journal of General Internal Medicine*, vol. 23, pp. 755–761, 2008.

[11] A. Boonstra, A. Versluis, and J. Vos, “Implementing electronic health records in hospitals: A systematic literature review,” *BMC Health Services Research*, vol. 14, pp. 1–10, 2014.

[12] D. Hoque, V. Kumari, M. Hoque, R. Ruseckaite, L. Romero, and S. Evans, “Impact of clinical registries on quality of patient care and clinical outcomes: A systematic review,” *PloS One*, vol. 12, e0183667, 2017.

[13] E. Vander Velde, J. Vriend, M. Mannens, C. Uiterwaal, R. Brand, and B. Mulder, “Concor, an initiative towards a national registry and dna-bank of patients with congenital heart disease in the netherlands: Rationale, design, and first results,” *European Journal Of Epidemiology*, vol. 20, pp. 549–557, 2005.

[14] C. Verheugt *et al.*, “Mortality in adult congenital heart disease,” *European Heart Journal*, vol. 31, pp. 1220–1229, 2010.

[15] G. Hickey *et al.*, “Clinical registries: Governance, management, analysis and applications,” *European Journal of Cardio-Thoracic Surgery*, vol. 44, pp. 605–614, 2013.

[16] P. Arjun, *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes*. London, UK: Apress, 2019.

[17] A. Gamal, S. Barakat, and A. Rezk, “Standardized electronic health record data modeling and persistence: A comparative review,” *Journal of Biomedical Informatics*, vol. 114, p. 103 670, 2021.

[18] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: Promise and potential,” *Health Information Science and Systems*, vol. 2, pp. 1–10, 2014.

[19] M. Badawy, N. Ramadan, and H. Hefny, “Big data analytics in healthcare: Data sources, tools, challenges, and opportunities,” *Journal of Electrical Systems and Information Technology*, vol. 11, p. 63, 2024.

[20] M. Islam, M. Hasan, X. Wang, H. Germack, and M. Noor-E-Alam, “A systematic review on healthcare analytics: Application and theoretical perspective of data mining,” *Healthcare*, vol. 6, p. 54, 2018.

[21] M. Badawy, N. Ramadan, and H. Hefny, “Healthcare predictive analytics using machine learning and deep learning techniques: A survey,” *Journal of Electrical Systems and Information Technology*, vol. 10, p. 40, 2023.

[22] B. Shivahare *et al.*, “Delving into machine learning’s influence on disease diagnosis and prediction,” *The Open Public Health Journal*, vol. 17, 2024.

[23] C. Melo Santos, A. Barbosa, and Â. Sant’Anna, “Performance measurement systems in primary health care: A systematic literature review,” *BMC Health Services Research*, vol. 25, p. 353, 2025.

[24] X. Fang, Y. Gao, and P. Hu, “A prescriptive analytics method for cost reduction in clinical decision making,” MIS Quarterly, Forthcoming, 2019.

[25] C. Bishop and N. Nasrabadi, *Pattern Recognition and Machine Learning*. Springer, 2006.

[26] T. Hastie, R. Tibshirani, and J. Friedman, *An Introduction to Statistical Learning*. 2009.

[27] R. Kohavi *et al.*, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *IJCAI*, vol. 14, 1995, pp. 1137–1145.

[28] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[29] N. Japkowicz and M. Shah, *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

[30] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. 2018.

[31] T. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” *Neural Computation*, vol. 10, pp. 1895–1923, 1998.

[32] G. Cawley and N. Talbot, “On over-fitting in model selection and subsequent selection bias in performance evaluation,” *The Journal Of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[33] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 2013.

[34] O. Bousquet and A. Elisseeff, “Stability and generalization,” *Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[35] S. Arlot and A. Celisse, “A survey of cross-validation procedures for model selection,” Tech. Rep., 2010, Technical report.

[36] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” *BMC Bioinformatics*, vol. 7, p. 91, 2006.

[37] H. He and E. Garcia, “Learning from imbalanced data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 1263–1284, 2009.

[38] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” *Intelligent Data Analysis*, vol. 6, pp. 429–449, 2002.

[39] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Information Processing & Management*, vol. 45, pp. 427–437, 2009.

[40] J. Kelleher, B. Mac Namee, and A. D’Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press, 2020.

[41] T. Fawcett, “An introduction to roc analysis,” *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[42] F. Provost and T. Fawcett, “Robust classification for imprecise environments,” *Machine Learning*, vol. 42, pp. 203–231, 2001.

[43] A. Bradley, “The use of the area under the roc curve in the evaluation of machine learning algorithms,” *Pattern Recognition*, vol. 30, pp. 1145–1159, 1997.

[44] D. Hand and R. Till, “A simple generalisation of the area under the roc curve for multiple class classification problems,” *Machine Learning*, vol. 45, pp. 171–186, 2001.

[45] C. Ferri, J. Hernández-Orallo, and M. Salido, “Volume under the roc surface for multiclass problems,” in *European Conference on Machine Learning*, 2003, pp. 108–120.

[46] M. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.

[47] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 233–240.

[48] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[49] E. Steyerberg, "Validation of prediction models," in *Clinical Prediction Models: A Practical Approach To Development, Validation, And Updating*, 2019, pp. 329–344.

[50] B. Van Calster, D. McLernon, M. Van Smeden, L. Wynants, E. Steyerberg, and A. Vickers, "Calibration: The achilles heel of predictive analytics," *BMC Medicine*, vol. 17, p. 230, 2019.

[51] M. Ribeiro, S. Singh, and C. Guestrin, *Model-agnostic interpretability of machine learning*, arXiv preprint arXiv:1606.05386, 2016.

[52] F. Doshi-Velez and B. Kim, *Towards a rigorous science of interpretable machine learning*, arXiv:1702.08608, 2017.

[53] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[54] J. Cook and G. Collins, "The rise of big clinical databases," *Journal of British Surgery*, vol. 102, e93–e101, 2015.

[55] G. Collins, J. Reitsma, D. Altman, and K. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement," *Journal of British Surgery*, vol. 102, pp. 148–158, 2015.

[56] E. Steyerberg *et al.*, "Prognosis research strategy (progress) 3: Prognostic model research," *PLoS Medicine*, vol. 10, e1001381, 2013.

[57] E. Steyerberg *et al.*, "Poor performance of clinical prediction models: The harm of commonly applied methods," *Journal of Clinical Epidemiology*, vol. 98, pp. 133–143, 2018.

[58] D. Kent *et al.*, “The predictive approaches to treatment effect heterogeneity (path) statement,” *Annals of Internal Medicine*, vol. 172, pp. 35–45, 2020.

[59] J. Paulus and D. Kent, “Predictably unequal: Understanding and addressing concerns that algorithmic clinical prediction may increase health disparities,” *NPJ Digital Medicine*, vol. 3, p. 99, 2020.

[60] R. Riley *et al.*, “Calculating the sample size required for developing a clinical prediction model,” *BMJ*, vol. 368, 2020.

[61] R. Little and D. Rubin, *Statistical Analysis with Missing Data*. John Wiley & Sons, 2019.

[62] S. Van Buuren, *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC Press, 2012.

[63] J. Sterne *et al.*, “Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls,” *BMJ*, vol. 338, 2009.

[64] E. Steyerberg, *Frank E. Harrell, Regression Modeling Strategies: With Applications, to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, Heidelberg: Springer.* Oxford University Press, 2016.

[65] E. Steyerberg, M. Eijkemans, F. Harrell Jr, and J. Habbema, “Prognostic modeling with logistic regression analysis: In search of a sensible strategy in small data sets,” *Medical Decision Making*, vol. 21, pp. 45–56, 2001.

[66] W. Samek, T. Wiegand, and K. Müller, “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models,” *arXiv preprint:1708.08296*, 2017.

[67] P. Domingos, “The role of occam’s razor in knowledge discovery,” *Data Mining and Knowledge Discovery*, vol. 3, pp. 409–425, 1999.

[68] K. Bruynseels, F. Sio, and J. Hoven, “Digital twins in health care: Ethical implications of an emerging engineering paradigm,” *Frontiers in Genetics*, vol. 9, p. 31, 2018.

[69] A. Fuller, Z. Fan, C. Day, and C. Barlow, “Digital twin: Enabling technologies, challenges and open research,” *IEEE Access*, vol. 8, pp. 108 952–108 971, 2020.

[70] F. Tao, H. Zhang, A. Liu, and A. Nee, “Digital twin in industry: State-of-the-art,” *IEEE Transactions on Industrial Informatics*, vol. 15, pp. 2405–2415, 2018.

[71] F. Tao and Q. Qi, “Make more digital twins,” *Nature*, vol. 573, pp. 490–491, 2019.

[72] I. Voigt, H. Inojosa, A. Dillenseger, R. Haase, K. Akgün, and T. Ziemssen, “Digital twins for multiple sclerosis,” *Frontiers In Immunology*, vol. 12, p. 669 811, 2021.

[73] S. Verma *et al.*, “Development of a semiautomated database for patients with adult congenital heart disease,” *Canadian Journal Of Cardiology*, vol. 38, pp. 1634–1640, 2022.

[74] E. Shortliffe and B. Buchanan, “A model of inexact reasoning in medicine,” *Mathematical Biosciences*, vol. 23, pp. 351–379, 1975.

[75] L. Lusted, “Introduction to medical decision making,” *American Journal Of Physical Medicine & Rehabilitation*, vol. 49, p. 322, 1970.

[76] E. Berner, *Clinical decision support systems*. Springer, 2007.

[77] E. Berner and T. La Lande, “Overview of clinical decision support systems,” in *Clinical Decision Support Systems: Theory And Practice*, 2016, pp. 1–17.

[78] R. Miotto, L. Li, B. Kidd, and J. Dudley, “Deep patient: An unsupervised representation to predict the future of patients from the electronic health records,” *Scientific Reports*, vol. 6, p. 26 094, 2016.

[79] A. Beam and I. Kohane, “Big data and machine learning in health care,” *JAMA*, vol. 319, pp. 1317–1318, 2018.

[80] K. Yu, A. Beam, and I. Kohane, “Artificial intelligence in healthcare,” *Nature Biomedical Engineering*, vol. 2, pp. 719–731, 2018.

[81] A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” *New England Journal Of Medicine*, vol. 380, pp. 1347–1358, 2019.

[82] S. Meystre, G. Savova, K. Kipper-Schuler, and J. Hurdle, “Extracting information from textual documents in the electronic health record: A review of recent research,” *Yearbook Of Medical Informatics*, vol. 17, pp. 128–144, 2008.

[83] S. Sheikhalishahi, R. Miotto, J. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, "Natural language processing of clinical notes on chronic diseases: Systematic review," *JMIR Medical Informatics*, vol. 7, e12239, 2019.

[84] A. Casey *et al.*, "A systematic review of natural language processing applied to radiology reports," *BMC Medical Informatics And Decision Making*, vol. 21, p. 179, 2021.

[85] K. Taha, "Machine learning in biomedical and health big data: A comprehensive survey with empirical and experimental insights," *Journal Of Big Data*, vol. 12, p. 61, 2025.

[86] Z. Chen *et al.*, "Harnessing the power of clinical decision support systems: Challenges and opportunities," *Open Heart*, vol. 10, e002432, 2023.

[87] R. Sutton, D. Pincock, D. Baumgart, D. Sadowski, R. Fedorak, and K. Kroeker, "An overview of clinical decision support systems: Benefits, risks, and strategies for success," *NPJ Digital Medicine*, vol. 3, p. 17, 2020.

[88] B. Abell *et al.*, "Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: A nasss framework-informed scoping review," *Implementation Science*, vol. 18, p. 32, 2023.

[89] H. Javed, S. El-Sappagh, and T. Abuhmed, "Robustness in deep learning models for medical diagnostics: Security and adversarial challenges towards robust ai applications," *Artificial Intelligence Review*, vol. 58, p. 12, 2024.

[90] E. Steyerberg and Y. Vergouwe, "Towards better clinical prediction models: Seven steps for development and an abcd for validation," *European Heart Journal*, vol. 35, pp. 1925–1931, 2014.

[91] C. Andrade, "Sample size and its importance in research," *Indian Journal Of Psychological Medicine*, vol. 42, pp. 102–103, 2020.

[92] M. Pourhoseingholi, M. Vahedi, and M. Rahimzadeh, "Sample size calculation in medical studies," *Gastroenterology And Hepatology From Bed To Bench*, vol. 6, p. 14, 2013.

[93] R. Riley *et al.*, "Uncertainty of risk estimates from clinical prediction models: Rationale, challenges, and approaches," *BMJ*, vol. 388, 2025.

[94] E. Štrumbelj and I. Kononenko, “Explaining prediction models and individual predictions with feature contributions,” *Knowledge And Information Systems*, vol. 41, pp. 647–665, 2014.

[95] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.

[96] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, “Interpretable machine learning: Fundamental principles and 10 grand challenges,” *Statistic Surveys*, vol. 16, pp. 1–85, 2022.

[97] P. Rothwell, “External validity of randomised controlled trials: “to whom do the results of this trial apply?”” *The Lancet*, vol. 365, pp. 82–93, 2005.

[98] S. Pocock, S. Assmann, L. Enos, and L. Kasten, “Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems,” *Statistics In Medicine*, vol. 21, pp. 2917–2930, 2002.

[99] D. Klaveren, M. Gönen, E. Steyerberg, and Y. Vergouwe, “A new concordance measure for risk prediction models in external validation settings,” *Statistics In Medicine*, vol. 35, pp. 4136–4152, 2016.

[100] S. Nijman *et al.*, “Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review,” *Journal Of Clinical Epidemiology*, vol. 142, pp. 218–229, 2022.

[101] I. White, P. Royston, and A. Wood, “Multiple imputation using chained equations: Issues and guidance for practice,” *Statistics In Medicine*, vol. 30, pp. 377–399, 2011.

[102] M. Alwateer, E. Atlam, M. Abd El-Raouf, O. Ghoneim, and I. Gad, “Missing data imputation: A comprehensive review,” *Journal Of Computer And Communications*, vol. 12, p. 53, 2024.

[103] M. Afkanpour, E. Hosseinzadeh, and H. Tabesh, “Identify the most appropriate imputation method for handling missing values in clinical structured datasets: A systematic review,” *BMC Medical Research Methodology*, vol. 24, p. 188, 2024.

[104] C. Ramspek, K. Jager, F. Dekker, C. Zoccali, and M. Diepen, “External validation of prognostic models: What, why, how, when and where?” *Clinical Kidney Journal*, vol. 14, pp. 49–58, 2021.

[105] E. Steyerberg and F. Harrell Jr, “Prediction models need appropriate internal, internal-external, and external validation,” *Journal Of Clinical Epidemiology*, vol. 69, p. 245, 2015.

[106] S. Bleeker *et al.*, “External validation is necessary in prediction research: A clinical example,” *Journal Of Clinical Epidemiology*, vol. 56, pp. 826–832, 2003.

[107] K. Moons *et al.*, “Risk prediction models: Ii. external validation, model updating, and impact assessment,” *Heart*, vol. 98, pp. 691–698, 2012.

[108] G. Collins *et al.*, “Evaluation of clinical prediction models (part 1): From development to external validation,” *BMJ*, vol. 384, 2024.

[109] R. Amarasingham, R. Patzer, M. Huesch, N. Nguyen, and B. Xie, “Implementing electronic health care predictive analytics: Considerations and challenges,” *Health Affairs*, vol. 33, pp. 1148–1154, 2014.

[110] D. Altman and P. Royston, “What do we mean by validating a prognostic model?” *Statistics In Medicine*, vol. 19, pp. 453–473, 2000.

[111] K. Moons, P. Royston, Y. Vergouwe, D. Grobbee, and D. Altman, “Prognosis and prognostic research: What, why, and how?” *BMJ*, vol. 338, 2009.

[112] A. Balendran *et al.*, “A scoping review of robustness concepts for machine learning in healthcare,” *NPJ Digital Medicine*, vol. 8, p. 38, 2025.

[113] C. Guo, G. Pleiss, Y. Sun, and K. Weinberger, “On calibration of modern neural networks,” in *International Conference On Machine Learning*, 2017, pp. 1321–1330.

[114] J. Nixon, M. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, “Measuring calibration in deep learning,” in *CVPR Workshops*, vol. 2, 2019.

[115] R. Sisk *et al.*, “Informative presence and observation in routine health data: A review of methodology for clinical risk prediction,” *Journal Of The American Medical Informatics Association*, vol. 28, pp. 155–166, 2021.

[116] R. Wolff *et al.*, “Probast: A tool to assess the risk of bias and applicability of prediction model studies,” *Annals Of Internal Medicine*, vol. 170, pp. 51–58, 2019.

[117] M. Alkan, I. Zakariyya, S. Leighton, K. Sivangi, C. Anagnostopoulos, and F. Deligianni, “Artificial intelligence-driven clinical decision support systems,” *arXiv preprint arXiv:2501.09628*, 2025.

[118] D. Hosmer Jr, S. Lemeshow, and R. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[119] T. Hastie, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

[120] Y. Hua, T. Stead, A. George, and L. Ganti, “Clinical risk prediction with logistic regression: Best practices, validation techniques, and applications in medical research,” *Academic Medicine & Surgery*, 2025.

[121] E. Zabor, C. Reddy, R. Tendulkar, and S. Patil, “Logistic regression in clinical studies,” *International Journal Of Radiation Oncology* Biology* Physics*, vol. 112, pp. 271–277, 2022.

[122] K. Olowe, N. Edoh, S. Zouo, and J. Olamijuwon, “Comprehensive review of logistic regression techniques in predicting health outcomes and trends,” *World Journal Of Advanced Pharmaceutical And Life Sciences*, vol. 7, pp. 16–26, 2024.

[123] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995.

[124] B. Schölkopf and A. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[125] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*, 2016, pp. 770–778.

[126] M. Taye, “Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions,” *Computation*, vol. 11, p. 52, 2023.

[127] H. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. Maros, and T. Ganslandt, “Transfer learning for medical image classification: A literature review,” *BMC Medical Imaging*, vol. 22, p. 69, 2022.

[128] A. Salehi *et al.*, “A study of cnn and transfer learning in medical imaging: Advantages, challenges, future scope,” *Sustainability*, vol. 15, p. 5930, 2023.

[129] C. Santos and J. Papa, “Avoiding overfitting: A survey on regularization methods for convolutional neural networks,” *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1–25, 2022.

[130] P. Petersen, *Riemannian geometry*. Springer, 2006.

[131] M. Congedo, P. Rodrigues, and C. Jutten, “The riemannian minimum distance to means field classifier,” in *BCI 2019-8th International Brain-Computer Interface Conference*, 2019.

[132] X. Pennec, P. Fillard, and N. Ayache, “A riemannian framework for tensor computing,” *International Journal Of Computer Vision*, vol. 66, pp. 41–66, 2006.

[133] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, “Classification of covariance matrices using a riemannian-based kernel for bci applications,” *Neurocomputing*, vol. 112, pp. 172–178, 2013.

[134] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, “Riemannian geometry applied to bci classification,” in *LVA/ICA*, vol. 10, 2010, pp. 629–636.

[135] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[136] D. Kingma and M. Welling, “Auto-encoding variational bayes,” Banff, Canada, 2013.

[137] G. Burch, “The history of vectorcardiography,” *Medical History*, vol. 29, pp. 103–131, 1985.

[138] G. Daniel, G. Lissa, D. Redondo, L. Vásquez, and D. Zapata, “Real-time 3d vectorcardiography: An application for didactic use,” in *Journal of Physics: Conference Series*, vol. 90, 2007, p. 012 013.

[139] F. Pedregosa *et al.*, *Scikit-learn: Machine learning in python*, 2011.

[140] J. Hoffman and S. Kaplan, “The incidence of congenital heart disease,” *Journal Of The American College Of Cardiology*, vol. 39, pp. 1890–1900, 2002.

[141] J. Hoffman, S. Kaplan, and R. Liberthson, “Prevalence of congenital heart disease,” *American Heart Journal*, vol. 147, pp. 425–439, 2004.

[142] M. Pierpont *et al.*, “Genetic basis for congenital heart defects: Current knowledge: A scientific statement from the american heart association congenital cardiac defects committee, council on cardiovascular disease in the young: Endorsed by the american academy of pediatrics,” *Circulation*, vol. 115, pp. 3015–3038, 2007.

[143] D. Van Der Linde *et al.*, “Birth prevalence of congenital heart disease worldwide: A systematic review and meta-analysis,” *Journal Of The American College Of Cardiology*, vol. 58, pp. 2241–2247, 2011.

[144] N. Jenkins and C. Ward, “Coarctation of the aorta: Natural history and outcome after surgical treatment,” *QJM*, vol. 92, pp. 365–371, 1999.

[145] H. Baumgartner *et al.*, “2020 esc guidelines for the management of adult congenital heart disease: The task force for the management of adult congenital heart disease of the european society of cardiology (esc),” *European Heart Journal*, vol. 42, pp. 563–645, 2021.

[146] K. Stout *et al.*, “2018 aha/acc guideline for the management of adults with congenital heart disease: A report of the american college of cardiology/american heart association task force on clinical practice guidelines,” *Journal Of The American College Of Cardiology*, vol. 73, e81–e192, 2019.

[147] S. Bowater and S. Thorne, “Management of pregnancy in women with acquired and congenital heart disease,” *Postgraduate Medical Journal*, vol. 86, pp. 100–105, 2010.

[148] T. Liu, A. Krentz, L. Lu, and V. Curcin, “Machine learning based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis,” *European Heart Journal - Digital Health*, vol. 6, pp. 7–22, 2025.

[149] J. Mayourian *et al.*, “Electrocardiogram-based deep learning to predict mortality in paediatric and adult congenital heart disease,” *European Heart Journal*, vol. 46, pp. 856–868, 2025.

[150] Z. Hoodbhoy, U. Jiwani, S. Sattar, R. Salam, B. Hasan, and J. Das, “Diagnostic accuracy of machine learning models to identify congenital heart disease: A meta-analysis,” *Frontiers in Artificial Intelligence*, vol. 4, p. 708 365, 2021.

[151] S. Tao *et al.*, “Development and validation of a clinical prediction model for detecting coronary heart disease in middle-aged and elderly people: A diagnostic study,” *European Journal of Medical Research*, vol. 28, p. 375, 2023.

[152] X. Liu *et al.*, “Applications of artificial intelligence-powered prenatal diagnosis for congenital heart disease,” *Frontiers in Cardiovascular Medicine*, vol. 11, p. 1 345 761, 2024.

[153] L. Liastuti and Y. Nursakina, “Diagnostic accuracy of artificial intelligence models in detecting congenital heart disease in the second-trimester fetus through prenatal cardiac screening: A systematic review and meta-analysis,” *Frontiers in Cardiovascular Medicine*, vol. 12, p. 1 473 544, 2025.

[154] C. Verheugt, C. Uiterwaal, D. Grobbee, and B. Mulder, “Long-term prognosis of congenital heart defects: A systematic review,” *International Journal of Cardiology*, vol. 131, pp. 25–32, 2008.

[155] A. Opotowsky *et al.*, “Clinical risk assessment and prediction in congenital heart disease across the lifespan: Jacc scientific statement,” *Journal of the American College of Cardiology*, vol. 83, pp. 2092–2111, 2024.

[156] F. Wang *et al.*, “Heart failure risk predictions in adult patients with congenital heart disease: A systematic review,” *Heart*, vol. 105, pp. 1661–1669, 2019.

[157] S. Cohen *et al.*, “Risk prediction models for heart failure admissions in adults with congenital heart disease,” *International Journal of Cardiology*, vol. 322, pp. 149–157, 2021.

[158] S. Mirjalili, S. Soltani, Z. Heidari Meybodi, P. Marques-Vidal, A. Kraemer, and M. Sarebanhassanabadi, “An innovative model for predicting coronary heart disease using triglyceride-glucose index: A machine learning-based cohort study,” *Cardiovascular Diabetology*, vol. 22, p. 200, 2023.

[159] G. Hu and M. Root, “Building prediction models for coronary heart disease by synthesizing multiple longitudinal research findings,” *European Journal of Cardiovascular Prevention & Rehabilitation*, vol. 12, pp. 459–464, 2005.

[160] Z. Jia *et al.*, “The importance of resource awareness in artificial intelligence for healthcare,” *Nature Machine Intelligence*, vol. 5, pp. 687–698, 2023.

[161] R. Deo, “Machine learning in medicine,” *Circulation*, vol. 132, pp. 1920–1930, 2015.

[162] G. Eisenberg and G. Stanley, “Congenital heart disease and the electrocardiogram,” *The Journal of Pediatrics*, vol. 19, pp. 452–469, 1941.

[163] V. Waldmann *et al.*, “Understanding electrocardiography in adult patients with congenital heart disease: A review,” *JAMA Cardiology*, vol. 5, pp. 1435–1444, 2020.

[164] X. Liu, H. Wang, Z. Li, and L. Qin, “Deep learning in ecg diagnosis: A review,” *Knowledge-Based Systems*, vol. 227, p. 107 187, 2021.

[165] K. Stone, R. Zwiggelaar, P. Jones, and N. Mac Parthaláin, “A systematic review of the prediction of hospital length of stay: Towards a unified framework,” *PLOS Digital Health*, vol. 1, e0000017, 2022.

[166] R. Bopche, L. Gustad, J. Afset, B. Ehrnström, J. Damås, and Ø. Nytrø, “In-hospital mortality, readmission, and prolonged length of stay risk prediction leveraging historical electronic patient records,” *JAMIA Open*, vol. 7, ooae074, 2024.

[167] E. Sandeep Kumar and P. Satya Jayadev, “Deep learning for clinical decision support systems: A review from the panorama of smart healthcare,” in *Deep Learning Techniques for Biomedical and Health Informatics*, 2020, pp. 79–99.

[168] J. Mayourian *et al.*, “Deep learning-based electrocardiogram analysis predicts biventricular dysfunction and dilation in congenital heart disease,” *Journal of the American College of Cardiology*, vol. 84, pp. 815–828, 2024.

[169] P. Shinde, M. Sanghavi, and T. Tran, “A survey on machine learning techniques for heart disease prediction,” *SN Computer Science*, vol. 6, p. 334, 2025.

[170] M. Ferdowsi, C. Goh, H. Liu, G. Tse, J. Ho Hui, and X. Wang, “Clinical application of artificial intelligence in the diagnosis, prediction, and classification of coronary heart disease,” *Cardiovascular Innovations and Applications*, vol. 10, p. 976, 2025.

[171] R. Vullings, “Fetal electrocardiography and deep learning for prenatal detection of congenital heart disease,” in *2019 Computing in Cardiology (CinC)*, 2019, Page–1.

[172] Y. Du, S. Huang, C. Huang, A. Maalla, and H. Liang, “Recognition of child congenital heart disease using electrocardiogram based on residual of residual network,” in *2020 IEEE International Conference on Progress in Informatics and Computing (PIC)*, 2020, pp. 145–148.

[173] M. Liu and Y. Kim, “Classification of heart diseases based on ecg signals using long short-term memory,” in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 2707–2710.

[174] Y. Du, C. Huang, S. Huang, and H. Liang, “Recognition of child congenital heart disease using cardiac cycle segment of electrocardiogram,” in *Computer Methods in Medicine and Health Care*, 2021, pp. 109–114.

[175] Y. Yuan, Y. Zhang, J. Wang, and P. Fang, “Classification of electrocardiogram of congenital heart disease patients by neural network algorithms,” *Scientific Programming*, 2021.

[176] G. Diller *et al.*, “Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: Data from a single tertiary centre including 10 019 patients,” *European Heart Journal*, vol. 40, pp. 1069–1077, 2019.

[177] B. Dhiyanesh, S. Ammal, K. Saranya, and K. Narayana, “Advanced cloud-based prediction models for cardiovascular disease: Integrating machine learning and feature selection techniques,” *SN Computer Science*, vol. 5, p. 572, 2024.

[178] P. Moreno-Sánchez, “Improvement of a prediction model for heart failure survival through explainable artificial intelligence,” *Frontiers in Cardiovascular Medicine*, vol. 10, p. 1 219 586, 2023.

[179] P. Pachiyannan, M. Alsulami, D. Alsadie, A. Saudagar, M. AlKhathami, and R. Poonia, “A novel machine learning-based prediction method for early detection and diagnosis of congenital heart disease using ecg signal processing,” *Technologies*, vol. 12, p. 4, 2024.

[180] H. Dhayne, R. Haque, R. Kilany, and Y. Taher, “In search of big medical data integration solutions-a comprehensive survey,” *IEEE Access*, vol. 7, pp. 91 265–91 290, 2019.

[181] X. Zhang *et al.*, “Improved prediction and risk stratification of major adverse cardiovascular events using an explainable machine learning approach combining plasma biomarkers and traditional risk factors,” *Cardiovascular Diabetology*, vol. 24, p. 153, 2025.

[182] P. Moreno-Sánchez *et al.*, “Ecg-based data-driven solutions for diagnosis and prognosis of cardiovascular diseases: A systematic review,” *Computers in Biology and Medicine*, p. 108 235, 2024.

[183] J. Zhang and Z. Zhang, “Ethics and governance of trustworthy medical artificial intelligence,” *BMC Medical Informatics and Decision Making*, vol. 23, p. 7, 2023.

[184] D. Ntiloudi, M. Gatzoulis, A. Arvanitaki, H. Karvounis, and G. Giannakoulas, “Adult congenital heart disease: Looking back, moving forward,” *International Journal Of Cardiology Congenital Heart Disease*, vol. 2, p. 100 076, 2021.

[185] M. Ladouceur *et al.*, “A new score for life-threatening ventricular arrhythmias and sudden cardiac death in adults with transposition of the great arteries and a systemic right ventricle,” *European Heart Journal*, vol. 43, pp. 2685–2694, 2022.

[186] K. Krishnathasan *et al.*, “Advanced heart failure in adult congenital heart disease: The role of renal dysfunction in management and outcomes,” *European Journal Of Preventive Cardiology*, vol. —, zwad094, 2023.

[187] J. Nelson *et al.*, “Development of a novel society of thoracic surgeons adult congenital mortality risk model,” *The Annals Of Thoracic Surgery*, 2023.

[188] L. Maessen *et al.*, “Short-term prognostic value of heart failure diagnosis in a contemporary cohort of patients with adult congenital heart disease,” *Canadian Journal Of Cardiology*, vol. 39, pp. 292–301, 2023.

[189] P. Kampaktsis *et al.*, “Machine learning-based prediction of mortality after heart transplantation in adults with congenital heart disease: A unos database analysis,” *Clinical Transplantation*, vol. 37, e14845, 2023.

[190] W. Sun *et al.*, “Towards artificial intelligence-based learning health system for population-level mortality prediction using electrocardiograms,” *NPJ Digital Medicine*, vol. 6, p. 21, 2023.

[191] C. Liu *et al.*, “Artificial intelligence–enabled model for early detection of left ventricular hypertrophy and mortality prediction in young to middle-aged adults,” *Circulation: Cardiovascular Quality And Outcomes*, vol. 15, e008360, 2022.

[192] R. Leur *et al.*, “Electrocardiogram-based mortality prediction in patients with covid-19 using machine learning,” *Netherlands Heart Journal*, vol. 30, pp. 312–318, 2022.

[193] L. Calò *et al.*, “The value of the 12-lead electrocardiogram in the prediction of sudden cardiac death,” *European Heart Journal Supplements*, vol. 25, pp. C218–C226, 2023.

[194] A. Ulloa-Cerna *et al.*, “Rechommend: An ecg-based machine learning approach for identifying patients at increased risk of undiagnosed structural heart disease detectable by echocardiography,” *Circulation*, vol. 146, pp. 36–47, 2022.

[195] E. Prifti *et al.*, “Deep learning analysis of electrocardiogram for risk prediction of drug-induced arrhythmias and diagnosis of long qt syndrome,” *European Heart Journal*, vol. 42, pp. 3948–3961, 2021.

[196] G. Diller *et al.*, “Prediction of prognosis in patients with tetralogy of fallot based on deep learning imaging analysis,” *Heart*, vol. 106, pp. 1007–1014, 2020.

[197] M. Khan, S. Aziz, M. Javeria, A. Shahjehan, Z. Mushtaq, and K. Iqtidar, “Ecg signal analysis for classification of congenital heart defects,” in *2020 International Conference On Computing And Information Technology (ICCIT-1441)*, 2020, pp. 1–5.

[198] S. Helman, E. Herrup, A. Christopher, and S. Al-Zaiti, “The role of machine learning applications in diagnosing and assessing critical and non-critical chd: A scoping review,” *Cardiology In The Young*, vol. 31, pp. 1770–1780, 2021.

[199] S. Mishra *et al.*, “Ecg paper record digitization and diagnosis using deep learning,” *Journal Of Medical And Biological Engineering*, vol. 41, pp. 422–432, 2021.

[200] M. Schäfer *et al.*, “Extraction and digitization of ecg signals from standard clinical portable document format files for the principal component analysis of t-wave morphology,” *Cardiovascular Engineering And Technology*, pp. 1–9, 2023.

[201] GE Healthcare, *Marquette 12sl ecg analysis program: Physician’s guide*, GE Healthcare: Chicago, IL, USA, 2008.

[202] R. Liu and J. McKie, *Pymupdf: Python bindings for mupdf’s rendering library*, May 2018.

[203] D. Makowski *et al.*, “Neurokit2: A python toolbox for neurophysiological signal processing,” *Behavior Research Methods*, pp. 1–8, 2021.

[204] J. Fortune, N. Coppa, K. Haq, H. Patel, and L. Tereshchenko, “Digitizing ecg image: A new method and open-source software code,” *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106 890, 2022.

[205] P. Dam, M. Boonstra, E. Locati, and P. Loh, “The relation of 12 lead ecg to the cardiac anatomy: The normal cineecg,” *Journal Of Electrocardiology*, vol. 69, pp. 67–74, 2021.

[206] B. Gopal, R. Han, G. Raghupathi, A. Ng, G. Tison, and P. Rajpurkar, “3kg: Contrastive learning of 12-lead electrocardiograms using physiologically inspired augmentations,” in *Machine Learning For Health*, 2021, pp. 156–167.

[207] A. Ribeiro *et al.*, “Automatic diagnosis of the 12-lead ecg using a deep neural network,” *Nature Communications*, vol. 11, p. 1760, 2020.

[208] F. Yger, M. Berar, and F. Lotte, “Riemannian approaches in brain-computer interfaces: A review,” *IEEE Transactions On Neural Systems And Rehabilitation Engineering*, vol. 25, pp. 1753–1762, 2016.

[209] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, “Xdawn algorithm to enhance evoked potentials: Application to brain–computer interface,” *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 8, pp. 2035–2043, 2009.

[210] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, “Multiclass brain–computer interface classification by riemannian geometry,” *IEEE Transactions On Biomedical Engineering*, vol. 59, pp. 920–928, 2011.

[211] M. Congedo, A. Barachant, and R. Bhatia, “Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review,” *Brain-Computer Interfaces*, vol. 4, pp. 155–174, 2017.

[212] G. Zoumpourlis and I. Patras, “Covmix: Covariance mixing regularization for motor imagery decoding,” in *2022 10th International Winter Conference On Brain-Computer Interface (BCI)*, 2022, pp. 1–7.

[213] L. Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, 2008.

[214] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings Of The 2019 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 1 (long And Short Papers)*, 2019, pp. 4171–4186.

[215] R. Kiros *et al.*, “Skip-thought vectors,” in *Advances In Neural Information Processing Systems*, vol. 28, 2015.

[216] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” In *Advances In Neural Information Processing Systems*, vol. 27, 2014.

[217] M. Huh, P. Agrawal, and A. Efros, “What makes imagenet good for transfer learning?” In *arXiv preprint arXiv:1608.08614*, 2016.

[218] C. Nadeau and Y. Bengio, “Inference for the generalization error,” in *Advances in Neural Information Processing Systems*, vol. 12, 1999.

[219] T. Reeves *et al.*, *Perioperative exercise testing and training society (poetts)*, Cardiopulmonary Exercise Testing (CPET) In The United Kingdom—A National Survey Of The Structure, Conduct, Interpretation And Funding. Perioper. Med.(Lond), 2, 2018.

[220] C. Wadey *et al.*, “The role of cardiopulmonary exercise testing in predicting mortality and morbidity in people with congenital heart disease: A systematic review and meta-analysis,” *European Journal Of Preventive Cardiology*, vol. 29, pp. 513–533, 2022.

[221] A. Hill and H. Lupton, “Muscular exercise, lactic acid, and the supply and utilization of oxygen,” *QJM: Quarterly Journal Of Medicine*, pp. 135–171, 1923.

[222] T. Reeves *et al.*, “Cardiopulmonary exercise testing (cpet) in the united kingdom—a national survey of the structure, conduct, interpretation and funding,” *Perioperative Medicine*, vol. 7, pp. 1–8, 2018.

[223] S. Nanas *et al.*, “Ve/vco2 slope is associated with abnormal resting haemodynamics and is a predictor of long-term survival in chronic heart failure,” *European Journal Of Heart Failure*, vol. 8, pp. 420–427, 2006.

[224] I. Abella, A. Tocci, C. Morós, and M. Grippo, “Cardiopulmonary exercise testing: Reference values in adolescent and adult patients with congenital heart diseases,” *Revista Argentina De Cardiologia*, vol. 88, 2020.

[225] Y. Shen *et al.*, “Ve/vco2 slope and its prognostic value in patients with chronic heart failure,” *Experimental And Therapeutic Medicine*, vol. 9, pp. 1407–1412, 2015.

[226] M. Hollenberg and I. B. Tager, “Oxygen uptake efficiency slope: An index of exercise performance and cardiopulmonary reserve requiring only submaximal exercise,” *Journal of the American College of Cardiology*, vol. 36, no. 1, pp. 194–201, 2000.

[227] X. Sun, J. Hansen, and W. Stringer, “Oxygen uptake efficiency plateau best predicts early death in heart failure,” *Chest*, vol. 141, pp. 1284–1294, 2012.

[228] P. Leczycki, M. Banach, M. Maciejewski, and A. Bielecka-Dabrowa, “Heart failure risk predictions and prognostic factors in adults with congenital heart diseases,” *Frontiers In Cardiovascular Medicine*, vol. 9, p. 692 815, 2022.

[229] B. Bongers, H. Hulzebos, A. Blank, M. Van Brussel, and T. Takken, “The oxygen uptake efficiency slope in children with congenital heart disease: Construct and group validity,” *European Journal Of Cardiovascular Prevention & Rehabilitation*, vol. 18, pp. 384–392, 2011.

[230] J. Otto, D. Levett, and M. Grocott, “Cardiopulmonary exercise testing for preoperative evaluation: What does the future hold?” *Current Anesthesiology Reports*, vol. 10, pp. 1–11, 2020.

[231] A. Kempny *et al.*, “Reference values for exercise limitations among adults with congenital heart disease. relation to activities of daily life—single centre experience and review of published data,” *European Heart Journal*, vol. 33, pp. 1386–1396, 2012.

[232] H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz, *Mixup: Beyond empirical risk minimization*, arXiv:1710.09412, 2017.

[233] M. Alkan, G. Veldtman, and F. Deligianni, “Riemannian prediction of anatomical diagnoses in congenital heart disease based on 12-lead ecgs,” in *2024 IEEE International Symposium On Biomedical Imaging (ISBI)*, 2024, pp. 1–5.

[234] E. Vander Velde, J. Vriend, M. Mannens, C. Uiterwaal, R. Brand, and B. J. Mulder, “Concor, an initiative towards a national registry and dna-bank of patients with congenital heart disease in the netherlands: Rationale, design, and first results,” *European Journal of Epidemiology*, vol. 20, no. 6, pp. 549–557, 2005.

[235] F. Ombelet *et al.*, “Creating the belgian congenital heart disease database combining administrative and clinical data (belcodac): Rationale, design and methodology,” *International Journal of Cardiology*, vol. 316, pp. 72–78, 2020.

[236] E. Steyerberg *et al.*, “Assessing the performance of prediction models: A framework for some traditional and novel measures,” *Epidemiology (Cambridge, Mass.)*, vol. 21, p. 128, 2010.

[237] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*. Chapman & Hall, 1995.

[238] R. Jaros, R. Martinek, and L. Danys, “Comparison of different electrocardiography with vectorcardiography transformations,” *Sensors*, vol. 19, p. 3072, 2019.

[239] G. Dower, H. Machado, and J. Osborne, “On deriving the electrocardiogram from vectorcardiographic leads,” *Clinical Cardiology*, vol. 3, pp. 87–95, 1980.