

Development of Artificial Intelligence Algorithms in Cardiovascular Research: Case Studies in Atrial Fibrillation and Myocardial Infarction

Vikram Rajapura Sreedhar

**A thesis submitted in partial fulfillment of the requirements of Liverpool John
Moores University for the degree of Doctor of Philosophy**

March 2024

Declaration

I, Vikram Rajapura Sreedhar, confirm that the work presented in this thesis is my own. Furthermore, I confirm this has been indicated in the thesis, where information has been derived from other sources.

Vikram Rajapura Sreedhar

Word count (excluding Declaration, acknowledgement, Table of contents, List of tables, figures, equations, appendices, and references):

48,594 words.

Acknowledgements

I express my heartfelt gratitude to my supervisory team, Prof. Ivan Olier-Caparroso and Prof. Sandra Ortega-Martorell, for their invaluable guidance, unwavering support, and insightful feedback throughout the course of this research. Their expertise and encouragement have been pivotal in shaping this thesis. I am also profoundly thankful to my third supervisor, Prof. Gregory Y.H. Lip, for his expert medical advice and support, which greatly enriched the quality and depth of this Study.

I am grateful to Liverpool John Moores University for providing an excellent learning and research environment that fostered my academic growth. A special mention goes to everyone at the School of Applied Mathematics and Computer Science, Faculty of Engineering and Technology, for creating a stimulating and supportive academic community. My sincere thanks to the teams at FET-Research-Admin and the Doctoral Academy for their constant assistance and encouragement. Their support has been indispensable in navigating the challenges of doctoral research.

The unwavering support and motivation from my family have been my stronghold. To my mother, whose wisdom, love, and blessings have been my guiding light; to my wife, whose belief in my capabilities knew no bounds; and to my daughter, who has been a constant source of joy and inspiration – I owe them my deepest gratitude. They have been the true heroes behind the scenes, making this journey not only possible but also memorable.

I would like to express my appreciation to the numerous mentors I have had the privilege of learning from during my research journey. Their guidance and insights have been invaluable in shaping my academic and professional path.

Lastly, I dedicate this thesis to my spiritual master and Guru, Bhagavan Sri Ramakrishna Paramhansa, whose teachings have been a source of strength and enlightenment in my life. His wisdom has been a guiding force, inspiring me to pursue knowledge with humility and dedication.

This journey has been a confluence of support, guidance, and inspiration from many, and for this, I am eternally grateful.

Abstract

This thesis explores the application of artificial intelligence techniques in the early detection and management of adverse cardiovascular events. The research is grounded in the quest to identify novel biomarkers using multi-modal data, including electrocardiograms, cardiac magnetic resonance imaging, and clinical records.

A comprehensive literature review reveals the evolution of segmentation approaches in cardiac health, highlighting the transition from traditional methods to advanced artificial intelligence driven techniques. The study then focuses on the development and validation of U-Net a neural network in segmenting cardiac magnetic resonance data of left atrium for Atrial Fibrillation fibrosis, incorporating explainability features like Gradient-weighted Class Activation Mapping and saliency maps to enhance transparency and interpretability.

Further, the research extends to analysing electrocardiogram data through machine learning and deep learning techniques, identifying critical gaps and limitations in existing approaches and suggesting areas for innovation. The integration of demographic and clinical records with electrocardiogram modalities is explored, emphasizing the potential for discovering multi-modal biomarkers that could revolutionize cardiac health management.

Generative Adversarial Networks are explored for their potential to enhance model robustness by augmenting datasets, addressing data scarcity and class imbalances prevalent in cardiac research. The implications of these findings are discussed in terms of their clinical relevance and potential applications.

Generative Adversarial Networks are investigated for their potential to enhance model robustness by augmenting datasets, addressing data scarcity and class imbalances prevalent in cardiac research. In the realm of electrocardiogram data, Earth Mover's Distance scores reveal the congruence between real and synthetic data distributions. Results indicate that Generative Adversarial Networks effectively capture the underlying pathology of myocardial infarction, with a mean Earth Mover's Distance score of 3.278 in principal component analysis and 5.544 in t-distributed stochastic neighbour embedding, suggesting a closer alignment in principal component analysis space for myocardial infarction signals. However, Generative Adversarial Networks posed challenges in recognizing healthy control instances, with a mean Earth Mover's Distance score of 4.576 in principal component analysis and 7.838 in t-distributed stochastic neighbour embedding. These findings underscore the need for careful consideration of the trade-offs introduced by synthetic data augmentation in medical diagnostics.

In further analysis, this study employs machine learning techniques, including Logistic Regression, Random Forest, and Extreme Gradient Boosting, to advance the classification of cardiovascular disorders. Interbeat intervals were extracted from electrocardiogram data and combined with clinical factors to enrich the dataset for binary and multi-class classifications of disorders such as myocardial infarction, bundle branch block, cardiomyopathy, dysrhythmia, and healthy heart conditions. The augmentation of this data with synthetic samples generated by Generative Adversarial Networks significantly enhanced model performance. Notably, in binary classification tasks, Random Forest model without data augmentation, demonstrated the highest accuracy and Area Under the Curve at 0.89 and 0.93 respectively, suggesting its effectiveness in accurately classifying and differentiating conditions. Meanwhile, models augmented with synthetic data, showed remarkable precision for the myocardial infarction class, with Extreme Gradient Boosting achieving the highest Area Under the Curve of 0.96. Precision for healthy controls detection increased from 0.69 to 0.99, while the

F1 score improves from 0.64 to 0.80. These results underscore the balanced performance across various classes, significantly enhanced by the integration of synthetic data, leading to improved diagnostic capabilities.

Furthermore, the study delves into multi-class classification of cardiac disorders, where data augmented by Generative Adversarial Networks significantly boosts model performance. Overall accuracy increased from 0.68 to 0.75, and precision and recall were notably improved across various classes, including bundle branch block (precision from 0.76 to 0.82, recall from 0.50 to 0.82), cardiomyopathy (precision from 0.67 to 0.75, recall from 0.50 to 0.70), and dysrhythmia (precision from 0.71 to 0.80, recall from 0.50 to 0.70). These results emphasize the importance of data quality in synthetic data generation, especially when faced with class imbalances.

In conclusion, this study showcases the efficacy of Generative Adversarial Networks -based data augmentation in enhancing the accuracy and reliability of predictive analytics in cardiovascular health. It provides valuable insights into the potential and challenges of synthetic data in medical diagnostics, with implications for improved patient care and disease detection.

Conclusively, this research contributes significantly to the field of cardiology, offering novel artificial intelligence driven solutions for early detection and management of adverse cardiovascular conditions. It underscores the transformative potential of artificial intelligence in healthcare, advocating for the development of ethically responsible artificial intelligence applications that align with clinical needs and enhance patient health outcomes.

Publication and Dissemination of Results

Peer-reviewed Journal Articles – In Preparation

Synthetic Data Generation for Improved Cardiac Event Prediction using Generative Adversarial Networks, Vikram R Sreedhar, Sandra Ortega-Martorell, Gregory Y. H. Lip, and Ivan Olier – to be submitted to Algorithms MDPI

Talks and Conferences

- LjMU Faculty Research Day, Online, Liverpool, UK. May 2023. Oral Presentation. Developing Artificial Intelligence Multi-modal Models for Atrial Fibrillation & Associated Adverse Cardiovascular Events (ACE)
- IEEE Cloud Computing 2022 Conference, December 2022. Oral Presentation. Generative - AI way of augmenting data and solving data imbalance issues.

Table of Contents

Declaration.....	2
Acknowledgements.....	3
Abstract.....	4
Publication and Dissemination of Results.....	6
Table of Contents.....	7
List of Figures	12
List of Tables	13
1. Introduction	13
1.1 Motivation	14
1.2 Study aims and research objectives.	15
1.3 Novelties of the Research.....	16
1.4 Summary of the thesis.....	16
1.5 Conclusion	18
2 Advanced Techniques in Cardiac Image Segmentation: A Literature Review	19
2.1 Introduction.....	19
2.2 Search criteria for the literature review	20
2.3 Segmentation Approaches	21
2.3.1 Conventional Segmentation Approaches	22
2.3.1.1 Thresholding.....	22
2.3.1.2 Edge-based.....	22
2.3.1.3 Region-based.....	22
2.3.1.4 Clustering	22
2.3.1.5 Watershed.....	23
2.3.2 Contemporary techniques in image segmentation	23
2.3.2.1 Graphcut.	23
2.3.2.2 Artificial Neural Networks -based.	23
2.3.3 Image segmentation approaches used in AF studies.....	24
2.4 Data augmentation approaches	25
2.4.1 Geometric transformations	26
2.4.2 Colour space or photometric transformations	26
2.4.3 Deep learning approaches for data augmentation.....	26

2.4.4	Data Augmentation Approaches used in AF studies.....	31
2.5	Loss Functions and Performance Metrics for Image Segmentation	31
2.5.1	Loss Functions for Segmentation.....	31
2.5.2	Performance Metrics for Segmentation	34
2.5.3	Loss functions and performance metrics used in AF studies.....	35
2.6	Other considerations – model explainability and transfer learning.....	36
2.6.1	Model interpretability and explainability	36
2.6.2	Transfer learning	37
2.7	Conclusion	37
3	Review of Artificial Intelligence models on Cardiac Diagnostics using Electrocardiogram	38
3.1	Introduction.....	38
3.2	Search criteria.....	39
3.3	Multimodal data, ECG Analysis and Cardiac Incident prediction – A Review	41
3.4	Limitation of Studies reviewed.....	46
3.4.1	Process Gaps	46
3.4.2	Methodology Gaps.....	46
3.5	Conclusion	47
4	Left Atrium segmentation using Neural Networks	48
4.1	Introduction.....	48
4.2	Data and pre-processing.....	48
4.2.1	Primary Dataset – Cardiac Atlas Study.....	48
4.2.2	Validation Dataset – Cardiac Delayed Enhancement Segmentation Challenge (cDEMRS)	49
4.3	Methods.....	51
4.3.1	Study Design.....	51
4.3.2	Convolutional neural networks.....	52
4.3.3	U-Net architecture for left atrium segmentation	52
4.3.4	Data Augmentation.....	55
4.3.5	Loss Function.....	57
4.3.6	Model Explainability Methods in Image Analysis	58
4.3.6.1	Vanilla Saliency.....	58
4.3.6.2	Smooth Gradient Saliency.....	58
4.3.6.3	Grad-CAM (Gradient-weighted Class Activation Mapping)	59
4.3.7	Ethics approval.....	60
4.3.8	Model Validation through Transfer Learning.....	60

4.4	Results	62
4.5	Model Metrics	64
4.6	Generalisation of the model on CDEMRS Database.....	64
4.7	Model explainability	66
4.8	Outcomes and Future Directions.....	68
4.9	Conclusion	69
5	In-depth Analysis of Electrocardiogram – Techniques and Applications in Artificial Intelligence	70
5.1	Background	70
5.2	Objective:.....	72
5.3	Data collection and description: PTB Diagnostic ECG Database	73
5.4	Data pre-processing.....	74
5.4.1	Noise Filtering – Baseline Wander	75
5.4.2	Noise Filtering - Band-Pass Filter	75
5.4.3	Noise Filtering - Notch Filter	75
5.4.4	Noise Filtering - Band-Stop Filter	76
5.4.5	Signal to Noise Ratio	76
5.5	Strategies for Data Augmentation in ECG Analysis – A focus on GANs.....	79
5.5.1	GAN based synthetic data creation based on raw signals.	79
5.5.2	GAN based synthetic data creation based on Inter beat intervals and metadata.	80
5.6	Methods.....	81
5.6.1	CNN for ECG raw signals classification.....	82
5.6.2	Inter Beat intervals (IBI) feature extraction.	84
5.6.2.1	Time- Domain HRV features [182]	85
5.6.2.2	Non-linear HRV features	85
5.6.2.3	Frequency-domain HRV features	86
5.6.2.4	Peak Detection – Mathematical Background.....	87
5.6.3	Binary Classification using IBI features and patient metadata.	88
5.6.4	Multilabel Classification using IBI features and patient metadata.	90
5.6.5	Optimal threshold for classification.....	91
5.6.5.1	Thresholding concepts – Mathematical background.....	91
5.6.6	Model Explanation – Shapley Additive method.....	92
5.7	Validation methods	93
5.8	Results	93
5.8.1	Classification Performance & Statistical Analysis - Binary Classification –Model 12.....	98

5.8.2	Classification Performance & Statistical Analysis - Multilabel Classification Model 8	100
5.9	Model Explainability and Discussion	104
5.9.1	Model Interpretation SHAP – Binary Classification Model	104
5.9.2	Top predictors from SHAP – feature distribution plots and interpretations.....	105
5.9.2.1	Systolic blood pressure	105
5.9.2.2	Frequency measure (HRV) – LF-absolute – ECG Lead V2.....	106
5.9.2.3	Age	107
5.9.2.4	Frequency measure (HRV) – HF-absolute – ECG Lead iii	107
5.9.2.5	Additional considerations	108
5.10	Summary of findings and Limitations	109
5.11	Conclusion	110
6	Enhancing Electrocardiogram Data with Generative Adversarial Networks – Approaches and Outcomes	111
6.1	Introduction.....	111
6.2	Objectives	112
6.3	Data Processing	112
6.4	Methods.....	113
6.4.1	GAN Architecture for ECG – Signals	113
6.4.2	Mathematical background for creating and training real and synthetic samples.....	116
6.4.3	GAN Architecture - Synthetic samples with IBI features and clinical factors for binary classification.....	117
6.4.4	Mathematical background tabular of GAN for synthetic samples in binary classification.....	120
6.4.5	GAN Architecture - Synthetic samples with IBI features and clinical factors for multilabel classification.....	122
6.4.5.1	Simple GAN with Binary Cross-Entropy loss function	122
6.4.5.2	Simple GAN with Custom loss function	125
6.4.5.3	Wasserstein GAN (WGAN) with Wasserstein loss	127
6.4.5.4	Wasserstein GAN (WGAN) with Custom loss function	130
6.4.6	PCA and t-SNE Analysis	132
6.4.7	Earth Mover Distance	133
6.5	Results	134
6.5.1	Real versus synthetic samples ECG – raw signals	147
6.5.2	Real versus synthetic samples – GAN augmented for binary classification	149
6.5.2.1	PCA and t-SNE analysis.....	149

6.5.2.2	Analysing the perplexity and EMD values.....	150
6.5.2.3	Binary Classification model performance with augmented synthetic data.....	150
6.5.3	Real versus synthetic samples – GAN augmented for multilabel classification	151
6.6	Summary of Findings and Limitations	152
6.7	Conclusion	154
7	Discussion: Bridging AI Insights with Clinical Applications	155
7.1	Research objectives restatement.....	155
7.2	Implications of the Findings.....	155
7.2.1	MRI Segmentation of model explainability.....	155
7.2.2	ECG Classification enhancements	157
7.2.3	GAN-Based Augmentation	158
8	Concluding Insights – The Future of AI in Cardiac Health.....	160
8.1	Summary of Research contributions	160
8.2	Impact on the field of ECG Analysis and AI.....	160
8.3	Future Horizons and Practical Applications.....	160
9	References	162
10	Appendices.....	179
10.1	GAN Architectures – Binary Classification – Raw Signals	179
10.2	Image Segmentation Experiments and results.....	184

List of Figures

FIGURE 2-1 - PRISMA FLOW DIAGRAM OF STUDIES INCLUDED IN REVIEW	21
FIGURE 2-2: REPRESENTATION OF THE GRAPH CUT TECHNIQUE. A) PIXELS ARE CONSIDERED AS A GRAPH. B) THE COST BETWEEN THE ADJACENT PIXELS IS DEFINED IN TERMS OF THE INVERSE PROPORTION OF INTENSITY DIFFERENCES BETWEEN THE PIXELS. C) THE IMAGE IS SPLIT USING THE MINIMUM CUT METHOD. D) THE BEST SEGMENT OF THE IMAGE IS FOUND.	23
FIGURE 2-3: REPRESENTATION OF A NEURAL NETWORK, INCLUDING THE INPUT LAYER, TWO HIDDEN LAYERS, THE OUTPUT LAYER, AND THE BLUE CONNECTIONS REPRESENTING THE WEIGHTS.	24
FIGURE 2-4: DATA AUGMENTATION APPROACHES	26
FIGURE 3-1 - MULTIMODAL DATA & ECG DATA WITH AI APPLICATIONS - STUDY REVIEW.....	40
FIGURE 4-1: MRI IMAGES OF LA CAVITY - CARDIAC ATLAS PROJECT DATASET	49
FIGURE 4-2: LGE CMR IMAGES - CDEMIRIS DATASET	50
FIGURE 4-3: U-NET ARCHITECTURE. ADAPTED FROM [32].....	53
FIGURE 4-4: EXAMPLE OF THE CONVOLUTION AND MAX-POOLING PROCESS	53
FIGURE 4-5: EXAMPLES OF UP SAMPLING	54
FIGURE 4-6: EXAMPLE OF THE CONCATENATION PROCESS	54
FIGURE 4-7: EXAMPLE OF CONVOLUTION.....	54
FIGURE 4-8: DATA AUGMENTATION AND PIXEL DISTRIBUTION	57
FIGURE 4-9: TWO EXAMPLES BEST AND WORST PREDICTIONS	63
FIGURE 4-10 : BEST MODEL METRICS.....	64
FIGURE 4-11 - OPTIMAL SEGMENTATION OF CDEMIRIS DATASET	65
FIGURE 4-12 - SUBOPTIMAL SEGMENTATION OF CDEMIRIS DATASET	66
FIGURE 4-13 : BEST MODEL METRICS.....	67
FIGURE 4-14 : VANILLA SALIENCY RESULTS	67
FIGURE 4-15 : SMOOTH GRAD VANILLA SALIENCY RESULTS	68
FIGURE 5-1 - ECG WAVEFORM OF VARIOUS CARDIOVASCULAR CONDITIONS AND HEALTHY HEART	71
FIGURE 5-2: MEDIAN SNR BETWEEN UNCLEANNED AND CLEANED TRAIN ECG DATA	78
FIGURE 5-3: MEDIAN SNR BETWEEN UNCLEANNED AND CLEANED TEST ECG DATA.....	78
FIGURE 5-4: MEDIAN SNR BETWEEN UNCLEANNED AND CLEANED VALIDATION ECG DATA.....	79
FIGURE 5-5: STUDY METHODOLOGY	81
FIGURE 5-6: CNN ARCHITECTURE FOR ECG SIGNALS BINARY CLASSIFICATION.....	83
FIGURE 5-7 : ROC CURVE OF BEST MODEL (MODEL 12)	99
FIGURE 5-8 : CONFUSION MATRIX OF BEST MODEL 12- BINARY CLASSIFICATION	99
FIGURE 5-9 : CONFUSION MATRIX OF BEST MODEL 8- MULTILABEL CLASSIFICATION – TRAIN DATA	101
FIGURE 5-10 : CONFUSION MATRIX OF BEST MODEL 8- MULTILABEL CLASSIFICATION - TEST DATA	102
FIGURE 5-11 : ROC CURVE OF BEST MODEL 8 - MULTILABEL CLASSIFICATION	103
FIGURE 5-12 : SHAP VALUE IMPACT - SUMMARY PLOT ON BEST MODEL 12 - BINARY CLASSIFICATION	104
FIGURE 5-13 : DISTRIBUTION PLOT - FEATURE SYSTOLIC BP	105
FIGURE 5-14 : DISTRIBUTION PLOT - FEATURE FREQ MEASURE LOW FREQUENCY ABSOLUTE - LEAD V2	106
FIGURE 5-15 : DISTRIBUTION PLOT - FEATURE AGE	107
FIGURE 5-16 : DISTRIBUTION PLOT - FEATURE FREQ MEASURE HIGH FREQUENCY ABSOLUTE - LEAD III.....	107
FIGURE 6-1 : GENERATOR MODEL FOR SYNTHETIC SAMPLES OF HC CLASS - LEAD I.....	114
FIGURE 6-2: DISCRIMINATOR MODEL FOR SYNTHETIC SAMPLES OF HC CLASS - LEAD I	115
FIGURE 6-3: POSITIVE AND NEGATIVE CLASS GENERATOR – IBI & METADATA.....	119
FIGURE 6-4: DISCRIMINATOR - IBI & METADATA	119
FIGURE 6-5: GENERATOR & DISCRIMINATOR - MULTILABEL CLASSIFICATION - SIMPLE GAN WITH BCE LOSS FUNCTION	124
FIGURE 6-6: GENERATOR & DISCRIMINATOR - MULTILABEL CLASSIFICATION - SIMPLE GAN WITH CUSTOM LOSS FUNCTION	126
FIGURE 6-7: GENERATOR & DISCRIMINATOR - MULTILABEL CLASSIFICATION - WGAN WITH WASSERSTEIN LOSS FUNCTION.....	129
FIGURE 6-8: GENERATOR & DISCRIMINATOR - MULTILABEL CLASSIFICATION - WGAN WITH CUSTOM LOSS FUNCTION	131
FIGURE 6-9: LEAD 0 TO 3 - PCA, T-SNE PLOTS AND EMD SCORES- REAL VS SYNTHETIC SAMPLES.....	134
FIGURE 6-10: LEAD 4 TO 7 - PCA, T-SNE PLOTS AND EMD SCORES- REAL VS SYNTHETIC SAMPLES.....	135
FIGURE 6-11 : LEAD 8 TO 11 - PCA, T-SNE PLOTS AND EMD SCORES- REAL VS SYNTHETIC SAMPLES.	136
FIGURE 6-12 : PCA, T-SNE PLOTS AND EMD SCORES- IBI FEATURES REAL VS SYNTHETIC SAMPLES – BINARY CLASSIFICATION (PERPLEXITY FACTOR 30).....	138

FIGURE 6-13 : PCA, T-SNE PLOTS AND EMD SCORES- IBI FEATURES REAL VS SYNTHETIC SAMPLES – BINARY CLASSIFICATION (PERPLEXITY FACTOR 225).....	138
FIGURE 6-14: PCA, T-SNE PLOTS AND EMD SCORES- REAL VS SYNTHETIC SAMPLES – MULTILABEL – SIMPLE GAN WITH BCE LOSS FUNCTION.....	140
FIGURE 6-15:PCA, T-SNE PLOTS AND EMD SCORES- REAL VS SYNTHETIC SAMPLES – MULTILABEL – SIMPLE GAN WITH CUSTOM LOSS FUNCTION.	141
FIGURE 6-16:PCA, T-SNE PLOTS AND EMD SCORES- REAL VS SYNTHETIC SAMPLES – MULTILABEL – WGAN WITH WASSERSTEIN LOSS FUNCTION.	142
FIGURE 6-17: PCA, T-SNE PLOTS AND EMD SCORES- REAL VS SYNTHETIC SAMPLES – MULTILABEL – WGAN WITH CUSTOM LOSS FUNCTION.	143
FIGURE 10-1:GENERATOR MODEL FOR SYNTHETIC SAMPLES OF HC CLASS - LEAD V6	180
FIGURE 10-2: DISCRIMINATOR MODEL FOR SYNTHETIC SAMPLES OF HC CLASS - LEAD V6	181
FIGURE 10-3: GENERATOR MODEL FOR SYNTHETIC SAMPLES OF MI CLASS - LEAD V6	182
FIGURE 10-4:DISCRIMINATOR MODEL FOR SYNTHETIC SAMPLES OF MI CLASS - LEAD V6	183

List of Tables

TABLE 2-1: MEAN ESTIMATES OF THE GLOBAL BURDEN OF DISEASE FOR AF IN 1990 AND 2019. [1][8]	19
TABLE 2-2: STUDIES INVOLVING THE SEGMENTATION OF ATRIA, VENTRICLES, AND OTHER AREAS OF INTEREST.	24
TABLE 2-3 - COMPARATIVE ANALYSIS OF DATA AUGMENTATION APPROACHES AND TECHNIQUES	27
TABLE 2-4: DATA AUGMENTATION TECHNIQUES USED IN THE AF STUDIES.	31
TABLE 2-5: LIST OF LOSS FUNCTIONS USED IN SEGMENTATION TASKS, ORGANIZED BY THEIR CATEGORY AND MAIN AIM.	31
TABLE 2-6: PERFORMANCE OR EVALUATIONS METRICS USED IN SEGMENTATION TASKS.	34
TABLE 2-7: LOSS FUNCTIONS, PERFORMANCE METRICS AND OTHER RELEVANT INFORMATION FROM THE AF STUDIES SURVEYED.	35
TABLE 3-1:RELATED WORK IN ECG DATA ANALYSIS USING ML AND DL APPROACHES	42
TABLE 4-1: COMPARISON OF SALIENCY METHODS.....	59
TABLE 4-2:SUMMARY OF THE TRAINING RESULTS OF THE BEST FOUR MODELS DEVELOPED.	62
TABLE 5-1 : PTB DATASET	74
TABLE 5-2: TIME DOMAIN HRV FEATURES	85
TABLE 5-3: NON-LINEAR HRV FEATURES	86
TABLE 5-4: FREQUENCY MEASURES - HRV	87
TABLE 5-5: LIST OF MODELS FOR BINARY CLASSIFICATION.....	89
TABLE 5-6 : COMPARISON OF RESULTS MODEL 1 TO 2 – MULTILABEL CLASSIFICATION.....	95
TABLE 5-7 : COMPARISON OF RESULTS MODEL 3 TO 4 – MULTILABEL CLASSIFICATION.....	96
TABLE 5-8 : COMPARISON OF RESULTS MODEL 5 TO 6 – MULTILABEL CLASSIFICATION.....	96
TABLE 5-9 : ON OF RESULTS MODEL 7 TO 8 – MULTILABEL CLASSIFICATION.....	97
TABLE 6-1: DOWNSAMPLING - PROS AND CONS	112
TABLE 6-2:LIST OF MODELS DESCRIPTION USED FOR BINARY CLASSIFICATION.....	145
TABLE 6-3: BINARY CLASSIFICATION MODELS PERFORMANCE WITHOUT AND WITH DATA AUGMENTATION VIDE GAN.....	145
TABLE 6-4: LIST OF MODELS DESCRIPTION USED FOR MULTI LABEL CLASSIFICATION.....	146
TABLE 6-5: MULTI LABEL CLASSIFICATION MODELS PERFORMANCE WITHOUT AND WITH DATA AUGMENTATION VIDE GAN.	146
TABLE 6-6: EMD SCORES LEAD WISE FOR SYNTHETIC SAMPLES.	147
TABLE 10-1 : SEGMENTATION EXPERIMENTS AND RESULTS PERFORMED ON LOCAL RESOURCE.....	184
TABLE 10-2: SEGMENTATION EXPERIMENTS AND RESULTS DONE ON SERVER.....	185

1. Introduction

The burden of cardiovascular diseases (CVDs) globally underscores a critical area of medical research, where innovative methodologies can significantly transform diagnosis, treatment, and prevention strategies. This thesis explores the integration of advanced artificial intelligence (AI) technologies to tackle cardiovascular health challenges, highlighting the potential of AI in revolutionizing the approach to heart-related ailments. As the leading cause of mortality worldwide, the study of non-communicable diseases, with a focus on cardiovascular conditions such as myocardial infarction, atrial fibrillation, bundle branch block, and dysrhythmia, offers a pivotal opportunity for significant advancements in medical technology and patient care. By leveraging multi-modal data analysis and predictive modelling techniques, this research aims to refine diagnostic accuracy and enhance clinical outcomes through the innovative use of AI-driven tools. The following sections detail the motivation behind the research, specific aims and objectives, the novelty of the approaches used, and a structured outline of the thesis content.

1.1 Motivation

Cardiovascular diseases (CVDs) are the foremost group of conditions causing death globally but are only one part of a larger spectrum of non-communicable diseases (NCDs). NCDs are chronic conditions that include not only cardiovascular diseases but also cancers, chronic respiratory diseases, and diabetes. These diseases are collectively responsible for almost 70% of all deaths worldwide, affecting health care systems and imposing significant economic burdens due to their chronic nature and the extensive care they often require.[1], [2]

Cardiovascular diseases (CVDs) remain a significant global health concern, with adverse events such as stroke, heart failure, coronary artery disease, hypertension, and mortality posing substantial burdens on healthcare systems worldwide. In addition to their direct impact on health, CVDs also have far-reaching implications for productivity and life expectancy. Individuals affected by CVDs often experience reduced productivity due to illness, hospitalization, and long-term disability, leading to economic losses for both individuals and societies. Furthermore, CVDs are a leading cause of premature death globally, significantly impacting life expectancy and quality of life for affected individuals and their families.

Among the cardiovascular conditions, myocardial infarction (MI), atrial fibrillation (AF) bundle branch block (BBB), and dysrhythmia (Dsy) are particularly significant. Myocardial infarction, commonly known as a heart attack, (ICD-10 code: I21), occurs when blood flow to a part of the heart is blocked for a long enough time that part of the heart muscle is damaged or dies. Atrial fibrillation (ICD-10 code: I48), is an irregular and often very rapid heart rhythm that can lead to blood clots in the heart. Atrial fibrillation increases the risk of stroke, heart failure and other heart-related complications. Bundle branch block (ICD-10 code: I44.4 for left and I45.1 for right), which involves delays or obstructions along the pathway that electrical impulses travel to make the heartbeat, can indicate underlying heart disease and increases the risk of cardiac events. Dysrhythmia (ICD-10 code: I49), encompassing various types of abnormal heart rhythms, can significantly impact the heart's ability to effectively pump blood, posing risks of severe health complications such as heart failure and sudden cardiac arrest.

The global burden of these diseases is immense. Cardiovascular diseases are responsible for more than 17 million deaths annually, which represents about 31% of all global deaths. These conditions not only lead to substantial mortality but also contribute to decreased quality of life due to heart failure, sudden cardiac arrest, and the long-term effects of stroke.[2]

Addressing these challenges requires a deeper understanding of the underlying mechanisms and predictive markers associated with adverse cardiovascular events. The primary motivation of this research is to leverage artificial intelligence (AI) methodologies to develop predictive models capable of identifying biomarkers and risk factors for adverse cardiovascular outcomes, thus contributing to improved risk assessment, early detection, and patient management strategies.

Adverse cardiovascular events has been studied from several angles such as detection, risk prediction and management, and its understanding involves the analysis of patient clinical records along with several data modalities, mainly electrocardiograms (ECG), although others such as cardiac magnetic resonance imaging (CMR), echocardiograms (Echo) and computerized tomography scans (CT) are also used.[3] Biomarkers are indicators of the presence or state of a medical condition[4], [5]. They are biological measures derived from bodily fluids, tissues, or imaging that indicate normal or pathogenic processes, or responses to therapeutic interventions. Examples include levels of particular proteins in the blood or changes detected on an MRI scan, ECG signals. Typically, clinicians rely on one or more of these to help with their decision making. It is known that combining biomarkers greatly reduce uncertainty in their decisions.[5] . Importantly, this process is guided by established clinical guidelines, such as those provided by the National Institute for Health and Care Excellence (NICE) in the UK, which standardize and enhance the reliability of clinical assessments.[6] For instance, biomarkers extracted from the left atrium using CMR are regarded as the most reliable. However, CMR is expensive, and its access is very limited. Although typically less reliable than CMR, ECG is significantly cheaper and far more accessible. With the support of AI modelling, fusing ECGs with clinical records could lead to the identification of novel biomarkers that could reduce the dependence of CMR.

1.2 Study aims and research objectives.

The overarching aim of this study is to develop advanced AI-driven predictive models for identifying biomarkers and risk factors associated with adverse cardiovascular events. To achieve this aim, the following research objectives were pursued:

Objective 1: Analyse and synthesize a comprehensive range of AI algorithms and methodologies utilized in modelling adverse cardiovascular events during the first year, focusing on multi-modal data analysis and predictive modelling techniques to identify gaps and formulate foundational knowledge for subsequent developments.

Objective 2: Create an AI model using Cardiac Magnetic Resonance (CMR) for segmentation tasks by the end of the first year to advance the precision of AI-driven segmentation in cardiac imaging analysis, thereby enhancing the accuracy and utility of diagnostic imaging.

Objective 3: Develop and evaluate AI models that integrate demographic and clinical factors, including electrocardiogram (ECG) data, by the second year to discover novel biomarkers and risk factors for adverse cardiovascular outcomes, assessing their validity against clinical benchmarks.

Objective 4: Investigate the feasibility and apply generative adversarial networks (GANs) for synthetic data generation and augmentation during the third year to enhance the robustness and data diversity for cardiovascular disease prediction, evaluating their impact on model performance.

Objective 5: Assess the predictive performance, interpretability, and clinical utility of the developed AI models in the third year by examining their effectiveness in identifying and stratifying patients at risk of adverse cardiovascular events, using established performance metrics and comparing outcomes.

1.3 Novelties of the Research

The novel aspects of this research include:

- Application of advanced AI methodologies for multi-modal data analysis in adverse cardiovascular event prediction, encompassing CMR, ECG, and clinical data integration.
- Emphasis on interpretability and transparency in AI-driven predictive models, ensuring clinical relevance and usability in cardiovascular risk assessment.
- Exploration of GANs for synthetic data generation and augmentation, addressing data scarcity and enhancing the robustness of predictive models in adverse cardiovascular event prediction.

1.4 Summary of the thesis

- Chapter 2: Conducts a comprehensive review of literature on segmentation approaches used in cardiac health, ranging from traditional thresholding methods to advanced region-based segmentation techniques employing AI models. This chapter sets a historical and methodological context for AI-driven segmentation in cardiac imaging.
- Chapter 3: Presents a detailed review of literature on ECG data analysis using machine learning and deep learning. It covers various techniques and AI methodologies, identifying relevant gaps and limitations in current approaches and setting the stage for innovations and novelties in our research.
- Chapter 4: Explores the development and validation of an AI model for the segmentation of the left atrium using CMR data. It emphasizes the explainability of AI decision-making processes in segmentation tasks, employing techniques like Grad-CAM and saliency maps to enhance model transparency and interpretability.
- Chapter 5: Discusses the advancement of AI models in integrating demographic and clinical records with ECG modalities, focusing on the discovery of multi-modal biomarkers and the implications for cardiac health.
- Chapter 6: Focuses on the utilization of Generative Adversarial Networks (GANs) for dataset augmentation. This chapter demonstrates how GANs can enhance the robustness of AI models and contribute to the diversity of medical datasets, particularly in the context of cardiac imaging.
- Chapter 7: Delves into the implications of the research findings, tying them back to the initial research objectives stated in Chapter 1. We evaluate the impact of the developed models and techniques on MRI segmentation, ECG classification enhancements, and the innovative use of GANs for data

augmentation in medical AI. The chapter restates the research objectives, reaffirming the original aims and scope of the research. Further, the broader implications of the research findings in the fields of AI, medical imaging, and cardiology are explored. Through this assessment, contributions of our research to advancing the understanding and application of AI in cardiovascular healthcare is highlighted.

- Chapter 8: Concludes the thesis by summarizing the research contributions and their impact on the field of ECG analysis and machine learning. This chapter also looks forward to future horizons and practical applications, discussing how this research can be applied in healthcare settings and the ethical considerations involved.

Table 1-1 will provide a clear linkage between the study objectives and the corresponding content in this thesis

Objective	Description	Reference to Chapters
1	Analyse and synthesize AI algorithms and methodologies for modelling adverse cardiovascular events, focusing on multi-modal data analysis and predictive modeling techniques.	Chapter 2 (AI methodologies and segmentation approaches), Chapter 3 (Review of AI algorithms and methodologies in multi-modal data analysis including ECG data)
2	Create an AI model using CMR for segmentation tasks, aiming to advance the precision of AI-driven segmentation in cardiac imaging, addressing explainability and model validation through transfer learning.	Chapter 4 (Development and validation of segmentation models, including addressing gaps found in model explainability and generalization)
3	Develop and evaluate AI models integrating demographic and clinical factors to discover novel biomarkers and risk factors for cardiovascular outcomes.	Chapter 5 (Development and assessment of AI models using clinical and demographic data)
4	Investigate the feasibility and application of GANs for synthetic data generation to enhance data diversity for cardiovascular disease prediction.	Chapter 6 (Application of GANs for data augmentation and enhancement of model robustness)
5	Assess the predictive performance, interpretability, and clinical utility of developed AI models in identifying patients at risk of adverse cardiovascular events.	Chapter 5 (5.9 Evaluation of model performance, interpretability, and clinical utility)

1.5 Conclusion

Thus, this introduction chapter provides a comprehensive overview of the global burden of cardiovascular diseases (CVDs), detailing their significant impact on health and economic systems worldwide. This chapter underscores the importance of using advanced artificial intelligence (AI) methodologies to tackle these challenges. Through the integration of AI in cardiovascular research, this study aims to improve diagnostic processes and patient care. The motivation and objectives outlined here pave the way for the subsequent chapters, which further explore the application of AI technologies in addressing the complexities of cardiovascular diseases. For a deeper exploration of specific AI methodologies used in cardiac health, Chapter 2 – Literature review delves into segmentation approaches and reviews the literature surrounding current technologies and practices.

2 Advanced Techniques in Cardiac Image Segmentation: A Literature Review

2.1 Introduction

Atrial fibrillation (AF), the most prevalent form of cardiac arrhythmia, poses significant challenges in medical diagnostics and treatment due to its sporadic nature and the complex interplay of factors influencing heart function. This chapter delves into various segmentation approaches critical in the analysis of cardiac magnetic resonance images, pivotal for understanding the structural dynamics and pathological changes in the heart associated with AF. The incorporation of advanced computational techniques such as machine learning, deep learning, and transfer learning has revolutionized the precision and efficiency of these methods. Through a systematic review of literature and exploration of state-of-the-art methodologies, this chapter aims to present a comprehensive examination of current practices and innovative advancements in the segmentation of cardiac images, providing a crucial foundation for subsequent diagnostic improvements.

AF is the most common type of cardiac arrhythmia. In AF, the atria contract randomly, sometimes hindering the heart muscle relaxation between contractions, hence reducing the heart's efficiency and performance. AF can be i) occasional (paroxysmal AF), lasting for a few minutes or hours and not always requiring treatment, ii) persistent, where the person requires treatment, iii) long-standing persistent, where the AF is continuous and lasts longer than 12 months, and iv) permanent, where the irregular heart rhythm cannot be restored. For many years, AF was regarded as a trivial condition, but it has been increasingly recognised as an independent contributor to mortality, morbidity and impaired quality of life[7]. In addition to the number of deaths, the burden of a disease is also measured by a metric called disability-adjusted life-years (DALYs), which is a time-based metric that combines years of life lost due to premature mortality and years of healthy life lost due to disability. According to the World health organization (WHO), the AF prevalence doubled from 1990 to 2019 (>28M versus >56M), the DALYs increased from 3.79M to 8.39M, and the mortality increased from 117K to 315K (see Table 2-1) [1][8]. This data also supports the view of AF as a new epidemic [2][9][10].

Table 2-1: Mean estimates of the Global Burden of Disease for AF in 1990 and 2019. [1][8]

Table 2-1: Mean estimates of the Global Burden of Disease for AF in 1990 and 2019. [1][8]

Metric	1990	2019
Prevalence	28,285,570.89	56,695,082.92
DALYs	3,787,838.26	8,393,635.01
Deaths	117,037.99	315,336.77

Treatment options for AF are limited as the mechanisms that drive and maintain AF are elusive [11]. The structure and function of the human atria, particularly the left atria, is closely related to AF clinical phenotypes, with direct implications in the level of success in treatment strategies [12][13][14]. In addition, the remodelling of the left atrium can determine the clinical course of AF [15]. Magnetic resonance imaging (MRI) is a modality

that has been widely used to study the human atria structure and remodelling [16]. Gadolinium contrast agents are used in MRI scans to enhance the clarity of the images of a patient's inner formations, such as the atria. Gadolinium-enhanced magnetic resonance imaging (GE-MRI) is widely used to study the amount of fibrosis or scars across the atria [17][18] and to help in creating an AF treatment plan. Usually, the quantification of the level of fibrosis in left atrium chamber of heart has been done manually. Due to left atrium cavity having a lesser volume, inhibited by a thin atrial wall added by neighbouring intricate anatomy, manual segmentation is very laborious and susceptible to errors. To assist with this task, automatic segmentation approaches have been developed, aiming to improve accuracy, speed, and reduced human intervention.

Often artificial intelligence (AI), machine learning (ML) and deep learning (DL) are interchangeably used. However, ML is a subset of AI and DL in turn is a subset of ML. Both ML and DL is component of the bigger universe of AI. AI is a process of making machines think like humans. It is a process of conveying data, inferences, information, and intelligence to machines. Through AI human tasks of classifying, recognition, translation, and decision making is automated.

ML is a process of creating computer programs which allows the system to automatically learn from the data. It focuses on building applications, programs and models which solve intricate problems in various fields. It uses both structured and unstructured data and can be used to predict an outcome related to an underlying clinical problem. DL is a component of ML which emphasizes on learning from voluminous data through multifaceted algorithms. Its fundamentals are inspired by the human brain and uses neural networks which enables to learn patterns and make decisions [120] [121].

2.2 Search criteria for the literature review

To obtain associated contributions, to the study a structured search strategy was employed to identify relevant papers and publications. The search was conducted using the following databases: PubMed and Scopus. The inclusion of Scopus, alongside PubMed, was aimed at broadening the scope of the search to capture a wider array of publications, including those not indexed in PubMed, thereby enhancing the inclusiveness and diversity of the studies reviewed.

The search was formulated using specific Medical Subject Headings (MeSH) terms and their synonyms to ensure consistency and comprehensiveness. The following MeSH terms along with their synonyms were used:

Atrial Fibrillation: ["AF", "Atrial Flutter", "Cardiac Arrhythmias"]

Cardiac Imaging: ["Cardiac MRI", "Magnetic Resonance Imaging, Cine", "Heart Imaging"]

Segmentation: ["Image Segmentation", "Region Segmentation", "Boundary Detection"]

Following search query Strings were used,

"((machine learning approaches) OR (artificial intelligence approaches) OR (deep learning approaches)) AND (atrial fibrillation OR atrial segmentation) AND (cardiac imaging OR MRI)"

"Left Atrium and Left Ventricle segmentation using (Thresholding OR Graph Cut OR Clustering OR Region-Based Segmentation) methods"

Conference proceedings from relevant IEEE conferences were also searched, focusing specifically on titles related to cardiac segmentation and imaging technologies. The dual approach of using structured database queries alongside manual searches of conference records ensures a thorough exploration of both peer-reviewed articles and significant grey literature.

Inclusion Criteria: Studies that focus on the application of AI, machine learning, or deep learning techniques for the segmentation of the left atrium and ventricle using cardiac imaging modalities. Only articles published in English with full texts available were considered.

Exclusion Criteria: Papers that do not focus primarily on segmentation tasks using MRI or cardiac imaging. Studies published before 1972 were excluded to maintain the relevance of the technological context. Additionally, non-peer-reviewed articles, opinion pieces, and editorials were excluded.

To ensure the studies included are unbiased and representative of the current understanding in the field, a rigorous peer-review filter was applied, and efforts were made to include a diverse range of methodologies and viewpoints by accessing multiple databases and incorporating both journal articles and conference papers.

The last update to the included studies was conducted on 12th April 2021, and all searches were restricted to materials published from 1972 to the present to focus on contemporary techniques and advancements in the field.

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram for studies included is shown in figure 2-1.

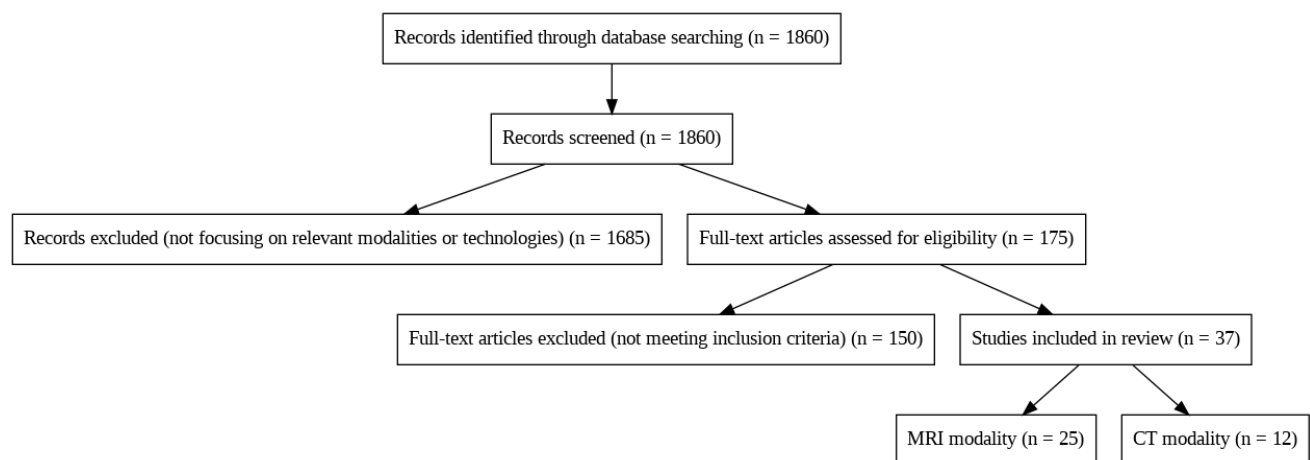


Figure 2-1 - PRISMA flow diagram of studies included in review

2.3 Segmentation Approaches

The segmentation of cardiac magnetic resonance images helps to analyse the heart's anatomical structure, study changes in the heart, and identify pathological regions. This is useful for surgical planning and other image-guided interventional procedures [19], [20]. Below there is a summary of the most commonly used segmentation techniques, grouped into two main segmentation approaches, conventional or modern.

2.3.1 Conventional Segmentation Approaches

2.3.1.1 Thresholding

This is the simplest and most used method of segmentation. Given a threshold T , the pixel located at lattice position (x, y) with greyscale value g_{xy} is allocated to category 0 if $g_{xy} \leq T$ or else to category 1. There are several ways to decide on T – this can be manually or using methods such as the intermeans or the minimum error. In the intermeans method, a guess is made initially at a likely value for the threshold. From this estimated threshold, the mean values of pixels are segregated into two cohorts, and the threshold is repositioned to lie exactly in between the two means. The process is iterated until the threshold stops changing its value. In the minimum error method, from an initial guess value of the threshold, the proportions, means and variances of pixel values in the two groups are calculated. The threshold is repositioned according to the above criterion, and proportions, means and variances are recalculated. The process is iterated until there are no changes in values between iterations [21].

2.3.1.2 Edge-based

In edge-based segmentation, an edge filter is applied to the image, where output pixels are characterised as edge or non-edge, with pixels that are not separated by an edge being assigned to the same category. Many edge detection methods use a gradient operator, along with a threshold identification on the gradient, to pinpoint the edges in the images. Areas of disjointedness in grey levels, colour, surface, etc. are detected using derivative operations, and derivatives can be calculated using differential operators.[21]–[23]

2.3.1.3 Region-based

Region-based segmentation works by repetitively grouping pixels that are neighbours and have similar values, and splitting pixels that are dissimilar in value into different groups. The thresholding in this method emphasises the difference in pixel intensity. The two subprocesses in this method are:

- 1) Region growing. This process clusters the pixels with comparable intensities. It begins with a smaller set of pixels and starts accruing based on certain pre-determined resemblance limits. It starts by choosing a random seed pixel in the image and comparing it with its adjacent pixels. If there is a match or similarity in adjoining/adjacent pixels, then they are added to the initial seed pixel, thus making the region larger. When the growth of the region cannot proceed further, the algorithm selects an alternative seed pixel, which does not belong to any existing region and iterates the process again. Region growing methods are useful for noisy images, where detecting edges becomes more challenging.
- 2) Region splitting and merging. This process uses two steps done together in combination – region severing and region integrating. The severing step involves iteratively splitting an image into regions having similar characteristics and the integrating step involves combining the neighbouring regions having similarities. As a result, the segmented image will have several comparable marked regions distributed across the image pixels, hence the intermittent segmentation output will contain dispersed groups of adjacent regions that have comparable properties. Finally, the scattered clusters are merged, which is comparable with adjacent regions, and based on similar gradations.[21]

2.3.1.4 Clustering

This family of methods can be used to segment images into clusters of pixels with similar or comparable features. Specifically, methods such as k-means and fuzzy c-mean (FCM) can be used for image segmentation. K-means aims to divide the image space into k groups/clusters. For this, the image pixels are allocated to 'K' cluster based on the closest cluster centroid (initially selected randomly), and the process iterates until the

centroids no longer shift or are considered to have converged. The image is then segmented into k groups or clusters where the constituent pixels show some degree of similarity, and each pixel is assigned to only one of the clusters. FCM has the same aim as k -means, but in FCM the image pixels can be represented by more than one cluster with varying levels of belongingness or similarity to each cluster/group.

2.3.1.5 Watershed

Watershed [23] is a region-based method that considers a topological representation of a landscape, where the elevation values are usually user-defined markers. The method then outlines “basins” (low levels) which are the minima points and hence, basins are swamped from the markers until basins meet on watershed lines. The watersheds will result in distinct basins from each other. Hence the picture gets segmented, with pixels assigned to each region or watershed.

2.3.2 Contemporary techniques in image segmentation

2.3.2.1 Graphcut.

The graph cut technique has its origin in graph theory, aiming to split a single connected graph into two or more separate graphs with the least separation cost. A graph is defined as a set of vertices connected by edges, which have a specific cost (or weight). Two graphs are defined as disjointed if no edge connects any of the vertices in one graph to any vertex in the other graph, by removing the edges with minimal cost.[24] The graph cut method is used in image segmentation using the image as a graph where the pixels relate to vertices (see Fig 2-2. A). The edges relate to the resemblance between adjacent pairs of pixels, and their costs (orange lines in figure 2-2. B) are proportional to the similarity between two pixels. The edges cost will be higher between comparable regions (or pixels). Thus, the minimum cut will ensure the best segmentation of the image. [25][26]

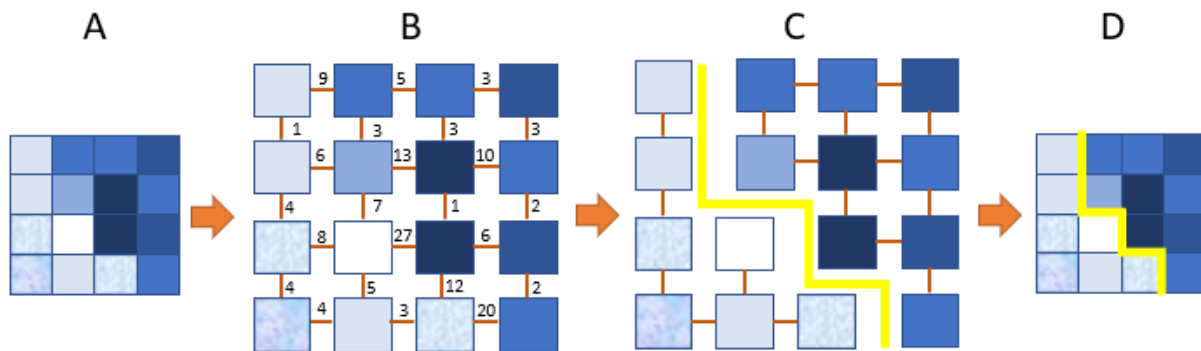


Figure 2-2: Representation of the graph cut technique. A) Pixels are considered as a graph. B) The cost between the adjacent pixels is defined in terms of the inverse proportion of intensity differences between the pixels. C) The image is split using the minimum cut method. D) The best segment of the image is found.

2.3.2.2 Artificial Neural Networks -based.

Artificial Neural Networks (ANNs), inspired by biological neural networks, are based on collections of artificial neurons which are connected (like synapses) to other neurons, with those connections having a weight that adapts as the learning occurs. ANNs are organised in layers, with an input layer for the data that enters the network, accompanied by one or more hidden layers, and an output layer that computes the predictions as per the Neural Network (NN) (see Figure 2-3) [27]–[31]. Different ANN-based methods have been successfully produced for image segmentation, especially using Convolutional Neural Networks (CNN) which are a type of ANNs. Examples of these methods are U-Net [32], ResNet[33] region-based CNN (R-CNN)[34] and recurrent neural network (RNN)[35]. More details on some of them are provided in the Methods section.

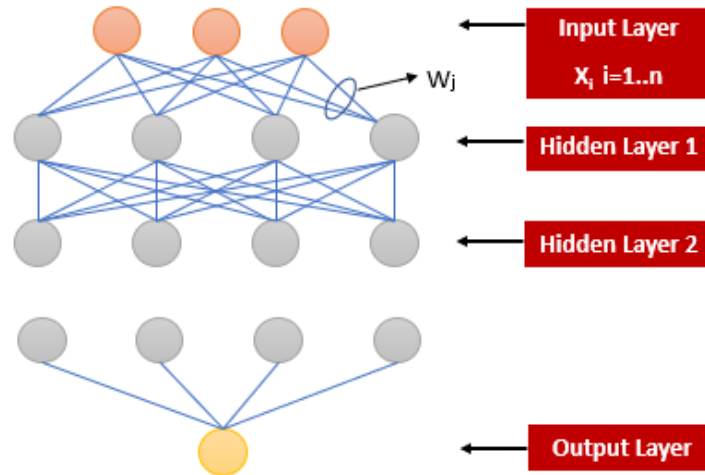


Figure 2-3: Representation of a neural network, including the input layer, two hidden layers, the output layer, and the blue connections representing the weights.

2.3.3 Image segmentation approaches used in AF studies.

Table 2-2 summarizes the list of studies that involved atrial and/or ventricles segmentation. It also includes the different segmentation approaches used, and the metrics used for the evaluation of the quality of the segmentation.

Table 2-2: Studies involving the segmentation of atria, ventricles, and other areas of interest.

Related work	Anatomy	Segmentation Approach
Karim et al.[36]	Left atrium	Graph cut
Positano et al.[37]	Left ventricle	Clustering, thresholding
Kolipaka et al.[38]	Left ventricle	Automatic thresholding
Hennemuth et al.[39]	Left ventricle	Clustering
Schmidt et al.[40]	Left ventricle	Thresholding - standard deviation
Yan et al.[41]	Left ventricle	Thresholding - standard deviation
Perry Daniel et al.[42]	Left atrium	K-means clustering
Lu et al.[43]	Left ventricle - myocardium tissue	Graph cut
Oakes et al.[44]	Left atrium	Thresholding - standard deviation
Knowles et al.[45]	Left atrium	Thresholding - maximum intensity Study
Ravanelli et al.[46]	Left atrium	Thresholding - standard deviation
Veni et al.[47][48]	Left atrium	Graph cut
Jin et al.[49]	Left atrium appendage	Graph cut
Ortuño et al.[50]	Mitral and aortic valves	Graph cut
Agarwal et al.[51]	Left atrium, mouse brain	Shape-based clustering
Kang et al. [52]	Left and right atria and ventricles	K-means clustering and watershed
Wachinger et al.[53]	Left atrium	Region-based, clustering and thresholding
Razeghi et al.[54]	Pulmonary vein	U-Net
Xiong et al.[55]	Left atrium	U-Net, RCNN
Lu et al.[56]	Right atrium	U-Net
Yang et al.[57]	Left atrium, pulmonary vein	U-Net
Du et al.[58]	Left atrium	CNN
Lei et al.[59]	Left atrium	Graph cut and CNN

Hung et al.[60]	Atrial fibrosis, animal	CNN
Xiong et al.[61]	Left atrium	U-net and Fully Convolutional Neural Network (FCN)
Yang et al.[62]	Left atrium, pulmonary Vein	U-Net
Chakravarty et al.[63]	Left atrium, Optic Disc and Optic Cup	RNN
Zhu et al.[64]	Left atrium	Region-based methods – Multiphase graph cut (shape prior method [65])
Kutra et al.[66]	Left atrium, pulmonary Vein	Region-based methods and clustering using support vector machines
Depa et al.[67]	Left atrium	Atlas-based segmentation, weighted thresholding (label fusion algorithm)
Koikkalainen et al.[68]	Left atrium	Region-based methods
Firouznia et al.[69]	Left atrium and pulmonary Vein	Clustering, feature extraction using Fractal analysis and Random Forest Classifier
Liu et al.[70]	Left atrium, pulmonary Vein	ResNet34
Li et al.[71]	Left atrium and pulmonary Vein	3D U-net with group attention and channel attention block
Chen et al.[72][73]	Left atrium	U-Net and transfer learning with encoder as ResNet50
Baskaran et al. [74][75]	Aorta, vena cavae, pulmonary artery, coronary sinus, right ventricular wall and left atrial wall	U-net
Jin et al.[76]	Left atrium appendage	FCN with VGG16 layer network

2.4 Data augmentation approaches

Recently, ANNs have proven to be able to perform image segmentation successfully. However, deep architectures like the ones implemented in such networks may require an extensive amount of data to achieve acceptable performance and be able to generalise. This, therefore, applies to the analysis of medical imaging modalities such as MRI and CT scans. Though MRI has various advantages in terms of soft tissue imaging, high-resolution images, and multi-channel images, it is marred by the inconsistent image quality. Similarly, CT is widely used for being less expensive, having short lead times for scans, high spatial resolution, and higher diagnosing capability in detecting sub-arachnoids haemorrhage and calcifications. However, CTs are inferior to MRI in terms of soft tissue imaging and have high radiation exposure leading to less consistent image quality [19].

Some types of image inconsistencies are related to their contrast, resolution, texture, signal to noise. It can also be related to the acquisition technology (Siemens, Philips, etc.). These inherent challenges in the data, together with the limited availability (due to sharing restrictions), are some of the major reasons for the poor generalisation of deep learning models [77]. To mitigate them, data augmentation strategies have been proposed and used in the literature.

The purpose of the data augmentation is to increase the size of training data, making the quality of the data more consistent and adding variety and variability [78], leading to increased levels of reproducibility and consistency in the results [79]. Data augmentation results in reducing stochastic uncertainties in deep learning algorithms. In a survey of reproducible & non-reproducible deep learning models [80], of the total of 91% of the non-reproducible works, 65% of them did not follow data augmentation strategies. In addition, class imbalance in data [81] has been addressed and mitigated through effective data augmentation approaches

[82]. Figure 2-4: Data Augmentation approaches includes the general classification and taxonomy of data augmentation techniques.

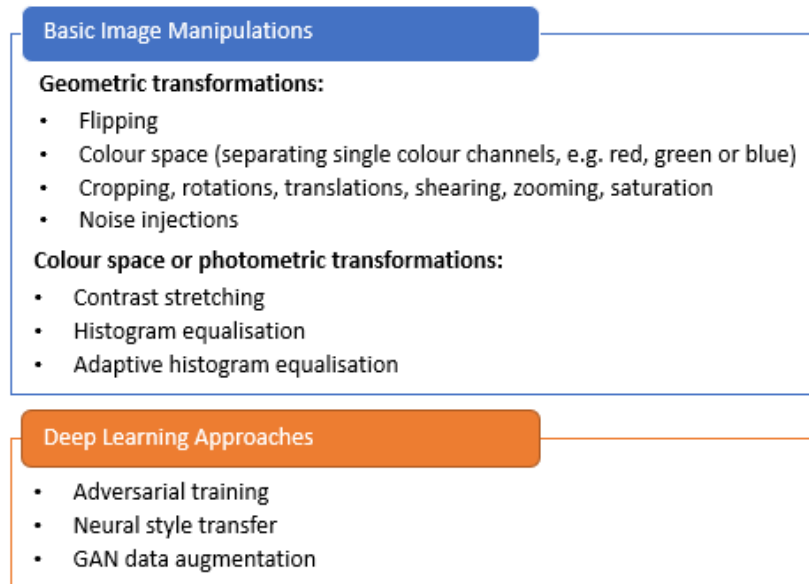


Figure 2-4: Data Augmentation approaches

2.4.1 Geometric transformations

Manipulations of image geometry can be defined as geometric transformations.

Geometric transformations augur well when positional bias exist in training image data. They are well suited for facial recognition and classification tasks. In [83], the authors found that geometric transformations, especially image rotations, are more effective than photometric transformations (explained below) in improving neural network performance.

2.4.2 Colour space or photometric transformations

Lighting and intensity variances are known to pose the biggest challenges in medical imaging segmentation and classification tasks. Photometric transformations aim to mitigate them, although are not able to completely overcome them [83][84]. Images are arrays of pixels, of which histogram can be produced. Manipulating the histogram to change the pixel information and features forms the core of photometric transformations.

2.4.3 Deep learning approaches for data augmentation

Deep learning approaches encompass a suite of advanced model-based techniques designed to modify or generate new data, thereby enhancing the diversity and richness of training sets. These methods are pivotal for introducing complex and varied image variations that may not be present in the original dataset, significantly enhancing the model's ability to generalize from its training. However, while these techniques are powerful, they come with notable challenges. They are computationally intensive, demanding significant resources, which can be a limiting factor in their application. Moreover, there is an inherent risk that the data generated could be non-representative or introduce biases, potentially leading to models that perform well in training but poorly in real-world applications.

Table 2-3 summarises techniques in each approach with their advantages and drawbacks.

Table 2-3 - Comparative analysis of Data augmentation approaches and techniques

Data Augmentation Approach	Data Augmentation Technique	Description	Advantages	Drawbacks
Geometric Transformations <i>These involve altering the spatial configuration of images, such as rotating, scaling, or cropping, to simulate different scenarios.</i>	Rotation	Images are rotated at various angles from 0 to 360 degrees, simulating how objects might appear from different viewpoints.	- Allows models to recognize objects in images regardless of their orientation. - Essential for applications requiring rotational invariance.	Significant rotations can obscure important image details
	Cropping	Sections of the image are selected and retained, while the rest are discarded, focusing on specific features.	- Focuses training on relevant image features, potentially enhancing feature learning. - Useful for extracting localized patterns.	Can miss contextual information outside the cropped area, which may be crucial for accurate interpretation.
	Noise Addition	Random noise is added to images, simulating real-world imperfections.	- Trains the model to be effective even with poor-quality inputs. - Mimics real-world scenarios where images may have digital noise.	Excessive noise can degrade model performance by obscuring subtle features.
	Translation	The image is shifted across the plane, altering the position of objects within the frame.	Develops spatial robustness by teaching the model to recognize objects regardless of their position within the image frame.	Can lead to incomplete images where important features are shifted out of view.
	Contrast Adjustment	The contrast of the image is adjusted for its color and luminance	- Enhanced feature visibility - Standardization by normalizing the images from different source of light conditions	Can introduce visual artifacts if not applied uniformly

	Flipping	Flipping the image either horizontally or vertically. Pixels are rearranged while retaining the features and textures of images.	Effectively increases the data size	Contextual inaccuracy where orientation of image plays a crucial role
	Scaling/resizing	Image is resized either larger or smaller than the original size	Size normalization where input images need to be uniform for algorithm processing	Can degrade image quality by blurriness and pixelation.
	Augmenting colour	Pixels from grayscale are changed to RGB (red, green, blue) and vice versa. This alters the color properties of an image	Enhances features by colour differentiation	Adding colour to original greyscale images can mislead algorithms by introduction non original hues that might be interpreted wrongly
	Zooming	Image is expanded or contracted for more feature and texture visibility	Allows focusing on specific parts of image	- Loss of context. - Over zooming beyond the resolution limit can lead to pixelation
	Shear Intensity	Image is transformed to slant form. Image is stretched which is not seen in rotation	Simulating motion or tilt	Distort the image making it less representative of real-world data
	Perspective Transformation	Image is studied from a different viewpoint. A perspective transformation done without knowing the positional coordinates may degenerate the dataset. But it merits emphasising on parts of the image that the network needs to learn.	- Enhanced depth perception by providing more three-dimensional feel from a single viewpoint. - Improved object recognition.	- Can degrade the dataset quality if transformation is applied incorrectly. - Computational intensity
<i>Photometric Transformations</i> <i>These adjustments</i>	Kernel Filters	To Sharpen or blur images	- Enhanced image detailing - Flexible image manipulation - Feature enhancements	- Potential information loss - Dependency or sensitivity to parameters. Incorrect parameters like kernel

<i>affect image color and intensity, such as adjustments to brightness, contrast, and color saturation.</i>				size can exacerbate the issues meant to mitigate.[85]
	Mixing Images[86]	a method by which the pixel values in an image is averaged	▪ Can lead to smoothing extreme features or outliers in single images ▪ Increases size and variability of training dataset	▪ Overfitting to averaged features rather than learning to recognize actual and distinct patterns ▪ Feature distortion leading to unnatural feature combinations.
	Random Erasing[87]	to delete certain areas in images	▪ Improved generalization and robustness to occlusion	▪ Potential for misleading training
	Histogram Equalisation	The distribution of pixel intensities is adjusted to enhance image contrast, making features more discernible.	▪ Improves visibility of subtle features, especially in medical imaging where detail is crucial. ▪ Can aid in better feature extraction.	▪ Can result in artificial appearance if applied too aggressively. ▪ May highlight noise along with features.
	Contrast limited adaptive histogram equalisation	Contrast Limited Adaptive Histogram Equalization (CLAHE) is an advanced form of histogram equalization that is used to improve the contrast in images. Unlike standard histogram equalization, which enhances the contrast of the entire image, CLAHE applies the histogram equalization to small data regions called tiles. These tiles are then combined using bilinear interpolation to eliminate artificially induced boundaries.	▪ Enhance local contrast ▪ Reduces noise overamplification ▪ Adaptive and enhances contrast in specified areas of image based on local content features	▪ Computationally intensive ▪ Parameter Sensitivity ▪ Risk of Artifacts
	Contrast Stretching	Enhances contrast by adjusting the luminance range within the image, particularly improving the visibility of darker areas.	▪ Makes images clearer, especially under varied lighting conditions.	▪ Can lead to overexposure of bright areas, losing detail in these regions.

			Facilitates the detection of features that are subtle or faint	
	Colour Space Transformation	Converts images between color models, such as from RGB to grayscale, affecting how colors are processed and interpreted.[88]	- Reduces computational demands by simplifying the colour information. - Can improve segmentation accuracy in certain contexts.[89][90]	Loss of colour information can be detrimental in contexts where colour differentiation is key.[91]
<i>Deep Learning Approaches</i> <i>Advanced model-based techniques that modify or generate new data to enhance training sets.</i>	Adversarial Training	Introduces small, calculated modifications to images to create "adversarial" examples that challenge the model during training.[92]	- Improves model robustness against subtle perturbations that could otherwise lead to misclassifications. - Tests model limits in controlled ways.[93]	- Complex to implement correctly. - May divert resources from learning genuine features.
	GAN Data Augmentation	Uses generative adversarial networks to create new images that expand the variability of the training dataset.[94]	- Generates realistic images that can fill gaps in training data. - Particularly useful for rare conditions or viewpoints.[95]	GAN-generated images may sometimes be too idealized or diverge from realistic scenarios.
	Neural Style Transfer	Applies the style of one image to the content of another, altering its aesthetic attributes while preserving structural details.[96]	- Allows creative recombination of image features, enhancing the dataset diversity. - Useful for cross-domain training scenarios.	May result in overly stylized images that do not contribute to practical training objectives.

2.4.4 Data Augmentation Approaches used in AF studies.

Table 2-4 summarizes the data augmentation techniques used for the segmentation tasks in the AF studies surveyed.

Table 2-4: Data augmentation techniques used in the AF studies.

Study	Data modality	Data augmentation techniques used
Razeghi et al. [54]	MRI	Rotations, shear, Gaussian noise
Xiong et al. [55]	MRI	CLAHE, rotation, elastic deformation, perspective scaling, flips
Lu et al. [56]	MRI	Horizontal and vertical flipping, multiple angle rotations, random cropping, Gaussian noise, colour jittering
Yang et al. [57]	MRI	Normalisation
Xiong et al. [61]	MRI	Balancing of positive and negative pixels, Elastic deformations, affine transformations, wrapping
Koikkalainen et al. [68]	MRI	Smoothing to images using covariance matrix, deformation modes using eigenvectors, partitioning, non-rigid scaling, principal component analysis
Liu et al. [70]	CT	Normalisation, rotation, zooming
Li et al. [71]	CT	Normalisation
Chen et al. [72], [73]	CT	Normalisation, rotation, zooming
Baskaran et al. [74], [75]	CT	Normalisation
Jin et al [76]	CT	Transformation of grey image to colour palettes, smoothing of images

MRI: magnetic resonance imaging. CT: computerised tomography.

2.5 Loss Functions and Performance Metrics for Image Segmentation

2.5.1 Loss Functions for Segmentation

Loss functions are used to measure how far the estimated value is from its true value. They are used as part of the training process to evaluate how well the algorithm is modelling the data in every iteration, as part of the optimization process. Loss functions differ significantly based on the area and context of the learning the machine learning algorithm is tasked for. Hence, we will focus the attention on the loss functions used on image segmentation tasks, particularly in the AF studies surveyed. Table 2-5 classifies the loss functions used in segmentation tasks according to its objective [97]. DICE and cross-entropy are the most commonly-used loss function used in many segmentation tasks [98][55][99][100].

Table 2-5: List of loss functions used in segmentation tasks, organized by their category and main aim.

Category	Aim of the loss function	Loss functions
Distribution-based	To decrease variation between two distributions	• Cross entropy • Focal loss • Top K loss • Distance map penalized cross entropy. • Weighted Cross Entropy
Region-based	Reduce the divergence or increase the intersection(overlap) between ground truth and predicted segmentation	• DICE Loss • Sensitivity-specificity Loss • IOU (Jaccard) Loss • Lovasz loss

		• Tversky loss • Generalized DICE loss • Focal Tversky loss • Asymmetric similarity loss • Penalty loss
Boundary-based	Reduce the distance between predicted segmentation and ground truth	• Boundary loss • Hausdorff distance loss
Combination	Weighted combinations of loss functions	• Combo loss (DICE + Cross entropy) • Exponential logarithmic loss • DICE loss with focal loss

Binary cross-entropy loss function

The cross entropy[101][32] is a degree of variance between the probability distribution of the actual outcome and the probability distribution of predictions. It increases exponentially as the forecast deviates from the real outcome. Cross entropy is also denoted as negative log-likelihood. The binary cross-entropy loss function is suitable for the classification of either one of the two available outcomes. The loss function is represented as

$$L_{CE} = -(y \log[\hat{y}_i] + (1 - y_i) \log(1 - \hat{y}_i)) \quad \text{Equation 2-1}$$

Where L is the loss function and y_i is the actual outcome and $\hat{y}_i$, read as y hat is the prediction of the algorithm. The overall cost function (loss value summed over each training epochs) can be represented as

$$C = -\frac{1}{N} \sum_{i=0}^n (y \log[\hat{y}_i] + (1 - y_i) \log(1 - \hat{y}_i)) \quad \text{Equation 2-2}$$

Weighted binary cross-entropy loss function.

Weighted cross-entropy is an extension of binary cross-entropy, where w is the weight applied to each class. w is inversely proportional to the frequency of each class. This loss function is useful whenever there is imbalanced data, and weights aim to penalize the majority classes. Alternatively, weight can be assigned based on each pixel which is maximum in representation. It can be represented as

$$L_{wCE} = -(w * y \log[\hat{y}_i] + (1 - y_i) \log(1 - \hat{y}_i)) \quad \text{Equation 2-3}$$

With the overall cost function represented as

$$C = -\frac{1}{N} \sum_{i=0}^n (w * y \log[\hat{y}_i] + (1 - y_i) \log(1 - \hat{y}_i)) \quad \text{Equation 2-4}$$

Categorical cross-entropy loss function

This loss function is used in multiclass classification tasks. Here, the same principle of binary cross-entropy applies but it is a summation over all the classes. If M is the number of classes, the loss function can be represented as

$$L_{CEcat} = \sum_{j=0}^M (y \log[\hat{y}_i]) \quad \text{Equation 2-5}$$

The overall cost function is then calculated as $C = -\frac{1}{N} \sum_{i=0}^n \sum_{j=0}^M (y \log[\hat{y}_i])$ Equation 2-6

Focal loss function

The focal loss [102] is a variation of the cross-entropy function used more frequently in classification and segmentation tasks having imbalanced data. Higher class frequency is down weighted, ensuring the model equally learns from the smaller class. It can be represented as

$$L_F \equiv \frac{-1}{N} \sum_{i=0}^n -\alpha(1-p)^\gamma \log(p) \quad \text{Equation 2-7}$$

Where $\gamma > 0$ and α ranges from $[0,1]$

DICE loss function

DICE loss function [103] is derived out of the DICE metric, which assigns a higher penalty for mismatched overlaps between the predicted segmentation and the ground truth or region of interest. As a result, the DICE loss function is highly responsive to segmentation errors [104]. It can be represented as

$$L_{DICE} = 1 - 2 \frac{\sum_{r=1}^n y_p^r \hat{y}_p^r}{\sum_{r=1}^n (y_p^r)^2 + \sum_{r=1}^n (\hat{y}_p^r)^2} \quad \text{Equation 2-8}$$

Where y_p^r is the true or ground truth binary value of pixel p belonging to class r and $\hat{y}_p^r$ is the predicted value of pixel p belonging to class r. The other version of DICE loss function is not the square the ground truth binary value and predicted value of pixel terms in denominator. [105]

Generalised DICE loss

The generalised DICE loss [106] is an improved version of DICE loss with a weight w being imposed to address class imbalance (the weights are inversely proportional to the class frequencies). It can be represented as

$$L_{wDICE} = 1 - 2 \frac{\sum_{r=1}^n w y_p^r \hat{y}_p^r}{\sum_{r=1}^n w (y_p^r)^2 + \sum_{r=1}^n w (\hat{y}_p^r)^2} \quad \text{Equation 2-9}$$

Where $w = \frac{1}{(\sum_{i=1}^n y_p^i)^2}$

Jaccard or Mean IOU loss

The Jaccard loss or mean IOU (Intersection over Union)[107] loss is similar to the DICE loss and can be represented as

$$L_{IOU} = 1 - \frac{\sum_{r=1}^n y_p^r \hat{y}_p^r}{\sum_{r=1}^n (y_p^r + \hat{y}_p^r - (y_p^r \hat{y}_p^r))} \quad \text{Equation 2-10}$$

Tversky index and focal Tversky loss

The Tversky index (TI) is an asymmetric similarity measure that is a generalization of the DICE coefficient and the Jaccard index. It can be represented as

$$TI = \frac{TP}{TP + \alpha FN + \beta FP}$$

Equation 2-11

where TP is True Positive, FN is False Negative, and FP is False Positive. The Tversky index adds two parameters, α and β , where $\alpha + \beta = 1$. In the case where $\alpha = \beta = 0.5$, it simplifies into the dice coefficient. It simplifies to the Jaccard index if $\alpha = \beta = 1$. By setting the value of $\alpha > \beta$, false negatives can be de-weighted even more.

The focal Tversky loss (FTL)[108] is a generalisation of the Tversky loss. The non-linear nature of the loss gives hold over how the loss behaves at different values of the Tversky index attained. It is represented as $FTL = (1 - TI)^\gamma$ where γ is a parameter that controls the non-linearity of the loss. As γ tends to positive ∞ , the gradient of the loss tends to ∞ as the Tversky index tends to 1. As γ tends to 0, the gradient of the loss tends to 0 as TI tends to 1.

Combined loss function

These loss functions are the combination of one of the loss functions discussed. Combo Loss [105], [109] can be represented as $L_{Combo} = L_{CE} + L_{DICE}$.

Equation 2-12

2.5.2 Performance Metrics for Segmentation

Table 2-6 summarises the performance or evaluations metrics used in segmentation tasks.

Table 2-6: Performance or evaluations metrics used in segmentation tasks.

Segmentation Accuracy metric	Description	Mathematical representation
Pixel Accuracy[101]	Ratio of correctly classified pixels against the total pixels	$PA = \frac{\sum_{i,j=0}^n p_{ii}}{\sum_{i,j=0}^n p_{ij}}$ Where P_{ii} is pixels of class i and P_{ij} is total pixels of class i and j
Intersection over Union (IOU) or Jaccard Index[107]	It is defined as the area of intersection between the predicted segmentation area and the ground truth area divided by the area of union between the predicted segmentation and ground truth area	$IOU = J(A, B) = \frac{ A \cap B }{ A \cup B }$ Where A and B represent ground truth and predicted segmented areas. Mean IOU is average IOU over all classes.
Precision/Recall and F1 score (DICE)[103]	Precision is the ratio of true positives over true and false positives. Recall is the ratio of True positives over true positive and false negative. F1 Score is a harmonic mean of precision and recall	$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}$ $F1 \text{ Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$ Where TP is true positive, FP is false positive, and FN is false negative. F1 Score is also known as DICE coefficient which can be defined as twice the overlap area of ground truth and predicted area, divided by total pixels. $DICE = \frac{2TP}{2TP + FN + FP} = \frac{2 A \cap B }{ A + B }$

2.5.3 Loss functions and performance metrics used in AF studies.

Table 2-7 contains the list of AF studies surveyed, together with the loss functions used, the metric for evaluation, their performance, and other relevant information.

Table 2-7: Loss functions, performance metrics and other relevant information from the AF studies surveyed.

Study	Loss function used	Metric for evaluation	Model performance	Other network parameters	Network architecture
Razeghi et al.[54]	DICE	DICE	DICE: 0.89 (Blood pool), 0.61 (Pulmonary veins), 0.73 (Mitral Valve)	Adam optimiser, LR: 0.001. Best results with U-net depth of 5 and feature map of 32	U-Net
Xiong et al.[55]	Combination function of cross-entropy and DICE	DICE, Mean-IOU/Jaccard index	DICE - 85.1% to 93.2% IOU - 74.4% to 87.4%	Kernel 5*5*5 and 3*3*3	U-Net, RCNN
Lu et al.[56]	Combination function of Mean IOU/Jaccard index + binary cross-entropy	DICE, IOU/Jaccard index, HD	DICE: 0.949; Jaccard index: 0.903; HD: 7.563	Adam Optimizer, LR: 1×10^{-4}	U-Net
Yang et al.[57]	Combination function of DICE + binary cross entropy	DICE	DICE- 92.11%	ReLU, Sigmoid activation function	U-Net
Du et al.[58]	Combination of loss functions. Cross entropy, DICE loss, Mean Absolute error and Overlap (IOU)	DICE; IOU/Jaccard index; HD; Concordance%;	DICE - 93.64%; Jaccard: 88.068%; Concordance: 86.377% and HD: 11.889 mm	N/A	CNN
Huang et al.[60]	Mean Square Error	Root Mean Square Error (RMSE)	Root mean square error (RMSE - 2.6%)	Adam Optimizer	CNN
Xiong et al. [61]	N/A	DICE, HD, Sensitivity, Specificity	DICE: 94.2%; HD: 10.25 mm	Adam optimizer, LR: 0.0001, 3*3 kernel filter	U-net and FCN
Yang et al.[62]	Mean Square Error	DICE	DICE - 0.897	ReLU, Adam Optimizer, LR: 0.001	U-Net
Chakravarty et al.[63]	DICE	DICE	DICE - 0.92	N/A	Recurrent Neural Network
Liu et al.[70]	N/A	AUROC, Accuracy, Sensitivity, Specificity	Accuracy: 82.4%. Sensitivity: 64.3%. Specificity: 88.4%. AUROC: 0.82	N/A	Neural Network – ResNet34
Li et al.[71]	N/A	DICE; Average Surface distance; HD.	DICE - 83.4%, ASD - 0.776 mm; HD: 7.459 mm	Adam Optimizer, LR: 0.001	Neural Network - 3D U-net with Group attention

					and channel attention block
Chen et al.[72], [73]	Binary cross-entropy	Mean IOU/Jaccard Index	Mean IOU - 91.42%	LR: 0.005	Neural Network - U-Net - Transfer learning with Encoder with ResNet50
Baskaran et al.[74][75]	DICE	DICE, median DICE scores	Overall DICE: 0.820 (0.782, 0.843).	4 layers architecture, 3*3 kernel, ReLU activation layer, batch normalization layer, Adam optimizer, LR: 0.001	U-net
Jin et al.[76]	Cross-entropy	DICE, Mean IOU/Jaccard index	DICE - 94.76%, Mean IOU: 91.10%	ReLU activation function, LR: 0.0001	FCN with VGG16 layer Net

N/A: Not Available. HD: Hausdorff distance. LR: learning rate, IOU: Intersection over Union, AUROC: Area under the Receiver Operating Characteristics Curve, DICE: Dice Similarity coefficient, ASD: Average Surface Distance, RMSE: Root mean square error, ReLU: Rectified Linear Unit, CNN: Convolutional Neural Network, FCN: Fully convolutional network, RCNN: Region based convolutional neural network, ResNet34: Residual Network with 34 layers.

2.6 Other considerations – model explainability and transfer learning

2.6.1 Model interpretability and explainability

In recent years, the development of interpretable and/or explainable models has attracted considerable interest within the AI community. Interpretability and explainability are slightly different in aim. The former refers to the development of models that are inherently interpretable, whereas the latter attempts to provide post hoc explanations for existing black-box models [110]. In medical applications, understanding the reasoning behind AI model decisions through explainable artificial intelligence (XAI) not only enhances their reliability but also holds potential for identifying novel predictive indicators, thereby contributing to advanced diagnostic processes.[111]. Post hoc explanations/attentions help in visualizing trained models highlighting relevant things more effectively. Post hoc attentions include heatmaps occlusion [112], saliency maps [113], Class Activation Maps (CAM) [114], Gradient CAM (Grad CAM)[115], and HiResCAM[116].

However, it is crucial to note that our literature review revealed a gap in the direct discussion and implementation of model interpretability methods specifically designed for segmentation task on AF. Model interpretation methods implementation on AF with ECG modality is discussed by Mousavi et al [117]. They have adopted intrinsic model interpretability by using a post hoc attention method drawing analysis from weights of wave lengths and weights of heartbeats causing the class of interest in model decision. Stiglic et al [118] in their paper on interpretability of ML models have given overview of model interpretability based on feature summary statistics, feature summary visualization and intrinsic model interpretability approaches based on model weights specific to health care.

2.6.2 Transfer learning

The lack of availability of comprehensive medical image datasets, the difficulty or expense of their annotation, and their limited quality, are some of the main challenges faced in medical image analysis. One way to address them is with the use of transfer learning. In our review of AF studies, Chenn et al. [72], [73] used Resnet50 as an encoder as part of their U-Net architecture and Jin C et al.[76] used VGG16 as embedded layer in their FCN architecture enabling transfer learning. Saiz-Viv et al.[119] used transfer learning model and showed how it improved the trained model. Puybareu et al.[120] used VGG16 as an embedded layer in FCN and applied it to the 2018 AF challenge dataset, improving not only the segmentation accuracy but also having lesser computational time.

Through the literature review, several key findings emerged regarding the utilization of transfer learning in medical image analysis for atrial fibrillation studies. Firstly, transfer learning approaches, such as utilizing pre-trained convolutional neural network (CNN) architectures like ResNet50 and VGG16, have shown promise in enhancing model performance and efficiency. These pre-trained models, initially trained on large-scale datasets for general image recognition tasks, can effectively capture and transfer meaningful features to medical image analysis tasks, thereby reducing the need for extensive manual annotation and accelerating model training. Secondly, the adoption of transfer learning techniques has enabled researchers to overcome challenges related to dataset scarcity and quality in medical image analysis. By leveraging pre-trained models and fine-tuning them on smaller medical image datasets specific to atrial fibrillation, researchers have achieved notable improvements in segmentation accuracy and computational efficiency.

Hence, our literature review underscores the utility of transfer learning in addressing data limitations in medical image analysis for atrial fibrillation. Utilizing pre-trained models not only speeds up the development process but also introduces a level of sophistication and accuracy that manual feature engineering might not achieve. It points to a growing trend where transfer learning is not just a workaround for data scarcity but a strategic approach to enhance the capabilities of medical image analysis systems.

2.7 Conclusion

This chapter has systematically reviewed the development and application of various segmentation techniques for the analysis of cardiac conditions, with a particular focus on atrial fibrillation. Our literature review has highlighted significant technological advancements and identified persistent challenges within the field of medical image analysis. The utilization of machine learning, deep learning, and transfer learning has been instrumental in improving the precision and efficiency of these segmentation techniques, contributing to better clinical outcomes. Additionally, transfer learning has been effective in mitigating issues related to data scarcity and quality, thereby improving the systems' predictive accuracy.

As this study progresses, Chapter 3 will build upon these foundations and continue literature review to explore a wider array of AI methodologies applied across different modalities of cardiovascular data. This will include a detailed examination of electrocardiogram data analysis using AI methodologies and multi-modal data analysis techniques and their application in predicting adverse cardiovascular events, thus expanding the scope of AI utility in cardiac health beyond imaging to encompass comprehensive patient assessments.

3 Review of Artificial Intelligence models on Cardiac Diagnostics using Electrocardiogram

3.1 Introduction

Building on the exploration of AI in cardiac imaging segmentation addressed in Chapters 2, this chapter shifts focus on another critical aspect of cardiovascular diagnostics: electrocardiography (ECG) analysis. This progression is rooted in Objective 1 of our research, which seeks to analyse and synthesize a comprehensive range of AI algorithms and methodologies utilized in modeling adverse cardiovascular events. Specifically, this chapter aims to bridge the gap identified in previous chapters by focusing on multi-modal data analysis and predictive modeling techniques, particularly those involving ECG data.

Electrocardiography (ECG) analysis offers a critical pathway for enhancing patient diagnostics and optimizing treatment strategies. The capabilities of AI to transform raw ECG data into actionable insights can significantly advance the understanding of cardiac abnormalities and arrhythmias. With cardiovascular diseases posing a significant health burden globally, the integration of advanced AI-driven models for ECG analysis holds the promise of identifying novel biomarkers and risk factors that are crucial for predicting adverse cardiovascular events. The aims of this chapter are twofold:

- To critically review the latest AI-driven methodologies in ECG analysis, assessing their effectiveness in classifying and predicting cardiac conditions. This review will explore the potential of various AI approaches, including deep learning, machine learning, and hybrid models, that have been recently applied in this field.
- To outline the research opportunities and potential advancements that could enhance the utility and accuracy of AI models in ECG analysis. This aligns with the overarching goal of developing predictive models that are not only effective but also clinically applicable and interpretable.

Objectives for this review being

- To provide a detailed examination of state-of-the-art AI techniques in ECG analysis and their ability to predict and classify cardiac events accurately.
- To identify existing gaps in current research and propose pathways that could potentially improve the predictive capabilities and clinical applications of AI in cardiovascular care.

This chapter, therefore, serves not only as a review but also as a critical analysis platform, linking the detailed methodologies of AI in imaging from previous chapters to the nuanced applications in ECG analysis. By doing so, it sets the stage for subsequent discussions on integrating these methodologies with broader clinical data to enhance diagnostic and predictive outcomes in cardiovascular care.

3.2 Search criteria

To ensure a comprehensive and structured approach to gathering relevant literature for the review of AI applications in ECG analysis, we conducted a meticulous search using the following databases: PubMed and Scopus. The inclusion of Scopus, alongside PubMed, was intended to broaden the research scope, capturing a diverse array of publications, including those not indexed in PubMed.

The search was strategically formulated using specific Medical Subject Headings (MeSH) terms and their synonyms to ensure both consistency and thoroughness. The MeSH terms employed include:

Electrocardiography (ECG): ["ECG", "Electrocardiogram", "Cardiac Electrical Activity", "Electrocardiographic Monitoring"]

Electroencephalography (EEG): ["EEG", "Electroencephalogram", "Brain Electrical Activity", "Neurological Monitoring"] Cardiac Arrhythmias: ["Arrhythmias", "Cardiac Irregularities", "Heart Rhythm Disorders"]

Cardiac Arrhythmias: ["Arrhythmias", "Cardiac Irregularities", "Heart Rhythm Disorders", "Dysrhythmias"]

Myocardial Infarction: ["Heart Attack", "Acute Myocardial Infarction", "coronary artery disease", "Cardiac Infarction"]

Cardiomyopathies: ["Cardiomyopathy", "Heart Muscle Disease", "Myocardial Diseases"]

Bundle Branch Block: ["Bundle Branch Block", "Heart Block", "Conduction Disorder"]

Atrial and Ventricular Disorders: ["Atrial Fibrillation", "Ventricular Dysfunction", "Heart Chamber Diseases"]

Artificial Intelligence in Medicine: ["AI in Healthcare", "Machine Learning in Cardiology", "Deep Learning for ECG", "Computational Cardiology"]

Machine Learning Algorithms: ["Support Vector Machines", "Decision Trees", "Neural Networks", "Deep Learning Techniques", "Recurrent Neural Networks"]

Multimodal Data Analysis: ["Data Fusion in Medicine", "Multimodal Imaging", "Integrated Diagnostics"]

Following search queries were used in PubMed and Scopus respectively:

("Artificial Intelligence"[MeSH] OR "Machine Learning"[MeSH] OR "Deep Learning"[MeSH] OR "Neural Networks"[MeSH] OR "Data Fusion"[MeSH] OR "Computational Cardiology"[MeSH]) AND ("Electrocardiography"[MeSH] OR "Electrocardiogram" OR "Cardiac Electrical Activity" OR "Electrocardiographic Monitoring" OR "Electroencephalography"[MeSH] OR "Electroencephalogram" OR "Brain Electrical Activity" OR "Neurological Monitoring") AND ("Cardiac Arrhythmias"[MeSH] OR "Arrhythmias" OR "Cardiac Irregularities" OR "Heart Rhythm Disorders" OR "Dysrhythmias" OR "Myocardial Infarction"[MeSH] OR "Heart Attack" OR "Acute Myocardial Infarction" OR "coronary artery disease" OR "Cardiac Infarction" OR "Cardiomyopathies"[MeSH] OR "Cardiomyopathy" OR "Heart Muscle Disease" OR "Myocardial Diseases" OR "Bundle Branch Block"[MeSH] OR "Heart Block" OR "Conduction Disorder" OR "Atrial Fibrillation"[MeSH] OR "Ventricular Dysfunction" OR "Heart Chamber Diseases") AND ("Support Vector Machines" OR "Decision Trees" OR "Neural Networks" OR "Deep Learning Techniques" OR "Recurrent Neural Networks" OR "Data Fusion in Medicine" OR "Multimodal Imaging" OR "Integrated Diagnostics")

(TITLE-ABS-KEY("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural networks" OR "data fusion" OR "computational cardiology") AND TITLE-ABS-KEY("ECG" OR "electrocardiogram" OR "cardiac electrical activity" OR "electrocardiographic monitoring" OR "EEG" OR "electroencephalogram" OR "brain electrical activity" OR "neurological monitoring") AND TITLE-ABS-KEY("arrhythmias" OR "cardiac irregularities" OR "heart rhythm disorders" OR "dysrhythmias" OR "heart attack" OR "acute myocardial infarction" OR "coronary artery disease" OR "cardiac infarction" OR "cardiomyopathy" OR "heart muscle disease" OR "myocardial diseases" OR "bundle branch block" OR "heart block" OR "conduction disorder" OR "atrial fibrillation" OR "ventricular dysfunction" OR "heart chamber diseases") AND TITLE-ABS-KEY("support vector machines" OR "decision trees" OR "neural networks" OR "deep learning techniques" OR "recurrent neural networks" OR "data fusion in medicine" OR "multimodal imaging" OR "integrated diagnostics"))

Additionally, conference proceedings, particularly those hosted by relevant professional bodies like the IEEE, were scrutinized for pertinent titles focusing on AI-driven ECG analysis. This dual-method approach of utilizing both structured database queries and manual searches of conference records ensured a comprehensive exploration of both peer-reviewed articles and significant grey literature.

Inclusion Criteria: Studies focusing on the application of AI, machine learning, or deep learning techniques to analyse ECG data. Only articles published in English and with full texts available were considered. Studies including participants from all sociodemographic clusters, without exclusion based on age, sex, or nativity, ensuring a broad representation in the analysed data

Exclusion Criteria: Papers not primarily focusing on AI applications in ECG analysis. Studies published before Papers published 2012 were also excluded for relevancy to maintain relevance to current technology. Non-peer-reviewed articles, opinion pieces, and editorials were also excluded.

A rigorous peer-review filter was applied to ensure that the studies included were unbiased and representative of the current understanding in the field. Efforts were made to include a diverse range of methodologies and viewpoints by accessing multiple databases and incorporating both journal articles and conference papers.

The last update to the included studies was conducted on 13th April 2023., with all searches restricted to materials published from 1990 to the present, focusing on contemporary methods and innovations in the field.

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram for studies included is shown in figure 3-1.

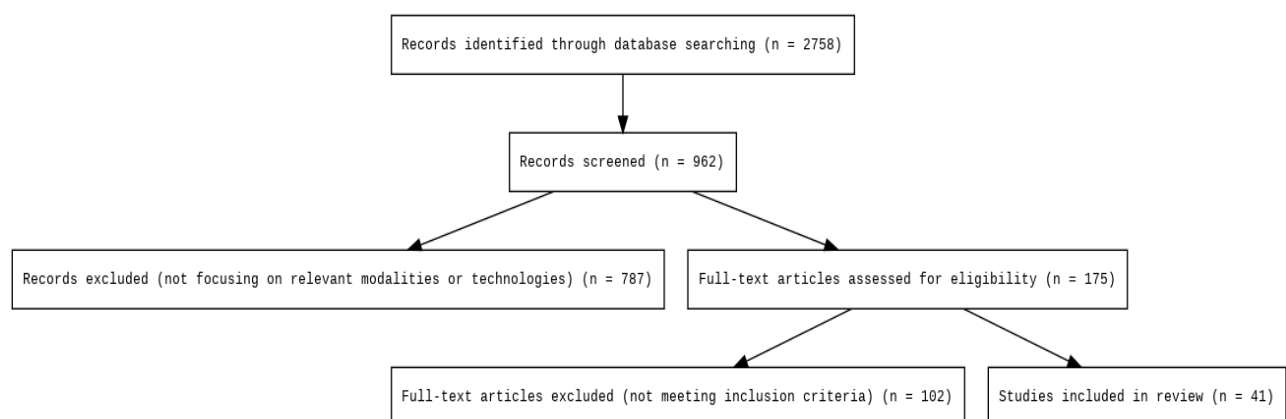


Figure 3-1 - Multimodal data & ECG data with AI Applications - Study review

3.3 Multimodal data, ECG Analysis and Cardiac Incident prediction – A Review

In a series of studies, research in machine learning (ML) and deep learning (DL) have substantially influenced the field of electrocardiogram (ECG) data analysis, demonstrating significant promise in enhancing diagnostic accuracies and exploring new methodologies for cardiac health monitoring. Central to these advancements is the use of diverse datasets such as PTB-XL and MIT-BIH, which have facilitated a variety of approaches ranging from convolutional neural networks (CNNs) to complex ensemble models. Studies consistently highlight the efficacy of DL techniques in not only boosting diagnostic precision but also in addressing specific challenges such as data imbalance, computational efficiency, and real-time processing capabilities.

Deep learning models have been pivotal in progressing ECG classification tasks. The use of multireceptive field CNNs, as explored by Ullah et al. [129] and Feyisa et al.,[148] showcases an evolution towards models that can effectively capture and analyse the intricate patterns within ECG signals across multiple classes of heart diseases. These studies underscore the models' abilities to handle large-scale data while maintaining high accuracy levels, illustrating the potential of DL in clinical applications. However, the need for extensive computational resources and the challenge of ensuring model robustness across diverse patient demographics remain pertinent concerns.

The introduction of Generative Adversarial Networks (GANs) by Seo HC et al. [127] represents a novel approach to synthesizing ECG data, which helps overcome limitations associated with single-lead ECG recordings. This method not only enhances the volume and variety of data available for training but also presents new avenues for testing and improving the generalizability of ECG analysis models. Nonetheless, the application of GANs in a clinical setting requires rigorous validation to establish their effectiveness and reliability in multi-lead ECG data synthesis.

While the reviewed studies present significant advancements, the transition from research to real-world application necessitates a focus on scalability, real-time applicability, and the robustness of AI models. The studies by Martin H et al. [131] and Chen X et al. [132] offer insights into real-time myocardial infarction detection using LSTM and RNN models, which are commendable for their high accuracy. However, these models need further validation in diverse clinical settings to confirm their effectiveness and reliability. Similarly, the integration of AI tools in clinical workflows must also consider user-friendliness and interpretability, aspects that are crucial for widespread adoption.

The cumulative insights from these studies illuminate the substantial progress in applying ML and DL techniques to ECG data analysis. However, they also delineate the critical areas needing further research and adaptation to meet clinical requirements fully. By focusing on enhancing model generalization, reducing computational demands, and addressing data imbalances, future research can pave the way for more robust and clinically viable solutions. This analysis not only aligns with our research directions but also enriches our understanding of the potential and limitations of current methodologies in ECG analysis.

Studies reviewed collectively contribute to the advancement of ECG analysis, encompassing machine learning, deep learning, and signal processing techniques, with each addressing specific aspects of ECG analysis. Table 3-1 summarizes the AI techniques, algorithm, performance metrics and key findings of the studies reviewed in ECG and multi modal data analysis regarding cardiovascular incidents.

Table 3-1: Related Work in ECG data Analysis using ML and DL approaches

Related Work	Dataset	Techniques	Algorithm	Performance Metrics	Key Findings
Wagner P et al.[125]	PTB	Introduction to PTB Dataset	Not Applicable	Not Applicable	This paper utilizes deep learning methodologies for the classification of ECG signals, focusing predominantly on elucidating the dataset characteristics
Pałczyński K et al.[126]	PTB-XL	Few-shot Learning (FSL)	Custom FSL Model	Accuracy: 93.2% to 89.2%	Focuses on few-shot learning for ECG classification with limited data. Literature Survey
Seo HC et al.[127]	China Dataset, PTB-XL	GAN	Custom GAN Model	Fréchet distance (FD) and mean squared error (MSE) scores: 7.237 and 0.024, 8.055 and 0.011	Usage of GAN in creating synthetic ECG's.
Śmigiel S et al.[128]	PTB-XL	Sincnet	Deep Neural Network – Sincnet CNN 1D	2-class classification Accuracy – 0.882 AUC 0.953 5-class classification Accuracy – 0.72 AUC 0.877 20-class classification Accuracy – 0.589 AUC 0.856	This study employs deep learning techniques for the classification of electrocardiogram (ECG) signals, with a specific focus on the PTB-XL dataset. Both entropy features and raw signals are utilized in the classification process, with the analysis revealing that entropy features yield the best results.
Ullah W et al.[129]	MIT-BIH and PTB	CNN, LSTM, Attention Model	CNN + LSTM with Attention	Accuracy: CNN 99.12%, CNN + LSTM 99.3%, CNN + LSTM + Attention 99.29%	The study employs deep learning, featuring an attention-based model, to classify arrhythmias in ECG heartbeats. They extract heartbeats using a sliding window approach and apply a convolutional neural network (CNN) followed by a self-attention mechanism to highlight key features. Another CNN is used for arrhythmia classification with a loss function addressing class imbalance.
Śmigiel S et al.[146]	PTB-XL	R Wave detection through Hamilton detector; Two average detector; CNN - 1 D Stationary Wavelet Transform detector; Christov detector;	CNN - 1D	2-class classification Accuracy – 90.9% to 89.6% AUC 96.6% to 95.5% Multi-class classification Accuracy – 79.1% to 74.9%	Utilizing deep learning techniques, this study focuses on the detection of R-peaks in electrocardiogram (ECG) signals. Multiple methods for detecting R-waves (QRS complexes) are explored, including models based on the raw signals and their entropies, as well as models specifically designed for analysing the QRS complexes of all 12 ECG leads and their respective entropies.

		Pan–Tompkins’ detector; Engzee detector with modification.		AUC 91.8% to 88.6%	
Sun Q et al.[147]	PTB Dataset	ECG Conversion to VCG and then feature extraction with spectral fitting exponent, Lyapunov exponent and LZ Complexity. Boosting, SVM, Random Forest for classifiers and finally an ensemble model	Ensemble Model	Average accuracy: 89.10%, sensitivity: 91.72%, specificity: 82.69%	Combines deterministic and ensemble learning for myocardial ischemia detection.
Feyisa DW et al.[148]	PTB XL	Multireceptive Field CNN	Multireceptive Field CNN	F1 score: 0.72, AUC: 0.93 for 5 super class	This study employs lightweight multireceptive field Convolutional Neural Networks (CNNs) for the classification of 12-lead electrocardiograms (ECG). In this context, a multireceptive field CNN is characterized by its convolutional layers with multiple kernel sizes, allowing it to capture a broad spectrum of features from the input data. While traditional definitions of multireceptive field CNNs may involve processing 3D data, in the case of ECG signals, which are typically 1D or 2D, this concept applies to the ability of the network to capture diverse features from the input data.
Brisk R et al.[130]	PTB XL	U-Net, ECG synthetic generator, Representative learning	U-Net	MI detection Sensitivity 0.91 Specificity 0.84 AF detection Sensitivity 0.65 Specificity 0.99	This study focuses on wave segmentation pretraining as a method for ECG analysis.
Martin H et al.[131]	PTB XL and PTB	LSTM	LSTM	Accuracy: 77.12%, recall/sensitivity: 75.85%, specificity: 83.02%	Concentrates on real-time detection of myocardial infarction using single-lead ECG. R peak detection from only lead 2 and then passing the feature through LSTM for myocardial and healthy control classification
Chen X et al.[132]	PTB XL	RNN	Neural Network	AUC: training 0.964, validation 0.944, testing 0.977; precision, sensitivity, and specificity, and F1 score also provided	Focuses on the real-time detection of myocardial infarction using single-lead electrocardiograms (ECG). The approach involves detecting R peaks exclusively from lead 2 and subsequently feeding these features through a Long Short-Term Memory (LSTM) network for classification between myocardial infarction and healthy control cases.

Strodtzoff N et al.[149]	PTB XL	Comparison of all models and methods in terms of classification	Comparative Study of DL Models	Not Applicable	Provides benchmarks and insights on ECG analysis using deep learning. Survey of all methods till date used for classification in PTBXL dataset and ICBE2018 challenge
Pham BT et al.[150]	MIT-BIH and PTB XL	CNN and Residual network	Residual Neural Network	Accuracy scores: 98.5% on the Physionet MIT-BIH dataset and 98.28% on the PTB database	Focuses on heartbeat classification for arrhythmias and myocardial infarction using deep learning techniques. The methodology involves detecting R peaks from all leads and extracting features such as the median of R-R intervals. These features are then processed using a proposed neural network architecture after padding to ensure equal length. The classification task entails binary categorization between normal and abnormal myocardial infarctions.
Tseng KK et al.[133]	TVGH and PTBD	Unsupervised learning, Clustering, Feature extraction and Neural network	Neural Network	Diagnostic accuracy: PTB dataset 80.58%, Wavelet feature dataset 75.63%	Feature extraction in terms of Reduced binary pattern and wavelet feature
Santos Rodrigues A et al.[134]	PTB XL	CNN1D for feature extraction and Regression network in output	CNN1D	Absolute mean and median errors of 11.4° and 7.3°	Utilizes deep learning for the estimation of spatial QRS-T angle from reduced-lead ECGs.
Chen H et al.[135]	PTBD	C Means clustering	Fuzzy C-Means Clustering	Average prediction time approximately 5 min, accuracy of 98.4%, sensitivity of 97.5%, and specificity of 99.1%	Utilizes phase-space reconstruction and fuzzy c-means clustering for ventricular arrhythmia prediction and classification.
Safdar MF et al.[151]	PTB XL	CNN - 2D for spectrograms and CNN-1D for raw Signals	CNN (2D and 1D)	Highest accuracy of 99.06%	Focuses on ECG classification using convolutional neural networks with denoising and Fourier transformation-based spectrograms.
Strodtzoff N et al.[152]	PTBD	CNN and Recurrent Network (LSTM) used for raw signals classification	Fully Convolutional Network + LSTM	Sensitivity: 93.3%, Specificity: 89.7%	Detects and interprets myocardial infarction using fully convolutional neural networks.
Chumachenko D et al.[153]	PTB XL	K-Nearest Neighbours, Radial basis function, Decision tree and Random Forest	Random Forest	Accuracy: 99.63%, root mean absolute error < 0.004	Utilizes machine learning methods for predicting patients with suspected myocardial infarction based on short-time HRV data.

Boudnik L et al.[154]	MIT-BIH PTB-XL	Temporal Convolutional Networks	Temporal convolutional networks, (TCN) Fully conventional denoising auto encoder, (FCDAE) Denoising GAN(DGAN)	Signal noise ratio index (in db) TCN 9.11 FCDAE 8.59 DGAN 8.84	Focuses on muscle artifact (noise) removal in single-channel electrocardiograms using temporal convolutional networks.
Lee J et al.[155]	PTBD	GAN	GAN for V-lead ECG synthesis	Mean structural similarity index (SSIM) 0.92, mean error rate of percent mean square difference (PRD) 7.21%	Utilizes GANs for synthesizing V-lead ECG signals from limb-lead measurements.
Liu J et al.[156]	PTBD	Tuneable Q - Factor wavelet transform	Wavelet Transform	Accuracy: 86.3%	Focuses on myocardial infarction detection using wavelet transform.
Rai HM et al.[136]	MIT-BIH PTBD	CNN, CNN & LSTM, Ensemble model	Ensemble model with CNN and LSTM	Overall highest accuracy: 99.02%	Addresses the prediction of cardiac abnormality and minority class accuracy enhancement in imbalanced ECG signals using modified deep neural network models. Uses SMOTE for balancing the imbalance problem. Modelled based on ECG beats extracted from ECG signals
Yang T et al.[137]	Telemetric and Holter warehouse	Neural network and random forest	Neural Network	Sensitivity: 91.7%, specificity: 85.6%, accuracy: 88.7%	Detects strict left bundle branch block using neural networks.
Sun Q et al.[138]	PTBD and Independent clinical dataset	Decision Tree, KNN, LR, SVM, ANN and RF	Ensemble Model	Maximum accuracy: 84.96%, sensitivity: 85.04%, and specificity: 84.80%	Employs dynamic feature extraction methods for myocardial infarction detection.
Jo YY et al.[139]	Sejong Hospital and PTB-XL	Convolutions Neural network after concatenating the feature extracted irregularity of heart rhythm and absence of p wave	CNN	AUCs using a 12-lead ECG: 0.997–0.999	Deep learning (DL) models, particularly those enabled for electrocardiogram (ECG) interpretation, are a primary focus. These models are designed to detect underlying hidden AF from sinus rhythms and interpret ECGs in innovative ways
Farag MM et al.[140]	MIT-BIH and PTBD	CNN	Tiny Matched Filter-Based CNN	Average accuracy: 98.18%, sensitivity: 91.90%, F1 scores: 92.17%	Introduces a Tiny Matched Filter-Based CNN for inter-patient ECG classification.

Geng Q et al.[141]	PTB-XL	FCN, LSTM, Resnet, Wavelet, NN, Bi-LSTM, SE-RESNet	Multi-task deep neural network	Average F1 score: 0.833 on the PTB-XL dataset	Employs multi-task learning and CoT (contextual transformer) attention mechanisms for ECG classification.
Zvuloni E et al.[142]	TNMG PTB-XL	CNN	CNN, Random forest + Ensemble	Mean AUPRC 0.90, F1 0.90, FPR 0.003, FNR 0.095	Focuses on merging feature engineering and deep learning for diagnosis, risk prediction, and age estimation based on the 12-Lead ECG.
Rath A et al.[143]	MIT-BIH PTBD	LR, SVM, Ada boost, Ensemble (Adaboost & LR)	Ensemble model with AdaBoost and LR	Accuracy: 0.946, F1-score: 0.949, AUC: 0.951 for the PTB-ECG dataset	Concentrates on imbalanced ECG signal-based heart disease classification using ensemble machine learning techniques.
Zhang H et al.[144]	PTB-XL	3D recurrence inception Resnet	Inception ResNet-v2	Average F1 score: 0.9254 for the 3D ECG-based case, 0.9350 for the 3D VCG-based case	Addresses cardiac arrhythmia classification based on 3D recurrence plot analysis and deep learning.
Zhang H et al.[157]	CPSC Georgia PTB-XL	Inception ResNet-v2	Inception ResNet-v2	Average F1-score: 0.8521, 0.8529, and 0.8862 across multiple databases	Utilizes Inception-ResNet-v2 and recurrence plot analysis for cardiac arrhythmia classification.
Shany Biton et al.[145]	TNMG	Random Forest	Random Forest	Area under the receiver operating characteristic curve (AUROC): 0.909	The study used a random forest classifier, capitalizing on features from demographics, clinical data, engineered features, and deep representation learning. Their model demonstrated high predictive performance, with an AUROC of 0.909, indicating that integrating multiple data modalities and leveraging deep learning techniques can effectively predict AF risk.

3.4 Limitation of Studies reviewed

3.4.1 Process Gaps

A significant gap identified in the literature on AI methods applied to ECG data involves the integration of clinical factors with Heart Rate Variability (HRV) variables across all ECG leads. Current studies, while exploring various machine learning models, fall short in thoroughly combining clinical information with HRV data for each lead. This gap highlights the potential for more sophisticated data pre-processing and feature engineering. Such an approach would more effectively harness the valuable insights from clinical and HRV data within the ECG dataset. Furthermore, there is a notable lack of diversity in the machine learning models employed. Expanding the range of models explored could enhance predictive accuracy and provide deeper insights in this field.

3.4.2 Methodology Gaps

The literature review also reveals a methodology gap in the underuse of Generative Adversarial Network (GAN) for data augmentation with synthetic samples, particularly in the context of ECG data and HRV measures. While

some studies have experimented in this area, their focus has predominantly been on single-lead ECG data, neglecting the potential of GANs to create synthetic samples incorporating multi-lead ECG features. The limited exploration of GANs for augmenting datasets with synthetic samples overlooks their demonstrated ability to generate realistic data, which could significantly improve model performance. This underutilization represents a lost opportunity to enhance dataset quality and diversity.

3.5 Conclusion

This chapter has meticulously reviewed the integration of artificial intelligence in electrocardiography (ECG) analysis, emphasizing the application of machine learning (ML) and deep learning (DL) techniques across diverse datasets. This analysis has highlighted significant advancements in predictive modeling and diagnostic accuracy, notably through innovative approaches such as multireceptive field CNNs and the use of GANs for synthetic data generation. However, it has also uncovered critical gaps, particularly in the integration of clinical factors with Heart Rate Variability (HRV) data across all ECG leads, and a methodological underuse of GANs for multi-lead ECG data generation. These findings underscore opportunities for future research to enhance data diversity and model robustness, setting a solid foundation for the subsequent chapters.

This thesis aims to address and bridge above identified gaps highlighted in 4.4. Chapter 5 will extensively explore the integration of clinical factors and Heart Rate Variability (HRV) variables across all leads in the PTB dataset. This in-depth investigation aims to leverage the valuable insights contained in clinical data and HRV information, potentially leading to more accurate and comprehensive predictive models. Additionally, this chapter also explores a broader array of machine learning models beyond those examined in the previous literature, providing a comprehensive overview of the potential models that can be employed to tackle the research problem.

Further, in Chapter 6, the study delves into the utilization of GAN augmentation techniques to enrich the dataset with synthetic samples, aiming to improve the diversity and quality of the data available for analysis. This approach is expected to contribute to enhanced model performance and more robust predictions, ultimately elevating the clinical applicability and interpretability of AI-driven ECG analysis.

The next Chapter 4 will explore the methodological frameworks that support segmentation tasks used in AI technologies. It will provide an in-depth examination of the computational models and algorithms essential for enhancing segmentation processes, while also addressing the identified gaps concerning explainability and transfer learning as addressed in Chapter 2. This exploration is intended to elucidate the technical details and enhance the operational effectiveness necessary for advancing medical imaging technologies.

4 Left Atrium segmentation using Neural Networks

4.1 Introduction

Building upon the methodologies discussed in chapter 2, this chapter delves into the application of medical image analysis for segmenting MRI images of the left atrium in patients with atrial fibrillation (AF). AF, noted for its rapid and irregular heart rhythms, poses significant challenges in both diagnostic accuracy and treatment planning. Recent advancements in deep learning, particularly the use of U-Net architectures, have shown promise in automating the segmentation process, which is crucial for precise anatomical delineation and subsequent diagnostic decision-making.

The literature review highlighted notable gaps in the explainability of models and the generalization of methodologies across different datasets. In response, this chapter addresses these gaps by implementing saliency maps and other explainability techniques to interpret the segmentation tasks more transparently. These methods aim to demystify the model's decision-making process, offering insights into the neural network's focus during the segmentation.

Moreover, the chapter explores the use of transfer learning to validate the model's effectiveness on an independent dataset, the CDEMRS database. This approach not only tests the robustness of the developed models but also ensures that the methodologies are adaptable and reliable across varying imaging conditions. By integrating advanced deep learning techniques with rigorous validation methods, this chapter contributes to the field of medical image analysis. By doing so, it contributes to improving diagnostic accuracy and understanding of cardiac imaging modalities, supporting more reliable treatment planning.

4.2 Data and pre-processing

4.2.1 Primary Dataset – Cardiac Atlas Study

The data used to address the research objective 2 was extracted from the Cardiac Atlas Study¹[121] an international effort to establish a large-scale standardized database of cardiac imaging. This dataset includes 154 three-dimensional MRIs from patients diagnosed with atrial fibrillation (AF), with an original resolution of $0.625 \times 0.625 \times 0.625 \text{ mm}^3$. A significant portion of this data was sourced from The University of Utah (NIH/NIGMS Centre for Integrative Biomedical Computing), with additional contributions from various other institutions. All clinical data has received institutional ethics approval. Each 3D MRI dataset consists of raw MRI scans and corresponding ground truth labels for the left atrial cavity, manually segmented by field experts. These grayscale MRIs and binary segmentation labels (255 = positive, 0 = negative) maintain a uniform depth across all scans, consisting of exactly 88 slices along the z-axis.

Data storage is organized by patient, with each patient's data stored in a designated folder containing the raw MRI ("lgemri. nrrd") and the left atrium(LA) cavity segmentation ("laendo. nrrd"). The ". nrrd" file format, a standard in medical imaging, supports various programming environments. For modelling purposes, the dataset was divided, allocating 100 patient datasets for training and 54 for testing and evaluation. Data conversion from NRRD to NIFTI format was performed using 3D Slicer, a Flexible Image Transport System (FITS). Subsequent manipulation and preparation for modelling were facilitated through the Python library 'nibabel'.

¹ <https://www.cardiacatlas.org/>

The images from MRI niftii format were floating point images with uint 8 datatype, with image intensity ranging between 0 to 255. These images were either 576-by-576 or in 640-by-640 pixels, with 88 slices, and in grayscale. The images that were in three dimensions were sliced on X, Y and Z planes. X and Y axis images were cropped and did not hold full information of binary segmentation of the left atrium cavity. Hence, a decision was taken to use only Z-axis information that had full image representation of the left atrium cavity and mask (ground truth) for the binary segmentation task. The images along the Z-axis were extracted in png format and saved patient wise along with respective binary segmentation data both for training and test data set.

Images were resized to uniform dimensions of 200-by-200, 256-by-256 and 192-by-192 to reduce computational load. Given the vast amount of data and the computational demands of high-resolution images, these specific dimensions were chosen to balance detail retention with efficiency. This resizing was carefully managed to maintain the integrity of the information crucial for accurate segmentation. Despite the reduction in resolution, the z-axis was preserved in its entirety to ensure no significant loss of detail in the left atrial cavity, which is critical for the study's objectives. The effectiveness of this approach was validated through preliminary models, which confirmed that the resizing did not adversely affect the model's accuracy or its ability to generalize from training to unseen data. Figure 3-1 shows the unprocessed MRI images of LA cavity pertaining to cardiac atlas project dataset.

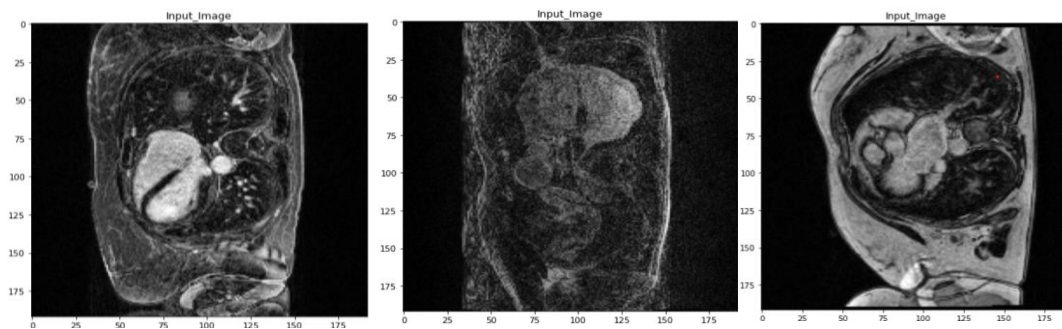


Figure 4-1: MRI Images of LA Cavity - Cardiac Atlas Project Dataset

4.2.2 Validation Dataset – Cardiac Delayed Enhancement Segmentation Challenge (cDEMRIS)

The CDEMRIS dataset, utilized as an independent validation dataset in this study, originates from the Cardiac Delayed Enhancement Segmentation Challenge organized during the IEEE International Symposium on Biomedical Imaging (ISBI) in 2012.[122] This dataset is crucial for assessing the efficacy of segmentation algorithms designed to identify and classify fibrotic and scarred tissues within the myocardium of the left atrium, using Late Gadolinium Enhancement (LGE) cardiovascular magnetic resonance (CMR) imaging.

LGE CMR is instrumental for the visualization of myocardial fibrosis and scarring, which are pivotal for determining the treatment strategy and monitoring post-treatment recovery in patients with atrial fibrillation (AF). The dataset is specifically designed to facilitate benchmarking of algorithms that classify fibrosis and scar tissues from LGE CMR images.

The CDEMRIS dataset comprises 60 sets of images categorized into 'pre-ablation' and 'post-ablation' conditions, with each category containing 30 datasets. These categories are crucial for understanding the

efficacy of radiofrequency catheter ablation (RFCA) treatments on patients. The data is organized into folders labelled as pXX, where XX denotes the case number, and includes:

- de_xx_nn: Delayed-enhancement (DE) scan images.
- la_seg_xx_nn: Binary segmentation of the left atrial endocardium, registered to the corresponding DE scan.
 - ❖ The 'xx' signifies the scan timing relative to the ablation ('a' for pre and 'b' for post),
 - ❖ 'nn' denotes the case number.

The dataset pools images from three major imaging centers, each contributing to a rich diversity of image quality and resolution: University of Utah School of Medicine, Beth Israel Deaconess Medical Center (BIDMC) and King's College London Imaging Sciences (KCL-IM).

Each center has provided images captured using state-of-the-art MRI scanners: Siemens Avanto 1.5T or Vario 3T and Philips Achieva 1.5T

The images exhibit variations in basic scanning parameters and resolutions, tailored to typical clinical acquisitions: Following is the resolution parameters of each source of data:

University of Utah: $1.25 \times 1.25 \times 2.5$ mm, TI 300 ms, TR 5.4 ms, TE 2.3 ms

BIDMC: $1.4 \times 1.4 \times 1.4$ mm, TI 280 ms, TR 5.3 ms, TE 2.1 ms

KCL-IM: $1.3 \times 1.3 \times 4.0$ mm, TI 280 ms, TR 5.3 ms, TE 2.1 ms

Where,

- mm: Millimeters, indicating the spatial resolution of the imaging.
- ms: Milliseconds, denoting the time duration for imaging parameters.
- TI: Inversion Time, the period between the inversion pulse and the central k-space line used in magnetic resonance imaging.
- TR: Repetition Time, the time between successive pulse sequences applied to the same slice.
- TE: Echo Time, the time between the delivery of the RF pulse and the peak of the echo signal.

Pre-ablation scans were taken within 7 days prior to the procedure. Post-ablation scans varied more significantly, ranging from 30 days to 3-6 months after the procedure, depending on the imaging center. For validating the segmentation models, we have used only pre-ablation scans. Figure 4-2 shows unprocessed LGE CMR images within myocardium of LA cavity of CDEMRIS dataset.

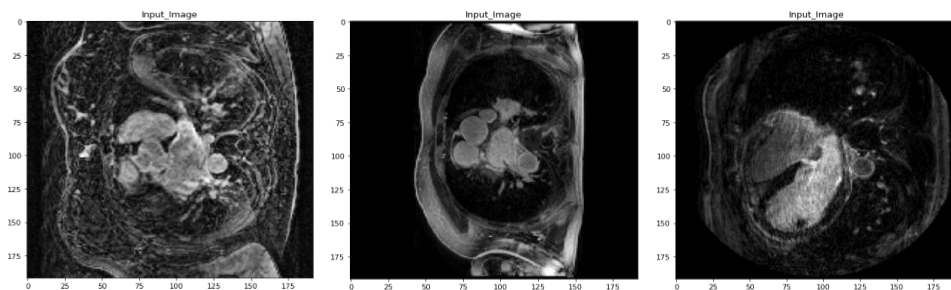


Figure 4-2:LGE CMR images - CDEMRIS dataset

4.3 Methods

In context to MRI segmentation task AI techniques are utilized for automating the complex task of segmenting MRI images. AI leverages pattern recognition capabilities to analyze complex visual data from MRI scans, supporting decision-making processes in medical imaging. ML, particularly through data augmentation methods like photometric transformations, enhances the dataset used for training models. This improves the model's ability to generalize from the training data to new, unseen MRI images. These techniques enable the model to better adapt to variations in image quality and anatomical differences. Specifically, Deep Learning, through the use of architectures like U-Net, allows for precise and efficient segmentation of MRI images. The U-Net model, with its deep network structure, excels in capturing the high-level features necessary for detailed segmentation of medical images, thus significantly improving the accuracy of the segmentation process.

4.3.1 Study Design

The study design for the segmentation of the left atrium from MRI images incorporates a systematic approach involving several key stages:

- **Objective Setting:** Defining clear research objectives based on identified gaps in the current understanding of MRI segmentation and its application to atrial fibrillation (AF).
- **Data Collection:** Utilizing the Cardiac Atlas Study for training data and the CDEMRS dataset for validation, ensuring a robust assessment of the segmentation models across different clinical conditions and imaging techniques.
- **Segmentation Model Development:** Employing deep learning architectures, specifically the U-Net model, optimized for the segmentation of medical images. This involves training the model on the primary dataset with a focus on high precision and accuracy in segmenting the left atrium.
- **Validation and Testing:** Implementing a rigorous validation process using an independent dataset to assess the generalizability and effectiveness of the segmentation models under varied conditions.
- **Ethical Considerations:** Ensuring all data used adhere to the highest ethical standards with proper approvals and anonymization to safeguard patient information.
- **Outcome Measurement:** Using performance metrics such as the Dice coefficient, Intersection over Union (IoU), and others to quantitatively evaluate the segmentation accuracy of the models.

By structuring the study around these components, the research aims to provide a comprehensive analysis of the potential and limitations of current deep learning approaches in medical image segmentation.

4.3.2 Convolutional neural networks

Convolutional neural networks (CNN) [123] is deep learning neural network algorithm. It consists of input layer, output layer and hidden layers further consisting of convolutional layers, pooling layers, and a fully connected layer. CNN's can take an input image, assign importance to various pixels through different aspects in images via weights and biases and be able to accomplish tasks of differentiate, classification and segmentation.

In image classification task, convolutional layers detect features, and pooling layers select features in an image. Output from convolutional and pooling layers are forward fed into fully connected layer for classification task which can be object detection, facial recognition, or image classification. In the case of segmentation tasks, CNN uses encoders and decoders. Encoders are convolutional networks and decoders are transposed convolutional networks. Encoders input the image representation and do feature extraction and decoders upsample the features extracted and represent final segmentation maps.

Convolutional Neural Networks (CNNs) are particularly effective in MRI image segmentation due to their ability to capture spatial hierarchies in image data. CNNs work by using a sequence of convolution and pooling layers to extract features at different scales. This is crucial in MRI segmentation where spatial relationships and structural details are vital. For instance, 3D CNNs, which extend the convolution and pooling operations into three dimensions, are capable of learning spatial features in all three dimensions. This enables them to analyse volumetric data more effectively, capturing inter-slice information and offering a more comprehensive understanding of the 3D structure present in MRI scans. The key advantage here is their ability to understand the spatial relationships and patterns within the MRI data, which is critical for accurate segmentation of complex anatomical structures[124].

4.3.3 U-Net architecture for left atrium segmentation

Recently, a CNN-based architecture known as U-Net[32] was proposed to address the specific image analysis problem of image segmentation. This makes U-Net useful to identify and characterise regions of interest within an image such as left atrium segmentation. As shown in Figure 4-3, U-Net has an "U"- shaped architecture, which gives it its name. The idea of the U-Net is to first obtain a lower-dimensional representation of the image through a traditional CNN, and then up-sample its low-dimensional representation to map an image that highlights the regions of interest (mask). The down-sampler section of the U-net is known as the contracting path, whereas the up-sampler section, the expansive path 4-3.

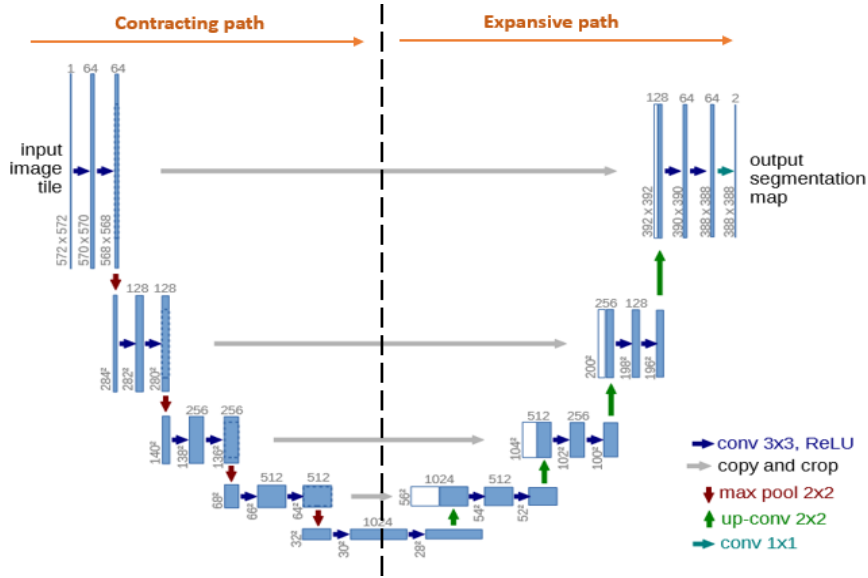


Figure 4-3:U-Net architecture. Adapted from [32]

The contracting path (also known as encoder) is a traditional CNN, which produces a low-dimensional representation. The expansive path (also known as decoder) up samples the representation to produce the final output segmentation map. Gray arrows represent copying operations, in which high-resolution feature maps from the contracting path are copied over and concatenated to feature maps in the expansive path to make it easier for the network to learn a high-resolution segmentation.

U-Net consists of 4 components: convolution, up sampling, concatenation and again convolution. These components are detailed in the four steps below:

Step 1 – Convolution + Maxpooling

$$\text{Output width} = \frac{W - Fw + 2P}{Sw} + 1 \quad \text{Output height} = \frac{H - Fh + 2P}{Sh} + 1$$

Where W is Width, Fw and Fh are Kernel shape, P is Padding and S is stride.

Let's construct model with kernel of convolution = 3*3*3, Stride = 1, Padding = 0, number of filters = 64 and same for Maxpooling (filter shape 3*3*64), Convolution output width will be $[(10 - 3 + 2*0)/1] + 1 = 8$ and after convolution we get an image tensor of 8*8*64, Applying Maxpooling we get $[(8 - 3 + 2*0)/1] + 1 = 6$ and an image tensor shape of 6*6*64. See the representation of this example in figure 4-4. The height and width are now reduced, but the depth of the three RGB channels of the image has now increased to 64 layers.

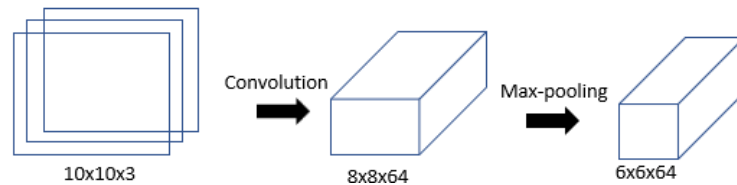


Figure 4-4:Example of the convolution and max-pooling process

Step 2 – Up sampling: a method of increasing an image's height and width. The example in Figure 4-5 uses the nearest neighbour method, where the pixels are simply duplicated.

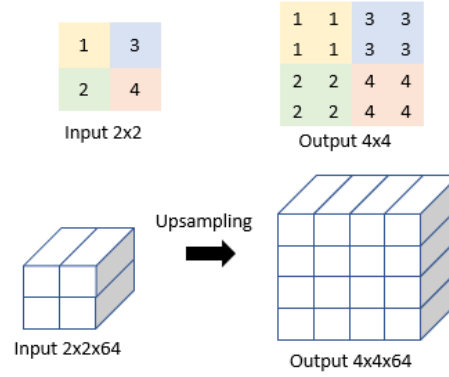


Figure 4-5: Examples of up sampling

Step 3 – Concatenation: the convolution layers have two outputs with shape 8x8x64, one which goes to max-pooling, and another goes to concatenation operation. The concatenation is done on depth (third axis) and not on height and width. Concatenating the output of the convolutional layer and the up-sampling layer we get an image tensor of 8x8x128 (see example in Figure 4-6).

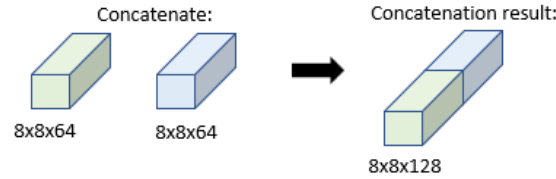


Figure 4-6: Example of the concatenation process

Step 4 – Convolution: in the convolution process, we increase the height and width of the image and try to reduce the depth. This can be achieved by applying convolutional layers without Maxpooling. First one: $F=3 \times 3 \times 128$, $P=2$, $S=1$, number of filters =64, we get an image tensor of 10x10x64. Second One where $F=3 \times 3 \times 64$, $P=1$, $S=1$, number of filters = 1, we finally, get an image tensor of 10x10x1.

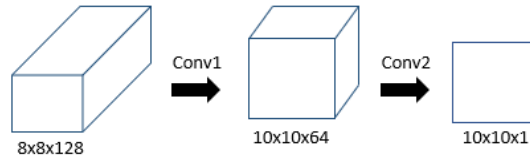


Figure 4-7: Example of convolution

The U-Net architecture is implemented to segment the left atrium from MRI scans, harnessing its symmetrical structure to process and combine multi-scale information. The network comprises a contracting path for feature extraction and an expansive path for spatial localization, as delineated below:

Contracting Path: Initially, the input image undergoes a series of convolutional operations (Conv2D), with each convolutional layer equipped with a ReLU activation function to introduce non-linearity. Subsequent pooling layers (MaxPool2D) reduce dimensionality, augmenting the network's focus on salient features.

Mathematically, this reduction can be expressed as,

$$W_{out} = SW_{in} - F_w + 2P + 1 \quad \text{Equation 4-1} \quad \text{and}$$

$$H_{out} = SH_{in} - F_h + 2P + 1 \quad \text{Equation 4-2}$$

Where W_{out} and H_{out} denote the width and height and F_w and F_h are kernel dimensions, P presents the padding and S is the stride. The dimensions of the feature maps decrease, whereas the depth increases to capture a richer representation of features within the MRI images, as illustrated by the progression of layers in the provided figures.

Expansive Path: The model subsequently employs transposed convolutions (Conv2DTranspose) for up sampling, effectively increasing the resolution of feature maps. This operation inversely mirrors pooling and is mathematically characterized by transposed convolution operations that study lower-dimensional feature representations back to higher-dimensional space. The expansive path reinstates the spatial dimensions essential for precise segmentation.

Concatenation: At each level of the expansive path, the model performs a concatenation of the corresponding feature maps from the contracting path with the up sampled features. This step, executed via the Concatenate operation, preserves high-resolution details necessary for accurate delineation of the anatomical structure.

Final Convolution: The culmination of the expansive path involves a series of convolutions that refine the segmentation output, with the last layer employing a sigmoid activation (Conv2D with activation='sigmoid') to map the probabilities to a binary mask indicating the presence of the Atrial Fibrillation. Detailed architecture is shown in appendices.

4.3.4 Data Augmentation

Data augmentation is a pivotal process in enhancing the robustness and performance of image recognition algorithms. It addresses intrinsic challenges within medical imaging datasets, such as variations in lighting, occlusion, viewpoint, background, and scale. The transformative procedures artificially expand the training dataset, improving the model's ability to generalize across diverse imaging conditions and thus ensuring that the training and validation error rates converge, indicative of a model well-suited for real-world application.

In the context of segmenting the left atrium from Cardiac Magnetic Resonance (CMR) images, our study leverages both geometric and photometric transformations. Geometric transformations alter the spatial structure of the image, including rotations, translations, scaling, and flips, thus simulating varying orientations and sizes of the anatomical structures as they might appear in clinical scenarios. Photometric transformations, on the other hand, manipulate the pixel values directly, adjusting image qualities such as contrast, brightness, and colour balance, thus mimicking the diverse imaging qualities encountered in practice.

Figure 4.8 demonstrates the applied augmentation techniques and their effects on the pixel intensity distributions within the CMR images. The figure elucidates the impact of each augmentation method on the visual and statistical properties of the images. Specifically, the augmentation strategies explored include contrast stretching, histogram equalization, and adaptive equalization.

Contrast stretching enhances the contrast in an image by stretching the range of intensity values it contains to span a desired range of values. This is particularly beneficial in medical imaging, where the contrast between different tissues can be pivotal for accurate segmentation.

Histogram equalization redistributes the image's pixel intensities, effectively spreading out the most frequent intensity values, thereby increasing the global contrast of images—especially when the usable data of the image is represented by close contrast values.

Adaptive histogram equalization, a variant of histogram equalization, computes several histograms, each corresponding to a distinct section of the image, and uses them to redistribute the lightness values of the image. This is particularly useful for improving the contrast of local regions of an image and is well-suited for CMR images, where regional contrast may vary.

During experimentation, we incorporated random noise into the training set images to evaluate its effect on model performance. However, it was observed that the inclusion of random noise did not yield an improvement. Instead, employing solely photometric transformations—specifically, contrast-limited adaptive histogram equalization (CLAHE)—resulted in the most significant enhancement of model performance across both training and test datasets.

The Python code provided utilizes the skimage library's functionality to apply the transformations to the CMR images. The matplotlib library is then used to plot both the original and augmented images alongside their respective histograms and cumulative distribution functions, providing a visual and quantitative analysis of the effects of data augmentation. The code demonstrates the utility of these transformations in preparing the dataset for effective model training. The transformation parameters were carefully selected to simulate realistic variations in image quality, ensuring that the model trained on this augmented dataset would possess the generalization capability necessary for accurate and reliable segmentation in clinical applications.

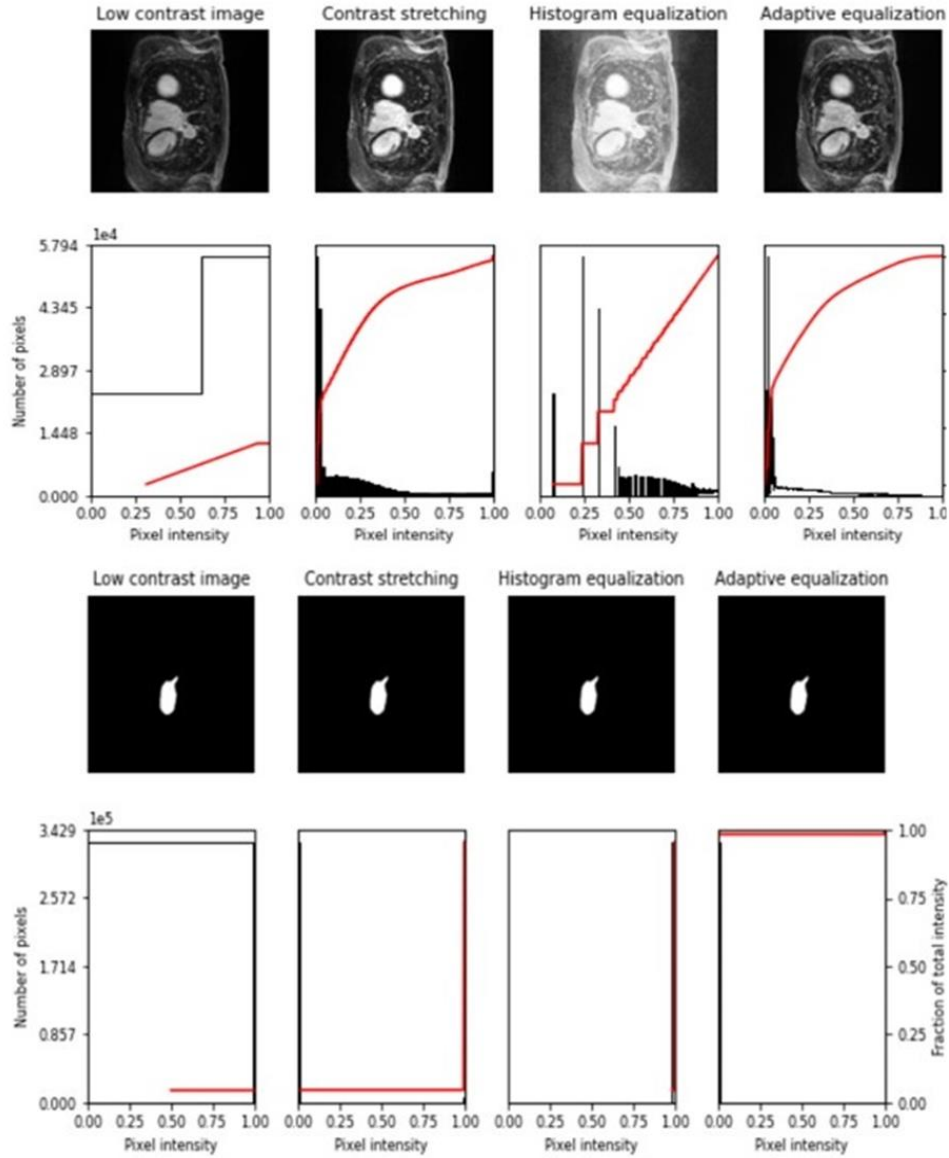


Figure 4-8: Data augmentation and pixel distribution

4.3.5 Loss Function

Binary cross-entropy, Dice Loss function, Tversky loss function and Focal loss function have been used in different model development experiments. The theoretical framework and mathematical background of these loss functions and their impact on the segmentation task is detailed in section 2.5. The best model performance was with Dice Loss function. The dice loss function is best suited for the nuanced task of left atrium segmentation, as it ensures that the model is not only precise in general but also accurate in the specific contours and edges of the atrial structure. The model's performance is quantified through metrics such as the Mean Intersection over Union (IoU) and the Dice coefficient, providing comprehensive feedback.

4.3.6 Model Explainability Methods in Image Analysis

In the realm of image analysis and deep learning, model explainability methods play a crucial role in understanding and interpreting the decisions made by complex neural network models. These methods provide insights into the inner workings of these models, shedding light on how they arrive at their predictions, particularly in tasks such as image segmentation. Several techniques have been developed to achieve model explainability in image analysis tasks. These include saliency methods, which highlight important regions in the input image, activation maximization, which generates synthetic images to maximize certain neuron activations, and class activation mapping (CAM), which localizes discriminative image regions for each class.

When deploying model explainability methods, it's crucial to consider various factors including the selected technique, model architecture, and task-specific demands. This section delves into the implementation of three specific methods: vanilla saliency, smooth gradient saliency, and gradient activation maps, elucidating their application within the scope of this study.

4.3.6.1 *Vanilla Saliency*

Vanilla Saliency is a method used for interpreting deep learning models by identifying which input features contribute most to the model's predictions. It operates by computing the gradient of the loss function with respect to the input features. Higher gradient values indicate features that strongly influence the model's output. The resulting saliency map highlights these influential features, providing insights into the model's decision-making process.

In this study, the initial step involved computing the saliency map through the integration of the loss function and input data. Following computation, normalization of saliency values was undertaken to maintain uniformity across various inputs. Subsequently, visualization of the saliency map was carried out using matplotlib, with pixel intensity reflecting its influence on the model's prediction. This approach facilitated the examination of saliency maps across different inputs, thereby enhancing our comprehension of the features influencing the model's decision-making process.

4.3.6.2 *Smooth Gradient Saliency*

Smooth Grad Saliency is an extension of the Vanilla Saliency method that aims to improve the stability and interpretability of saliency maps. It addresses the issue of high-frequency noise in the gradients by adding random noise to the input data and averaging the resulting saliency maps. This smoothing process helps to reduce the impact of noise and provides more reliable interpretations of the model's behaviour.

In this study, the initial step consisted of computing the saliency map using the loss function and input data, along with additional parameters to control the level of smoothing was applied. Following this, random noise was introduced to the input data, and the saliency map was iteratively computed for each perturbed input. These individual saliency maps were then averaged to generate a smoothed saliency map, which underwent normalization and subsequent visualization using matplotlib. The integration of smoothing techniques ensured the creation of more robust and interpretable saliency maps, offering accurate insights into the model's decision-making process.

4.3.6.3 Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM is a technique used for visualizing and interpreting deep learning models, particularly convolutional neural networks (CNNs). Unlike saliency methods, which focus on individual input features, Grad-CAM provides insights into the spatial locations within an image that contribute most to the model's predictions. It achieves this by computing the gradient of the target class score with respect to the feature maps of the last convolutional layer. The resulting gradients are then used to generate a heat map highlighting the important regions in the input image.

The first step involved utilizing the Grad-CAM library to generate heat maps for visualizing the important regions in the input images. This starts with specifying the target class of interest and selecting the pre-trained CNN model for analysis. Grad-CAM operates to compute the gradients of the target class score concerning the feature maps extracted from the last convolutional layer in the model. These gradients are then weighted by the global average pooling of the corresponding feature maps to produce the final heat map. The heat map is overlaid onto the input image to visually identify the regions that are most relevant for the model's predictions. Through Grad-CAM visualization, we gain insights into the model's decision-making process and understand which image regions contribute most significantly to its predictions.

Comparison of saliency methods for model explainability is summarised in Table 4-1

Table 4-1: Comparison of Saliency Methods

Saliency Method	Description	Advantages	Drawbacks
Vanilla Saliency	Computes the gradient of the loss function with respect to the input features to identify influential features.	▪ Straightforward to implement ▪ Provides direct insight into features influencing the model's predictions	▪ Can be noisy ▪ Often highlights edges or high-contrast areas, which may not be relevant
Smooth Gradient Saliency	Averages the effects of multiple saliency maps with added noise to improve interpretability	▪ Reduces noise and artifacts compared to vanilla saliency ▪ Provides smoother and more interpretable maps	▪ Computationally more intensive due to multiple computations
Grad-CAM	Utilizes gradient information flowing into the last convolutional layer to produce a heatmap of important regions.	▪ Highlights important spatial locations in an image ▪ Provides context by showing relevant regions for predictions	▪ Limited to convolutional layers ▪ May not capture fine-grained importance in dense layers

4.3.7 Ethics approval

In conducting this study, it is crucial to address the ethical considerations associated with the use of medical imaging data. Both the primary dataset from the Cardiac Atlas Study and the validation dataset, CDEMRIS, are open datasets. These datasets are publicly available for research purposes and did not require additional ethical approval from my side for their use in this study.

The primary data utilized in this research, sourced from the Cardiac Atlas Study, was collected under stringent ethical standards. The organizers of this international collaboration ensured comprehensive ethical oversight. Institutional ethics approval was secured by each contributing entity, adhering to both local and international ethical standards for medical research involving human subjects. This approval process included a thorough review by institutional review boards (IRBs) or ethics committees at each participating institution, ensuring that all data collection methods complied with ethical requirements concerning patient consent, privacy, and data security.

The validation data from the Cardiac Atlas Study was anonymized to protect patient confidentiality, ensuring that no personal identifiers were linked to the imaging data. This anonymization process is a critical component of ethical data handling, minimizing the risk of privacy breaches while allowing for valuable scientific research.

Similarly, the CDEMRIS dataset, used as the independent validation set, adheres to ethical guidelines established by the challenge organizers. This dataset was specifically designed for scientific research and benchmarking purposes, collected with proper ethical approval from relevant authorities. The ethical framework for CDEMRIS included obtaining informed consent from all participants involved in the study, ensuring their right to privacy and the confidentiality of their medical data.

The organizers of the CDEMRIS challenge have publicly documented the ethical approval process, providing transparency and reassurance that the data collection and usage comply with high ethical standards. The use of these datasets in our study does not involve any direct interaction with human subjects or any intervention, thereby exempting it from additional ethical review under my research protocol.

4.3.8 Model Validation through Transfer Learning

Transfer learning is employed as a pivotal model validation technique in this study to test the adaptability and robustness of the pre-trained models on the Cardiac Atlas Study data when applied to the CDEMRIS dataset. This approach is essential for the following reasons:

- **Generalization Testing:** It allows for the evaluation of how well the segmentation models can perform under different imaging conditions, which is critical for their use in diverse clinical settings.
- **Model Optimization:** Transfer learning helps in fine-tuning the models on a new dataset with minimal requirement for retraining from scratch, thus saving resources and improving efficiency.
- **Benchmarking:** By applying the models to a standard challenge dataset (CDEMRIS), it provides a benchmark against other segmentation techniques, offering insights into the strengths and weaknesses of the developed models.

The validation of the segmentation models using transfer learning involves a sequential process designed to test the adaptability and effectiveness of the models when applied to a new dataset. The procedure begins with the Model Pre-training, where models are initially trained on the primary dataset, the Cardiac Atlas Study. This phase is crucial as it allows the models to learn generic features that are relevant to MRI image segmentation, setting a solid foundation for robust model performance.

Following the initial training, the models undergo an Adaptation phase. In this phase, the pre-trained models are adapted to the CDEMRS dataset by fine-tuning them on pre-ablation scans. This step is essential as it involves slight adjustments to the model parameters to accommodate the differences in imaging characteristics found in the CDEMRS dataset, compared to the original training data. This ensures that the models are not only specialized to the conditions of the initial dataset but are also flexible enough to handle variations arising from new data sources.

The final phase is the Evaluation of the adapted models. Here, the performance of the models is thoroughly evaluated using the same metrics applied in the primary study phase, such as the Dice coefficient and the Intersection over Union (IoU). This consistent application of evaluation standards is critical to ensure that the performance enhancements and adaptations made during the transfer learning phase hold up under rigorous testing conditions.

By incorporating transfer learning as a validation strategy, the study enhances the reliability of the segmentation models. This approach bridges the gap between experimental accuracy and clinical utility, ensuring that the models developed are practically viable in diverse clinical environments.

4.4 Results

Overall, a total of 63 experiments were conducted to develop and evaluate various models for cardiac magnetic resonance imaging (CMR) segmentation. These experiments were integral in identifying the most effective model configurations and preprocessing techniques to enhance segmentation accuracy. Table 3.1 provides a concise summary of the training results of the top four performing models derived from these comprehensive evaluations. These models were selected based on their superior performance metrics and their relevance to the study's objectives. For a detailed review of the methodologies, performance metrics, and outcomes of the remaining experiments, please refer to Appendices 10.2. This comprehensive documentation in the appendices ensures that all experimental data is accessible for review and further research, while this section focuses on discussing the best four results that are most critical for understanding the advancements in CMR segmentation achieved in this study. Table 3.2 provides a summary of the training results of the best four models derived from these experiments.

Table 4-2: Summary of the training results of the best four models developed.

Description	Model 1	Model 2	Model 3	Model 4
U-Net Levels	3	5	5	5
Loss Function	DICE	Tversky	Binary cross-entropy	Binary cross-entropy
Mean IOU/Jaccard index	0.88	0.88	0.73	0.80
DICE	0.89	0.88	0.88	0.89
Tversky Index	0.88	0.87	-	-
Image Resized	256 x 256 x 1	192 x 192 x 1	256 x 256 x 1	192 x 192 x 1
Trainable Parameters	918,725	15,339,333	15,339,333	15,339,333
Training Time	605 MS per step 133 Sec per epoch	555 MS per Step 126 sec per epoch	275 MS per Step 76 sec per epoch	551 MS per Step 121 sec per epoch
Output Activation	Sigmoid	Sigmoid	Sigmoid	Sigmoid
Image Pre-processing	CLAHE Clip limit 0.01	CLAHE Clip limit 0.01	Contrast Stretching	CLAHE Clip limit 0.0125

The models were evaluated based on several metrics, including the Dice coefficient, the Jaccard index (also known as the Intersection over Union or IOU), and the Tversky index. These metrics are indicative of the segmentation accuracy and overlap between the predicted and ground truth segmentations.

Among the four best-performing models, Model 1 and Model 2 emerged as particularly noteworthy. Model 1 utilized a U-Net architecture with three levels, employing the DICE loss function and sigmoid activation for output, while Model 2 employed a similar architecture with five levels and the Tversky loss function. Model 1 demonstrated impressive segmentation performance, achieving a Mean IOU (Intersection over Union) of 0.88 and a DICE coefficient of 0.89. Additionally, Model 1 exhibited a Tversky Index of 0.88, indicating its robust performance across multiple evaluation metrics. In comparison, Model 2 achieved a Mean IOU of 0.88 and a DICE coefficient of 0.88, with a slightly lower Tversky Index of 0.87. Despite the comparable segmentation accuracy between the two models, Model 1 showcased superior computational efficiency, with a relatively lower number of trainable parameters (918,725) and shorter training time. Furthermore, Model 1 employed Contrast Limited Adaptive Histogram Equalization (CLAHE) with a clip limit of 0.01 for image pre-processing, enhancing its ability to capture subtle details in CMR images.

The segmentation predictions of both best and worst are visualized in Figure 4.9, for model 1 providing further visual insights into the segmentation performance and model predictions.

Detailed model metrics on model 1 is presented under Figure 4.10, providing insights into their performance and loss. Overall, these results highlight the effectiveness of models in accurately segmenting CMR images and their potential for clinical applications in cardiac imaging analysis.

These results underscore the effectiveness of the developed models in accurately segmenting CMR images for atrial fibrillation analysis, with implications for improving diagnostic accuracy and clinical decision-making in cardiovascular healthcare settings.

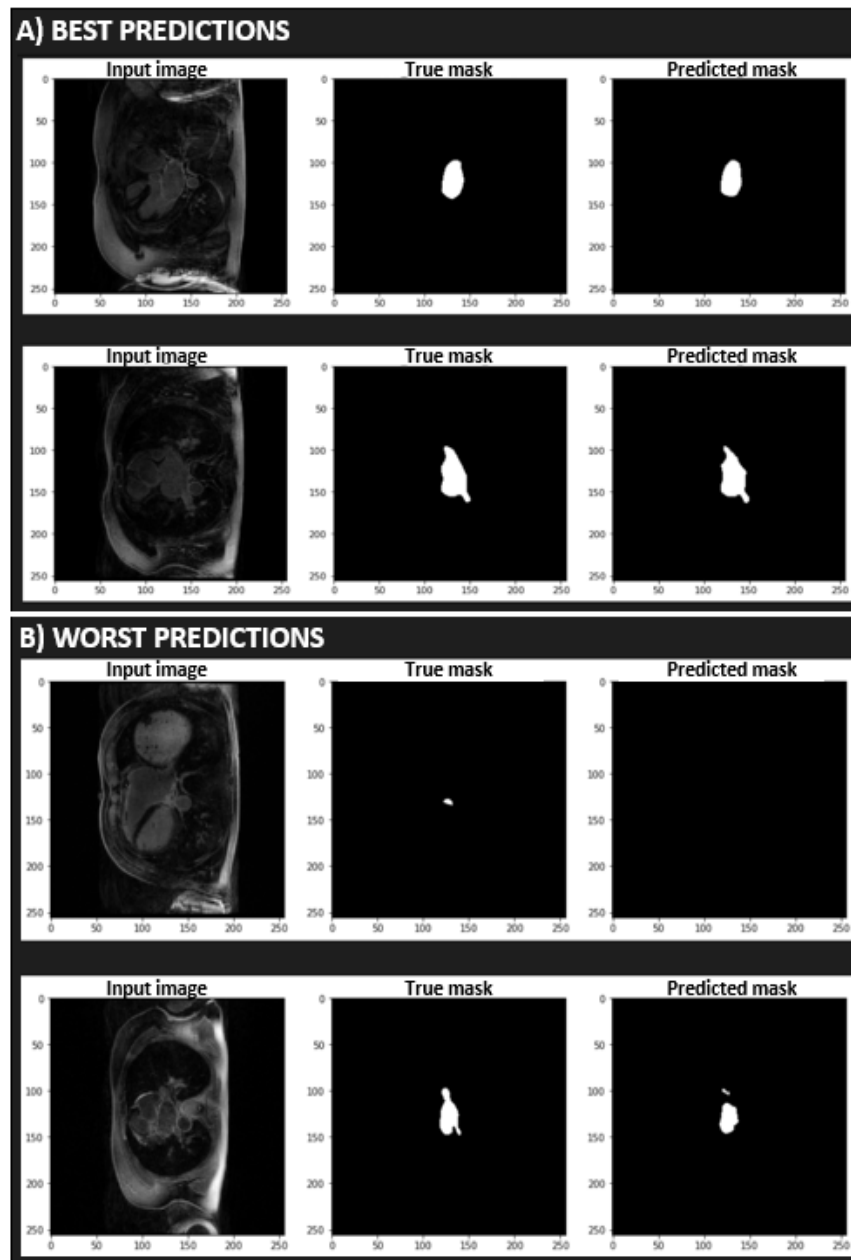


Figure 4-9: Two examples best and worst predictions

4.5 Model Metrics

Dice Loss and model metrics of Mean IOU (Jaccard Index), Dice Coefficient and Tversky index are summarized in figure 4-10.

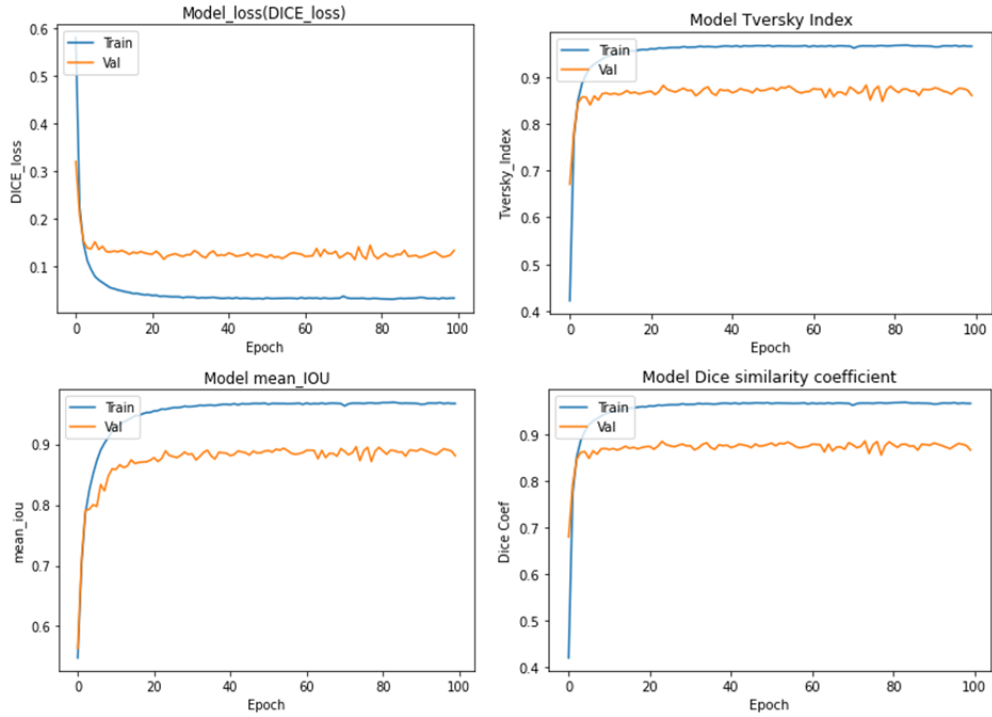


Figure 4-10 : Best model metrics

Based on the performance metrics obtained for validation and training data, Model 1 demonstrates robust performance in segmenting cardiac magnetic resonance imaging (CMR) scans. For the validation dataset, the model achieves a mean Intersection over Union (IOU) of approximately 0.88, indicating a high degree of overlap between the predicted and ground truth segmentation masks. Additionally, the Dice coefficient, a measure of segmentation accuracy, reaches approximately 0.86, further confirming the model's performance in delineating cardiac structures. The Tversky index, which quantifies the balance between true positives and false positives in segmentation tasks, is around 0.86, reflecting the model's ability to accurately capture relevant anatomical features while minimizing false positives. Notably, the model demonstrates similar performance trends on the training dataset, with slightly higher metrics, including a mean IOU of approximately 0.97 and a Dice coefficient of approximately 0.97. These results indicate strong agreement between the predicted and ground truth segmentations.

4.6 Generalisation of the model on CDEMIRIS Database

The best model 1 was evaluated on an independent dataset from the CDEMIRIS 2013 challenge dataset, where it returned a Dice coefficient of 0.84 and a Mean IOU of 0.79, indicating its robustness and generalizability across different datasets.

In the final phase of evaluation, Model 1, which demonstrated the highest efficacy during initial testing phases, was rigorously tested on the independent CDEMRIS dataset from the 2013 Cardiac Delayed Enhancement Segmentation Challenge. This dataset, distinct from the training dataset in terms of imaging characteristics and clinical settings, provided a robust platform for assessing the model's generalizability and adaptability to new, unseen data.

The evaluation metrics returned a Dice coefficient of 0.84 and a Mean Intersection over Union (IoU) of 0.79. These metrics are crucial as they indicate the precision and accuracy of the segmentation results in medical imaging. A Dice coefficient closer to 1 signifies a high degree of similarity between the predicted segmentation and the ground truth, and an IoU above 0.75 is typically considered excellent in clinical applications.

To visually substantiate the quantitative assessments, segmentation results from the CDEMRIS dataset were compiled. The images displayed below illustrate the segmentation performance of Model 1, highlighting areas where the model excelled and where it faced challenges:

Figure 4.11: Example of Successful Segmentation on CDEMRIS Dataset

This figure shows a side-by-side comparison of the ground truth (left) and the model's segmentation output (right). The images demonstrate precise delineation of the left atrial cavity, closely matching the ground truth.

Figure 4.12: Example of Suboptimal Segmentation on CDEMRIS Dataset

This figure contrasts the ground truth (left) with a less accurate model output (right). It identifies specific regions where discrepancies occur, such as under-segmentation or over-segmentation, which are critical for refining future iterations of the model.

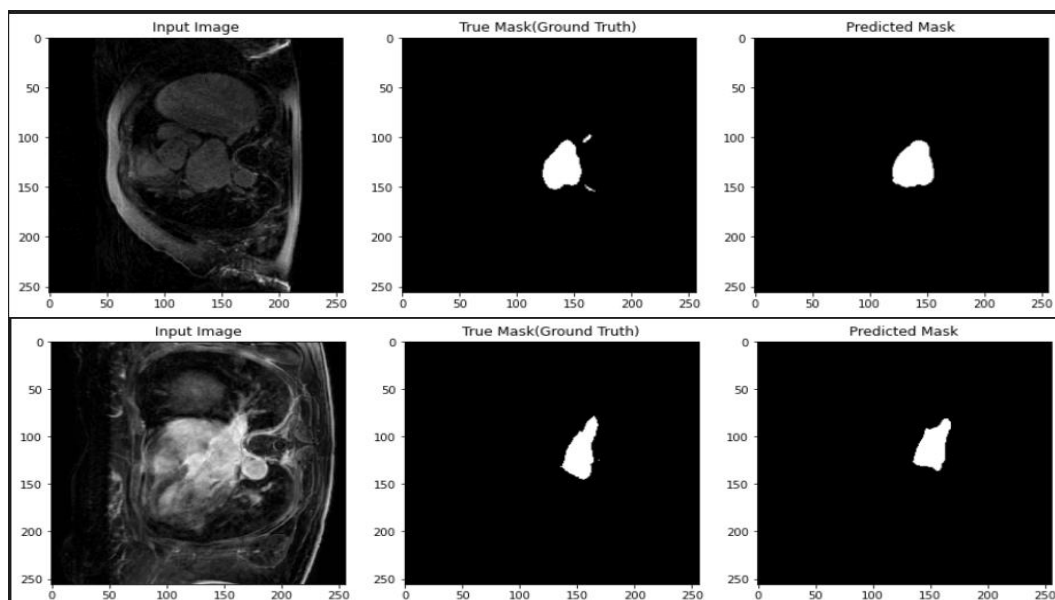


Figure 4-11 - Optimal Segmentation of CDEMRIS dataset

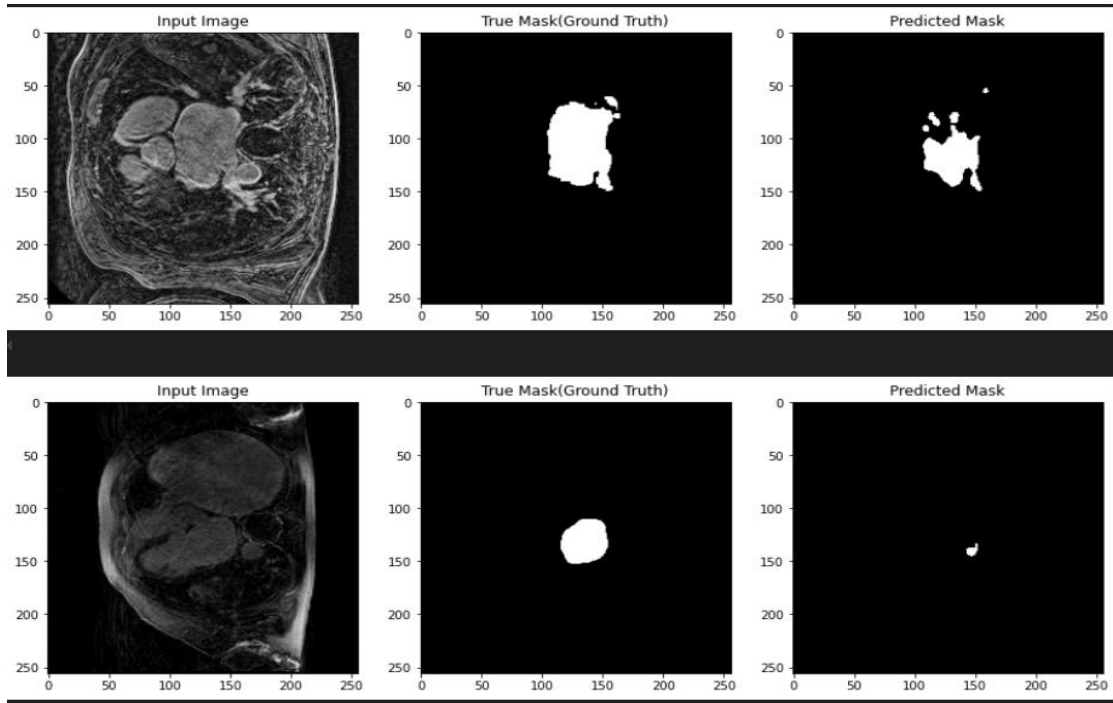


Figure 4-12 - Suboptimal segmentation of CDEMRS dataset

These visual comparisons serve as a powerful tool for understanding the practical implications of the model's performance metrics, offering clear insights into its strengths and areas for improvement. They also provide a transparent method for communicating the effectiveness of the model to clinical professionals, ensuring that the technological advancements are grounded in real-world applicability.

The robust performance of Model 1 on the CDEMRS dataset underscores its potential for clinical application across various settings. The findings from this evaluation will contribute to continuous improvements in the model, focusing on areas identified as problematic in the visual assessments. Further studies will aim to refine the model's accuracy, particularly in challenging cases as revealed by the suboptimal segmentation examples.

By integrating both quantitative metrics and qualitative visual assessments, this section not only confirms the model's capabilities but also sets the stage for future enhancements, ensuring that the models developed are not only theoretically effective but also practically reliable for clinical use.

4.7 Model explainability

Model explainability is critical in clinical settings, ensuring trust and interpretability of AI-based decisions. In this study, Grad-CAM and Saliency maps are employed to visualize the model's reasoning by identifying the image pixels most influential for the segmentation tasks.

Grad-CAM provides insight into the model's focus by highlighting the areas contributing most significantly to the predictions. As shown in Figure 4-13, the red areas signify regions with higher importance in determining the segmentation mask. This visualization not only pinpoints the critical pixels within the target area but also

identifies other regions in the image that the model deemed relevant, which could correspond to similar anatomical structures or patterns recognized by the neural network. It shows the model's capacity to localize the features essential for decision-making without any pixel-level annotations, ensuring the model's attentiveness to relevant patterns in the image data.

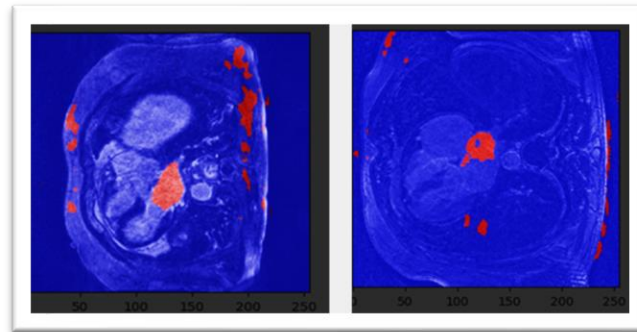


Figure 4-13 : Best model metrics

In contrast, the Vanilla Saliency maps (Figure 4-14) offer a more granular perspective by illuminating individual pixels based on their contribution to the model's output. The bright spots indicate where slight changes would most affect the segmentation decision, essentially capturing the gradient of the prediction score with respect to the input image. These maps, however, can be noisy and less interpretable, often highlighting edges or high-contrast areas regardless of their relevance to the task.

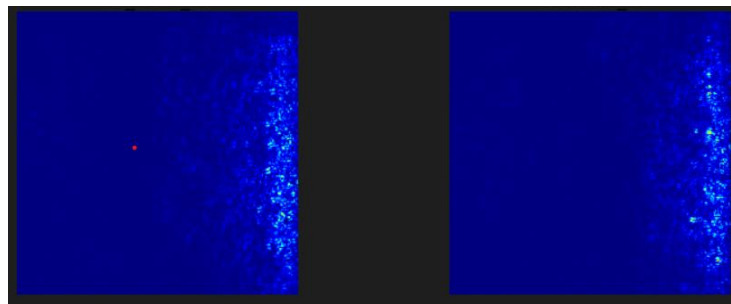


Figure 4-14 : Vanilla Saliency results

The Smooth Grad Vanilla technique (Figure 4-15) seeks to address the noisiness of Vanilla Saliency maps by averaging the results of many saliency maps with added noise, leading to smoother and potentially more interpretable visualizations. This method can sometimes provide a clearer indication of the important regions by reducing the visual clutter and highlighting consistent patterns across multiple perturbations of the input image.

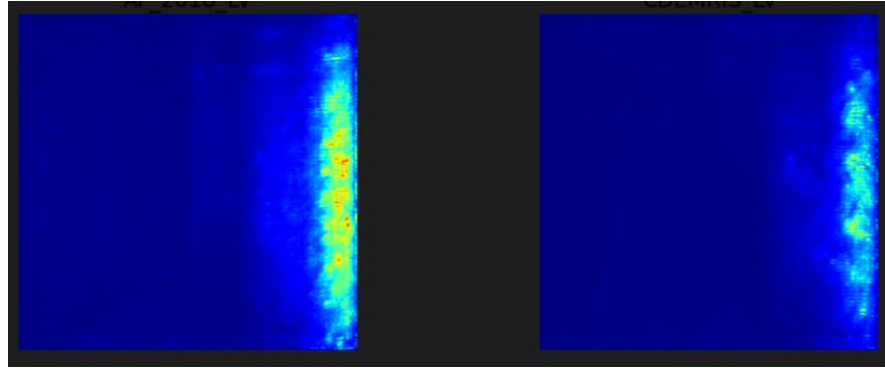


Figure 4-15 : Smooth Grad Vanilla Saliency results

The preliminary results from both Saliency map techniques, while less specific than Grad-CAM, still offer valuable insights. Comparing these with Grad-CAM's more focused regions, it's evident that Saliency maps could benefit from refinement to match the specificity of Grad-CAM. They currently act as broad indicators of influential areas and may be complemented by techniques like Integrated Gradients, which can provide pixel-wise attribution scores by considering the entire path of the input image to its current state.

For further improvement, it's proposed to explore Integrated Gradients to deepen the interpretability aspect. Such methods could provide a more nuanced understanding of the model's behaviour and decisions, making them more transparent and clinically reliable. Collaborating with clinicians to review these explainability visualizations will be vital to align the AI model's segmentation choices with expert medical knowledge and practical diagnostic reasoning. The aim will be to iterate on these methods to achieve a more intuitive and clinically relevant explanation of the AI model's functionality.

4.8 Outcomes and Future Directions

The findings from this study affirm the efficacy of advanced deep learning models, notably U-Net architectures, in segmenting the left atrium from CMR scans for atrial fibrillation patients. Despite the challenges inherent in medical imaging datasets, strategic data pre-processing and augmentation techniques, such as CLAHE, have proven to be instrumental in augmenting model performance. The best models demonstrated superior segmentation capabilities, as reflected by high Dice coefficients and Mean IOU values. Model explainability was significantly advanced through the use of Grad-CAM and Saliency maps, with further refinements suggested to improve interpretability in clinical settings. The successful application of these models to an independent dataset underscores their robustness and generalizability, marking a step forward in the diagnostic accuracy and decision-making in cardiovascular healthcare.

Future work will focus on enhancing explainability through integrated gradients and clinician collaboration, ensuring the AI models not only perform with high accuracy but also align with the nuanced realities of clinical practice.

4.9 Conclusion

In conclusion, the methodologies and findings presented in this chapter underscore the capabilities of advanced deep learning models, specifically U-Net architectures, in enhancing the segmentation accuracy for CMR scans of patients with atrial fibrillation. The strategic application of data pre-processing and augmentation techniques, like CLAHE, has significantly bolstered model performance, as evidenced by high Dice coefficients and Mean IOU values. Moreover, the use of explainability methods such as Grad-CAM and Saliency maps has improved the interpretability of these models, making the AI's decision-making process more transparent.

The successful validation of these models against an independent dataset not only demonstrates their robustness and generalizability but also highlights the potential for these techniques to advance diagnostic accuracy and decision-making in cardiovascular healthcare. However, despite these advancements, further work is needed to enhance model explainability and to ensure that these AI solutions are in alignment with clinical realities, through methods such as integrated gradients and ongoing collaboration with clinical experts.

As this study progresses, Chapter 5 will build upon these foundations to explore a wider array of AI methodologies applied across different modalities of cardiovascular data. This will include a detailed examination of multi-modal data analysis techniques and their application in predicting adverse cardiovascular events, thus expanding the scope of AI utility in cardiac health beyond imaging to encompass comprehensive patient assessments.

5 In-depth Analysis of Electrocardiogram – Techniques and Applications in Artificial Intelligence

5.1 Background

Building upon the methodologies discussed in Chapter 3, this chapter advances the analysis by focusing on the extraction and fusion of heart rate variability (HRV) features with clinical factors to diagnose a spectrum of cardiac conditions, including myocardial infarction, dysrhythmia, bundle branch block, cardiomyopathy, and general heart health. This integrative approach addresses the previously identified gap of effectively combining HRV data with relevant clinical inputs to enhance the predictive accuracy of machine learning models. It includes an exploration of machine learning algorithms for feature extraction and their impact on model performance and prediction accuracy for cardiovascular events. Furthermore, this chapter evaluates the preliminary effects of augmenting our dataset with synthetic samples generated through Generative Adversarial Networks (GANs), aiming to improve model robustness and prediction accuracy. Detailed methodologies for GAN augmentation, various models of GANs in generating synthetic data, and their effect on model performance are thoroughly explored in Chapter 6. This step is crucial in refining our multimodal diagnostic models and showcases the potential of sophisticated AI techniques in advancing cardiac healthcare.

Electrocardiogram (ECG) signals serve as a cornerstone in cardiac health diagnostics. The high prevalence of cardiac diseases, particularly myocardial infarction, underscores the necessity for developing reliable and efficient classifiers for ECG signals. This endeavour is pivotal in the overall research landscape, serving as a crucial step toward harnessing machine learning methodologies to advance healthcare outcomes and management strategies for cardiac diseases.

Myocardial infarction represents one of the most severe and life-threatening cardiac diseases, responsible for significant morbidity and mortality worldwide. Atrial fibrillation (AF) is increasingly recognized as a significant risk factor for myocardial infarction, due to its role in creating irregular heart rhythms that can lead to decreased cardiac efficiency and the formation of blood clots, which may precipitate an MI. Incorporating the identification of AF into diagnostic models could enhance the prediction and prevention strategies for myocardial infarction. In this context, the use of Inter Beat Intervals (IBIs) as a feature variable in machine learning models offers a fresh perspective. By capturing subtle temporal patterns and variations in heart activity, including those induced by AF, IBIs can provide a more nuanced characterization of myocardial infarction. This not only enhances diagnostic precision but also guides treatment strategies. Coupled with machine learning methodologies, this approach has the potential to revolutionize diagnosis and pave the way for personalized, timely, and effective treatment modalities.

Figure 5-1 illustrates the ECG representations of lead II of various heart conditions, including a healthy heart, myocardial infarction, bundle branch block, dysrhythmia, and conditions with cardiomyopathy. Each trace highlights the distinct electrical patterns that AI models must learn to classify, emphasizing the complexity of accurate diagnosis and the potential of AI to assist in such tasks. The study of ECG signals often extends beyond the mere interpretation of raw waveforms.

Incorporating IBIs provides a more comprehensive viewpoint. IBIs, defined as the time intervals between successive heartbeats, offer unique insights into the heart's electrical activity and overall health. The utility of extracting IBIs from raw ECG signals lies in its potential to capture temporal patterns and variations that may

be less discernible in raw signals alone. This layered approach enriches the feature set, enhancing the effectiveness of classification algorithms in diagnosing myocardial infarction and other cardiac anomalies.

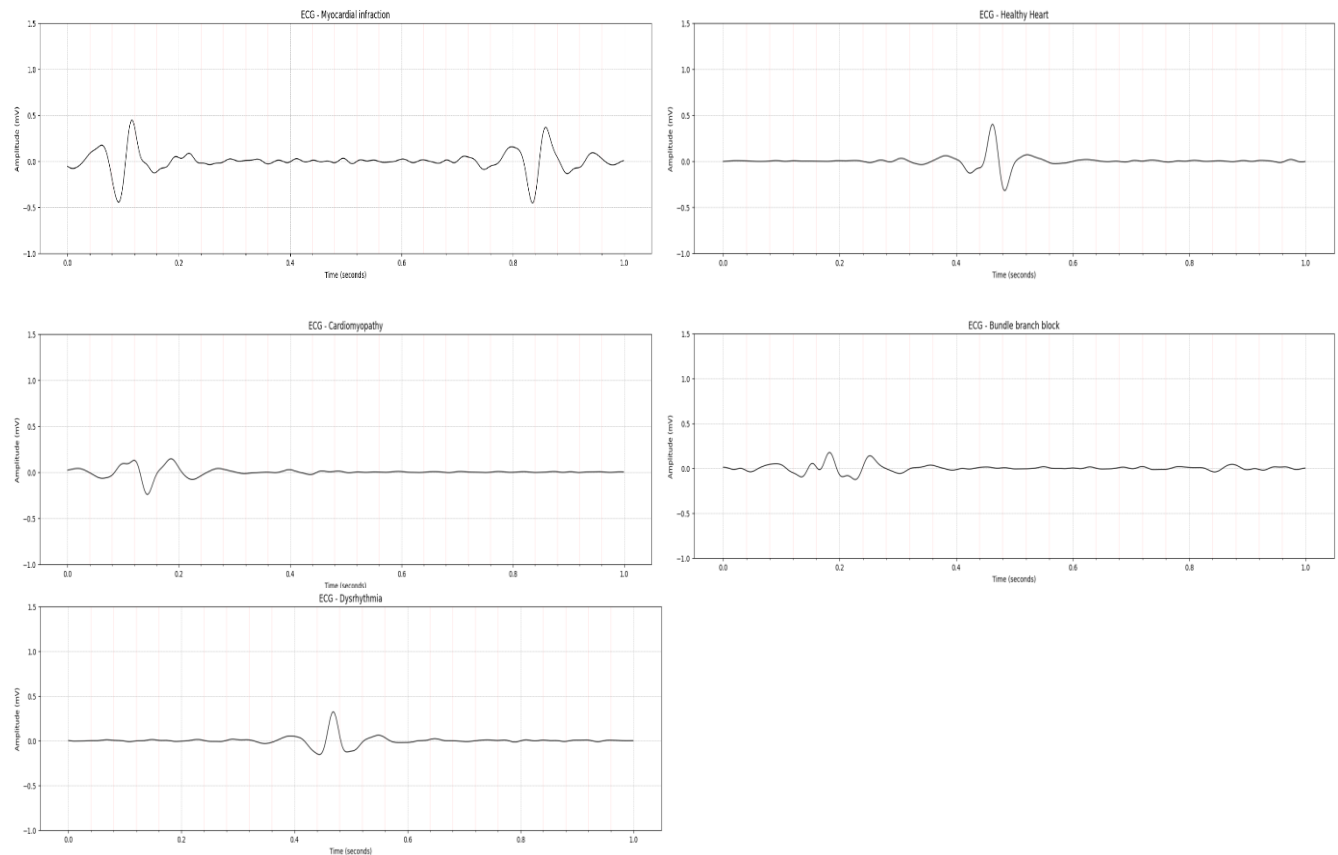


Figure 5-1 - ECG Waveform of various cardiovascular conditions and Healthy Heart

Inter Beat Intervals can be extracted from ECG signals through multiple techniques, ranging from simple peak detection algorithms to complex wavelet transformations. While simpler methods are computationally efficient, Peak detection is a common method for finding the R-peaks in ECG signals. The basic idea is to identify the points in the ECG signal where the amplitude is the highest. There are many different peak detection algorithms, but they all work on this basic principle. On the other hand, more complex methods calculated from the wavelet coefficients, or extract the instantaneous frequency of the ECG signal offer greater accuracy but at the cost of computational efficiency. In this study, have leveraged the capabilities of the HeartPy package and used simple peak detection algorithm to achieve a balance between accuracy and computational load."

Incorporating ECG data, particularly IBIs, along with clinical factors serves a dual purpose in cardiac health diagnostics: cost-effectiveness and expedited diagnosis. Traditional diagnostic methods often require invasive procedures and multiple tests, incurring higher costs and prolonging the diagnostic timeline. By contrast, leveraging machine learning models trained on ECG data and clinical variables offers a non-invasive and quicker alternative for identifying adverse cardiac outcomes.

Various machine learning algorithms have been explored for ECG classification, each with its own set of advantages and limitations. While Support Vector Machines offer high accuracy, they often require extensive computational resources. Neural Networks offer the advantage of learning complex patterns but are prone to overfitting. In this study, algorithms such as CNN, logistic regression, Extreme Gradient Boosting (XGBoost), and random forest were chosen for their balance of accuracy and computational efficiency.

Initial attempts were made using a Convolutional Neural Network (CNN) to build a binary classification model based solely on raw ECG signals. The model aimed to discriminate between healthy hearts and those exhibiting myocardial infarction. Utilizing the PTB open database, the model achieved reasonable metrics, specifically an accuracy rate of 81% and an Area Under the Curve (AUC) score of 79%. However, the model faced limitations in terms of precision and specificity, which were largely attributed to data noise and imbalance.

To improve the model's performance, an exhaustive feature extraction methodology focusing on Inter Beat Intervals (IBI) was adopted, leveraging the capabilities of the HeartPy package in Python. Features pertaining to time-domain, amplitude-based linear, and frequency-domain measures were extracted from the ECG waveform data. Despite offering a comprehensive understanding of cardiac health, initially, noise interference and poor data quality inhibited complete feature extraction across samples.

To counteract these limitations, a series of rigorous pre-processing techniques were implemented. High-pass filters were used to eliminate baseline wander induced by patient movement, notch filters eradicated power line interference noise, and band-stop filters removed muscle contraction noise. This culminated in the successful extraction of features for all but three of the 445 patient samples.

Further development led to the creation of a robust binary classification model employing machine learning algorithms such as logistic regression, Extreme Gradient Boosting (XGBoost) and random forest. Balanced partitioning into training, testing, and validation subsets addressed the issue of overfitting. Despite achieving AUC and accuracy scores ranging from 81% to 84%, challenges still existed, particularly concerning incorrect classification due to weak signals in the training dataset.

To surmount these obstacles, extracted features of IBI and clinical metadata were incorporated into the model. Additionally, data scarcity was addressed by employing Generative Adversarial Networks (GANs) to augment the dataset with synthetic samples, enriching the feature space. This strategy significantly improved the model's performance, aligning it with our overarching research goal of advancing cardiac healthcare through machine learning methodologies.

5.2 Objective:

The primary objective of this chapter is to elucidate the development, validation, and optimization of a machine learning-based model for the efficient and precise classification of Electrocardiogram (ECG) signals. This model specifically focuses on diagnosing adverse cardiovascular events such as myocardial infarction (MI), cardiomyopathy, bundle branch block, and dysrhythmia, as well as assessing general heart health. These conditions represent a spectrum of potential cardiac issues that can be detected and analysed through advanced AI methodologies. This work forms a crucial component of the broader thesis, which seeks to enhance healthcare outcomes related to these adverse cardiac outcomes by leveraging AI technologies.

5.3 Data collection and description: PTB Diagnostic ECG Database

The dataset used in this research is sourced from the PTB Diagnostic ECG Database [158], available on Physionet. This database is a well-regarded resource for ECG research and contains a diverse collection of 549 records from 294 subjects, including healthy volunteers and patients with a variety of cardiac conditions.

Dataset Characteristics

- **Number of Samples:** The database comprises 549 records.
- **Types of Subjects:** The dataset is diverse, encompassing both healthy individuals and patients diagnosed with a wide range of cardiac anomalies such as myocardial infarction, cardiac arrhythmia, and heart murmurs. Detailed subjects in each diagnosis class is presented in Table 5-1
- **Data Features:** Each record contains 15 simultaneously measured signals: the standard 12-lead ECG as well as three Frank lead ECGs. The signals are sampled at a frequency of 1000 Hz.
- **Data Duration:** Each ECG record in the database varies in length but generally captures 10 minutes of ECG data.

Relevance to the Study

The PTB Diagnostic ECG Database was chosen for this research due to its diversity, comprehensive nature, and the credibility it enjoys in the medical research community. Its wide variety of diagnostic categories makes it suitable for developing a robust machine learning model capable of classifying myocardial infarction alongside other cardiac conditions. Moreover, the dataset contains raw ECG signals, enabling both time-domain and frequency-domain analysis, which aligns well with our research objectives of exploring various feature extraction methods and machine learning models. The richness of clinical summary data was also one of the factors contributing to research objectives for choice of data set.

The PTB Diagnostic ECG Database comprises 549 records, collected from 290 individuals who range in age from 17 to 87, with an average age of 57.2 years. Of these subjects, 209 are men with an average age of 55.5, and 81 are women with an average age of 61.6. The ages of 15 subjects are not documented. Each participant in the database is represented by one to five separate records. It should be noted that there are no records for subjects numbered 124, 132, 134, or 161.

For each record, the dataset offers 15 signals that are measured simultaneously. This includes the standard 12-lead ECG (i, ii, iii, avr, avl, avf, v1, v2, v3, v4, v5, v6) and an additional three Frank lead ECGs (vx, vy, vz). The signals are digitized at a rate of 1000 samples per second and offer 16-bit resolution within a range of ± 16.384 mV.

Further enhancing the dataset's value, most records come with a comprehensive clinical summary in their header files. These summaries frequently include details such as age, gender, diagnostic information, and where relevant, additional data concerning medical history, medication, interventions, coronary artery pathology, ventriculography, echocardiography, and hemodynamic. However, it's worth mentioning that clinical summaries are not provided for 22 of the subjects. Diagnostic categorizations are available for the remaining 268 subjects.

Table 5-1 : PTB Dataset

Diagnostic class	Number of subjects[158]
Myocardial infarction	148
Cardiomyopathy/Heart failure	18
Bundle branch block	15
Dysrhythmia	14
Myocardial hypertrophy	7
Valvular heart disease	6
Myocarditis	4
Miscellaneous	4
Healthy controls	52

In the present study, for binary classification task the dataset was divided into training, testing, and validation sets to ensure a robust analysis. The dataset consists of two categories: Healthy Control (HC) and Myocardial Infarction (MI). The training set comprises 268 samples, of which 49 belong to the HC category and 219 are classified as MI. The test set consists of 90 samples, with 15 from the HC group and 75 from the MI group. Similarly, the validation set also includes 90 samples, with 16 categorized as HC and 74 as MI.

For multilabel classification task the dataset encompasses five different categories: Bundle Branch Block (BB), Cardiomyopathy (CM), Dysrhythmia (Dsy), Healthy Control (HC), and Myocardial Infarction (MI). The training set consists of 300 samples, broken down as follows: 11 samples for BB, 11 for CM, 10 for Dsy, 49 for HC, and 219 for MI. For the testing set, a total of 99 samples are used, distributed among BB (3), CM (3), Dsy (3), HC (15), and MI (75). Similarly, the validation set is composed of 99 samples, with BB, CM, and Dsy each having 3 samples, HC having 16, and MI having 74. The dataset's grand total stands at 498 samples. This stratification ensures a balanced representation of each category across all subsets, thereby aiding in a more robust and generalizable analysis.

5.4 Data pre-processing

Data preprocessing is a crucial step in the analysis of biomedical signals, such as electrocardiogram (ECG) data, particularly when dealing with datasets like the PTB Diagnostic ECG Database, which is characterized by noise. Noise in ECG signals can arise from various sources, including baseline wander, power line interference, and other artifacts introduced during signal acquisition. To ensure accurate and reliable downstream analysis, it's imperative to mitigate these sources of noise through preprocessing techniques.

The PTB Diagnostic ECG Database was acquired from the Physionet website. Metadata and raw ECG signals were stored in two initialized data structures, which were populated iteratively by looping through each record in the PTB database.

Metadata extraction involved utilizing a specialized utility within a widely used package for handling physiological waveform data. This utility, designed for parsing raw ECG signals, extracted essential information such as patient demographics, recording parameters, and signal characteristics. The extracted metadata was then stored in a list, with patient and record IDs as keys.

Subsequently, raw ECG data arrays were converted to 32-bit floating-point numbers for computational efficiency and memory optimization. These transformed data arrays were stored in a dictionary with unique keys generated by concatenating patient and record IDs. Following this, the transformed data arrays were converted into CSV files for storage and accessibility. This approach established a structured and optimized data pipeline, facilitating subsequent feature extraction and machine learning modelling.

The initial phase of data pre-processing for noise removal in ECG signals commences by establishing the working directory, which grants access to the relevant ECG data files. Subsequently, these files are enumerated to facilitate subsequent processing steps. Each individual file within the enumerated list undergoes a defined series of pre-processing steps implemented through an iterative loop.

5.4.1 Noise Filtering – Baseline Wander

Baseline wander, primarily caused by low-frequency fluctuations introduced by factors such as respiration or electrode motion, was addressed using a high-pass filter with a cutoff frequency of 0.05 Hz. This filter effectively treated frequencies below 0.05 Hz as noise, removing baseline wander from the ECG signals. Specifically, a third-order Butterworth filter was employed for noise reduction. By selectively removing baseline wander, the accuracy of cardiac signal analysis is improved, as the true ECG signal is better isolated from unwanted fluctuations.

Advantages: The high-pass filter's computational efficiency and adaptability, coupled with the Butterworth filter's frequency response characteristics, ensures optimal noise reduction while maintaining signal fidelity.[159]–[162]

5.4.2 Noise Filtering - Band-Pass Filter

In contrast to baseline wander removal, which specifically addresses low-frequency fluctuations, the band-pass filter operates differently. While the baseline wander removal focuses on attenuating low-frequency noise below 0.05 Hz, the band-pass filter selectively permits frequencies within a predefined range, typically set between 0.5 Hz to 40 Hz, while attenuating frequencies outside this range. This filtering technique effectively isolates the frequency band associated with cardiac activity, allowing essential signal components within this range to pass through, thereby enhancing the clarity and interpretability of the ECG signals. The selection of the Butterworth filter in the band-pass technique is deliberate, as this filter is mathematically engineered to maintain a maximally flat frequency response within its passband. Consequently, when applied to the signal, the filter ensures that no unwanted fluctuations or ripples are introduced in the frequency domain within the specified passband.

Advantages: This filter excels in retaining all frequencies of interest while effectively eliminating other noise. Its use here ensures that the most important frequencies relevant to cardiac activity are retained, making it highly effective for ECG data.[163]–[166]

5.4.3 Noise Filtering - Notch Filter

A Notch filter is utilized to remove specific frequency components, commonly encountered as power line interference. The selection of the 50 Hz frequency for the Notch filter is deliberate and corresponds to the frequency typically associated with power line interference. This choice allows the filter to effectively nullify the amplitude of signals at this unwanted frequency, thereby removing power line noise from the ECG data.

The Notch filter serves a distinct purpose from baseline wander removal and band-pass filtering. While baseline wander removal targets low-frequency fluctuations, and band-pass filtering selectively permits frequencies within a predefined range, the Notch filter specifically addresses the elimination of interference at a particular frequency, such as power line noise. This targeted approach ensures the removal of unwanted frequency components while preserving the integrity of the rest of the signal.

Advantages: The notch filter is highly selective and only removes the specified frequency, thereby preserving the integrity of the rest of the data. This makes it highly effective for eliminating singular interference frequencies without affecting the overall signal quality.[167], [168]

5.4.4 Noise Filtering - Band-Stop Filter

Finally, the band-stop filter is used to remove a band of frequencies around a centre frequency, in this case, around 250 Hz. This is technically achieved by combining a high-pass and a low-pass filter, allowing all frequencies except those in the stop band to pass through.

Advantages: Band-stop filters are highly effective in removing a range of unwanted frequencies while allowing the rest of the spectrum to pass through. This is particularly useful in scenarios where noise occurs in a specific frequency band, thereby preserving the integrity of frequencies outside of this range.[169], [170]

These filtering techniques work synergistically, enabling a comprehensive noise reduction strategy tailored to the specific challenges posed by ECG data. By employing these methods, the dataset was substantially cleaned, making it more conducive for accurate and reliable downstream analyses.

5.4.5 Signal to Noise Ratio

Signal-to-Noise Ratio (SNR) is a pivotal metric in signal processing and data analysis that quantifies the clarity of the desired signal amidst the backdrop of background noise. It is expressed in decibels (dB) and is computed by comparing the power of the meaningful signal to the power of the noise interfering with it. A higher SNR indicates a clearer, more distinguishable signal, which is particularly crucial in applications such as biomedical signal processing where the integrity of the data is vital. The logarithmic scale used in SNR not only compresses a wide range of power ratios into a more interpretable range but also provides a dimensionless unit that enables comparability across different signals and systems.[171]–[173]

The methodology employed for segregating signal and noise operates under the assumption that every column in the input data matrix represents an individual ECG lead. In this process, noise estimation for each lead involves calculating the arithmetic mean of all other leads, excluding the current one being analysed. Consequently, the signal for each lead is defined as the data contained within its corresponding column. This methodology results in the creation of a list of tuples, denoted as signal noise list, where each tuple encapsulates the estimated signal and noise components for every lead.

While this approach streamlines the calculation of Signal-to-Noise Ratio (SNR) by treating the remaining leads as noise, it's essential to acknowledge that this assumption may not universally apply. In practice, all leads

typically contain significant signal information, and assuming statistical independence between noise and signal in each lead may not always be accurate.

SNR is calculated based on the power of the signal P_s and P_n the power of the noise as follows:

$$P_s = \sum_{i=1}^N s_i^2 \quad \text{Equation 5-1}$$

$$P_n = \sum_{i=1}^N n_i^2 \quad \text{Equation 5-2}$$

$$\text{SNR (dB)} = 10 \cdot \log_{10} \left(\frac{P_s}{P_n} \right) \quad \text{Equation 5-3}$$

Where s_i and n_i are the signal and noise values at the i th sample, and N is the total number of samples. The logarithm (base 10) of the ratio of P_s to P_n is calculated to express the SNR in decibels (dB). The log transformation serves to normalize the potentially extensive range of signal and noise power levels, making the resulting SNR more interpretable and dimensionless.

The methodology outlined above iterates through each ECG file in a designated directory. For each file, it applies the preceding methods to calculate the SNR for every lead. The resulting DataFrame contains filenames, estimated signal, and noise, SNRs, and statistical summaries such as the mean and median SNRs across all leads. A Boolean flag is also incorporated to indicate if all leads have an SNR falling below a pre-determined threshold, which is defaulted to 6 decibels (dB). Decibels (dB) are a logarithmic unit used to express the ratio of two values, such as power levels or amplitudes. In the context of signal-to-noise ratio (SNR), a higher value in decibels indicates a stronger signal relative to the noise. Conversely, a lower value in decibels suggests that the noise is relatively stronger compared to the signal. Therefore, a threshold of "6 dB" means that the SNR should be at least 6 decibels for the signal to be considered adequately distinguishable from the noise.

To evaluate the effectiveness of our data cleaning methodology on the Signal-to-Noise Ratio (SNR), we employ box plots (Figures 5-2 to 5-4) to visualize the distribution of median SNR values across two groups—namely, the uncleaned and cleaned ECG training datasets. These box plots serve as a robust statistical graphical representation that provides insights into the data's central tendency, dispersion, and overall distribution.

In the first subplot, the box plot depicts the distribution of median SNR values for the uncleaned ECG training data. The y-axis represents the median SNR, and the plot is scaled to start from -2 units. The box extends from the lower quartile (Q1) to the upper quartile (Q3), encapsulating the interquartile range (IQR), which is a measure of statistical dispersion. The horizontal line inside the box indicates the median value of the dataset.

Similarly, the second subplot illustrates the median SNR values for the cleaned ECG training data. The y-axis is scaled to start from approximately 20.828 units, and it extends up to a value slightly greater than the maximum observed median SNR value in the cleaned data. As in the first plot, the box represents the IQR, and the horizontal line inside the box indicates the median value.

By juxtaposing these two box plots, we can clearly observe a significant improvement in the SNR for the cleaned dataset compared to the uncleaned dataset. This reinforces the efficacy of our adopted pre-processing techniques in enhancing the quality of ECG signals for subsequent analyses.

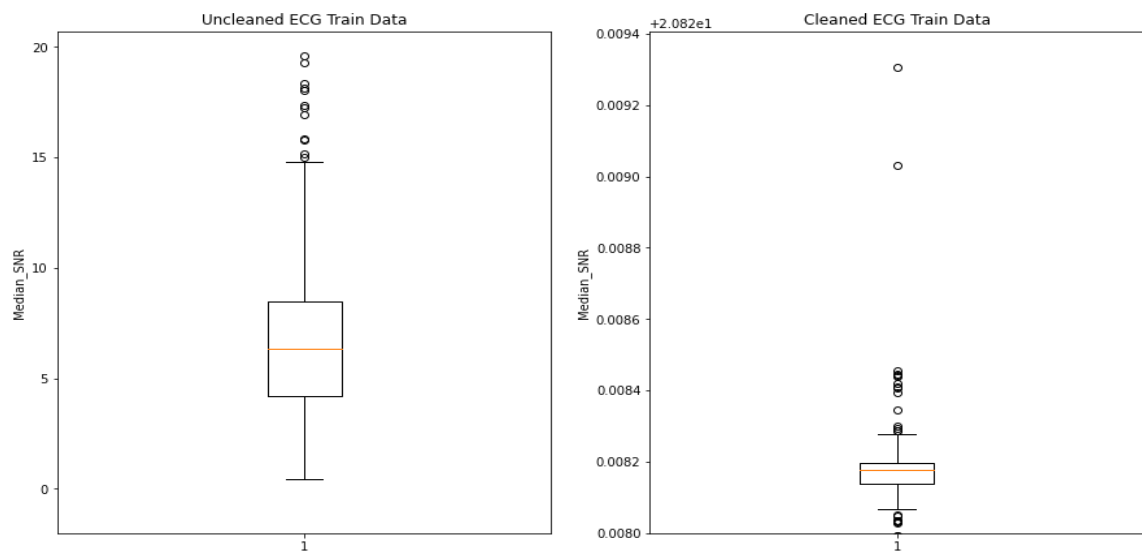


Figure 5-2: Median SNR between uncleaned and cleaned Train ECG Data

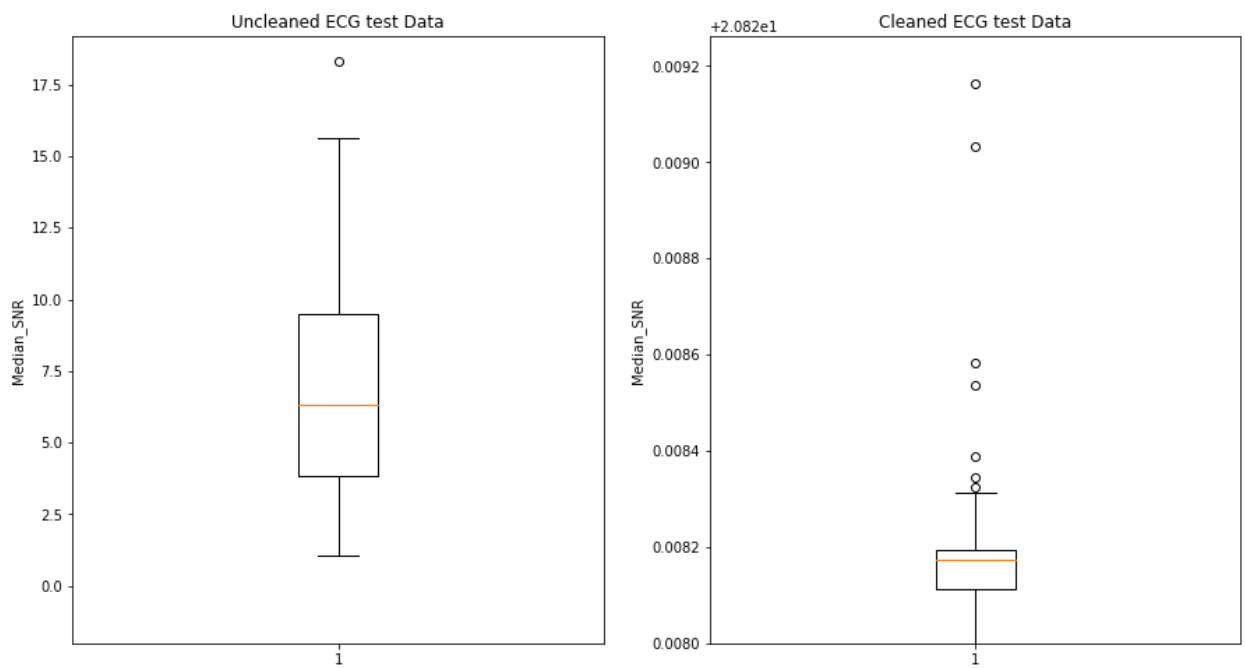


Figure 5-3: Median SNR between uncleaned and cleaned Test ECG Data

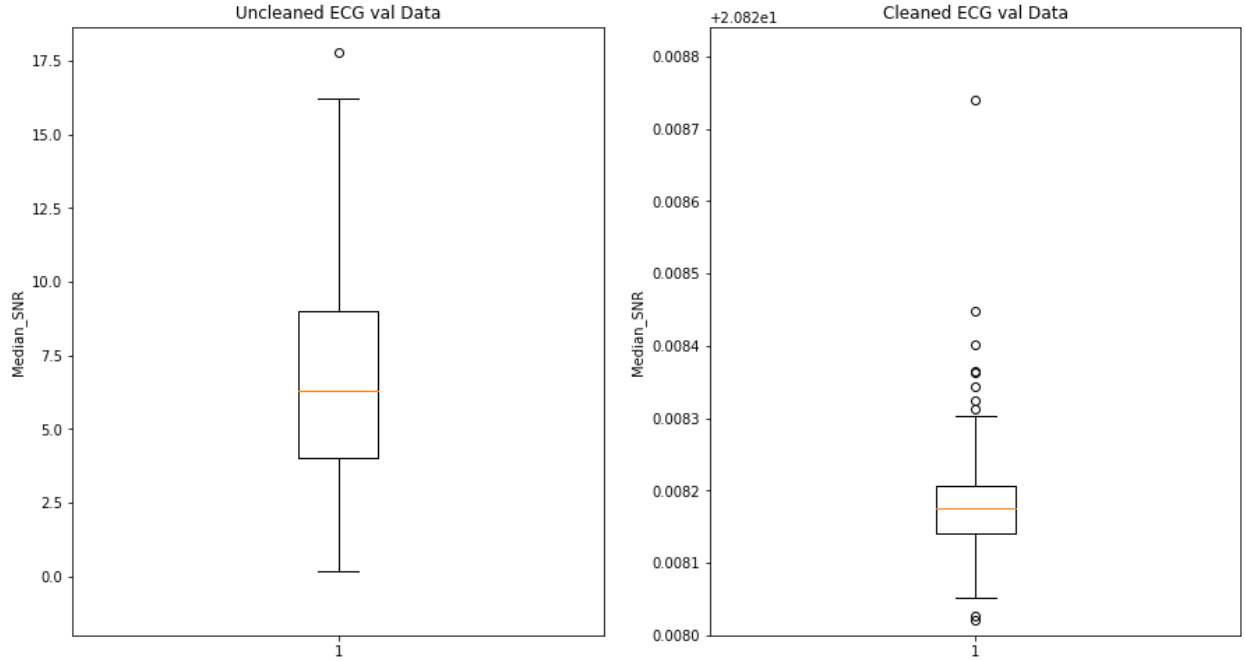


Figure 5-4: Median SNR between uncleaned and cleaned validation ECG Data

5.5 Strategies for Data Augmentation in ECG Analysis – A focus on GANs

5.5.1 GAN based synthetic data creation based on raw signals.

Data augmentation plays a pivotal role in enhancing the robustness and generalization capability of the developed model. In the realm of data augmentation, we have employed a Generative Adversarial Network (GAN) to produce synthetic ECG images, effectively boosting our dataset. The GAN framework consists of a discriminator and a generator, both designed with specific architectural configurations that were optimized through hyperparameter tuning.

For the discriminator, the architecture involves multiple Conv1D layers with varying filters—ranging from 32 to 256—and kernel sizes fixed at 16. The activation functions used are ReLU and Leaky ReLU, coupled with MaxPool1D layers for down-sampling. The model is compiled using the Adam optimizer with a learning rate of 0.0005, and it employs binary cross-entropy as the loss function. On the other hand, the generator is initiated with a latent dimension of 100 and is composed of Dense layers and bidirectional LSTM layers, followed by a series of Conv1D layers. Activation functions like Leaky ReLU are again utilized to introduce non-linearity.

The GAN was trained using a two-step iterative process. First, the discriminator was trained on both real and fake ECG samples—real samples being directly fetched from our ECG dataset and fake samples being generated by the generator. Second, the generator was trained to fool the discriminator, aiming to improve its capability to produce synthetic ECG signals that are indistinguishable from real ones. This training is facilitated through a custom function that considers parameters like the number of epochs, batch size, and evaluation steps. The performance of the generator and the discriminator is evaluated at regular intervals by checking their classification accuracy on both real and synthetic samples.

Synthetic data generation was executed for each lead from 0 to 11 in both Healthy Control (HC) and Myocardial Infarction (MI) categories, which allowed us a broad spectrum of data for analysis. The synthetic ECG signals

were further subjected to dimensionality reduction techniques like PCA and t-SNE for validation. Once validated, these synthetic records were saved as CSV files and used for the subsequent stages of model building.

5.5.2 GAN based synthetic data creation based on Inter beat intervals and metadata.

The methodology incorporates a Generative Adversarial Network (GAN) to produce synthetic data samples aimed at balancing a skewed binary classification dataset. The architecture comprises two distinct generators for the positive and negative classes, and a single discriminator that validates the quality of synthetic samples.

The generators for both the positive and negative classes are initialized with identical configurations. Each comprises three layers with a specified number of neurons and uses the ReLU activation function. The discriminator is similarly structured but uses the sigmoid activation function in its output layer to enable binary classification.

A grid search mechanism is employed to identify the optimal set of hyperparameters, including learning rate, dropout rate, and batch size. These are chosen based on their efficacy in minimizing the discriminator's loss function.

The generators and discriminators undergo simultaneous training. For each iteration, the generators create synthetic samples for both classes. The discriminator then evaluates these against real samples. Loss functions for both the generator and discriminator are calculated, and model weights are updated through backpropagation.

The generators are trained to produce synthetic samples that maintain a predetermined prevalence ratio between the positive and negative classes. This is crucial to ensure that the model generalizes well to real-world scenarios, where class distribution is seldom balanced.

After training the GAN, batches of synthetic samples, along with their class labels, are produced by the generators. These are then integrated into the original dataset, thereby augmenting it.

The model's training progress is gauged through key evaluation metrics including loss and accuracy for both real and synthetic samples. This helps to determine the point of convergence and can inform decisions about potential model refinements.

Evaluation of Synthetic Samples with PCA and t-SNE: The quality of the synthetic samples generated by the GAN is rigorously evaluated using Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE) techniques. Both methods are applied for dimensionality reduction and are implemented using machine learning libraries. Here's how the evaluation unfolds:

PCA: Initially, a PCA model is configured to reduce the data to two principal components. This model is first fitted to the synthetic data and then used to transform the original validation set. The transformed synthetic and original instances are subsequently visualized on a 2D scatter plot. These plots help in differentiating between the two classes—Healthy Control and Myocardial Infarction—and in assessing how well the synthetic instances approximate the original ones.

t-SNE: A t-SNE model is configured with two components and a perplexity setting. The synthetic and original data are independently subjected to t-SNE. Again, the results are visualized using 2D scatter plots. The plots

are configured to differentiate between Healthy Control and Myocardial Infarction categories. This serves to visually inspect the quality of synthetic samples in the context of the original ones.

In summary, 2D scatter plots of both PCA and t-SNE transformed data are used to facilitate a visual inspection of class distributions among the synthetic and original instances. This methodology serves as an indirect metric to assess the performance of the GAN in generating synthetic samples that closely mimic the original data.

An elaborate discussion on the GAN methodology is discussed in Chapter 6

5.6 Methods

This study employs a structured approach to the analysis of Electrocardiogram (ECG) signals for the detection of various cardiac conditions. The process begins with the acquisition of raw ECG data, which is divided into distinct sets for training and testing purposes. This segregation is crucial to ensure the robustness and generalizability of the machine learning models developed later in the study. The initial step in the workflow involves data preprocessing, where techniques are applied to reduce signal noise— an essential process that enhances the quality and clarity of the ECG readings, making subsequent analyses more precise and reliable.

Following preprocessing, the study emphasizes the extraction of heart rate variability (HRV) features from the ECG signals. HRV measures are indicative of autonomic nervous system dynamics and provide valuable insights into cardiac health. The extracted features are then integrated with multimodal data, including clinical factors, to form a comprehensive feature set that feeds into the machine learning models. This integration is pivotal as it leverages both physiological and clinical data, enhancing the predictive capability of the models. To further augment the dataset and improve model robustness against overfitting and under-representation of certain conditions, Generative Adversarial Networks (GANs) are used to generate synthetic samples. These samples enrich the training data, providing a diverse array of scenarios for the model to learn from. The final stages of the study involve training and evaluating two types of classifiers: a binary classifier that discerns between normal and abnormal conditions, and a multilabel classifier that identifies specific conditions such as myocardial infarction, bundle branch block, and dysrhythmia. Figure 5-5 illustrates the study methodology

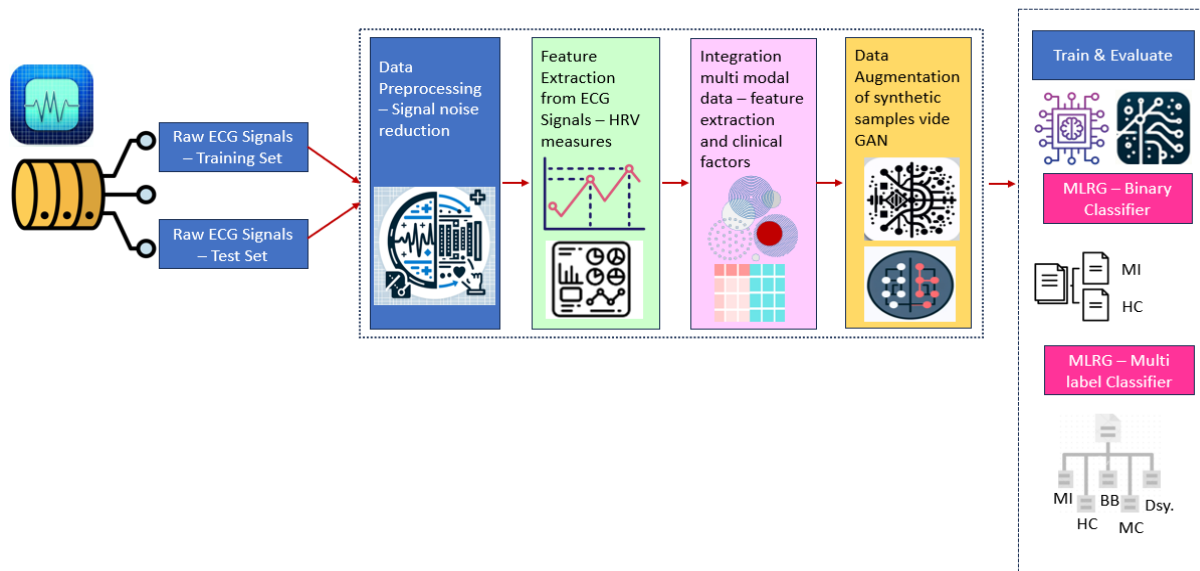


Figure 5-5: Study Methodology

5.6.1 CNN for ECG raw signals classification

The Convolutional Neural Network (CNN) designed for binary classification of raw ECG signals comprises multiple layers tailored to extract relevant features from normalized ECG data. Input data, represented as two-dimensional arrays capturing both time sequence and ECG feature channels, is fed into the network. The initial convolutional block incorporates 32 filters with an 8-sized kernel, applying L1 and L2 regularization and employing Parametric Rectified Linear Unit (PReLU) activation followed by batch normalization. A skip connection, facilitating gradient flow and mitigating vanishing gradient issues, is established between this layer and a subsequent convolutional layer.

In the second convolutional block, 30 filters with a 6-sized kernel are employed, using Leaky ReLU activation and batch normalization. Like the first block, a skip connection is introduced to preserve essential features. Max-pooling with a pool size of 2 and a stride of 1 is applied, followed by a dropout layer with a rate of 0.3 for regularization. The flattened output from the convolutional blocks is directed to fully connected layers, with a dense layer of 10 units utilizing a 'tanh' activation function, followed by a final dense layer with 2 units and SoftMax activation for binary classification.

In the context of the model's architecture, the skip connection serves a specific purpose. Skip connections, or shortcut connections, are used to allow the gradient to back-propagate to earlier layers without much hindrance or vanishing. Essentially, they create an alternate pathway for the gradient to flow through during back-propagation.

Mathematically, a skip connection performs element-wise addition of the output from one layer with the output from a subsequent layer. In our specific architecture, we have:

$$Y_{skip} = X * W + B + X \quad \text{Equation 5-4}$$

Here, Y_{skip} is the output after the skip connection, X is the input to the layer, $*$ denotes convolution, and W and B are the weight matrix and bias for this layer, respectively.

By doing this, the skip connection helps to combat the vanishing gradient problem, enabling the model to learn from the activation maps of previous layers. This results in a more robust feature hierarchy in the deep network and often leads to better performance and faster convergence.[174]–[178]

In the specific context of this study, the skip connection ensures that even during deep convolution operations, the essential features from the raw ECG signals are not lost, thereby making the model efficient in identifying myocardial infarction and healthy heart conditions.

To optimize model training, Stochastic Gradient Descent (SGD) with a learning rate of 0.002 is employed, considering its effective convergence. Cost-sensitive training is implemented to address class imbalance, where class weights are computed inversely proportional to class frequencies and incorporated into the training process. Early stopping, monitored via validation loss with a patience of 10 epochs, prevents overfitting by

restoring the best weights when validation loss increases. The model is compiled using Binary Cross-Entropy loss with logits and monitored with key performance metrics such as AUC, Precision, Recall, and Binary Accuracy during training.

Training is conducted with a batch size of 64 for a maximum of 100 epochs, ensuring sequence integrity of time-series data by avoiding shuffling. Validation against a separate dataset, along with the application of class weights and early stopping, contributes to model robustness and generalization. Through these integrated elements, the CNN architecture aims to effectively capture both low-level and high-level features in raw ECG signals for accurate binary classification. By integrating these elements, the CNN architecture aims to capture both low-level and high-level features in the raw ECG signals for accurate binary classification. Figure 5-5 shows the architecture of CNN employed for CNN raw signals binary classification.

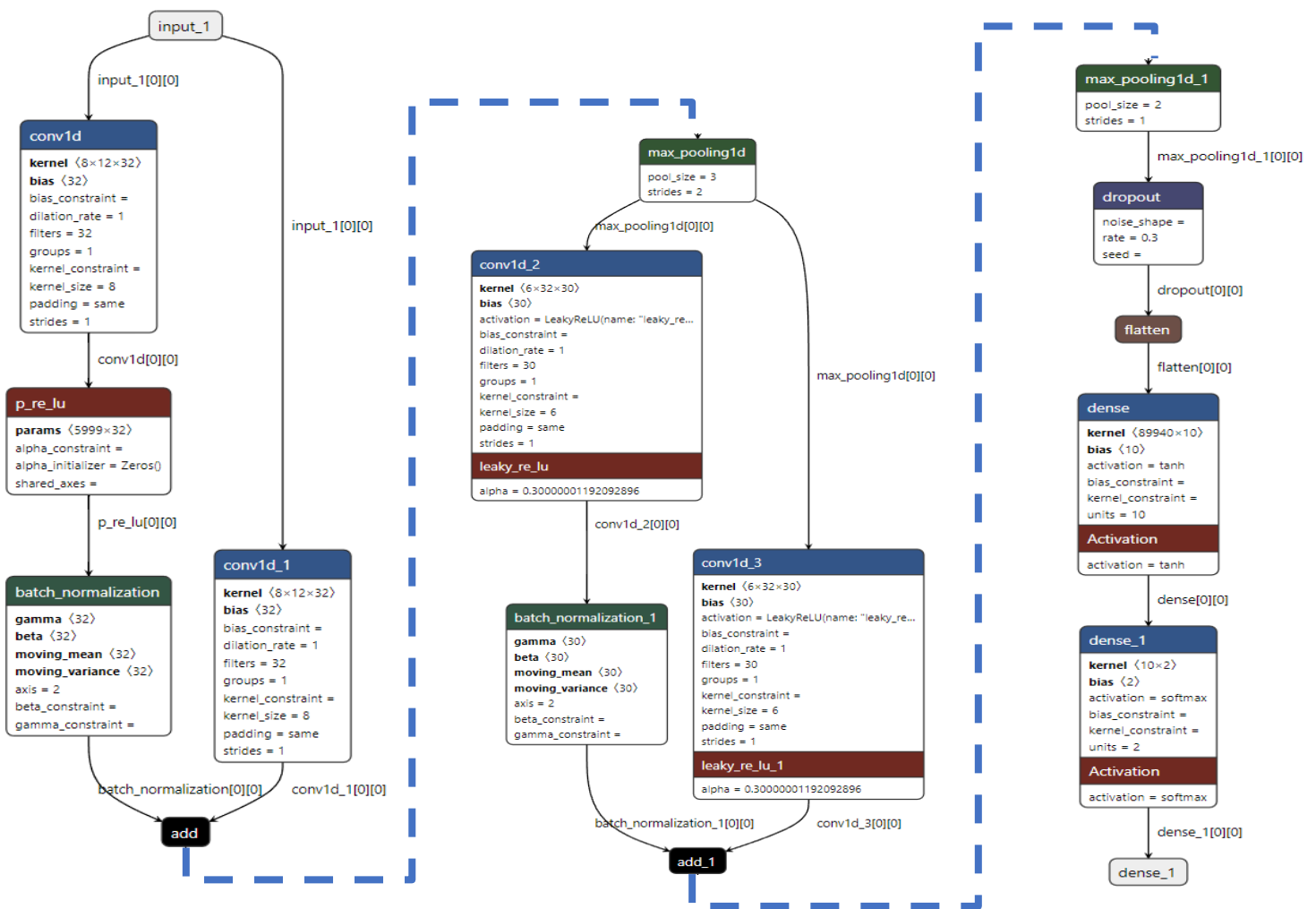


Figure 5-6: CNN Architecture for ECG Signals Binary Classification

5.6.2 Inter Beat intervals (IBI) feature extraction.

In the domain of ECG signal analysis, we employ an array of variables to develop a comprehensive understanding of cardiac health. These variables can be categorized into Time Measures, Linear Measures, and Frequency Measures. Time Measures, derived from time-domain analysis, include the Inter-beat Intervals (IBI), which is the time span between consecutive R-peaks, and the P-R Interval, representing the time between the onset of the P-wave and the QRS complex. Other variables in this category include QRS Duration, Q-T Interval, Heart Rate measured in beats per minute (BPM), and statistical metrics like standard deviation of normal-to-normal heart beat intervals (SDNN), Root Mean Square of Successive Differences (RMSSD), Percentage of NN50 intervals (pNN50), Percentage of NN20 intervals (pNN20), and Heart Rate Mean Absolute Deviation (HR_mad), which collectively offer nuanced insights into heart rate variability (HRV).

Linear Measures focus on amplitude-based features of the ECG signal. This category includes the amplitude of the P-wave, QRS complex, and T-wave, as well as metrics related to the ST-segment and T-wave Duration. These amplitude-based variables are crucial for identifying anomalies in the electrical patterns of the heart.

Frequency Measures are extracted from frequency-domain analysis and include variables such as Total Power, which encapsulates the overall power of the ECG signal. Sub-divisions of this power into Very Low Frequency (VLF), Low Frequency (LF), and High Frequency (HF) components offer additional perspectives on heart function. A key variable here is the LF/HF Ratio, which gives a comparative measure of LF and HF powers and is indicative of the balance between sympathetic and parasympathetic activity.

Together, these variables form a robust framework for the in-depth analysis of ECG signals, allowing us to understand the complex electrical activities of the heart and their implications for health.[179]–[182]

The methodology employed for inter-beat interval (IBI) feature extraction from raw ECG signals involved several sequential steps applied to each lead, resulting in the consolidation of features across all 12 leads. Initially, the raw ECG data was imported from the CSV file, followed by noise and artifact removal to obtain a clean signal. Subsequently, R-peaks, indicative of each heartbeat, were detected utilizing the Pan-Tompkins algorithm from the Hearty PY package. This algorithm identifies peaks in the signal, thereby pinpointing the occurrence of each heartbeat. The RR intervals, representing the time durations between successive R-peaks, were then calculated.

From the RR intervals, a plethora of heart rate variability (HRV) features were derived. Time-domain features such as heart rate, mean NN interval, standard deviation of NN intervals, nn50, pnn50, RMSSD, and SDSD were computed, providing insights into the variability of heartbeats. Additionally, nonlinear features like SD1, SD2, and the SD1/SD2 ratio were extracted to capture the complexity of HRV. Furthermore, frequency-domain features were obtained using the Welch method, encompassing low-frequency (LF), high-frequency (HF), and total power spectral density (PSD) within various frequency bands, along with the LF/HF ratio.

This comprehensive feature extraction process was iteratively applied to all 12 leads, ensuring the integration of time-domain, nonlinear, and frequency-domain features across all observations. The resulting dataset, consolidated into a single CSV file, encapsulated the extracted features for further analysis and modelling.

5.6.2.1 Time- Domain HRV features [182]

Following time-domain HRV (Heart Rate Variability) features and its mathematical representation as per table 5-2

Feature	Description	Mathematical representation
Heart Rate (HR)	Heart rate in beats per minute (bpm)	$HR = \frac{60}{mean_nni}$ Equation 5-5
Mean NN interval (mean_nni)	Average duration between consecutive heartbeats	$mean_nni = \frac{sum\ of\ all\ NN\ intervals}{total\ number\ of\ NN\ intervals}$ Equation 5-6
Standard Deviation of NN Intervals (SDNN)	Measure of variability in NN intervals	$SDNN = \sqrt{\frac{\sum (NN_i - mean_nni)^2}{N-1}}$ Equation 5-7
Number of NN50 (NN50)	Count of NN intervals that differ by more than 50 milliseconds	NN50
Percentage of NN50 (pNN50)	Percentage of NN intervals that differ by more than 50ms	$pNN50 = \frac{NN50}{N} \times 100$ Equation 5-8
Number of NN20 (NN20)	Count of NN intervals that differ by more than 20 milliseconds	NN20
Percentage of NN20 (pNN20)	Percentage of NN intervals that differ by more than 20ms	pNN20 (like equation 5-8)
Root Mean Square of Successive Differences (RMSSD)	Measure of short-term HRV	$RMSSD = \sqrt{\frac{\sum (NN_{i+1} - NN_i)^2}{N-1}}$ Equation 5-9
Standard Deviation of Successive Differences (SDSD)	Measure of short-term variability	$SDSD = \sqrt{\frac{\sum (SD_i - means_SD)^2}{N-1}}$ Equation 5-10

Table 5-2: Time Domain HRV features

5.6.2.2 Non-linear HRV features

In the context of Heart Rate Variability (HRV) analysis, SD1 and SD2 are derived from the Poincare plot, which is a graphical representation of the correlation between successive RR intervals. The plot is a scatter plot where each point (RR_n, RR_{n+1}) represents two successive RR intervals. In this plot, SD1 and SD2 represent the standard deviation of the points perpendicular and along the line of identity, respectively. Table 5-3 presents the nonlinear HRV features and its mathematical representation.

Feature	Description	Mathematical representation
SD1	It quantifies the dispersion of points perpendicular to the line of identity in the Poincare plot. SD1 reflects short-term variability in the RR intervals and is mainly influenced by high-frequency variations. Indicates short-term variability, and it is generally considered to reflect the parasympathetic activity.	$SD1 = \sqrt{\frac{1}{N} \sum_{i=1}^N (RR_{i+1} + RR_i - \overline{\Delta RR})^2}$ Equation 5-11 Where $\overline{\Delta RR}$ is the mean of the differences between adjacent RR intervals.
SD2	It quantifies the dispersion of points along the line of identity in the Poincare plot. SD2 represents long-term variability and reflects both sympathetic and parasympathetic influences and indicates long-term variability	$SD2 = \sqrt{\frac{1}{N} \sum_{i=1}^N ((RR_{i+1} + RR_i)/2 - \overline{RR})^2}$ Equation 5-12 Where $\overline{RR}$ is the mean of all the RR intervals.
SD1/SD2 Ratio	It is used as a measure to understand the balance between short-term and long-term HRV components.	$\frac{SD1}{SD2}$

Table 5-3: Non-linear HRV features

Overall, these metrics offer valuable insights into the autonomic nervous system's control over heart rate, often used in medical diagnostics and research.

5.6.2.3 Frequency-domain HRV features

Frequency domain measures are an important aspect of Heart Rate Variability (HRV) analysis. These measures provide insights into how the frequency of the RR intervals varies, which can be indicative of different kinds of physiological phenomena. Frequency domain features are calculated using the Fast Fourier Transform (FFT) method. In this study Welch's method is used, which is a variation of FFT, where the signal is divided into overlapping segments to calculate the PSD (Power Spectral Density). Frequency measures at its mathematical representation is show in Table 5-4.

Feature	Description	Mathematical representation
Power Spectral Density	This includes the power spectral densities for VLF (Very Low Frequency), LF (Low Frequency), and HF (High Frequency) bands	$Total\ Power\ (PSD_{Total}) = VLF\ Power + LF\ Power + HF\ Power$ Equation 5-13
Absolute Power	The power spectral density (PSD) within each frequency band (VLF, LF, and HF). Variables used in power domain measures like vlf_fft_abs, lf_fft_abs, and hf_fft_abs hold these values.	$PSD_{VLF} = \int_{0.0033}^{0.04} PSD(f)df$ Equation 5-14
		$PSD_{LF} = \int_{0.04}^{0.15} PSD(f)df$ Equation 5-15
		$PSD_{HF} = \int_{0.15}^{0.4} PSD(f)df$ Equation 5-16
Relative Power	The ratio of the absolute power of a frequency band to the total power. Variables like vlf_fft_rel, lf_fft_rel, and hf_fft_rel hold these values	$Relative\ Power_{LF} = \frac{PSD_{LF}}{PSD_{Total}}$ Equation 5-17
LF/HF Ratio	The ratio of LF power to HF power, sometimes used as an indicator of the balance between sympathetic and parasympathetic activity.	$LF/HF\ Ratio = \frac{PSD_{LF}}{PSD_{HF}}$ Equation 5-18

Table 5-4: Frequency measures - HRV

The frequency where the highest power is observed within individual frequency bands is referred as peak frequency. In interpreting frequency measures of HRV, the VLF band, physiological interpretation is ambiguous and considered to be associated with long-term regulatory mechanisms like the renin-angiotensin system. Regarding the LF band it is commonly linked with sympathetic modulation, yet it also reflects both sympathetic and parasympathetic activity. HF band reflects parasympathetic (vagal) activity and synchronizes with the respiratory cycle, a phenomenon recognized as respiratory sinus arrhythmia.[183]–[185]

5.6.2.4 Peak Detection – Mathematical Background

Peak detection in this study for extracting HRV from raw ecg signals utilizes Pan-Tompkins's algorithm consisting of several sequential steps[186], [187]. Firstly, the raw ECG signal denoted as $x[n]$ undergoes bandpass filtering to isolate the frequency band where R-peaks are prominent. This filtering process typically involves a combination of high pass denoted as $h_{hp}[n]$ and low pass denoted as $h_{lp}[n]$ filters. It is represented mathematically as,

$$y_{bp}[n] = (x[n] * h_{hp}) * h_{lp}[n] \quad \text{Equation 5-19}$$

Subsequently, the bandpass-filtered signals now denoted as $y_{bp}[n]$ are differentiated to accentuate high-frequency components to calculate the slope information. It is mathematically, represented as,

$$y_{diff}[n] = y_{bp}[n] - y_{bp}[n - 1] \quad \text{Equation 5-20}$$

Following differentiation, the signals are squared $y_{square}[n] = (y_{diff}[n])^2$ Equation 5-21 , to further emphasize the high frequency components. Finally, an integration process involves applying an averaging window to obtain integrated signal $y_{int}[n]$ facilitating peak detection. Mathematically, it is represented as,

$$y_{int}[n] = \frac{1}{N} \sum_{k=0}^{N-1} y_{square}[n - k] \quad \text{Equation 5-22}$$

Thresholding is then applied to $y_{int}[n]$ to detect R-Peaks with a dynamic thresholding technique commonly employed for this purpose.

5.6.3 Binary Classification using IBI features and patient metadata.

Three machine learning algorithms – Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB) – were employed for binary classification tasks. LR models the log odds of event occurrence based on independent features, RF combines decision trees to output class modes, and XGB iteratively improves weak learners to form a robust predictor. Chosen for their effectiveness with imbalanced datasets and interpretability, each algorithm underwent hyperparameter tuning for optimal performance.

To maintain uniformity across feature scales, scaling and normalization techniques are applied to the dataset. Scaling adjusts the range of features to a consistent scale, ensuring that no single feature dominates due to its larger magnitude. Normalization, on the other hand, centres the data around zero and scales it based on its standard deviation, ensuring that features are comparable and aiding in faster convergence during model training.

Furthermore, to address multicollinearity, correlation-based feature elimination is employed to identify features with correlation coefficients exceeding 0.8, mitigating the risk of unstable model estimates. Following this, the Variance Inflation Factor (VIF) is calculated as,

$$VIF_j = \frac{1}{1 - R_j^2} \quad \text{Equation 5-23}$$

evaluates the degree of multicollinearity in the remaining features. Features with a VIF greater than 5 are deemed to exhibit severe multicollinearity and are consequently removed from the dataset.

The Synthetic Minority Over-Sampling Technique (SMOTE) was employed to address the class imbalance inherent in the dataset. This algorithm generates synthetic samples in the feature space. By interpolating between existing instances of the minority class, SMOTE enhances the quantity of minority samples without introducing excessive variance, thereby mitigating the risk of model overfitting.

The logistic regression model underwent hyperparameter tuning via randomized search, considering penalty type, regularization strength (C), solver type, intercept fitting, maximum iterations, and warm start option. Evaluation metrics including ROC curve, confusion matrix, precision, recall, F1 scores, and accuracy were utilized to assess model performance on both training and test datasets. Additionally, 10-fold nested cross-validation ensured robustness and generalizability. Feature importance was visualized via a horizontal bar chart, and coefficients were extracted for interpretability, providing insights into the model's bias and directionality of features.

The hyperparameter tuning process for the Random Forest Classifier model employed random search to explore the parameter space efficiently. Various hyperparameters, including the number of estimators, maximum depth, minimum samples split, minimum samples leaf, maximum features, bootstrap, and criterion, were considered in the random search, with ranges specified for each parameter. After the search, optimal parameters were determined, yielding a mean test score of approximately 0.982. A max depth of 10 indicated that the trees are deep enough to capture complex relationships in the data and balancing for overfitting. The number of estimators set at 484 suggests that the model complexity was adequate for capturing underlying data patterns but without introducing significant computational inefficiency.

The hyperparameter tuning for the XGBoost model utilized RandomizedSearchCV to optimize parameters for achieving the highest ROC-AUC score. Parameters such as Alpha, Lambda, Learning Rate, Max Depth, Min Child Weight, N Estimators, Subsample Ratio, ColSample by Tree, and Gamma were considered in the search. After the process, optimal parameters were determined, resulting in a mean test score of approximately 0.988. The model trained with these parameters demonstrated robustness and accuracy, validating the effectiveness of the RandomizedSearchCV approach for similar predictive analytics tasks. The model, once trained with the optimal hyperparameters, yielded an overall test set score that affirmed its robustness and accuracy. **Error! Reference source not found.** lists various models used in binary classification.

Table 5-5: List of Models for Binary Classification

Model	Parameters/Model Description
1	CNN - Raw ECG signals of 12 Leads
2	CNN - Raw ECG signals of 12 Leads GAN Augmentation
3	LR - IBI features drawn from ECG Signals
4	RF - IBI features drawn from ECG Signals
5	XGB - IBI features drawn from ECG Signals
6	LR - IBI features drawn from ECG Signals along with patient metadata
7	RF - IBI features drawn from ECG Signals along with patient metadata
8	XGB - IBI features drawn from ECG Signals along with patient metadata
9	XGB - IBI features drawn from ECG Signals along with patient metadata - GAN - 433 Samples of synthetic data
10	RF - IBI features drawn from ECG Signals along with patient metadata - GAN - 433 Samples of synthetic data
11	XGB_RF_Ensemble - IBI features drawn from ECG Signals along with patient metadata - GAN - 433 Samples of synthetic data_Ensemble
12	XGB - IBI features drawn from ECG Signals along with patient metadata - GAN - 2503 Samples of synthetic data_Balanced
13	RF - IBI features drawn from ECG Signals along with patient metadata - GAN - 2503 Samples of synthetic data_Balanced
14	XGB_RF_Ensemble - IBI features drawn from ECG Signals along with patient metadata - 2503 Samples of synthetic data_Balanced

5.6.4 Multilabel Classification using IBI features and patient metadata.

For multilabel classification, a gradient boosting model was chosen due to its high accuracy and ability to handle both high-dimensional and imbalanced data sets. The objective was to predict multiple medical conditions (e.g., Bundle Branch Block, Cardiomyopathy, Dysrhythmia, Healthy Control, Myocardial Infarction) based on a set of clinical features. The Scikit-Learn library in Python was used to implement the algorithm.

In our investigation into the multilabel classification of various medical conditions, we opted to use the XGBoost algorithm exclusively. This decision was informed as in experiments involving binary classification tasks, XGBoost outperformed both Logistic Regression (LR) and Random Forest (RF) in terms of accuracy and other relevant performance metrics. Given its superior performance in binary classification, we deemed XGBoost to be the most suitable algorithm for extending our study to a multilabel classification framework. Experiments were conducted both with and without the incorporation of synthetic data to examine the model's performance under different conditions. For the generation of synthetic data, we utilized various Generative Adversarial Network (GAN) architectures, including Simple GAN, Wasserstein GAN (WGAN), and Conditional GAN. These techniques were employed to diversify our training dataset, improve model generalizability, and address potential class imbalances.

The model was fine-tuned using Randomized Search CV to obtain optimal hyperparameters. The hyperparameters explored were learning rate, maximum depth of the trees, number of trees, and regularization terms (alpha and lambda). The optimal hyperparameters were identified to be as follows, Learning rate: 0.187, Maximum depth: 5, Number of trees: 64, Alpha: 8.32, Lambda: 6.99

Optimal threshold selection was conducted to maximize the F1-score. Precision-Recall curves were generated, and the F1-score was calculated at each point on these curves. The threshold that maximized the F1-score was then identified and applied to the test data. Alternatively, Youden's index was also used for optimal threshold selection. The index was calculated by maximizing the difference between the true positive rate and the false positive rate. This method prioritizes both sensitivity and specificity.

After thresholding, multilabel confusion matrices, sensitivity, specificity, and AUC-ROC curves were used to assess the performance of the model. Both training and test datasets were evaluated, and a comprehensive classification report was generated. **Error! Reference source not found.**6 lists various models used in multilabel classification

Table 5-6: List of Models for Multilabel Classification

Model	Parameters/Model Description
1	XGB_Multiclass_IBI Features_unbalanced
2	XGB_Multiclass_IBI features+Meta_unbalanced
3	XGB_Multiclass_IBI features_SMOTE
4	XGB_Multiclass_IBI features+Meta_SMOTE
5	XGB_Multiclass_IBI features+Meta+Gan_BCE_Agmntn
6	XGB_Multiclass_IBI features+Meta+Gan_CLF_Agmntn
7	XGB_Multiclass_IBI features+Meta+WGAN_WL_Agmntn
8	XGB_Multiclass_IBI features+Meta+WGAN_CLF_Agmntn

5.6.5 Optimal threshold for classification

The optimal threshold for classification was determined using two key methodologies: the Youden Index and the F1 Score. The Youden Index, commonly referred to as Youden's J statistic, serves as a unified metric designed to quantify the performance of a binary classification model by striking a balance between sensitivity (true positive rate) and specificity (true negative rate). Mathematically, the Youden Index is formulated as

$$J = \text{Sensitivity} + \text{Specificity} - 1 \quad \text{Equation 5-24}$$

Optimization of the Youden Index provides the optimal decision threshold that maximizes this balance, thereby offering an insightful measure of a model's discriminatory power. A higher value of the Youden Index is indicative of superior model performance in terms of its ability to differentiate between classes effectively[188]–[190].

The F1 Score is an evaluative metric that amalgamates precision and recall providing a comprehensive view of a binary classification model's accuracy. Formulated as the harmonic mean of precision and recall, the F1 Score is particularly useful in contexts involving imbalanced class distributions where conventional accuracy measures might be misleading. The score ranges from 0 to 1, with 1 denoting ideal precision and recall, and 0 indicating the worst-case scenario.

Both the Youden Index and F1 Score are threshold-dependent metrics tailored for binary classification tasks. However, they differ fundamentally in the facets of classification performance they emphasize. While the Youden Index is concerned with the concurrent optimization of true positive and true negative rates, the F1 Score specifically focuses on the harmonic balance between precision and recall.

5.6.5.1 Thresholding concepts – Mathematical background

The Youden Index, J, is derived from the concepts of sensitivity (true positive rate, TPR) and specificity (true negative rate, TNR). In terms of basic components from the confusion matrix, sensitivity and specificity are defined as follows:

$$\text{Sensitivity (TPR)} = \frac{TP}{TP+FN} \quad \text{Equation 5-25}$$

$$\text{Specificity (TNR)} = \frac{TN}{TN+FP} \quad \text{Equation 5-26}$$

The Youden Index is then formulated as:

$$J = TPR + TNR - 1 = \frac{TP}{TP+FN} + \frac{TN}{TN+FP} - 1 \quad \text{Equation 5-27}$$

Optimizing J means finding the threshold that provides the best balance between sensitivity and specificity, making it a suitable metric for assessing the discriminatory ability of a model.

The F1 Score is rooted in the notions of precision (P) and recall (R), which are themselves formulated based on components of the confusion matrix:

$$\text{Precision (P)} = \frac{TP}{TP+FP} \quad \text{Equation 5-28}$$

$$\text{Recall (R)} = \frac{TP}{TP+FN} \quad \text{Equation 5-29}$$

(Recall is synonymous with Sensitivity)

The F1 Score is the harmonic mean of precision and recall and is defined as:

$$F1\ Score = \frac{2 \cdot P \cdot R}{P + R} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad \text{Equation 5-30}$$

The F1 Score is especially useful when you have uneven class distributions, as it takes both false positives and false negatives into account, providing a balanced measure of a model's performance.

The selection of an appropriate performance metric is contingent upon the unique attributes of the problem under investigation as well as the specific objectives that the analysis aims to fulfil [191]– [193].

5.6.6 Model Explanation – Shapley Additive method

SHapley Additive exPlanations (SHAP) provide a unified measure of feature importance, attributing the change in prediction for a particular data point to each feature. Especially in complex ensemble models like XGBoost, interpretability can be a challenge due to the black-box nature of the algorithm. SHAP values, grounded in cooperative game theory, break down the prediction to show the impact of each feature. They provide local interpretability—how each feature contributes to each individual prediction—as well as global interpretability—how each feature contributes overall across the dataset. Utilizing SHAP values with XGBoost thus facilitates a fine-grained understanding of model decision-making, making it invaluable for scientific investigations and clinical applications. This kind of interpretability is paramount, particularly in fields requiring a high degree of trust and understanding, such as healthcare.

At its core, SHAP is grounded in cooperative game theory, specifically using Shapley values to fairly distribute "payouts" or contributions among features. For a prediction model f that takes an input $x = \{x_1, x_2, \dots, x_n\}$ composed of n features, the Shapley value ϕ_i of the i th feature is computed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad \text{Equation 5-31}$$

Here, N is the set of all features, and S is a subset of N that does not include feature i . The term $f(S \cup \{i\})$ is the prediction with feature i added to subset S , and $f(S)$ is the prediction without feature i .

The Shapley value ϕ_i quantifies the average marginal contribution of feature i over all possible combinations of features. By distributing the contributions fairly among features, SHAP provides an equitable representation of feature importance, which is especially critical for complex models like XGBoost.

XGBoost models, being ensemble models based on decision trees, offer a natural setting for applying SHAP as they inherently consist of additive components. SHAP uses Tree Explainers to make the calculations efficient for tree-based models like XGBoost, exploiting the properties of decision trees to calculate Shapley values in polynomial time.

This rigorous mathematical foundation makes SHAP particularly robust and interpretable, making it an excellent tool for scientific research and clinical decision-making[191]–[193].

5.7 Validation methods

For both binary and multilabel classification tasks, the Confusion Matrix was employed to gain an intuitive yet granular understanding of the model's performance. This approach allowed us to quantify true positives, true negatives, false positives, and false negatives across all classes. Complementary to this, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was also evaluated. The ROC curve plots the true positive rate against the false positive rate, while the AUC metric quantifies the overall ability of the model to discriminate between the classes. The curves and AUC values were calculated for both training and test datasets, providing a robust measurement of model reliability and generalization capability.

In addition to confusion matrices and AUC-ROC curves, a detailed classification report was generated using Scikit-Learn's classification report function. This report offered an in-depth look at metrics such as precision, recall, and the F1-score for each class. These metrics are particularly beneficial for evaluating model performance in situations with imbalanced classes or when certain types of misclassifications are more costly than others.

ensure the robustness of our model, we employed a Nested Cross-Validation strategy with 10-folds for the outer loop. Nested Cross-Validation mitigates the bias introduced due to the overlap between the training and validation set by providing an unbiased assessment of how well the model is likely to perform on unseen data. The average and standard deviation of the outer CV scores were calculated, offering insights into both the consistency and reliability of the chosen models, which included Logistic Regression (LR), Random Forest (RF), Convolutional Neural Networks (CNN), and XGBoost, for binary classification. A similar process was also followed for multilabel classification, confirming the robustness of the chosen approaches.

5.8 Results

This section provides a comprehensive analysis of the performance of both binary and multilabel classification models developed to identify adverse cardiovascular events through Electrocardiogram (ECG) signals. The results are systematically presented, starting with an overview of the overall effectiveness of the models, followed by detailed evaluations segmented by model type and classification strategy. Each model's performance is scrutinized using a variety of metrics including precision, recall, F1 score, overall accuracy, and AUC (Area Under the Curve), enabling a thorough assessment of their diagnostic capabilities.

Binary Classification Results: The results for binary classification models are presented in Tables 5-7 and 5-8, illustrating the comparison across different models ranging from traditional machine learning approaches to those incorporating advanced data augmentation techniques. The evaluation highlights significant findings on the effectiveness of each model, with particular attention to their ability to distinguish between HC and MI conditions.

Multilabel Classification Results: Following the binary classification analysis, the multilabel classification results are detailed. Tables 5-9 to 5-12 summarize the outcomes of models designed to categorize multiple cardiac

conditions (HC, MI, BB, CM and Dsy) simultaneously. This section emphasizes the complexity of multilabel classification and the role of innovative techniques such as SMOTE and GANs in enhancing model performance.

Results of binary classification are populated in below table 5-7 and 5-8

Table 5-7: Comparison of Results Model 1 to 7 – Binary Classification

Model	1		2		3		4		5		6		7	
Class	HC	MI	HC	MI	HC	MI	HC	MI	HC	MI	HC	MI	HC	MI
Precision	0.17	0.84	0.00	0.82	0.26	0.88	0.38	0.90	0.50	0.92	0.36	0.87	0.73	0.91
Recall	0.33	0.68	0.00	0.92	0.53	0.69	0.53	0.83	0.60	0.88	0.33	0.88	0.53	0.96
F1 Score	0.23	0.75	0.00	0.87	0.35	0.78	0.44	0.86	0.55	0.90	0.34	0.87	0.62	0.94
Overall Accuracy	0.62		0.83		0.67		0.78		0.83		0.79		0.89	
Overall AUC	0.45		0.50		0.70		0.79		0.79		0.73		0.93	
Macro Avg Precision	0.50		0.41		0.57		0.64		0.71		0.61		0.82	
Macro Avg Recall	0.51		0.46		0.61		0.68		0.74		0.61		0.75	
Macro Avg F1 Score	0.49		0.43		0.56		0.65		0.72		0.61		0.78	
Wgted Avg Precision	0.73		0.68		0.78		0.81		0.85		0.78		0.88	
Wgted Avg Recall	0.62		0.77		0.67		0.78		0.83		0.79		0.89	
Wgted Avg F1 Score	0.66		0.72		0.70		0.79		0.84		0.79		0.88	

Table 5-8: Comparison of Results Model 8 to 14 – Binary Classification

Model	8		9		10		11		12		13		14	
Class	HC	MI	HC	MI	HC	MI	HC	MI	HC	MI	HC	MI	HC	MI
Precision	0.69	0.92	0.70	0.90	1.00	0.86	0.86	0.89	0.70	0.99	1.00	0.89	0.80	0.91
Recall	0.60	0.95	0.47	0.96	0.20	1.00	0.40	0.99	0.93	0.92	0.40	1.00	0.53	0.97
F1 Score	0.64	0.93	0.56	0.93	0.33	0.92	0.55	0.94	0.80	0.95	0.57	0.94	0.64	0.94
Overall Accuracy	0.89		0.88		0.87		0.89		0.90		0.90		0.90	
Overall AUC	0.93		0.96		0.92		0.95		0.96		0.87		0.94	
Macro Avg Precision	0.81		0.80		0.93		0.88		0.85		0.95		0.86	
Macro Avg Recall	0.77		0.72		0.60		0.70		0.93		0.70		0.75	
Macro Avg F1 Score	0.79		0.75		0.63		0.74		0.88		0.76		0.79	
Wgted Avg Precision	0.88		0.87		0.88		0.89		0.94		0.91		0.89	
Wgted Avg Recall	0.89		0.88		0.87		0.89		0.92		0.90		0.90	
Wgted Avg F1 Score	0.89		0.87		0.83		0.87		0.93		0.88		0.89	

Upon examining the tables 5-5, 5-7 and 5-8 depicting performance metrics across various models, following is a summarized comparison of their effectiveness. The models are evaluated based on precision, recall, F1 score, overall accuracy, AUC (Area Under the Curve), and weighted averages of these metrics for two classes, HC (Healthy Control) and MI (Myocardial Infarction).

In the table 5-7, models 1 to 7 are depicted with a noticeable improvement in Model 7 (RF model without data augmentation), which achieves the highest overall accuracy and AUC at 0.89 and 0.93, respectively. This suggests that Model 7 is highly effective at correctly classifying both classes and differentiating between the conditions without augmenting synthetic samples.

In the table 5-8, presenting models 8 to 14, there's a general trend of high performance in models augmented with synthetic data, with models 9 and 12 standing out. Model 9 exhibits exceptional precision for the MI class

(0.99) and high overall accuracy (0.96). Model 12, however, not only has high precision (0.99 for MI) but also has the highest AUC (0.96), which indicates its superior ability in discriminating between the Healthy Control and Myocardial Infarction classes.

The weighted averages across these models reflect the balance between classes in terms of the prevalence of each class within the dataset. Model 7 again performs admirably with the highest weighted average F1 score in the first table (0.89). Model 12 follows suit in the second table, with the highest weighted average F1 score of 0.93, which suggests a balanced performance across classes, making it an ideal model considering both the sensitivity and precision of the test outcomes.

Collectively, these results demonstrate the progression in model optimization, with later models (7, 9, and particularly 12) showing enhanced predictive capabilities, marked by high AUC and F1 scores, indicating robust performance in classifying myocardial infarction from healthy controls in the dataset. More analysis on the models is reported in section 5.8.1

Error! Reference source not found.6 lists various models used in multilabel classification and results of the same are populated in below tables 5-9 to 5-12:

Table 5-6 : Comparison of Results Model 1 to 2 – Multilabel Classification

Model	1					2				
Parameters/Model Description	XGB_Multiclass_IBI Features_unbalanced					XGB_Multiclass_IBI features+Meta_unbalanced				
Class	BB	CM	Dsy	HC	MI	BB	CM	Dsy	HC	MI
Precision	0.01	0.02	0.02	0.16	0.76	0.01	0.03	0.02	0.16	0.76
Recall (Sensitivity)	0.33	0.67	0.67	1.00	1.00	0.33	1.00	0.67	1.00	1.00
F1 Score	0.02	0.04	0.05	0.27	0.86	0.02	0.06	0.05	0.27	0.86
AUC	0.22	0.29	0.48	0.72	0.69	0.25	0.28	0.47	0.85	0.80
Specificity	0.13	0.07	0.17	0.04	-	0.11	0.02	0.15	0.05	-
Overall Accuracy	0.36					0.36				
Macro Avg Precision	0.19					0.20				
Macro Avg Recall	0.73					0.80				
Macro Avg F1 Score	0.25					0.25				
Wgted Avg Precision	0.60					0.60				
Wgted Avg Recall	0.96					0.97				
Wgted Avg F1 Score	0.70					0.70				

Table 5-7 : Comparison of Results Model 3 to 4 – Multilabel Classification

Model	3					4				
Parameters/Model Description	XGB_Multiclass_IBI features_SMOTE					XGB_Multiclass_IBI features+Meta_SMOTE				
Class	BB	CM	Dsy	HC	MI	BB	CM	Dsy	HC	MI
Precision	0.01	0.03	0.03	0.15	0.76	0.03	0.03	0.03	0.16	0.76
Recall (Sensitivity)	0.33	1.00	0.67	1.00	1.00	0.67	1.00	0.67	1.00	1.00
F1 Score	0.03	0.07	0.06	0.27	0.86	0.05	0.07	0.05	0.27	0.86
AUC	0.25	0.42	0.50	0.67	0.54	0.26	0.43	0.41	0.77	0.82
Specificity	0.29	0.14	0.40	0.02	-	0.25	0.13	0.24	0.04	-
Overall Accuracy	0.36					0.53				
Macro Avg Precision	0.20					0.20				
Macro Avg Recall	0.80					0.87				
Macro Avg F1 Score	0.26					0.26				
Wgted Avg Precision	0.60					0.60				
Wgted Avg Recall	0.97					0.98				
Wgted Avg F1 Score	0.70					0.70				

Table 5-8 : Comparison of Results Model 5 to 6 – Multilabel Classification

Model	5					6				
Parameters/Model Description	XGB_Multiclass_IBI features+Meta+Gan_BCE_Agmntn					XGB_Multiclass_IBI features+Meta+Gan_CLF_Agmntn				
Class	BB	CM	Dsy	HC	MI	BB	CM	Dsy	HC	MI
Precision	0.02	0.02	0.02	0.15	0.76	0.01	0.03	0.01	0.15	0.76
Recall (Sensitivity)	0.67	0.67	0.67	1.00	1.00	0.33	1.00	0.33	1.00	1.00
F1 Score	0.04	0.04	0.05	0.26	0.86	0.02	0.06	0.02	0.27	0.86
AUC	0.51	0.24	0.52	0.82	0.81	0.17	0.52	0.32	0.87	0.82
Specificity	0.09	0.10	0.15	-	-	0.11	-	0.19	0.01	-
Overall Accuracy	0.72					0.75				
Macro Avg Precision	0.19					0.19				
Macro Avg Recall	0.80					0.73				
Macro Avg F1 Score	0.25					0.25				
Wgted Avg Precision	0.60					0.60				
Wgted Avg Recall	0.97					0.96				
Wgted Avg F1 Score	0.69					0.70				

Table 5-9 : on of Results Model 7 to 8 – Multilabel Classification

Model	7					8				
Parameters/Model Description	XGB_Multiclass_IBI features+Meta+WGAN_WL_Agmntn					XGB_Multiclass_IBI features+Meta+WGAN_CLF_Agmntn				
Class	BB	CM	Dsy	HC	MI	BB	CM	Dsy	HC	MI
Precision	0.02	0.03	0.01	0.15	0.76	0.02	0.03	0.04	0.15	0.76
Recall (Sensitivity)	0.67	1.00	0.33	1.00	1.00	0.67	1.00	1.00	1.00	1.00
F1 Score	0.04	0.06	0.03	0.26	0.86	0.04	0.06	0.07	0.26	0.86
AUC	0.17	0.50	0.23	0.85	0.88	0.52	0.57	0.63	0.83	0.84
Specificity	0.07	0.02	0.23	-	-	0.13	-	0.22	-	-
Overall Accuracy	0.79					0.75				
Macro Avg Precision	0.19					0.20				
Macro Avg Recall	0.80					0.93				
Macro Avg F1 Score	0.25					0.26				
Wgted Avg Precision	0.60					0.60				
Wgted Avg Recall	0.97					0.99				
Wgted Avg F1 Score	0.69					0.70				

Summarizing the performance of various models on a multilabel classification task, as shown in tables 5-9 to 5-12 we can draw the following quick insights.

Model 1 and 2 (unbalanced and with and without patient clinical data) display relatively low overall performance, with overall accuracy peaking at 0.36. The use of unbalanced data is reflected in the poor precision and recall rates across multiple classes, and AUC values also remain modest, suggesting limited discrimination capability.

Model 3 and 4 (SMOTE balanced and with and without patient clinical data) shows that the implementation of SMOTE, a technique to balance class distribution, shows an improvement in the overall accuracy, particularly for Model 4 which reaches to 0.53. This indicates a better ability to differentiate between classes than Models 1 and 2, despite overall accuracy remaining constant at 0.36.

Model 5 which incorporates GAN with Binary Cross-Entropy (BCE) for data augmentation, displays a notable enhancement in AUC for some classes, reaching up to 0.82. The specificity and overall accuracy improve significantly, indicating an improved balance in class prediction.

Model 6 with a similar setup to Model 5 but using a custom loss function shows comparable performance, with a high AUC for some classes and an overall accuracy of 0.75. This model maintains a consistent weighted average F1 score compared to previous models.

Model 7 utilizes Wasserstein GANs (WGAN) with Wasserstein loss does not show marked improvement, with some AUC values decreasing compared to Model 5 and 6. The overall accuracy sees a slight increase, but the model struggles with certain classes, as indicated by lower AUC and F1 scores.

Model 8 utilizing WGAN and custom loss function reflects a balance of high AUC values across most classes, comparable overall accuracy to Model 6, and a consistent weighted average F1 score. This suggests that WGAN with custom loss function could be beneficial for handling multilabel classification tasks by augmenting synthetic data to solve imbalance of classes.

In summary, the models incorporating GAN-based data augmentation (Models 5, 6, 7, and 8) generally demonstrate better performance metrics, especially in AUC and overall accuracy, than the initial models (Models 1 and 2) that do not utilize such techniques. Models leveraging SMOTE (Models 3 and 4) also indicate an improvement in discrimination capability, as shown by the AUC. The results imply that sophisticated data augmentation and class balancing strategies can significantly benefit multilabel classification performance. Detailed analysis on the model performance is discussed in section 5.8.2.

5.8.1 Classification Performance & Statistical Analysis - Binary Classification –Model 12.

In the evaluation of our best-performing binary classification model, the model 12 (Refer Table 5 -7) exhibited a high level of accuracy and discriminative power. The overall accuracy was 0.90, and the area under the receiver operating characteristic curve (AUC) stood at an impressive 0.96. (Figure 5-6) When examining class-specific metrics, the model achieved a precision of 0.70 for the Healthy Control (HC) class and 0.99 for the Myocardial Infarction (MI) class. The recall for both classes was comparably high, at 0.93 for HC and 0.92 for MI. These results translated into F1 scores of 0.80 and 0.95 for HC and MI, respectively. At a macro level, the model's average precision was 0.85, the average recall was 0.93, and the average F1 score was 0.88. Weighted averages of these metrics, which account for class imbalance in test data set, were also robust: the weighted average precision was 0.94, the weighted average recall was 0.92, and the weighted average F1 score was 0.93.

In nested cross-validation analysis, the average area under the receiver operating characteristic curve (ROC AUC) was approximately 0.997, with a 95% confidence interval ranging from 0.990 to 1.000. This high average AUC score indicates that the binary classification model performs well in discriminating between the two classes across different subsets of the data. The narrow confidence interval suggests that the model's performance is consistent across different folds of cross-validation, reinforcing its generalizability. Thus, the practical interpretation remains that the model demonstrates almost perfect classification capabilities with a high level of statistical confidence.

In the healthcare context, the high precision of 0.99 for Myocardial Infarction (MI) implies that when the model predicts an MI event, there is a very high likelihood that this is a true positive case, crucial for preventing false alarms and unnecessary medical interventions. The precision of 0.70 for Healthy Control (HC), while lower, still indicates a substantial ability to correctly identify non-MI events. The recall of 0.93 for HC and 0.92 for MI is particularly noteworthy. In clinical terms, a high recall for MI means that the model is effective in capturing most actual MI events, a crucial characteristic to avoid missed diagnoses. Similarly, the high recall for HC means that a significant majority of true non-MI cases are identified, minimizing the risk of wrongly subjecting patients to further, more invasive testing. The F1 score, which balances precision and recall, is 0.80 for HC and 0.95 for MI, suggesting that the model offers a harmonious blend of sensitivity and specificity for both classes.

In addition to the metrics provided, it's worth noting that the model's decision threshold has been fine-tuned to achieve optimal performance. Based on F1 thresholding, an optimal threshold of approximately 0.895 was selected. This means that any instance with a predicted probability above this threshold will be classified as a case of Myocardial Infarction (MI), while probabilities below this value will be categorized as Healthy Control (HC). This thresholding approach aims to harmonize the trade-off between precision and recall, yielding a

balanced model capable of both high sensitivity and specificity. The threshold was selected to maximize the F1 score, which offers a robust single metric that encapsulates both false positives and false negatives in the classification task. Thus, this finely tuned threshold adds an additional layer of precision to our model, enhancing its applicability in a clinical setting.

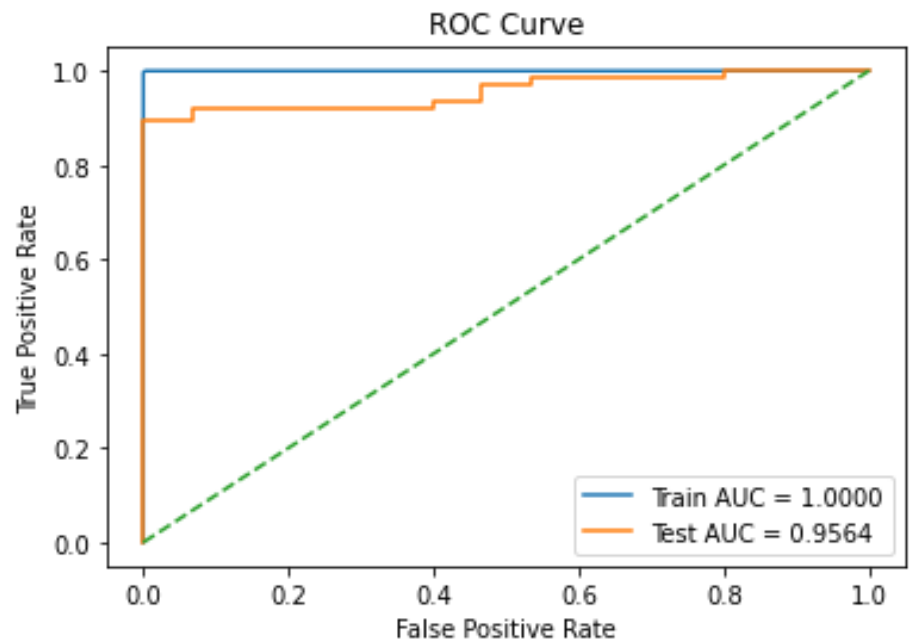


Figure 5-7 : ROC Curve of Best Model (Model 12)

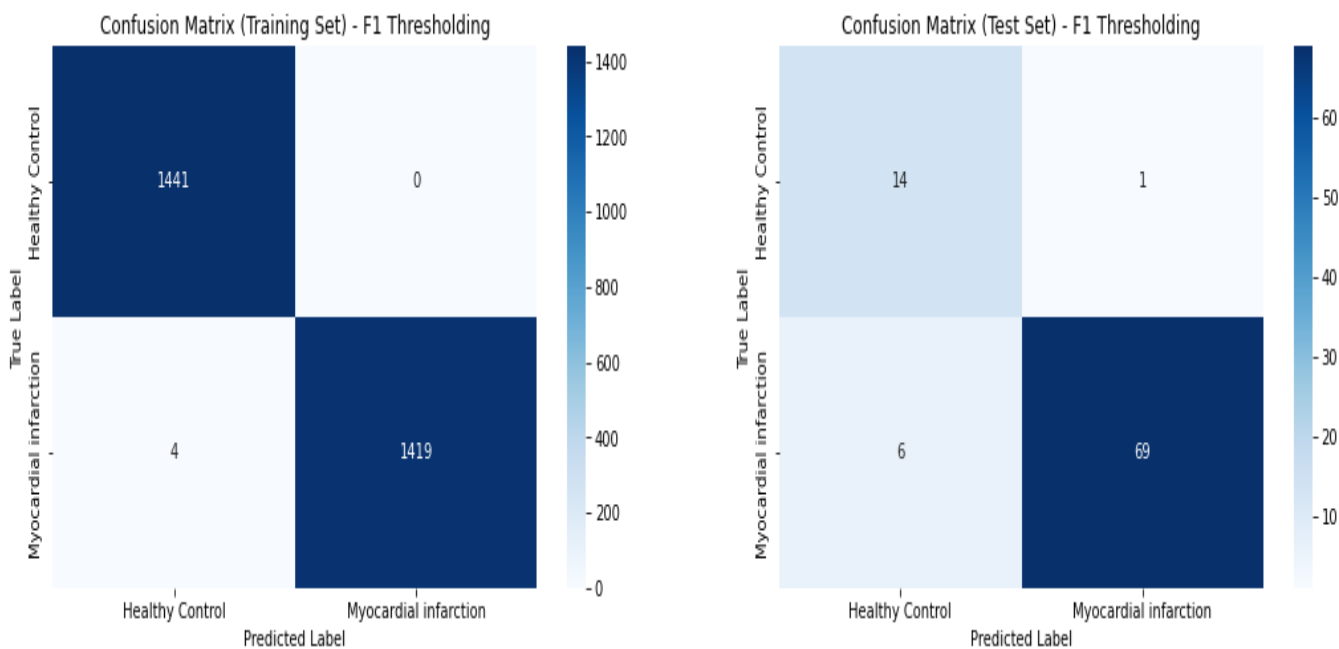


Figure 5-8 : Confusion Matrix of Best Model 12- Binary Classification

In evaluating the classification model, the Confusion Matrix (Figure 5-7) serves as an essential tool for understanding its performance on distinct classes—Healthy Control (HC) and Myocardial Infarction (MI). For the training data, the model displayed an exceptional ability to distinguish between the two classes, yielding 1441 True Positives (HC correctly classified as HC) and 1419 True Negatives (MI correctly classified as MI). Remarkably, there were only 4 False Negatives (MI incorrectly classified as HC) and zero False Positives (HC incorrectly classified as MI), showcasing both high sensitivity and specificity. On the test data, the model maintained its robustness with 14 True Positives and 69 True Negatives. However, the model produced 1 False Positive and 6 False Negatives, suggesting a slight margin of error in both Type I and Type II errors. While the training data nearly reaches an ideal Confusion Matrix, the test data exhibits minor disparities, indicating the need for further optimization. Nonetheless, the high accuracy rates observed affirm the model's clinical utility and reliability in distinguishing HC from MI.

5.8.2 Classification Performance & Statistical Analysis - Multilabel Classification Model 8

In the realm of multilabel classification, the challenges multiply due to the need for a model to predict multiple labels for each instance. Our chosen model 8 (Refer Table 5-8 and 5-12) is a Gradient Boosting algorithm trained on data augmented through a Wasserstein Generative Adversarial Network (WGAN) with a customized loss function. Despite the limitations imposed by data scarcity, the incorporation of GAN-augmented synthetic data proved superior to traditional SMOTE-based augmentation (model 3, Refer Table 5-10), enhancing the model's performance.

The present study evaluates the efficacy of various Generative Adversarial Networks (GANs) in improving the classification performance of a multi-class problem using limited training data. Four different architectures were analysed: a) a model with Wasserstein GAN (WGAN) and a custom loss function combining Wasserstein loss and binary cross-entropy (BCE), b) a WGAN with just Wasserstein loss, c) a simple GAN with a custom loss function combining Wasserstein loss and BCE, and d) a simple GAN with BCE.

The model 8 (Refer Table 5-12) shows varying levels of precision across classes. For instance, the MI class has a notably high precision of 0.76, whereas the precision for minority classes like BB, CM, and Dsy is significantly lower. Remarkably, the model shows a high recall score for minority classes—BB, CM, and Dsy—indicating that the model is sensitive to these underrepresented classes. For the majority classes, the recall is also high, standing at 1.00. Given the lower precision but higher recall for minority classes, the F1 Score somewhat balances these metrics, although it remains lower than for majority classes. The Area Under Curve (AUC) values are moderate for minority classes, ranging from 0.52 to 0.63, suggesting the model's reasonable ability to distinguish between classes. The specificity scores are relatively low for all classes, particularly for minority ones, indicating a relatively high false-positive rate. With an overall accuracy of 0.75, the model shows a balanced performance across classes, which is particularly noteworthy given the class imbalance. The Macro Average F1 Score is 0.26, and Macro Average Precision and Recall are 0.20 and 0.93, respectively. These metrics indicate that the model is relatively more sensitive (higher recall) but less precise (lower precision). The weighted average metrics are more skewed towards the majority classes, with an F1 Score of 0.70.

The high recall for minority classes is statistically significant given the limited training data for these classes. This is likely attributable to the effective data augmentation using WGAN. The low specificity and precision scores for minority classes suggest that the model might be more cautious, erring on the side of over-predicting the minority classes, which could be considered a reasonable trade-off in imbalanced scenarios.

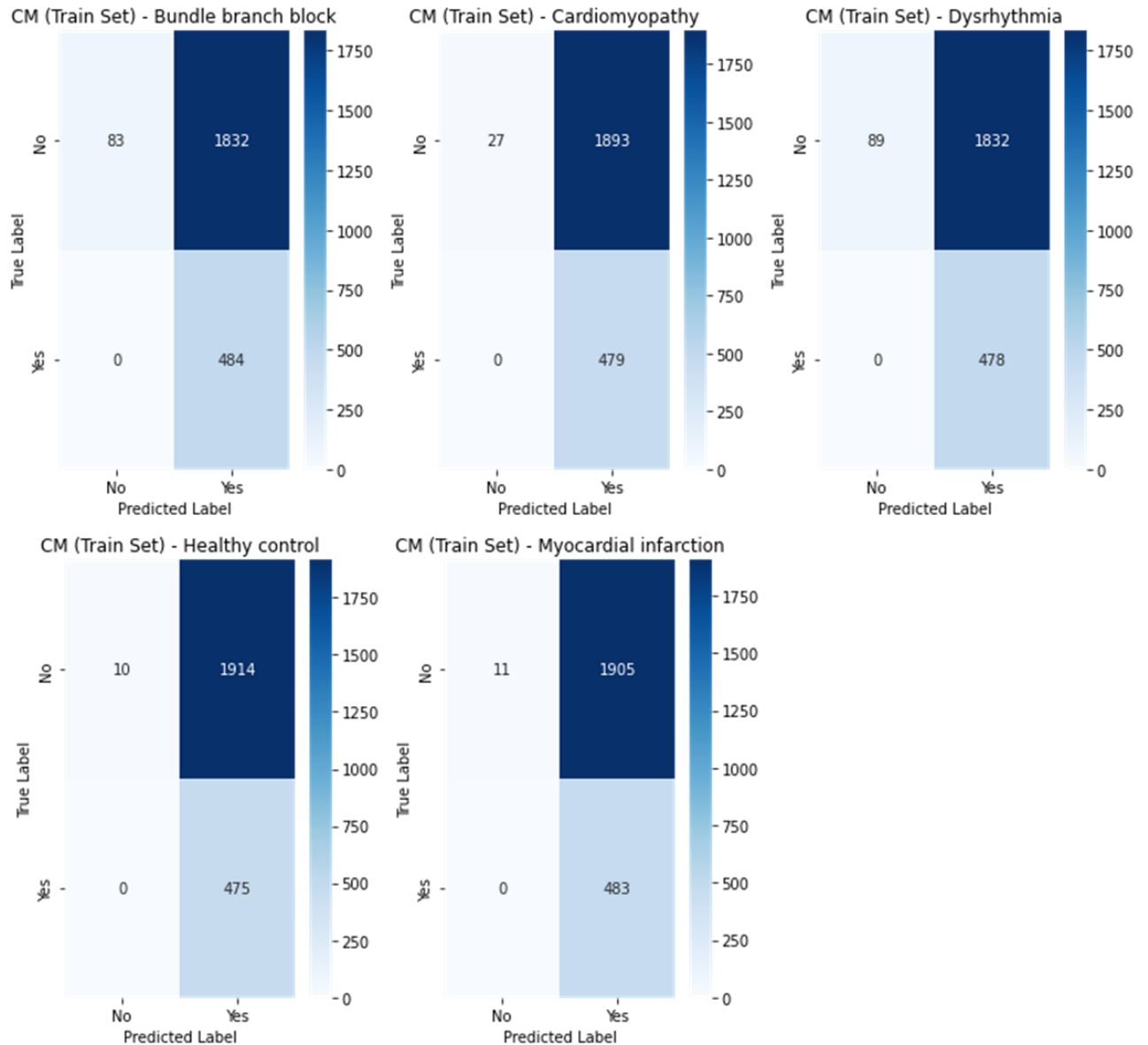


Figure 5-9 : Confusion Matrix of Best Model 8- Multilabel Classification – Train data

Interpreting the metrics derived from the training data set confusion matrix as shown in figure 5-8 reveals notable findings. Sensitivity analysis demonstrates the model's robustness in identifying positive instances across all classes, with zero false negatives and substantial true positives for each class. Specifically, the model accurately identifies 484, 479, 478, 475, and 483 true positives for Classes BB, CM, Dsy, HC, and MI, respectively. Conversely, the specificity analysis indicates a contrasting trend, with considerably fewer true negatives observed: 70 for Class BB, 17 for Class CM, 87 for Class Dsy, 3 for Class HC, and 7 for Class MI. Moreover, high counts of false positives ranging from 1845 to 1921 across classes emphasize the model's tendency to misclassify negative instances.

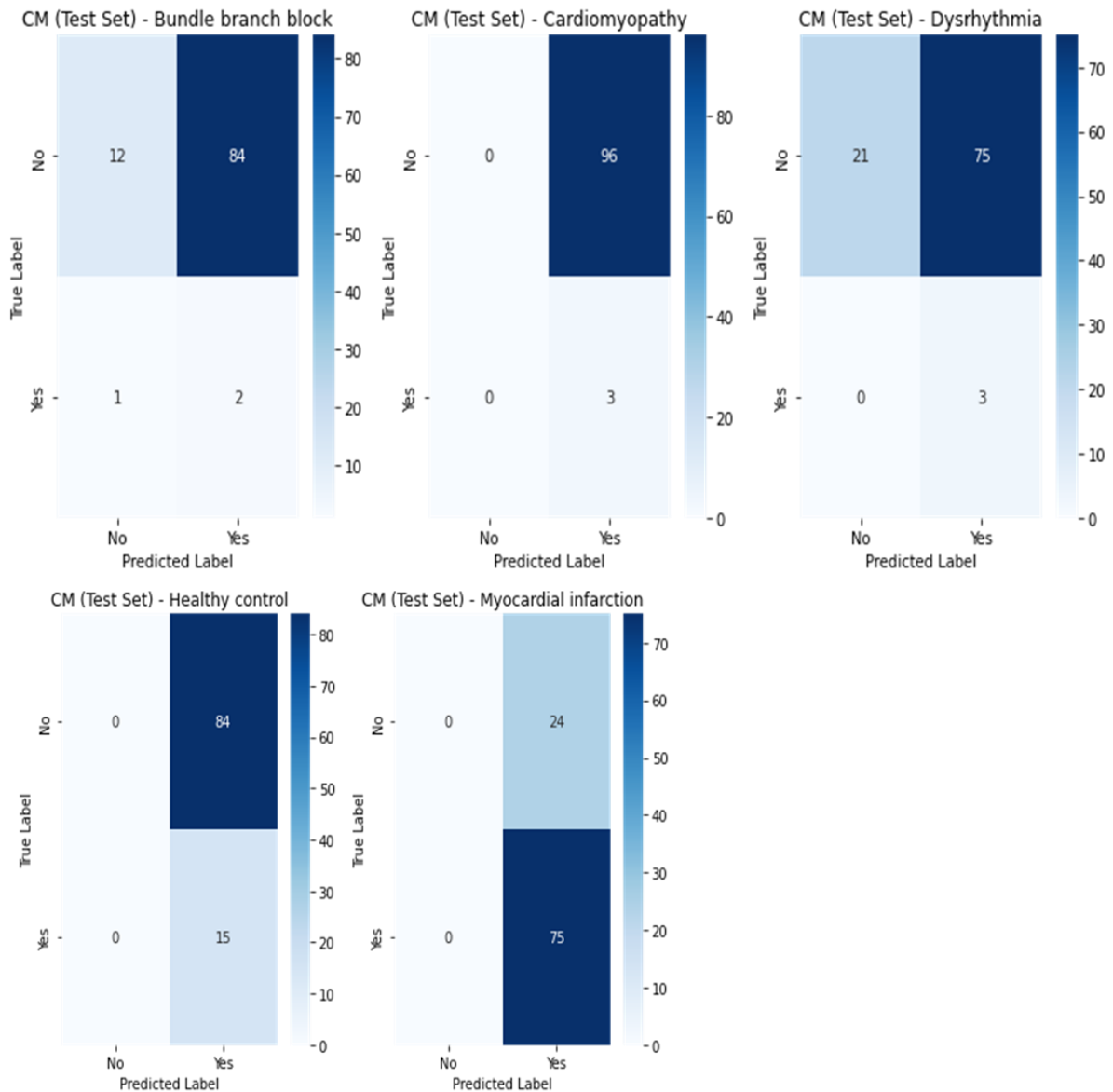


Figure 5-10 : Confusion Matrix of Best Model 8- Multilabel Classification - Test Data

Analysing the metrics derived from the test data set confusion matrix (Figure 5-9) reveals insightful observations. Sensitivity assessment demonstrates the model's adeptness, except for Class BB, where it registers only 2 true positives and a solitary false negative. Notably, Classes CM, Dsy, HC, and MI exhibit robust sensitivity with high true positive counts of 3, 3, 15, and 75, respectively, and no false negatives, indicating the model's proficiency in recognizing positive instances across diverse classes. However, specificity assessment unveils persistent challenges, particularly in Classes CM, Dsy, and HC, where true negative counts are either minimal or absent. Although Class MI shows a moderate improvement with 7 true negatives, the presence of false positives, ranging from 24 to 96 across classes, remains a notable concern, highlighting the model's ongoing struggle with misclassifying negative instances.

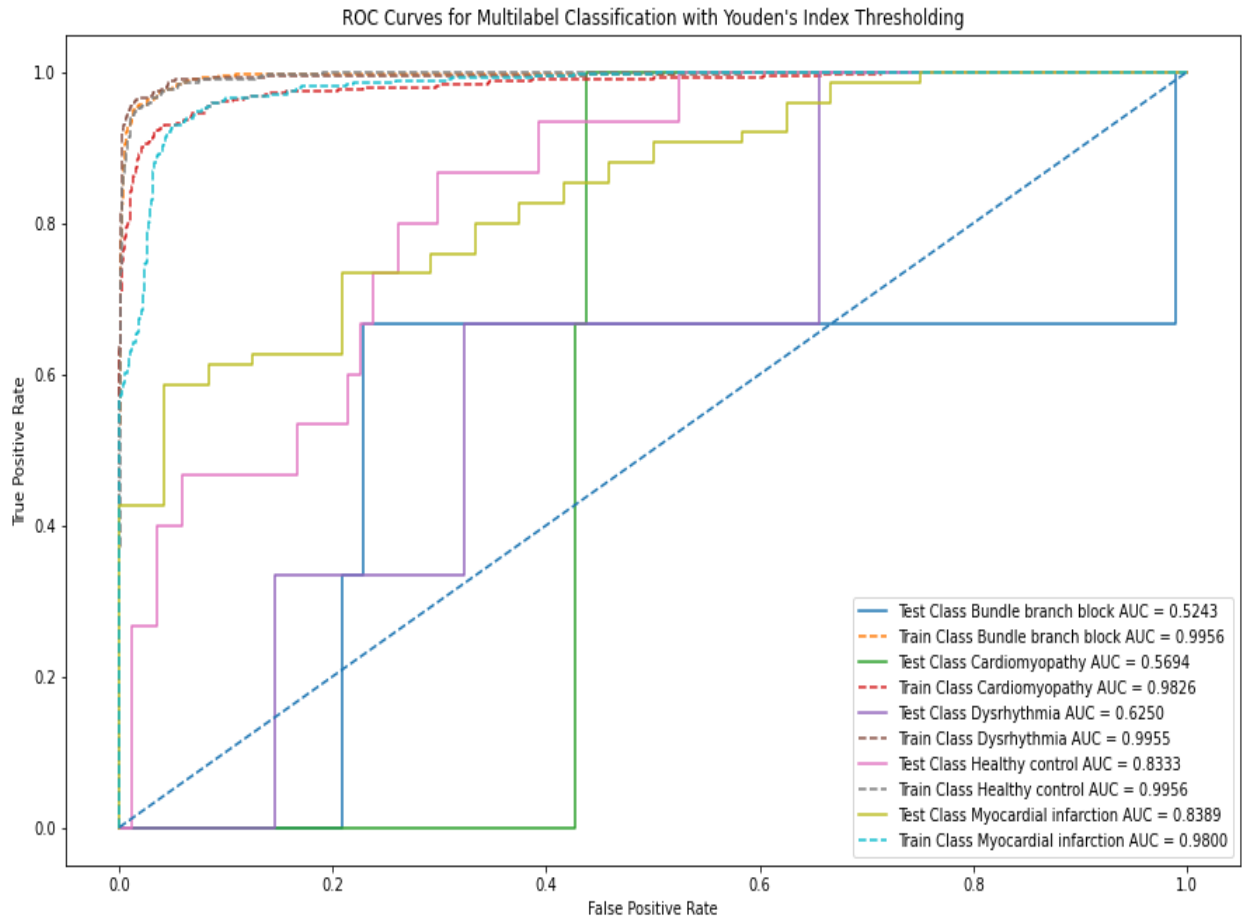


Figure 5-11 : ROC Curve of Best Model 8 - Multilabel Classification

The integration of WGAN for data augmentation notably improved the model's generalization capacity, as evidenced by the enhanced test AUC values (Figure 5-10). Despite indications of overfitting due to the disparity between training and test AUCs, the model's performance on the test set surpasses that of models without augmentation or with simpler GAN augmentation. This suggests that WGAN-generated synthetic samples possess a more realistic distribution, aiding the model in learning generalized features rather than memorizing training data. However, challenges persist in classifying the 'Bundle branch block' condition, as reflected by its relatively low-test AUC (0.5243), potentially stemming from class complexity or insufficient representative features post-augmentation. 'Cardiomyopathy' and 'Dysrhythmia' classes exhibit moderate test AUCs (0.5694 and 0.6250, respectively), showcasing performance enhancement attributed to synthetic data. Conversely, the 'Healthy control' and 'Myocardial infarction' classes demonstrate high test AUCs (0.8333 and 0.8389, respectively), indicating the model's proficiency in identifying healthy controls and accurately predicting myocardial infarction. Utilizing the Youden Index for threshold selection optimizes the model's discriminative ability by balancing sensitivity and specificity, crucial for imbalanced data and augmentation scenarios. Although ROC curves hint at overfitting, the WGAN augmentation strategy significantly bolsters test data performance, offering promise for addressing imbalanced datasets. Further model refinements may involve adjusting WGAN architecture or loss functions and exploring additional augmentation techniques to augment model generalization.

5.9 Model Explainability and Discussion

5.9.1 Model Interpretation SHAP – Binary Classification Model

Summary Plot

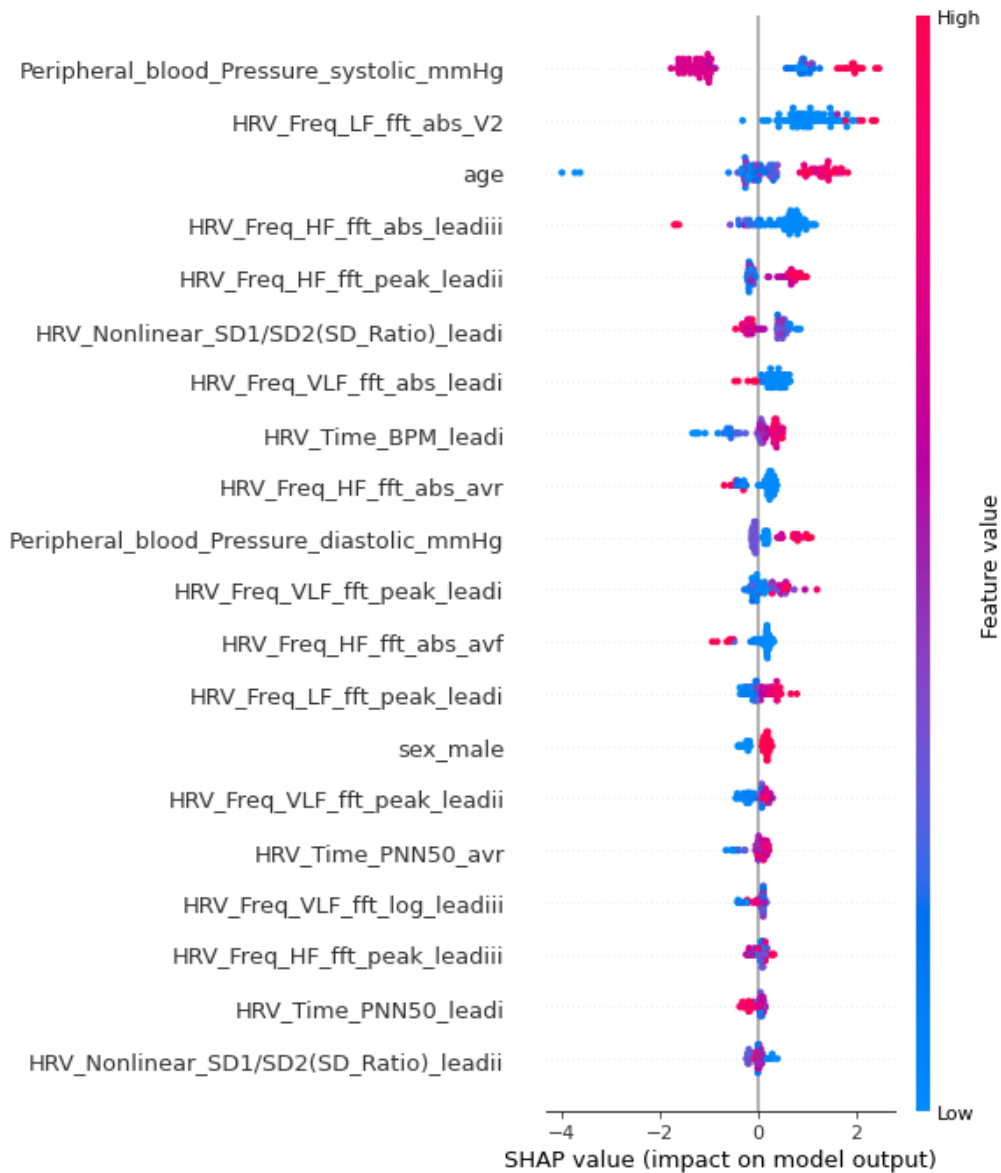


Figure 5-12 : SHAP Value Impact - Summary Plot on Best Model 12 - Binary Classification

The SHAP (SHapley Additive exPlanations) summary plot (Fig 5-11) visualizes the impact of features in an XGBoost model (model -12 – Refer table 5-7) used for predicting myocardial infarction (MI). In this plot, each dot represents a feature's SHAP value for a single instance, with the colour representing the feature's value (high or low).

From the analysis provided and the plot, it's clear that the top features influencing the model's predictions for MI risk are systolic blood pressure (SBP), the low-frequency component of heart rate variability of lead V2 (HRV Freq LF_fft abs V2), age, the high-frequency component of HRV of lead iii (HRV Freq HF_fit_abs_leadiii), and

sex - male. These features not only have the highest mean SHAP values but also exhibit a wide distribution, indicating a significant and variable impact on MI prediction across the dataset.

The clinical relevance of these findings aligns well with established medical knowledge. Systolic blood pressure is acknowledged as a critical risk factor for MI, with higher values increasing the risk. This is reflected in the plot by the positive SHAP values for SBP. Heart rate variability, represented by HRV Freq LF_fft abs V2 and HRV Freq HF_fit_abs_leadiii, is a measure of cardiac autonomic function and its association with cardiovascular diseases like MI is well documented. The SHAP plot suggests an increased risk of MI with certain values of HRV, corresponding to changes in autonomic control.

Age being a negatively correlated feature indicates that the risk of MI typically increases with age, which is consistent with the fact that older adults are more susceptible to MI. The sex feature (male) having a strong positive SHAP value suggests that being male is a risk factor for MI in the model, which aligns with clinical observations that men are at a higher risk for MI, especially at younger ages.

While the SHAP plot provides a robust framework for interpreting the XGBoost model's predictions and deriving clinical insights, it is essential to integrate these findings with the broader clinical context. For instance, a high SHAP value for SBP in a patient should be assessed alongside other risk factors, such as lifestyle, family history, and other health conditions, to evaluate MI risk comprehensively.

It is also critical to remember that the SHAP plot and the model it represents are based on statistical correlations, which may not capture the causal relationships inherent in clinical conditions fully. Therefore, while the SHAP plot is a powerful explanatory tool, its insights should be considered as part of a larger clinical assessment, and care should be taken to avoid over-reliance on model-based predictions without corroborative clinical evidence. In the next section we will analyse top four feature (predictors) distribution plots in detail from the SHAP summary plot and analyse its statistical and clinical relevance.

5.9.2 Top predictors from SHAP – feature distribution plots and interpretations

5.9.2.1 Systolic blood pressure

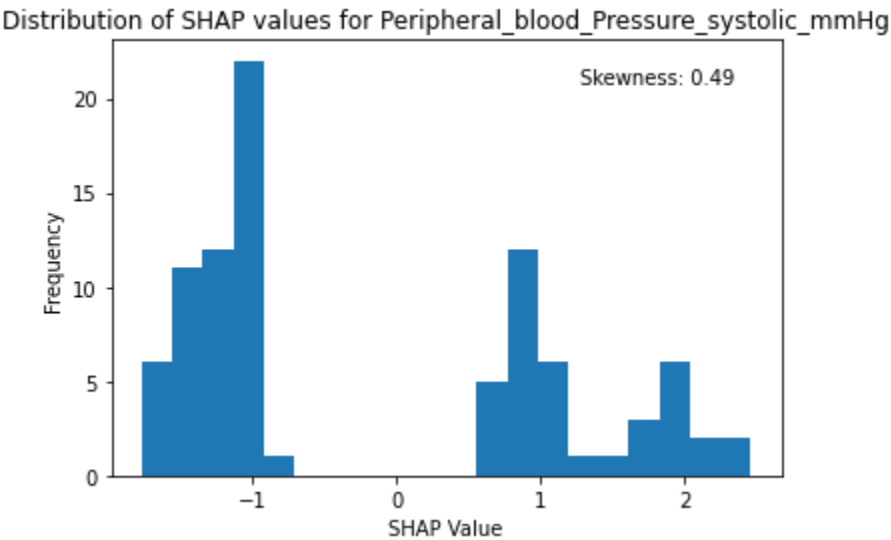


Figure 5-13 : Distribution Plot - Feature systolic BP

The skewness of the SHAP value distribution for the feature systolic blood pressure is 0.49 as shown in figure 5-12, which indicates that the distribution is moderately right skewed or right tailed. This means that there is a small number of data points with very high SHAP values, which are pulling the distribution to the right.

Statistically, the right-skewness of the SHAP value distribution for the systolic blood pressure feature suggests that the feature has a non-linear relationship with the predicted probability of MI. This means that the impact of the feature on the predicted probability of MI varies depending on the value of the feature. For example, if the feature is a measure of blood pressure, then the feature may have a greater impact on the predicted probability of MI for people with very high blood pressure than for people with moderately elevated blood pressure.

Clinically, the fact that the SHAP value distribution for the systolic blood pressure is right skewed suggests that the feature may be a useful target for intervention for a small subset of high-risk patients. For example, if a patient has a very high SHAP value for the systolic blood pressure, this may suggest that they are at high risk of MI and that they may benefit from more aggressive interventions to reduce their risk.

5.9.2.2 Frequency measure (HRV) – LF-absolute – ECG Lead V2

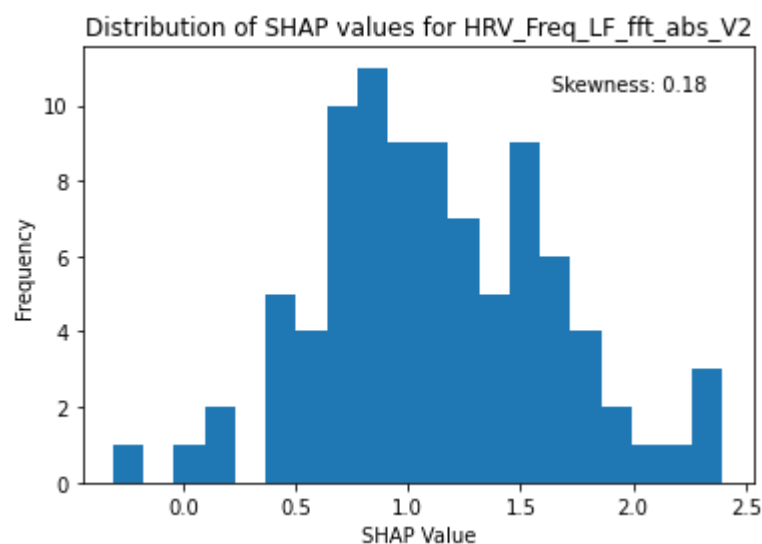


Figure 5-14 : Distribution Plot - Feature Freq Measure Low Frequency absolute - Lead V2

The SHAP value distribution for the feature "Frequency measure of low frequency absolute measure of lead v2 from ECG signal" (Figure 5-13) exhibits a slight right skew, evidenced by a skewness coefficient of 0.18 in figure 5-12. This indicates that there are outliers with higher SHAP values, pulling the distribution's tail towards the right. The relatively low skewness suggests only a modest non-linear relationship between this feature and the probability of a myocardial infarction (MI) diagnosis.

Clinically, this right-skewed pattern implies that higher values of this feature contribute more significantly towards predicting MI, albeit to a limited extent due to the low skewness coefficient. As a biomarker, this feature is related to cardiac autonomic function, a known risk factor for MI. However, its moderate impact on the prediction model advises against using it as a standalone criterion for MI risk assessment. It should be incorporated within a multi-factorial clinical evaluation for a comprehensive understanding of MI risk.

5.9.2.3 Age

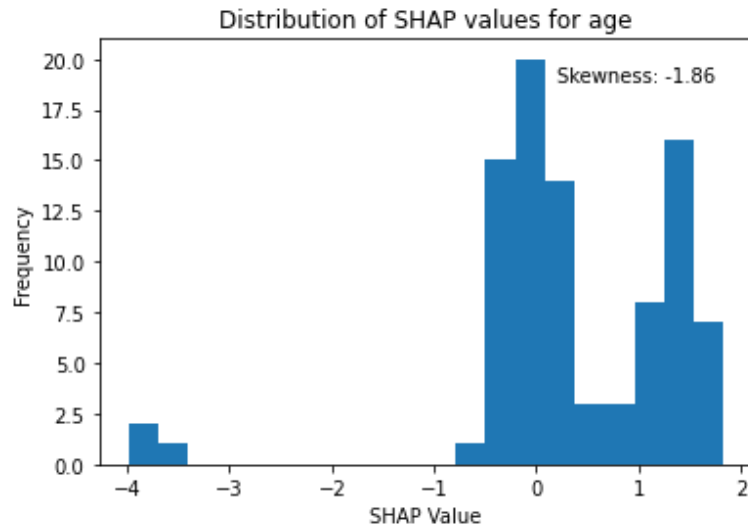


Figure 5-15 : Distribution Plot - Feature Age

Statistically, the negative skewness of -1.86 in the SHAP value distribution for the feature "age" (Figure 5-14) implies a concentration of higher values, with a tail extending towards the lower end. This suggests that, in most cases, a higher age is correlated with a greater likelihood of predicting Myocardial Infarction (MI), consistent with established clinical observations.

From a clinical standpoint, this aligns with the conventional wisdom that the risk of Myocardial Infarction increases with age. However, the distribution's long left tail points to exceptions where Myocardial Infarction occurs at a younger age. These outlier cases demand nuanced clinical attention and could be attributable to other risk factors such as hereditary predispositions, lifestyle choices, or comorbid conditions. Therefore, while age remains a significant predictor, it should not be the sole criterion for assessing MI risk, advocating for a more comprehensive diagnostic approach.

5.9.2.4 Frequency measure (HRV) – HF-absolute – ECG Lead iii

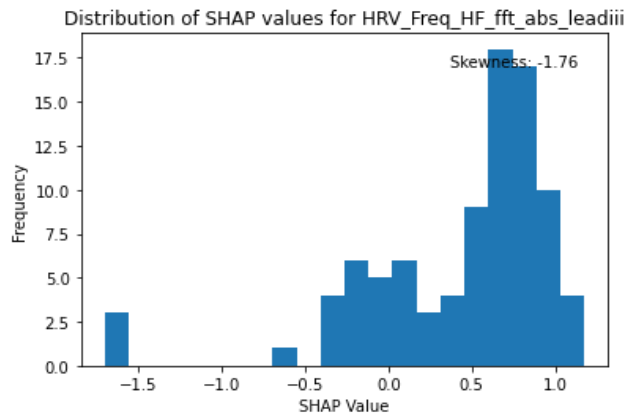


Figure 5-16 : Distribution Plot - Feature Freq Measure High Frequency absolute - Lead iii

The left-skewness of the SHAP value distribution plot (Figure 5-15) for the feature HF-absolute – ECG Lead iii (fourth feature) suggests that the feature has a non-linear relationship with the predicted probability of MI, but in the opposite direction of the first three features. This means that the impact of the feature on the predicted probability of MI decreases as the value of the feature increases.

Clinically, fact that the SHAP value distribution plot for the fourth feature is left-skewed suggests that people with lower high frequency absolute measure of lead iii are at higher risk of MI. High frequency absolute measure of lead iii is a measure of the high-frequency power of the ECG signal in lead III. High-frequency power is thought to be associated with cardiac autonomic function. Studies have shown that high-frequency power is reduced in patients with heart failure, myocardial infarction, and other cardiovascular diseases.

5.9.2.5 Additional considerations

It is important to note that the SHAP value distribution plot for the fourth feature is just one tool for interpreting the XGBoost model. It is important to consider other factors as well, such as the clinical context of the individual patient and the overall performance of the model.

For example, if a patient has a very low SHAP value for the fourth feature, this does not necessarily mean that they are at high risk of MI. Other clinical factors, such as their family history of MI and their overall cardiovascular health, also need to be considered. Similarly, it is important to remember that the SHAP value distribution plot is based on a statistical model. The model may not be perfectly accurate, and it is important to exercise caution when interpreting the results.

Overall, the SHAP value distribution plot for the fourth feature is a valuable tool for understanding the clinical factors that are associated with an increased risk of MI. However, it is important to use it in conjunction with other information, such as the clinical context of the individual patient and the overall performance of the model. Potential explanations for the non-linear relationship between high frequency absolute measure of lead iii and MI risk

There are several potential explanations for the non-linear relationship between high frequency absolute measure of lead iii and MI risk. One possibility is that the risk of MI decreases more rapidly at lower high frequency absolute measure of lead iii. This may be since people with lower high frequency absolute measure of lead iii are more likely to have other cardiovascular risk factors, such as high blood pressure, high cholesterol, and diabetes.

Another possibility is that the risk of MI plateaus at lower high frequency absolute measure of lead iii. This may be since people with lower high frequency absolute measure of lead iii are more likely to have undergone lifestyle changes to reduce their risk of MI, such as quitting smoking and exercising regularly. It is also possible that the non-linear relationship between high frequency absolute measure of lead iii and MI risk is due to a combination of these factors.

Clinically, the non-linear relationship between high frequency absolute measure of lead iii and MI risk has several implications for clinical practice. First, it suggests that people with lower high frequency absolute measure of lead iii should be screened for other cardiovascular risk factors and managed aggressively to reduce their risk of MI. Second, it suggests that people with lower high frequency absolute measure of lead iii may be

able to reduce their risk of MI by making lifestyle changes, even if they already have other cardiovascular risk factors.

Finally, the non-linear relationship between high frequency absolute measure of lead iii and MI risk suggests that the optimal approach to MI prevention may vary depending on the patient's high frequency absolute measure of lead iii. For example, people with lower high frequency absolute measure of lead iii may benefit from more aggressive interventions to reduce their risk, such as statin therapy and lifestyle counselling. People with higher high frequency absolute measure of lead iii may benefit from less aggressive interventions, such as monitoring their blood pressure and cholesterol levels.

The SHAP value distribution plot for analysis of IBI features extracted for modelling is a valuable tool for understanding the clinical factors that are associated with an increased risk of MI. However, it is important to use it in conjunction with other information, such as the clinical context of the individual patient and the overall performance of the model.

5.10 Summary of findings and Limitations

Summary of Findings

- **Model Architecture and Metrics:** Both multilabel and binary classification models were built using XGBoost. The binary classification model notably showed better performance metrics than the multilabel one.
- **Importance of Features:** The most important features for predicting myocardial infarction (MI) in the binary classification model were identified using SHAP values. These features include systolic blood pressure, heart rate variability frequency measures, age, and sex.
- **Clinical Relevance:** The SHAP summary plot unveiled features that are clinically relevant in assessing the risk for MI. The high importance of systolic blood pressure reaffirms its role as a significant risk factor, corroborating existing medical literature.
- **Non-Linear Relationships:** The skewness in the SHAP value distributions of features like systolic blood pressure and age revealed non-linear relationships with MI risk. These findings suggest that the model captures more nuanced aspects of the data.
- **GAN Augmentation:** The multilabel classification model, although limited by the scarcity of data, demonstrated an improvement in performance when augmented with data generated through Generative Adversarial Networks (GANs). This indicates that GANs hold promise for enhancing model training, particularly in data-scarce environments.

Limitations

- **Data Scarcity for Multilabel Classification:** The multilabel classification model suffered from limited data, affecting its robustness and predictive accuracy.
- **Model Interpretability:** While SHAP values provide a degree of interpretability, they are based on a statistical model. The clinical context and other factors need to be considered alongside these values.
- **Non-Linearity and Outliers:** The skewness in the SHAP value distributions indicates that the features have non-linear impacts on the prediction, which may complicate clinical interpretation and intervention strategies.

- **Generalizability:** Both models were trained on a specific dataset and may not generalize well to other populations or healthcare settings.
- **Reliability of GAN Augmentation:** Although GANs showed promise in augmenting the dataset for the multilabel model, the true efficacy of such synthetic data in clinical scenarios needs further validation.

Future Directions

- **More Comprehensive Data:** Acquiring more labelled data, especially for the multilabel classification model, could improve predictive accuracy and robustness.
- **Validation in Clinical Settings:** The findings should be validated using data from diverse clinical settings to ensure broader applicability.
- **Further Exploration of GANs:** Additional studies could probe deeper into the effectiveness of GAN-augmented data for training machine learning models in healthcare.

5.11 Conclusion

Chapter 5 has systematically addressed and bridged the gaps identified in Chapter 4 by integrating clinical factors and Heart Rate Variability (HRV) measures across all leads of the PTB dataset. Through the sophisticated use of machine learning techniques, this chapter has significantly advanced the capabilities of ECG analysis, focusing on the extraction and fusion of HRV features with clinical factors to diagnose a spectrum of cardiac conditions. This has included myocardial infarction, dysrhythmia, bundle branch block, cardiomyopathy, and general heart health. The results obtained underscore the enhanced predictive accuracy of the machine learning models developed, which could not only detect but also differentiate between various cardiac anomalies with higher precision.

Furthermore, Chapter 5 has set the stage for the subsequent exploration of data augmentation techniques in Chapter 6, specifically through the use of Generative Adversarial Networks (GANs). This approach aims to enrich the dataset with synthetic samples, thereby addressing the limitations of data scarcity and improving the diversity and quality of the data available for model training. By enhancing model robustness and predictive accuracy, these methodologies promise to elevate the clinical applicability and interpretability of AI-driven ECG analysis. The transition into Chapter 6 will delve deeper into the specific methodologies of GAN augmentation, exploring various models of GANs in generating synthetic data and evaluating their effect on model performance. This progression reflects a logical extension of the work completed in this chapter and aligns with the overarching thesis goal of harnessing cutting-edge AI techniques to advance cardiac healthcare outcomes. Thus, the foundation laid in Chapters 4 and 5 provides a robust basis for the innovative explorations and potential breakthroughs anticipated in Chapter 6.

6 Enhancing Electrocardiogram Data with Generative Adversarial Networks – Approaches and Outcomes

6.1 Introduction

Building on the innovative groundwork laid in Chapter 5, where the integration of clinical factors with Heart Rate Variability (HRV) data was meticulously explored through a variety of machine learning models, Chapter 6 shifts the focus towards addressing the persistent challenges of data sparsity and class imbalance through the application of Generative Adversarial Networks (GANs). The sophisticated data augmentation techniques discussed in Chapter 5 set a robust platform for delving deeper into the capabilities of GANs in Chapter 6. This transition is pivotal as it explores the potential of GANs to generate high-quality synthetic data that closely mirrors real datasets, enhancing the training process and thereby improving the robustness and accuracy of the predictive models. Chapter 6 aims to extend the models' diagnostic efficacy by providing a richer, more balanced data environment, thus building directly on the enhancements in model performance introduced in the previous chapter.

Generative Adversarial Networks (GANs) were introduced by Ian Goodfellow and his colleagues in 2014, revolutionizing the field of machine learning [31], [94]. Consisting of two neural networks—the Generator and the Discriminator—GANs have become a cornerstone in unsupervised learning. These networks work in tandem to perform a two-player minimax game where the Generator creates synthetic data, and the Discriminator tries to differentiate between real and synthetic samples.

The unique architecture of Generative Adversarial Networks (GANs) plays a pivotal role in their ability to generate high-quality synthetic data. In a GAN setup, the Generator is tasked with crafting realistic data, while the Discriminator evaluates the authenticity of this data by comparing it to actual samples. Through iterative training, the synthetic data quality is refined, making it indistinguishable from real data [194].

GANs are particularly crucial for data augmentation in machine learning studies that deal with limited or imbalanced datasets. Traditional augmentation techniques like flipping and rotating are less effective for complex data types, such as medical images or time-series data [195], [196]. The versatility of GANs extends beyond computer vision to sectors such as healthcare, finance, and natural language processing, leading to specialized variants like conditional GANs and cycle-GANs for problem-specific applications [197], [198].

The success of a machine learning model is often contingent on the quality and volume of the training data. Synthetic data generated by GANs can address the limitations of small or skewed datasets, thereby boosting model performance [199]–[203]. When models are trained with data supplemented by synthetic samples, they generally exhibit enhanced generalization, performing better with unseen data [197], [198].

One of the persistent challenges in machine learning is class imbalance, which can result in biased models. GANs can rectify this by producing synthetic samples for underrepresented classes, equalizing the data distribution and reducing bias. However, the generation of synthetic data, especially in sensitive areas like healthcare, raises ethical considerations. It's imperative to ensure that the use of synthetic data upholds privacy standards and does not perpetuate existing biases in the training datasets [204]. In summary, GANs have emerged as a powerful tool for generating high-quality synthetic data. They offer an alternative and often superior means of data augmentation that can be particularly useful in contexts where data is scarce, imbalanced, or both.

6.2 Objectives

The primary objective of employing Generative Adversarial Networks (GANs) in this study is to tackle the challenges of data sparsity and class imbalance inherent to our ECG and IBI feature datasets. GANs have shown significant promise in generating synthetic but realistic data that maintains the essential characteristics of the original data[205]. By augmenting the minority classes, we aim to enhance the model's ability to generalize well across different classes [206], [207] [214]. Also, critically compare the efficacy of GAN-based data augmentation with other popular techniques such as SMOTE and Cost-Sensitive Learning (CST), particularly in the context of the given datasets which has complex, high-dimensional structure of ECG and IBI data [211]–[213].

6.3 Data Processing

Chapter 5 elaborates on the data pre-processing techniques applied to electrocardiogram (ECG) signals. The initial step involves setting up the working directory and listing the ECG data files, which is accomplished by employing libraries for data manipulation and numerical computing. Subsequently, multiple filters, including baseline wander removal and band-pass, are implemented to mitigate various noise types. Signal-to-Noise Ratio (SNR) serves as the quantifiable metric for noise removal efficacy. A DataFrame assembles the SNR metrics for each lead, facilitating global signal quality assessment. Refer to Chapter 5.4 for in-depth analysis.

Initially, the ECG data sets comprising 69 patient files were accessed from a specified directory. Each data file, which contained dimensions that varied from one another, was processed to represent ECG data corresponding to a duration of 6 seconds. This equated to 6000 rows of data points per patient file. To ensure this uniformity, excess rows beyond the 6000-row limit were systematically skipped during the data-loading phase. The raw ECG data was subjected to a down sampling process using a decimation technique. Specifically, the 'decimate' function from the SciPy library was applied, with a downsampling factor of 8. This reduced the original data dimensions, along axes specified, by a factor of 8, effectively reducing the complexity of the dataset. Advantages and disadvantages of downsampling[208]–[210] are listed in Table 6-1.

Table 6-1: Downsampling - Pros and Cons

Advantages
• Computational Efficiency: Down sampling reduces the computational burden significantly, making it faster to train complex models like GANs. • Noise Reduction: The process also serves as a filtering mechanism, potentially removing high-frequency noise that may not be relevant for the classification task. • Focus on Salient Features: Down sampling can bring into focus more salient, lower-frequency features of the ECG signal that are usually more important for diagnostic tasks.
Disadvantages
• Loss of Information: Down sampling may result in the loss of some high-frequency components that could be significant in certain diagnostic scenarios. • Aliasing: If not done carefully, down sampling can introduce aliasing artifacts, where higher-frequency components masquerade as lower-frequency components, thereby distorting the signal. • Risk of Over-smoothing: The decimation process, if excessively applied, can smooth out the signal to the point where important transient features may be lost.

By judiciously choosing the decimation factor and considering the above-mentioned trade-offs, the study aimed to achieve a balance between computational efficiency and preservation of diagnostically relevant signal features. With down sampling effectively reducing the complexity of the data, the next logical step was to normalize these datasets. A specialized 3D Minimax scaling technique was applied to maintain the structural integrity of the multivariate time series data. This scaled the data within a range of zero to one.

6.4 Methods

6.4.1 GAN Architecture for ECG – Signals

In this section, we focus on the utilization of GANs for augmenting ECG signals [199]–[203] in binary classification tasks aimed at distinguishing between myocardial infarction and healthy heart classes. Specifically, the study employs a lead-wise GAN augmentation approach, considering 12 different leads (ranging from lead 0 to lead 11) for both myocardial infarction and healthy heart classes. While the core methodology remains consistent across these leads, some leads required minor architectural adjustments during hyperparameter tuning. These alterations were predominantly aimed at optimizing the loss functions for the discriminator and generator components of the GANs.

The architecture of the Generative Adversarial Network (GAN) was specifically designed to suit the task at hand. The discriminator part of the GAN was crafted to include a series of 1D convolutional layers, each followed by a Leaky ReLU activation function and a Max Pooling layer. The architecture was compiled using Adam optimizer with specified learning rates and clipping norms. Conversely, the generator part of the GAN started with an input layer set to a latent space dimensionality of 100. This was followed by a Dense layer, which reshaped the input into the desired dimensions. Bidirectional Long Short-Term Memory (LSTM) layers were employed to capture the time-dependent features of the ECG signals, followed by a series of 1D convolutional layers interspersed with Leaky ReLU activation functions.

The Generative Adversarial Network (GAN) model comprises two main components: a generator and a discriminator. The generator aims to produce synthetic ECG signals that are indistinguishable from real ECG signals, whereas the discriminator's goal is to distinguish between real and generated signals. To create a GAN model, the generator and discriminator models are sequentially connected, forming a combined architecture. During the training process, the discriminator's weights are frozen to ensure that only the generator is updated.

The Adam optimizer is employed with a learning rate of 0.0005, clipnorm set to 1, and clipvalue at 0.5. The binary cross-entropy loss function is used for model compilation, optimizing the generator's ability to deceive the discriminator. For training the discriminator, a balanced dataset comprising real and fake ECG signals is prepared. Real ECG signals are randomly sampled from the training set. Fake ECG signals are generated by the generator model using points in the latent space as input.

Performance evaluation is conducted intermittently during the training process. The accuracy of the discriminator on real and fake examples is computed and summarized. Additionally, graphical plots of generated and real ECG signals are saved for visual assessment. The training process follows an iterative approach. For each epoch, half of the batch comprises real ECG signals while the other half consists of generated signals. The discriminator is first trained on both real and fake signals. Subsequently, the GAN model is trained to update the generator, using the discriminator's current state as a measure of how convincing the generated signals are.

Following an exhaustive hyperparameter tuning process utilizing a randomized search methodology within the model development framework, optimization of the architecture was achieved. Key hyperparameters, including a latent space dimensionality of 100, a batch size of 32, and an extensive training duration spanning 1750 epochs, were established. Continuous evaluations were performed at regular intervals, precisely every 100 epochs, to monitor the model's progression and facilitate necessary adjustments. By following this methodology, the GAN model is iteratively trained until the generator produces ECG signals that the discriminator finds increasingly difficult to distinguish from real ones. Figure 6-1 and 6-2 shows schematic representation of the generator and discriminator architecture for raw signals for HC class. For few other leads, class wise, GAN architectures please refer section 9.1.1 in appendices.

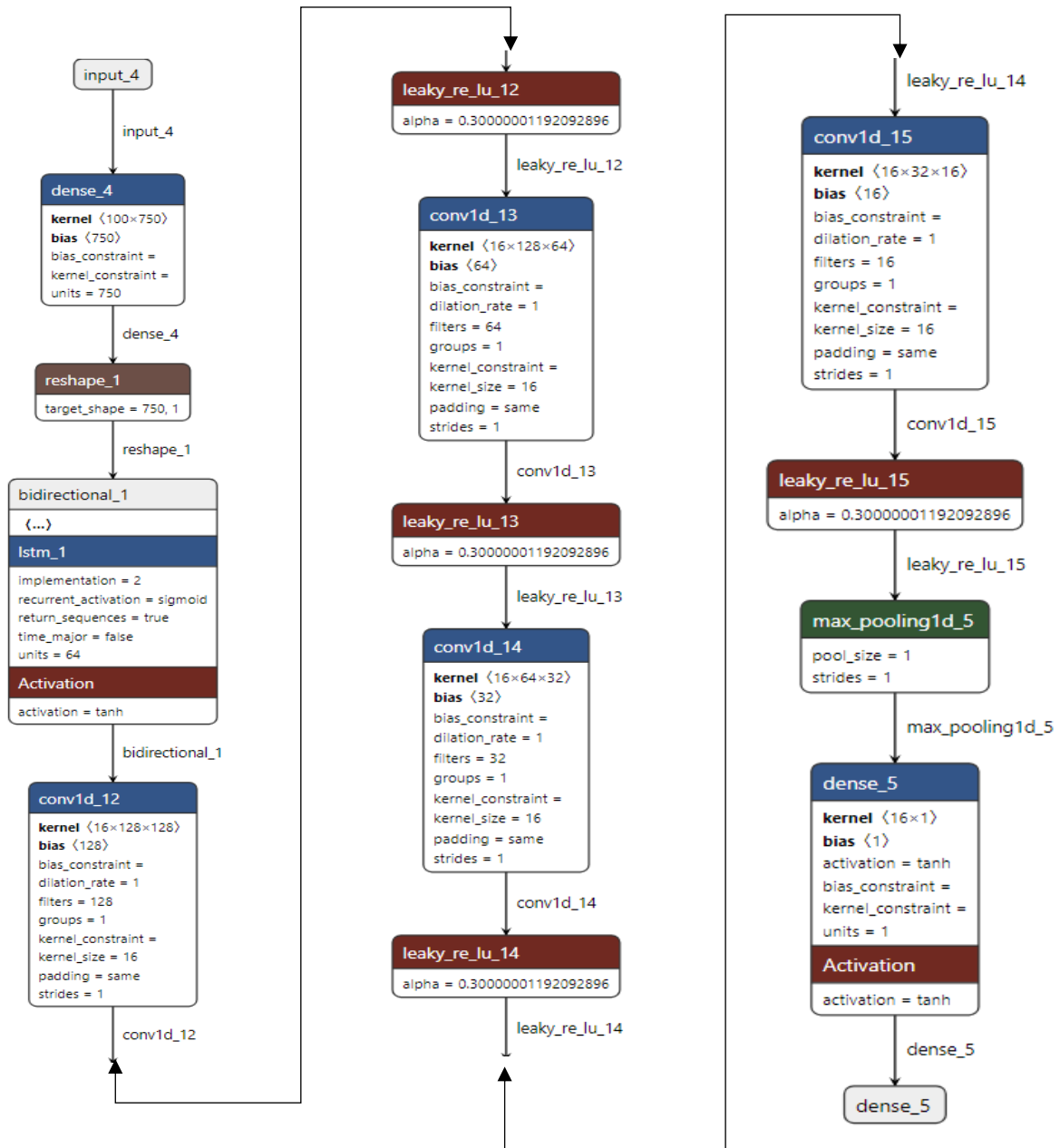


Figure 6-1: Generator Model for synthetic samples of HC Class - Lead i

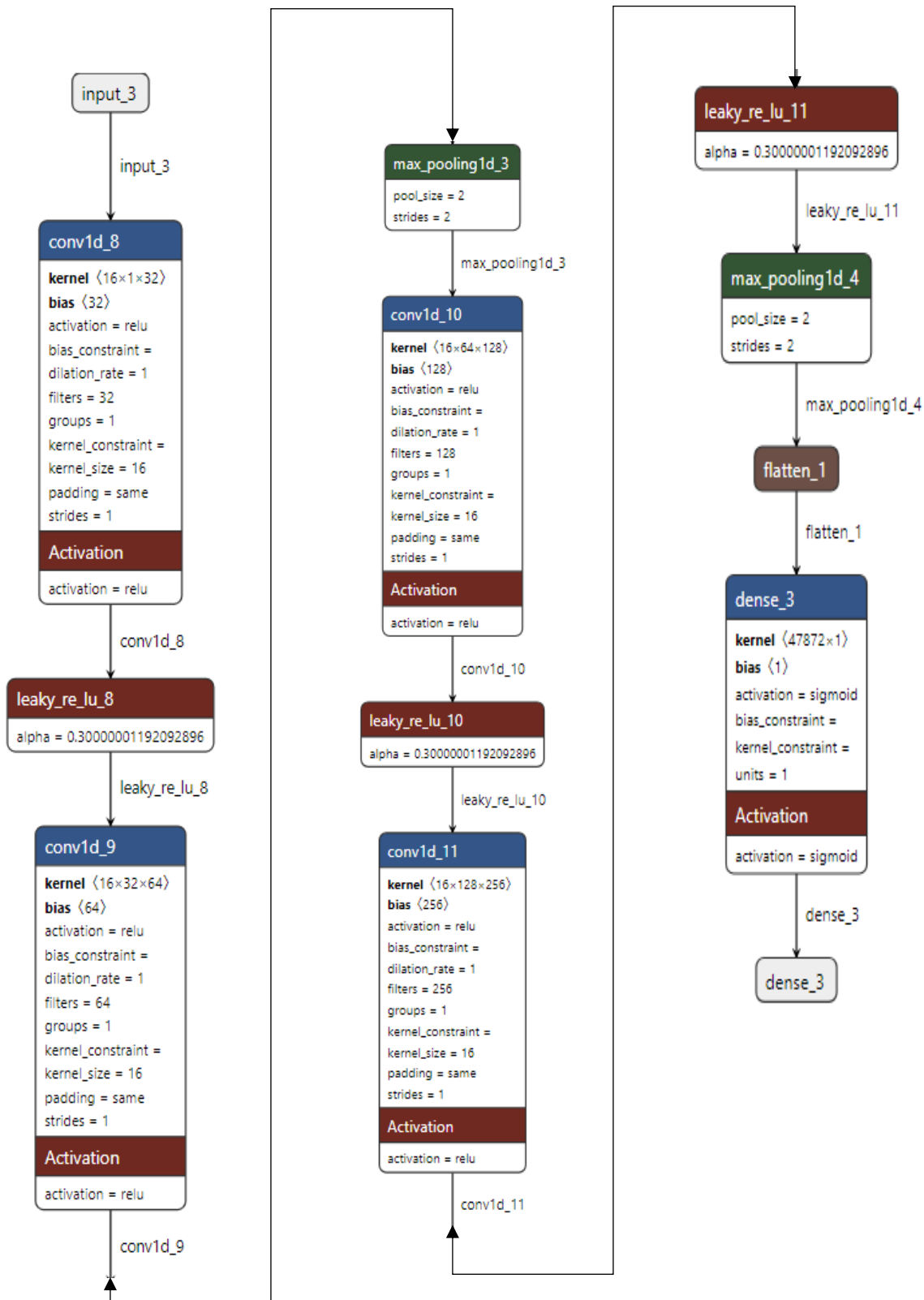


Figure 6-2: Discriminator Model for synthetic samples of HC Class - Lead i

6.4.2 Mathematical background for creating and training real and synthetic samples.

In the realm of Generative Adversarial Networks (GANs), the primary objective is to train a generator that can create synthetic samples that are nearly indistinguishable from real samples. To achieve this, we need mathematical functions that deal with real samples, synthetic samples, and latent space inputs[211]–[214].

This function picks n random real samples from a pre-existing dataset (ecg signal). These serve as the 'ground truth' during the training of the discriminator. The discriminator learns to identify these as real samples. This is done through random sampling process where,

$$S' = \{s_1, s_2, \dots, s_n\} \text{ where } s_i \in S \quad \text{Equation 6-1}$$

The function also labels these as real (1), resulting in a set:

$$L' = \{(s_1, 1), (s_2, 1), \dots, (s_n, 1)\} \quad \text{Equation 6-2}$$

Next, through random vector generation we generate n random points in a latent space dimension. These points serve as the "seed" for the generator to create synthetic samples. This is represented as if L is the latent space, then each point p is a vector in R^{latent_dim}

$$P = \{p_1, p_2, \dots, p_n\} \text{ where } p_i \in R^{latent_dim} \quad \text{Equation 6-3}$$

Subsequently, through transformation mapping latent points are taken and pushed to generator to transform them into synthetic samples. These are labelled as 0 and discriminator being trained to identify them as fake which is essentially a function $G: R^{latent_dim} \rightarrow R^{data_dim}$

$$F = \{G(p_1), G(p_2), \dots, G(p_n)\} \quad \text{Equation 6-4}$$

These generated samples are labelled as synthetic or fake 0.

$$L'' = \{(G(p_1), 0), (G(p_2), 0), \dots, (G(p_n), 0)\} \quad \text{Equation 6-5}$$

The following function performs the evaluation metrics for real and fake samples. Let D be the discriminator function. $D: R^{data_dim} \rightarrow [0,1]$

For real samples x_{real} accuracy metric acc_{real} is calculated as:

$$acc_{real} = \frac{1}{n} \sum_{i=1}^n D(x_{real_i}) \quad \text{Equation 6-6}$$

For synthetic samples x_{fake} accuracy metric acc_{fake} is calculated as:

$$acc_{fake} = \frac{1}{n} \sum_{i=1}^n D(x_{fake_i}) \quad \text{Equation 6-7}$$

The function also generates plots of real and fake samples, which can be seen as a visual representation of the functions G and D .

Finally, for encapsulating the training process, it is a series of optimization problems where the objective is to minimize the loss functions of the discriminator D and the generator G.

For the discriminator the loss function L_D is

$$L_D = -\frac{1}{n} \sum_{i=1}^n [y_i \log(D(x_i)) + (1 - y_i) \log(1 - D(x_i))] \quad \text{Equation 6-8}$$

For the generator the loss function L_G when using the GAN model is

$$L_G = -\frac{1}{n} \sum_{i=1}^n \log(D(G(z_i))) \quad \text{Equation 6-9}$$

where z_i are the points in the latent space. The optimizer then iteratively updates the parameters of G and D to minimize these loss functions.

6.4.3 GAN Architecture - Synthetic samples with IBI features and clinical factors for binary classification.

In the realm of biomedical data analysis, generating synthetic samples is essential for dealing with class imbalance and for enhancing the model's generalization ability. This section presents a Generative Adversarial Network (GAN) architecture tailored for generating synthetic samples that are characterized by Inter-beat Intervals (IBI) features and other clinical factors. Our GAN model aims to contribute to binary classification tasks by generating synthetic samples and balancing the data set and thereby improving the classification task metrics [215], [216].

The objective was to tune the hyperparameters of the GAN model to find the optimal settings for both the generator and discriminator networks. We aimed to achieve a balanced trade-off between generator and discriminator losses for tabular data augmentation.

We employed a custom estimator class for the GAN model. This class was designed with initialization parameters for learning rate, batch size, number of neurons per layer, number of layers, activation function, and optimization algorithm.

For hyperparameter tuning, we utilized RandomizedSearchCV from Scikit-learn. The model's hyperparameters were meticulously selected from predefined ranges to ensure optimal performance. These parameters included the learning rate, batch size, number of neurons, number of layers, activation function, and optimizer. The learning rate was chosen from a range of options spanning 0.0001 to 0.01, while the batch size was selected from 16, 32, or 64. The number of neurons ranged from 8 to 24, and the number of layers varied between 1 and 3. Activation functions such as 'ReLU', 'sigmoid', and 'tanh' were explored, along with optimizers including Adam, RMSprop, and SGD. Additionally, a custom scoring function was devised to combine both discriminator and generator losses, ensuring an effective optimization strategy throughout the model's training process. The hyperparameter tuning provided a set of optimal parameters for the learning rate, batch size, number of neurons, number of layers, activation function, and optimizer. These parameters were later used to build the final GAN model.

Then next step was to generate synthetic samples for both positive and negative classes of the tabular data using the optimized GAN model. Two generator models were initialized for the positive and negative classes, respectively. Additionally, a discriminator model was initialized. All the models used the optimal

hyperparameters obtained from the previous tuning stage. The discriminator model was compiled using binary cross-entropy as the loss function. Then, both generator models were combined with the discriminator to form the GAN model, which was also compiled using binary cross-entropy.

A custom function was employed to generate synthetic samples. This function took the real data, its labels, the number of synthetic samples, and a customized prevalence ratio for the classes as inputs. The GAN model was trained using the combination of real and synthetic samples for a pre-defined number of epochs and batch size, both obtained from the tuning stage. During training, several metrics were tracked for both the generator and discriminator, including their losses and accuracies on real and synthetic samples. These metrics were logged at specific epochs to observe the model's performance over time. The synthetic samples generated by the trained GAN model successfully augmented the existing dataset. The losses for both the generator and discriminator stabilized, indicating the model's efficacy.

In summary, the hyperparameter tuning and synthetic data generation processes were meticulously performed to ensure the quality and relevance of the synthetic samples. These steps make a significant contribution to the field of data augmentation for tabular data using GANs. Figures 6-3 and 6-4 represents generator and discriminator architecture for GAN's for creating synthetic samples.



Figure 6-3: Positive and Negative class Generator – IBI & Metadata

Figure 6-4: Discriminator - IBI & Metadata

6.4.4 Mathematical background tabular of GAN for synthetic samples in binary classification.

Conceptually the following functions were used in implementing GAN for synthetic data augmentation in tabular data set[217], [218].

The GAN consists of two primary networks—a generator G and a discriminator D. They are trained in a zero-sum game described by the following objective function:

$$\min_G \max_D V(D, G) = E_{X \sim P_{data}(x)}[\log D(x)] + E_{Z \sim P_Z(Z)}[\log(1-D(G(z)))] \quad \text{Equation 6-10}$$

Where $D(x)$ represents the probability that x is the real sample according to the discriminator. $G(z)$ is synthetic sample generated to fool the discriminator. E is the expectation which is effectively an average over many samples. $X \sim P_{data}(x)$ is the sample from the true data distribution. $Z \sim P_Z(Z)$, z is sampled from the input Gaussian noise distribution.

The term $\max_D V(D, G)$ represents the optimization objective for the discriminator. It aims to maximize $V(D, G)$ so that $D(x)$ is close to 1 (indicating real samples) and $D(G(z))$ is close to 0 (indicating fake or synthetic samples).

On the other hand, the term $\min_G V(D, G)$ represents the generator minimizing $V(D, G)$ effectively trying to make $D(G(z))$ as large as possible, ideally close to 1, to fool the discriminator into thinking that the generated samples are real.

The term $E_{X \sim P_{data}(x)}[\log D(x)]$ in the objective function encourages the discriminator to assign a high probability (close to 1) to real samples. When $D(x)$ is high, $\log D(x)$ is also increases, indicating the discriminator's confidence in classifying real data.

The term $E_{Z \sim P_Z(Z)}[\log(1-D(G(z)))]$ in objective function encourages the discriminator to assign a low probability (close to 0) to fake samples. The generator tries to minimize this term, thereby trying to make $D(G(z))$ large, which $\log(1-D(G(z)))$ small.

In summary, the objective function creates a sort of "game" where the discriminator tries to distinguish real samples from fake ones, while the generator tries to produce fake samples that are indistinguishable from real samples. The training process involves both the discriminator and the generator adapting and improving until either one outperforms the other or they reach a state of equilibrium.

Both the generator and discriminator are multi-layer perceptrons (MLPs). We have designed two separate generators: one for generating positive class samples and another for negative class samples [219]–[222].

The Generator architecture is designed as follows; it begins with an input layer consisting of number of neurons equal to the number of features in X_{train_val} , denoted by n . The hidden Layers totalling $L-1$ layers each contain k neurons, where L is the number of layers and k is the number of neurons per layer. Activation on functions ϕ such as ReLU, sigmoid, or tanh are applied throughout the hidden layers. Finally, the output layer employs a linear activation function to produce n synthetic features.

Mathematically, the generator can be represented as:

$$G(x) = W_L \phi(W_{L-1} \phi(\dots \phi(W_1 x + b_1) \dots) + b_{L-1}) + b_L \quad \text{Equation 6-11}$$

Where $W_1, W_2, \dots, W_L$ are the weight matrices for each layer in the network. $b_1, b_2, \dots, b_L$ are the bias terms for each layer in the network. ϕ is the activation function (e.g., ReLU, Sigmoid, Tanh). L is the number of layers in the generator.

We start with an input vector x (which is often random noise). This input is passed through the first layer, $W_1 x + b_1$, and then through an activation function ϕ . The output of this activation function becomes the input to the next layer. This process is repeated for each layer up to L , the final layer. The final output is a synthetic (or "fake") sample.

The discriminator architecture comprises an input layer that takes in n features. Hidden layers are structured with $L-1$ hidden layers each containing k neurons. Throughout the hidden layers the activation function ϕ as specified, is applied. The output layer consists of a single neuron with a sigmoid activation function, predicting the probability of a sample being real.

Mathematically, the discriminator can be represented as:

$$D(x) = \sigma(W_L \phi(W_{L-1} \phi(\dots \phi(W_1 x + b_1) \dots) + b_{L-1}) + b_L) \quad \text{Equation 6-12}$$

Where, σ is the Sigmoid activation function. This is used because the output of the discriminator needs to be a probability (i.e., a number between 0 and 1). The discriminator takes in a sample x (either real or fake). This sample is passed through multiple layers, like the generator. However, the output goes through a Sigmoid function σ because the discriminator's goal is to output a probability that the input sample is real.

In the data augmentation process, generators create synthetic samples $G_{positive}(x)$ and $G_{negative}(x)$ using input features. These synthetic samples are then labelled according to their generator positive or negative. Finally, the synthetic samples are aggregated with the original data, resulting in $X_{combined}$ and $y_{combined}$.

In the training and evaluation phase, the discriminator and combined GAN model undergo alternating training using the binary cross-entropy loss function and are optimized using RMSprop. Metrics such as loss and accuracy are tracked for the discriminator on both real and synthetic samples. The training process continues for a specified number of epochs, with metrics logged at regular intervals to monitor convergence. This framework lays the mathematical groundwork for GAN architecture and training in generating synthetic samples from tabular data.

6.4.5 GAN Architecture - Synthetic samples with IBI features and clinical factors for multilabel classification

Generative Adversarial Networks (GANs) have shown great promise in various domains such as image synthesis, style transfer, and data augmentation. In medical research, the use of synthetic data to augment real-world data sets has become an area of increasing interest. Specifically, for multilabel classification problems in cardiovascular health, which often suffer from limited or imbalanced datasets, synthetic data could be a game-changer.

In this section, we delve into the application of four different GAN architectures for generating synthetic samples, with a particular focus on Inter-Beat Intervals (IBI) features and clinical factors. The objective is to bolster multilabel classification tasks where each sample could belong to more than one class (e.g., different types of arrhythmias).

In generating synthetic multilabel data, the choice of Wasserstein GAN (WGAN) is primarily informed by its stability and effectiveness in handling complex distributional relationships. Traditional GANs often suffer from mode collapse or unstable training dynamics where the generator starts to produce only a limited variety of samples, ignoring other possible modes of the data distribution. WGANs are less prone to this, offering a more stable training process. In contrast, WGAN introduces the concept of Wasserstein distance to measure the discrepancy between real and synthetic data distributions, providing a smoother and more meaningful loss landscape. This makes it particularly suited for generating multilabel synthetic samples, as it can accommodate the complex class boundaries and interdependencies involved[223]–[227].

In WGAN, the loss correlates to the quality of the generated samples, unlike in a vanilla GAN where the loss does not necessarily provide an intuition about how well the generator is doing. Multilabel data typically involves complex relationships between different labels. WGANs can better model these nuanced relationships than standard GANs[228]–[231]. WGANs are generally better at dealing with situations where the real and fake data distributions do not overlap, which is often the case in multilabel data.

The four architectures explored are:

- Simple GAN with Binary Cross-Entropy Loss Function: A basic GAN structure leveraging the Binary Cross-Entropy (BCE) loss function for optimization.
- Simple GAN with Custom Loss Function: This architecture employs a hybrid loss function that combines the Wasserstein loss and Binary Cross-Entropy to enhance synthetic data quality.
- Wasserstein GAN (WGAN) with Wasserstein loss: Here, the WGAN framework is combined with the Wasserstein loss function for potentially better stability and performance.
- WGAN with Custom Loss Function: The most advanced of the four, this WGAN utilizes a custom loss function which is a blend of Wasserstein loss and Binary Cross-Entropy.

We will now evaluate the performance, stability, and data fidelity of these architectures through various experiments and metrics, aiming to offer a comprehensive understanding of their utility in the specific context of multilabel classification in cardiovascular health.

6.4.5.1 Simple GAN with Binary Cross-Entropy loss function

The GAN model architecture consists of two main components: the Generator and the Discriminator. The Generator employs a Sequential architecture, comprising layers including a Dense layer with 128 units, LeakyReLU activation, and BatchNormalization, followed by additional Dense layers with 256 units each, and a

final layer matching the input data's feature size. Similarly, the Discriminator also follows a Sequential architecture, featuring Dense layers with 256 and 128 units, both with LeakyReLU activation, and culminating in a Dense layer with a sigmoid activation function to produce binary output. Additionally, a custom loss function is defined, incorporating Binary Cross-Entropy loss, as depicted by the equation below.

$$BCE_{(y-\hat{y})} = -(y \log(\hat{y}) + (1 - y)\log(1 - \hat{y})) \quad \text{Equation 6-13}$$

The real samples for training are randomly selected from the existing dataset. For the synthetic data, noise vectors are generated. These noise vectors are not random but rather are generated based on predefined class proportions to ensure class balance.

Class proportions are predefined to ensure the synthetic data is class balanced. A mapping exists between the indices of the real classes and the synthetic classes, ensuring compatibility. Labels are one-hot encoded based on this mapping. For generating synthetic samples, the noise vector is generated according to predefined class proportions $p_1, p_2, \dots, p_n$ and batch size for class $i = \text{Batch size} \times p_n$. In scenario of n classes, each class i is associated with one-hot encoded vector $\rightarrow_{e_i}$, where all the elements are zeros except for the i -th position which is 1. Then for each synthetic sample generated, it is labelled with $\rightarrow_{e_i}$ based on the class proportions and class mapping.

Hyperparameter tuning played a crucial role in optimizing the GAN model. The learning rate of the RMSprop optimizer was meticulously set to 0.00005, ensuring efficient convergence during training. Additionally, the architecture of both the Generator and Discriminator underwent rigorous refinement through a combination of grid search and random search techniques, ensuring optimal performance across various hyperparameters.

During the training loop, a batch training approach was adopted to efficiently train the GAN model. In each epoch, the Discriminator and Generator were trained sequentially. Initially, the Discriminator underwent training on a combination of real and synthetic samples, enhancing its ability to distinguish between genuine and generated data. Subsequently, the Generator was trained to produce synthetic samples that could effectively deceive the Discriminator, thereby improving its generative capabilities. This structured training regimen facilitated the iterative refinement of both components, ultimately leading to the convergence of the GAN model. Figure 6-5 show the architecture used for generator and discriminator along with architecture attributes.

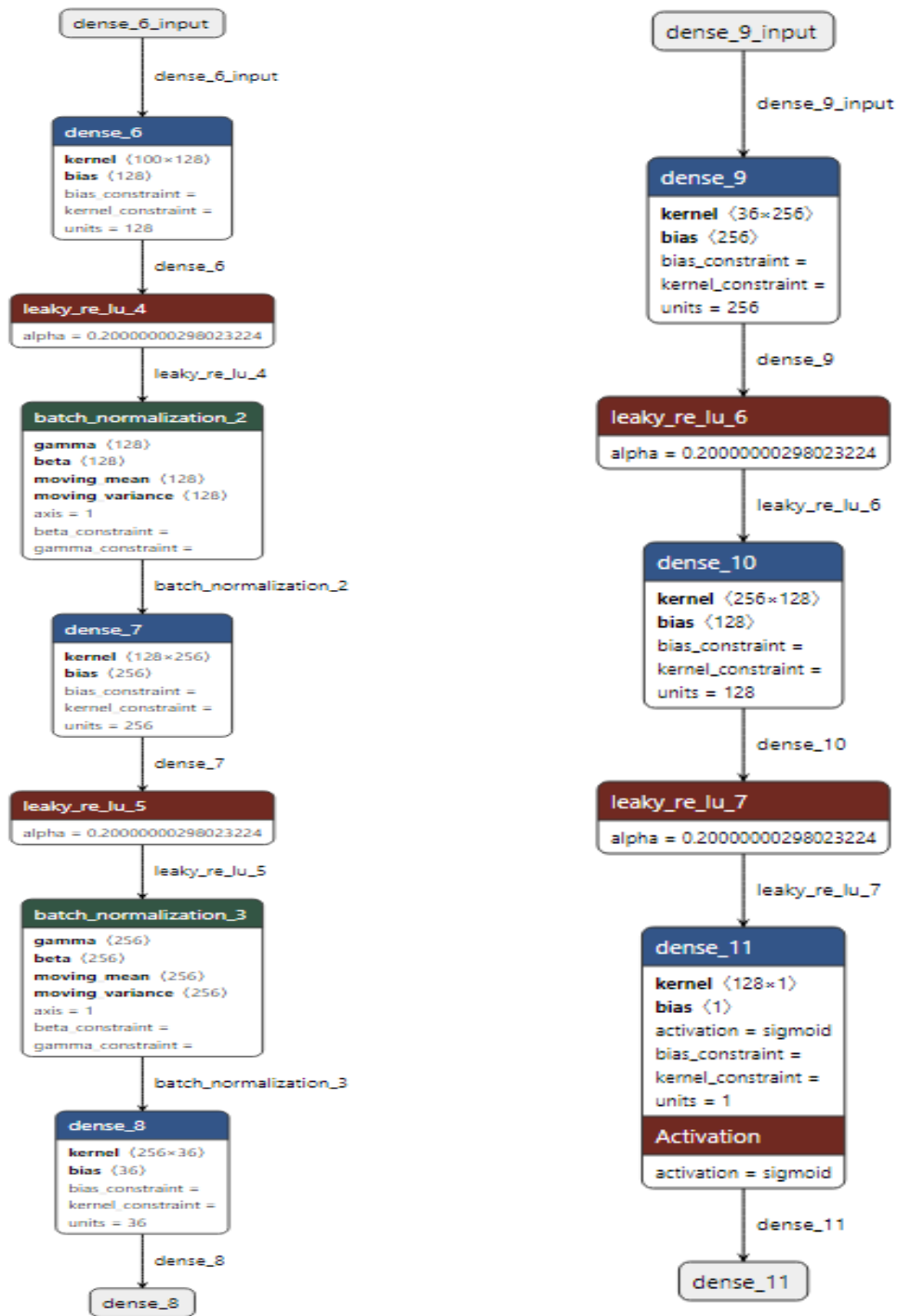


Figure 6-5: Generator & Discriminator - Multilabel Classification - Simple GAN with BCE loss function

6.4.5.2 Simple GAN with Custom loss function

While the architecture of the GAN model remains consistent with the prior section, the primary point of divergence lies in the loss function. In this section, the standard Binary Cross-Entropy loss function is substituted with a custom loss function designed to adapt to the specific requirements of the problem space.

The custom loss function can be defined as a linear combination of the Wasserstein Loss and the Weighted Binary Cross-Entropy. The equation is given by:

$$\text{Custom Loss} = \alpha \times \text{Wasserstein Loss} + \beta \times \text{Weighted BCE} \quad \text{Equation 6-14}$$

Here, α and β are constants set to 0.5, indicating equal importance between the two loss components. The Wasserstein Loss is calculated as the mean of the product of the true labels and the predicted labels. Mathematically expressed as,

$$\text{Wasserstein Loss} = \text{Mean}(y_{\text{true}} \times y_{\text{pred}}) \quad \text{Equation 6-15}$$

Additionally, the Weighted Binary Cross-Entropy extends the regular Binary Cross-Entropy loss, by weighting each term according to the class distribution. The equation is expressed as,

$$\text{Weighted BCE} = -\sum_{i=0}^k w_i (y_{\text{true}}^i \log(y_{\text{pred}}^i) + (1 - y_{\text{true}}^i) \log(1 - y_{\text{pred}}^i)) \quad \text{Equation 6-16}$$

Here, $w_0, w_1, \dots, w_k$ are the weights for each class, calculated based on the inverse of the class proportions $p_0, p_1, \dots, p_k$ given by equation 6-17 ensuring normalization such that their sum equals 1, effectively addressing class imbalance.

$$w_i = \frac{\frac{1}{p_i}}{\sum_{j=0}^k \frac{1}{p_j}} \quad \text{Equation 6-17}$$

This normalization ensures that the weights sum up to 1 and mitigates the issue of class imbalance. The GAN model is trained with the same architecture and hyperparameters as previously described in the section on GAN with Binary Cross-Entropy loss function, employing the custom loss function detailed here.

By incorporating this custom loss function, the model adapts more effectively to the class imbalance, contributing to improved training dynamics and, potentially, more accurate predictions which is shown in results section and further inferred in discussion section. Figure 6-6 show the architecture used for generator and discriminator along with attribute values.

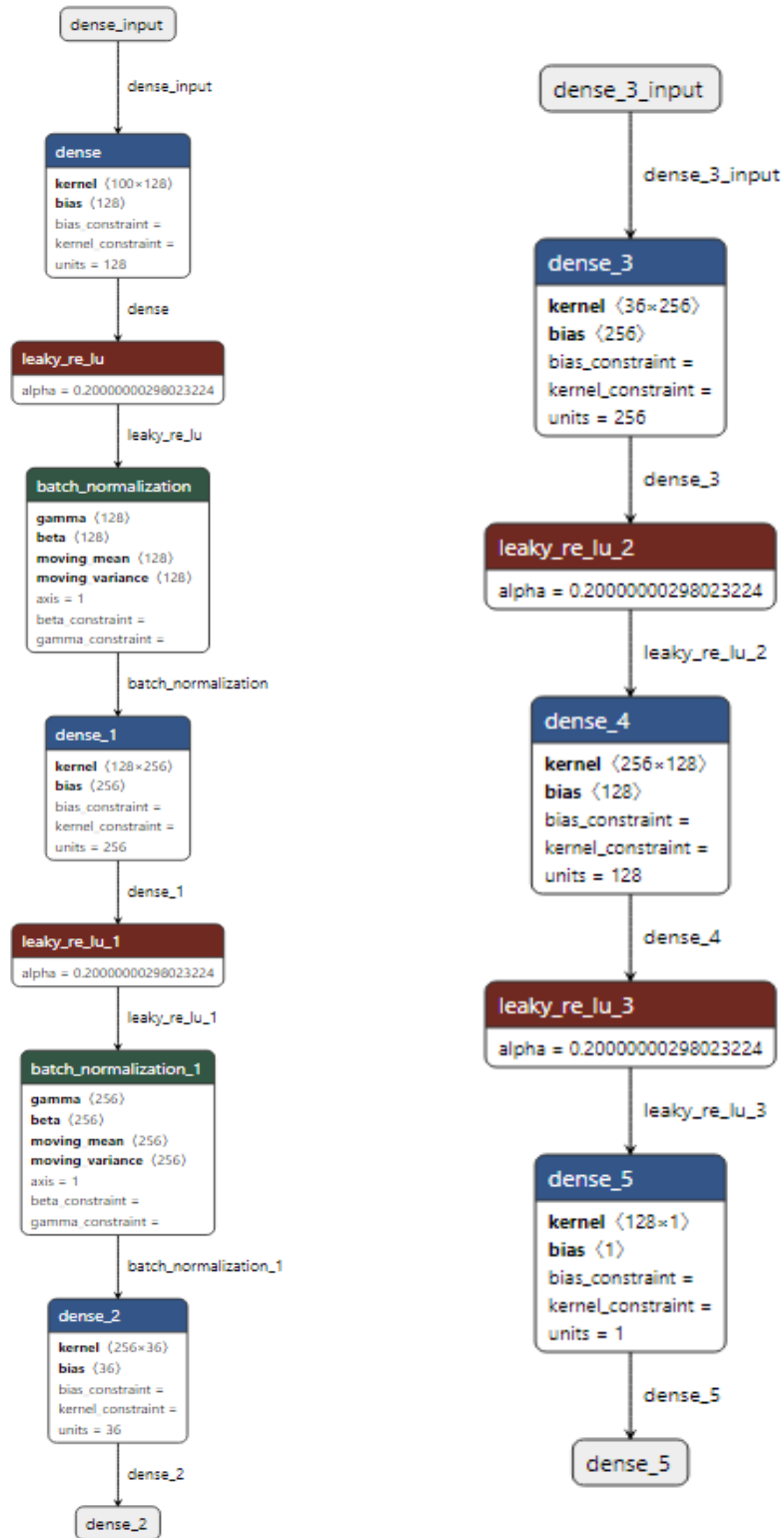


Figure 6-6: Generator & Discriminator - Multilabel Classification - Simple GAN with custom loss function

6.4.5.3 Wasserstein GAN (WGAN) with Wasserstein loss

In generating synthetic multilabel data, the choice of Wasserstein GAN (WGAN) is primarily informed by its stability and effectiveness in handling complex distributional relationships. Traditional GANs often suffer from mode collapse or unstable training dynamics where the generator starts to produce only a limited variety of samples, ignoring other possible modes of the data distribution. WGANs are less prone to this, offering a more stable training process. In contrast, WGAN introduces the concept of Wasserstein distance to measure the discrepancy between real and synthetic data distributions, providing a smoother and more meaningful loss landscape. This makes it particularly suited for generating multilabel synthetic samples, as it can accommodate the complex class boundaries and interdependencies involved.

In WGAN, the loss correlates to the quality of the generated samples, unlike in a vanilla GAN where the loss does not necessarily provide an intuition about how well the generator is doing. Multilabel data typically involves complex relationships between different labels. WGANs can better model these nuanced relationships than standard GANs. WGANs are generally better at dealing with situations where the real and fake data distributions do not overlap, which is often the case in multilabel data.

The WGAN comprises a generator and a discriminator, both implemented using neural networks. Generator has an architecture starting with a dense layer of 128 units taking a 100-dimensional noise vector. It employs LeakyReLU activation functions and BatchNormalization layers to stabilize the internal statistics. Finally, it outputs synthetic data matching the feature dimensions of the training set.

Discriminator network begins with a 256-unit dense layer and scales down through layers with LeakyReLU activations. It terminates in a single output, tasked to evaluate the Wasserstein distance. Both the generator and discriminator are compiled using the RMSprop optimizer with a learning rate of 0.00005, designed to optimize the Wasserstein loss function.

A function is said to be Lipschitz continuous if there exists a constant $L \geq 0$ such that for all pairs of points x and y , the following inequality holds:

$$|f(x) - f(y)| \leq L \cdot |x - y| \quad \text{Equation 6-18}$$

In other words, there's a bound on how fast the function can change. The constant L is referred to as the Lipschitz constant. In the context of WGANs, the Lipschitz continuity is applied to the critic function (or discriminator). The goal is to ensure that the critic function doesn't change too rapidly, avoiding extreme gradients that can lead to issues like mode collapse or training instability.

The original WGAN paper ensured Lipschitz continuity by clipping the weights of the critic. However, this method can be problematic, and this is where Gradient Penalty (GP) comes into play [232]–[235].

To enforce Lipschitz continuity, the Gradient Penalty (GP) technique is applied as a regularization method to regulate the gradients of the discriminator. The penalty is calculated involves following steps:

First, Interpolated samples $\hat{x}$ are calculated between real and fake data defined as

$$\hat{x} = \epsilon \cdot x_{real} + (1-\epsilon) \cdot x_{fake} \quad \text{Equation 6-19}$$

Where ϵ is sampled uniformly from $[0,1]$.

Next, the gradients of the discriminator's output with respect to $\hat{x}$ denoted as $\nabla_{\hat{x}} D(\hat{x})$ are determined. Subsequently, the penalty is calculated as the squared difference between the norm of the gradient and 1, expressed as

$$GP = (\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2 \quad \text{Equation 6-20}$$

Finally, the discriminator's ultimate loss function includes both the original Wasserstein loss and this gradient penalty, weighted by a hyperparameter λ

$$\text{Loss} = \text{Wasserstein Loss} + \lambda \cdot GP \quad \text{Equation 6-21}$$

These concepts are implemented in the gradient penalty loss function within the code, leveraging TensorFlow's automatic differentiation capabilities.

The introduction of gradient penalty ensures the Lipschitz continuity of the discriminator function. This constraint stabilizes the gradients during backpropagation, helping the network to learn more effectively. By constraining the gradient, it avoids the critic (discriminator) from becoming too steep, helping to stabilize training and produce more meaningful gradients for updating the generator. It is particularly beneficial in the context of multilabel classification, where class boundaries may be less distinct, requiring a model that can learn nuanced representations without overfitting or oscillating wildly during training.

In summary, the use of WGAN with a gradient penalty in the context of synthetic multilabel sample generation offers a robust and stable training mechanism, better capturing the complexities inherent in multilabel data. Figure 6-7 shows the architecture of generator and discriminator.

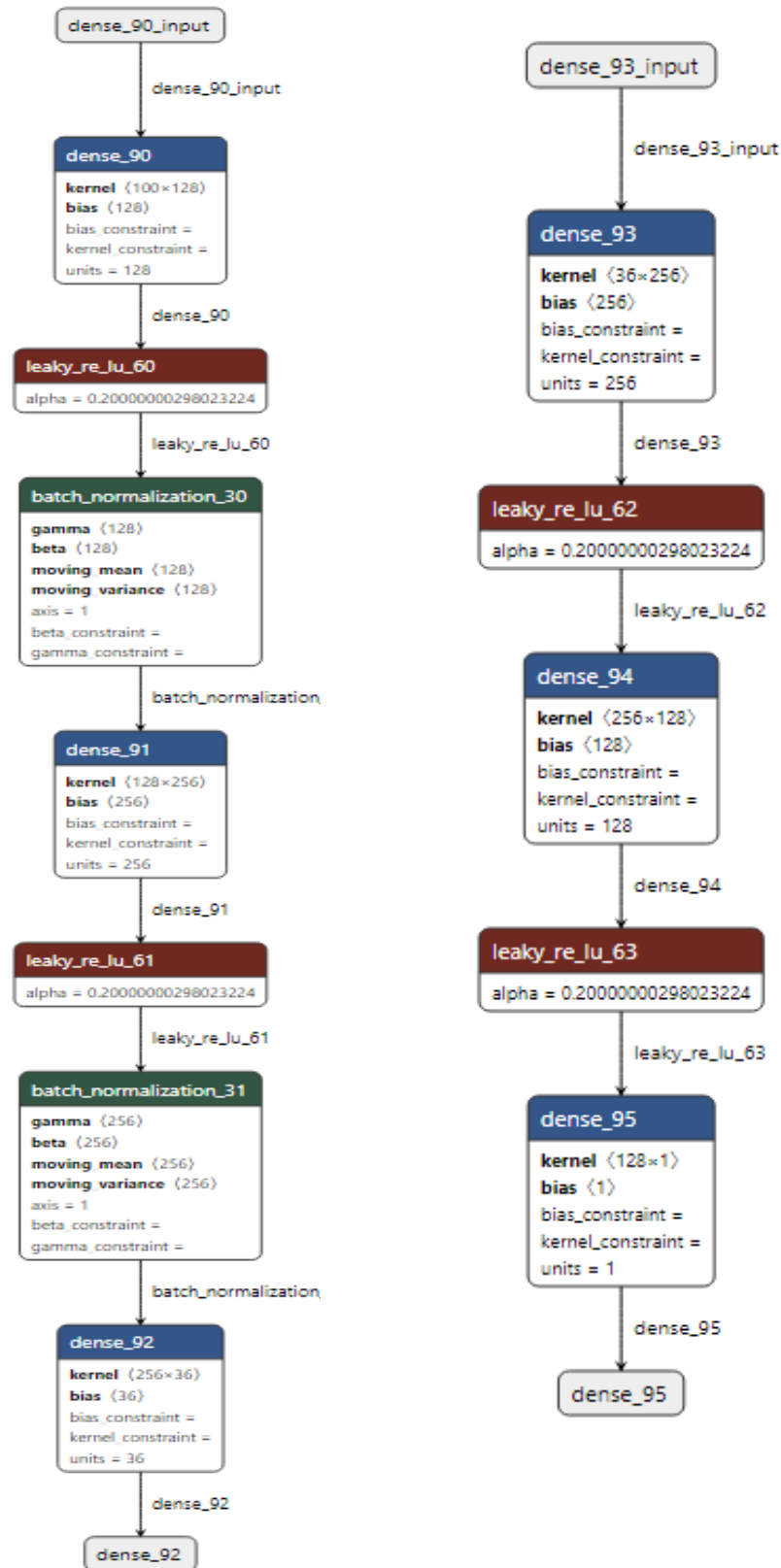


Figure 6-7: Generator & Discriminator - Multilabel Classification - WGAN with Wasserstein loss function

6.4.5.4 Wasserstein GAN (WGAN) with Custom loss function

The variant of Wasserstein GAN (WGAN) with a custom loss function presents a tailored solution to the challenges associated with generating synthetic multilabel data. This approach integrates the Wasserstein loss with a weighted Binary Cross-Entropy (BCE) loss, with the latter addressing class imbalance concerns. By combining these two objectives, the model achieves a balance between robustness and stability inherent in WGANs, while also effectively handling class imbalance, rendering it well-suited for generating intricate multilabel synthetic samples. Specifically, class weights are derived from the inverse proportions of multilabel classes to mitigate class imbalance. The custom loss function orchestrates the fusion of the Wasserstein loss with the weighted BCE, while the gradient penalty loss function facilitates the computation of the gradient penalty. Both the generator and discriminator models are compiled using this custom loss function. Notably, the weighted BCE incorporates class weights directly within the BCE computation process to address class imbalance more directly. The train function is employed to train the WGAN with the custom loss and gradient penalty, ensuring the model learns effectively from the defined loss objectives.

Both the generator and discriminator models are implemented using feedforward neural networks. The generator model comprises three dense (fully connected) layers. The first dense layer has 128 neurons and uses an input vector of size 100 (this is often referred to as the noise vector). It employs a Leaky ReLU activation function with an alpha of 0.2 to introduce non-linearity. Batch normalization is applied to stabilize the learning process and to reduce internal covariate shift. The second dense layer further expands the model with 256 neurons and uses a LeakyReLU activation with an alpha of 0.2, followed by another batch normalization. Finally, the output layer produces an output of the same shape as the training data effectively generating synthetic samples.

The discriminator also consists of dense layers but is simpler in structure compared to the generator. The first dense layer uses 256 neurons and takes an input of shape matching the shape of the training data or the synthetic data generated by the generator. It uses a LeakyReLU activation with an alpha of 0.2. The second dense layer has 128 neurons, followed again by a LeakyReLU activation with an alpha of 0.2. The output layer has just one neuron, which estimates the Wasserstein distance. The training function uses the custom loss to train the discriminator on both real and fake samples. It also incorporates the gradient penalty, ensuring that the function adheres to Lipschitz continuity, thus making the training stable and robust.

In summary, the WGAN with custom loss offers a specialized and efficient model for generating synthetic multilabel samples. The combined loss function leverages the strengths of both Wasserstein loss and weighted BCE to tackle the nuances involved in multilabel classification. Figure 6-8 shows the architecture of generator and discriminator.

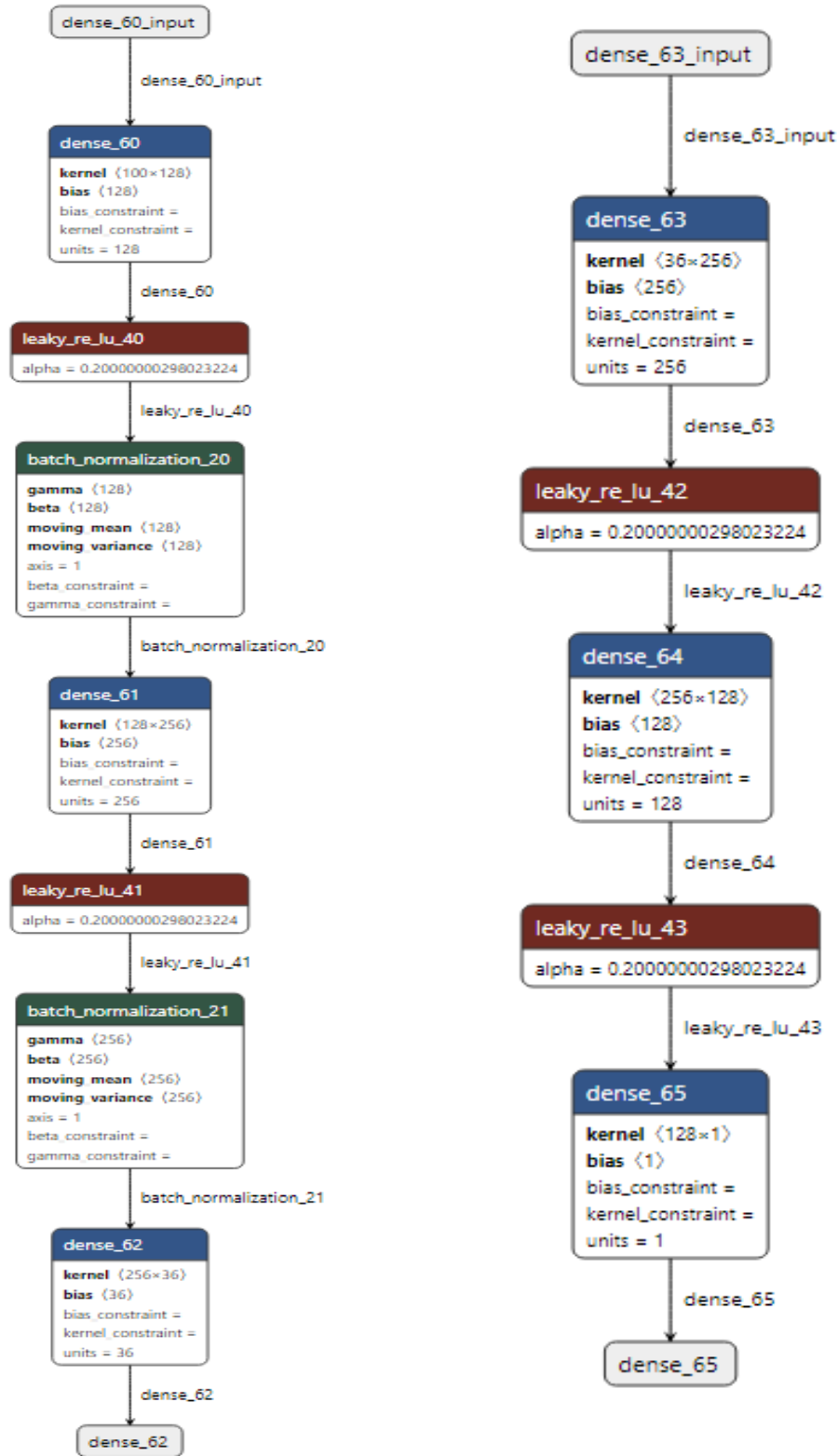


Figure 6-8: Generator & Discriminator - Multilabel Classification - WGAN with custom loss function

6.4.6 PCA and t-SNE Analysis

The evaluation of generative models like GANs often extends beyond conventional loss metrics to include domain-specific measures and visualizations. For the evaluation of synthetic multilabel data generated from our Wasserstein GAN (WGAN) model, PCA (Principal Component Analysis) and t-SNE (t-Distributed Stochastic Neighbour Embedding) are employed.

PCA is a dimensionality reduction technique that transforms the original variables into a new set of uncorrelated variables called principal components. Mathematically, PCA computes the eigenvalues and eigenvectors of the data's covariance matrix. The first principal component accounts for the maximum variance, the second (which is orthogonal to the first) accounts for the second most, and so on. For this analysis, we reduce the data dimensions to two principal components.

In the context of evaluating synthetic data, PCA allows us to visually compare the distribution of the real and synthetic datasets in a 2D feature space. In our implementation, two sets of PCA plots are generated for each class— one representing real training data and the other representing synthetic data.

While PCA is linear, t-SNE is a non-linear dimensionality reduction method well-suited for embedding data in two or three dimensions for visualization. It works by minimizing the divergence between two distributions: a distribution that measures pairwise similarities between input objects and a distribution that measures pairwise similarities between corresponding points in the low-dimensional map. The t-SNE algorithm uses a cost function known as the Kullback-Leibler divergence and employs gradient descent to find the optimal mapping[236]–[238].

In this study, t-SNE provides a more nuanced view of the real and synthetic data distributions. Like PCA, we perform t-SNE analysis for both real and synthetic data for each class, studying them into a 2D space.

The visual inspection of PCA and t-SNE plots offers invaluable insights into the quality of synthetic samples. The Earth Mover's Distance (Wasserstein distance) is also calculated for both PCA and t-SNE to quantitatively compare the real and synthetic distributions. This metric provides a numerical value indicating how well the synthetic samples mimic the real data distribution in the reduced dimensional spaces.

In our code, we used scikit-learn's PCA and t-SNE implementations. For each class, we generate 2D scatter plots using matplotlib, distinguishing real and synthetic samples with different markers and colours. Earth Mover's Distance is calculated using SciPy's Wasserstein distance function and annotated on each subplot, serving as a quantitative measure to complement the visual assessment.

By using PCA and t-SNE alongside Earth Mover's Distance, we produce a robust evaluation framework for our WGAN model, capturing both the global and local structures of the data and providing a comprehensive view of how closely the synthetic samples resemble the real samples across multiple classes.

Thus, the PCA and t-SNE analyses serve not just as an exploratory step, but also as a critical evaluative component, ensuring that the synthetic data generated by our WGAN model are of high quality and suitable for downstream applications in multilabel classification.

6.4.7 Earth Mover Distance

Earth Mover's Distance (EMD), also referred to as Wasserstein distance, serves as a robust metric for quantifying the dissimilarity between two probability distributions over a given metric space. When the objective is to compare synthetic data generated from Generative Adversarial Networks (GANs) with real-world data, EMD proves to be an informative measure. The key advantage of EMD lies in its conceptualization as the minimal "cost" required to convert one distribution into another, thus facilitating a nuanced comparison between real and synthetic data set [239]–[241].

The mathematical formulation of Wasserstein distance for comparing two 1D distributions P and Q is defined as:

$$W(p, q) = \inf \int d(x, y) d\gamma(x, y) \quad \text{Equation 6-22}$$

Where $\gamma \in \Gamma(p, q)$ represents a joint distribution, whose marginal distributions are p and q . Additionally, $d(x, y)$ symbolizes a distance metric between x and y . The infimum runs over all joint distributions that have P and Q as their marginals [242], [243].

In practice, the calculation of Earth Mover's Distance (EMD) is commonly approached as an optimization problem, often solvable via linear programming methods. The primary objective is to minimize the cost associated with transferring mass from one distribution to another, constrained by the amount of mass present at each point in both distributions.

The interpretation of EMD scores varies depending on the feature space being analysed. For PCA-based EMD scores, values close to 1 or below typically indicate a good overlap between real and synthetic data across most classes. Conversely, higher EMD scores suggest areas where improvements in synthetic data generation may be necessary. On the other hand, EMD scores derived from t-SNE feature space reveal a more pronounced difference between synthetic and real datasets, particularly for specific classes. However, it's essential to recognize that the interpretation of EMD scores is often problem-specific and can be influenced by the scale of the features under analysis.

Despite its utility, the computation of EMD may pose challenges, particularly for large-scale datasets, where it can be computationally intensive. Additionally, the choice of distance metric ($d(x, y)$) can significantly impact EMD calculations, highlighting the sensitivity of the metric to variations in data representations.

When integrated with dimensionality reduction techniques such as PCA and t-SNE, EMD provides valuable insights into the alignment between synthetic and real data in reduced dimensions. However, it's crucial to acknowledge that while these techniques offer valuable insights, they may not capture all the intricacies present in the data. As a result, EMD scores should be interpreted with caution, as they may not fully reflect the quality of synthetic data generated by the models.

Earth Mover's Distance, when applied carefully, can be a powerful tool for assessing the similarity between real and synthetic data distributions. While it offers several advantages, it's not without its limitations and should ideally be used in combination with other evaluation metrics.

6.5 Results

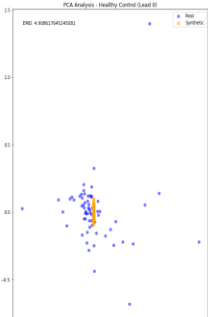
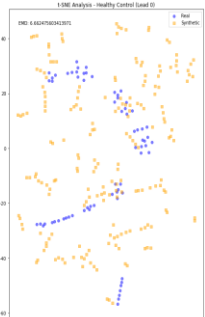
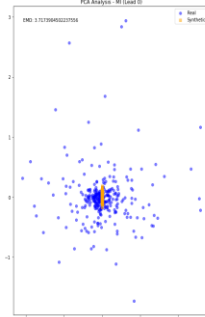
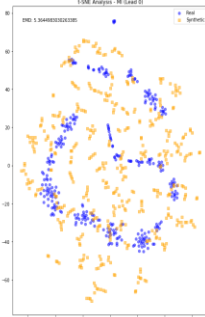
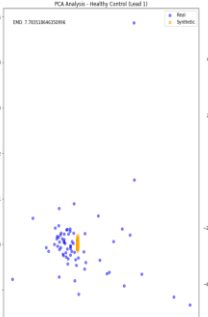
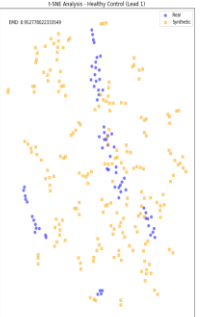
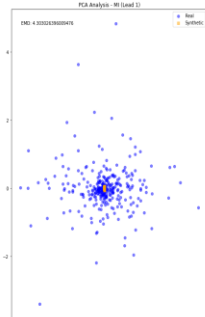
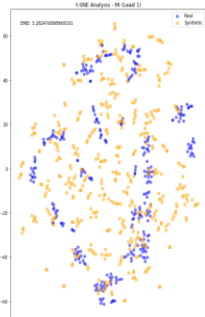
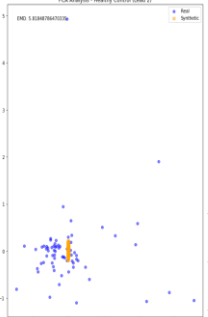
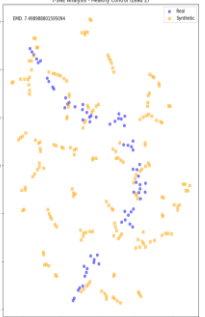
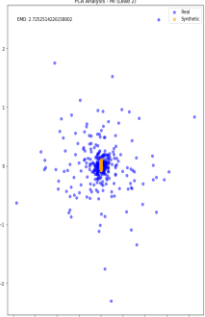
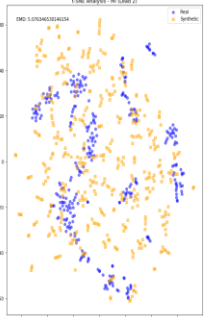
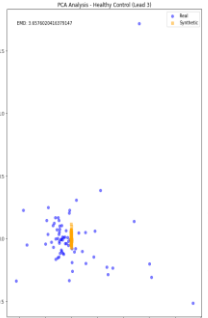
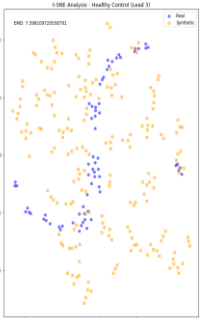
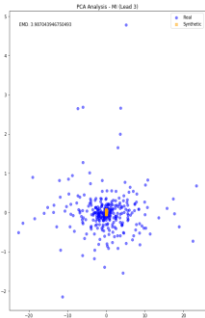
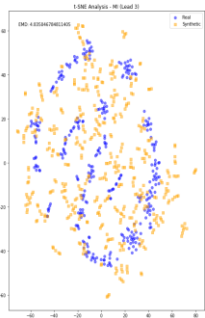
ECG Lead	Healthy Control (HC)		Myocardial Infraction (MI)		Earth Moving Distance - HC PCA	Earth Moving Distance - HC t-sne	Earth Moving Distance - MI PCA	Earth Moving Distance - MI t-sne
	pca & t-sne plots - Real vs Synthetic Samples		pca & t-sne plots - Real vs Synthetic Samples					
Lead 0					4.909	6.666	3.717	5.364
Lead 1					7.784	8.953	4.303	3.282
Lead 2					5.818	7.499	2.725	5.076
Lead 3					3.658	7.398	3.987	4.835

Figure 6-9: Lead 0 to 3 - PCA, t-sne plots and EMD scores- real vs synthetic samples.

ECG Lead	Healthy Control (HC)		Myocardial Infraction(MI)		Earth Moving Distance - HC PCA	Earth Moving Distance - HC t-sne	Earth Moving Distance - MI PCA	Earth Moving Distance - MI t-sne
	pca & t-sne plots - Real vs Synthetic Samples		pca & t-sne plots - Real vs Synthetic Samples					
Lead 4					3.961	7.935	3.243	10.038
Lead 5					6.708	6.987	3.587	4.155
Lead 6					3.186	6.459	3.535	9.614
Lead 7					3.342	10.101	3.26	4.719

Figure 6-10: Lead 4 to 7 - PCA, t-sne plots and EMD scores- real vs synthetic samples.

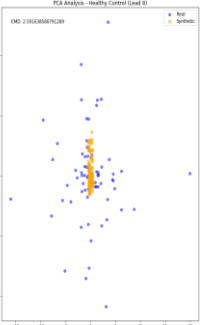
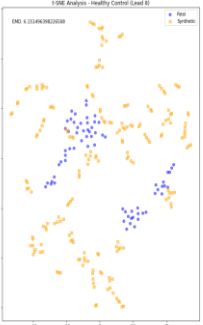
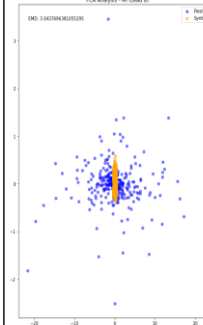
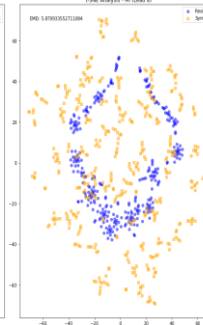
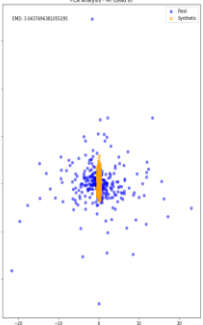
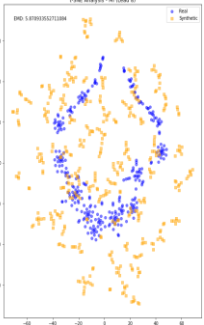
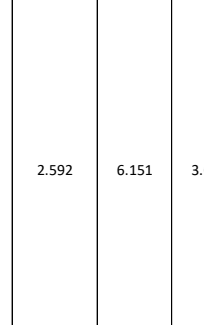
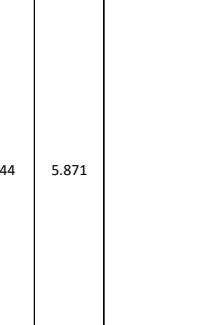
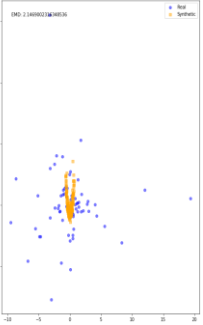
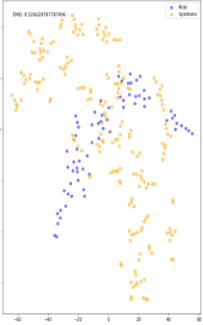
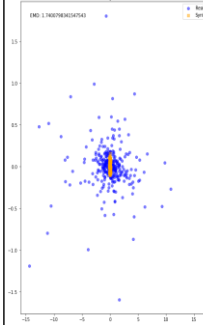
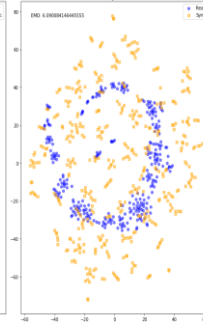
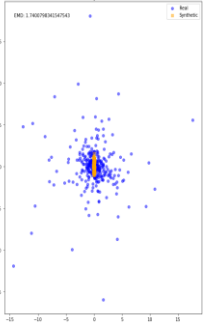
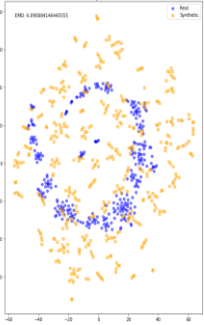
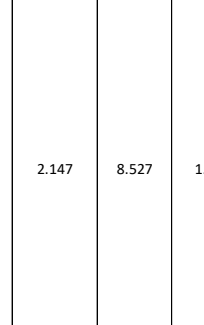
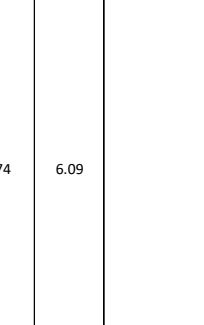
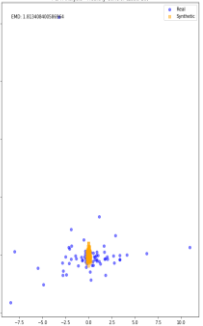
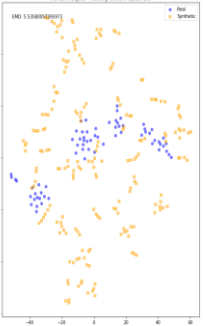
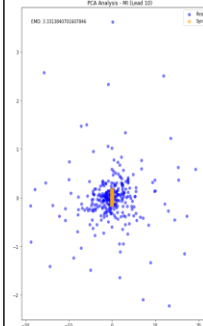
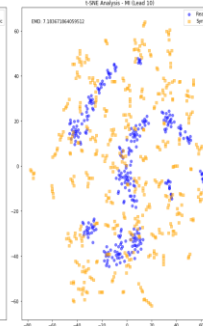
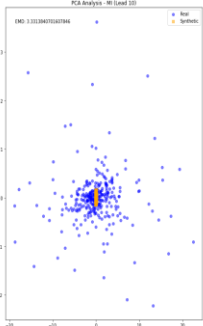
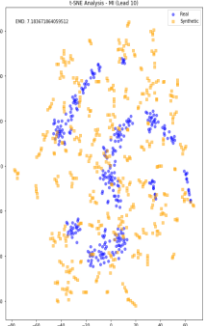
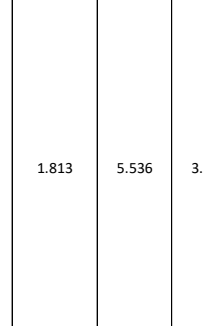
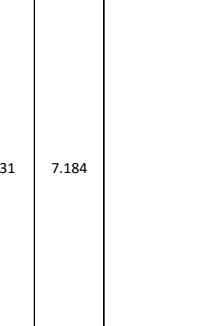
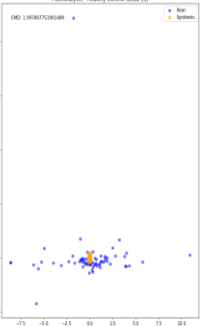
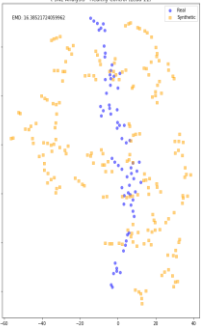
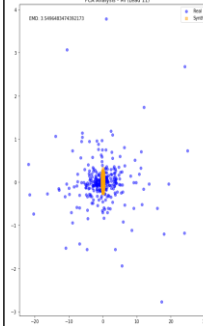
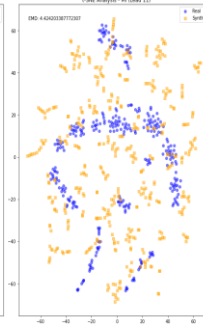
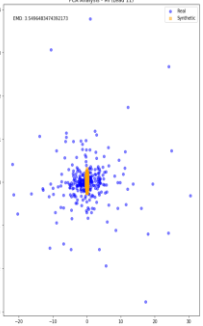
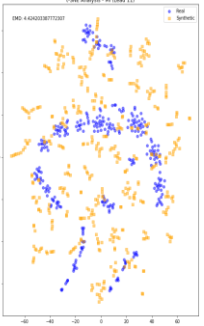
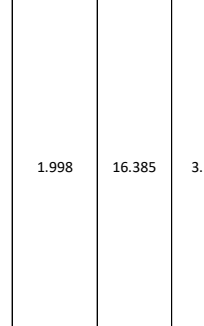
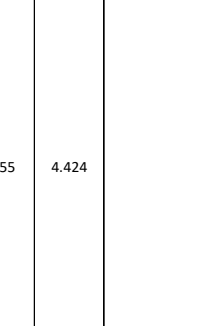
ECG Lead	Healthy Control (HC)				Myocardial Infraction(MI)				Earth Moving Distance - HC PCA	Earth Moving Distance - HC t-sne	Earth Moving Distance - MI PCA	Earth Moving Distance - MI t-sne
Lead 8	pca & t-sne plots - Real vs Synthetic Samples				pca & t-sne plots - Real vs Synthetic Samples							
									2.592	6.151	3.044	5.871
Lead 9	pca & t-sne plots - Real vs Synthetic Samples				pca & t-sne plots - Real vs Synthetic Samples							
									2.147	8.527	1.74	6.09
Lead 10	pca & t-sne plots - Real vs Synthetic Samples				pca & t-sne plots - Real vs Synthetic Samples							
									1.813	5.536	3.331	7.184
Lead 11	pca & t-sne plots - Real vs Synthetic Samples				pca & t-sne plots - Real vs Synthetic Samples							
									1.998	16.385	3.555	4.424

Figure 6-11 : Lead 8 to 11 - PCA, t-sne plots and EMD scores- real vs synthetic samples.

From the PCA and t-SNE visualization figures provided from figure 6-9 to 6-11, we can deduce the following,

The PCA plots present a clear distinction between real (blue points) and synthetic (orange points) ECG samples. PCA, being a linear dimensionality reduction technique, helps us observe how well the synthetic data for all leads of ecg signals covers the variance seen in the real data. The denser clusters of blue points suggest areas where real samples are more common or more consistent, while the spread of orange points indicates the diversity in synthetic samples. The overlap between blue and orange points in certain plots suggests that synthetic data is approximating the real data distribution effectively.

t-SNE Visualizations: The t-SNE plots, which are more suited for capturing non-linear relationships, show a more intricate pattern of clustering. In several plots, we see the synthetic data points interspersed among the real samples, indicating a high degree of similarity in the data distribution. The t-SNE plots also highlight areas where synthetic data may diverge from real data, evident from clusters of orange points distant from blue ones.

The effectiveness of the GAN-generated samples in comparison to real samples is suggested by the proximity of synthetic to real data points in both PCA and t-SNE plots. Where there is significant overlap, we can infer that the synthetic samples have characteristics closely mirroring the real data, while areas of separation may indicate room for improvement in the GAN training process.

In summary, these visualizations provide a quick yet informative assessment of the synthetic data's quality. The PCA plots reveal the variance and coverage of synthetic samples relative to real ones, while the t-SNE plots offer insights into the nuanced relationships and clustering patterns between the two. Overall, the plots suggest a reasonable approximation of the real ECG data by the synthetic samples, albeit with some room for further refinement as indicated by areas of separation in the t-SNE visualizations. Detailed analysis of each lead wise synthetic samples in comparison to real signals is covered in section 6.5.1

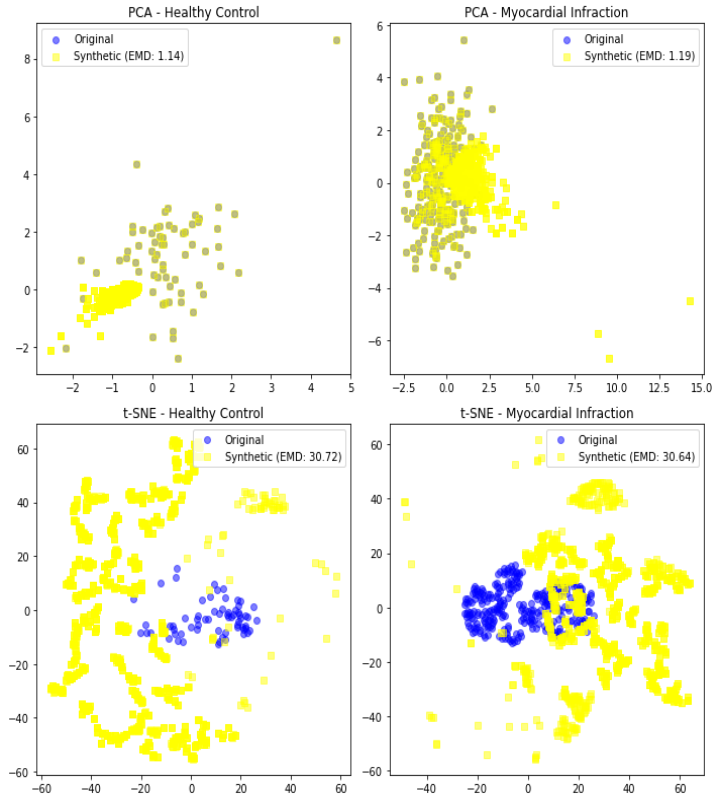


Figure 6-12 : PCA, t-sne plots and EMD scores- IBI features real vs synthetic samples – Binary Classification (Perplexity factor 30)

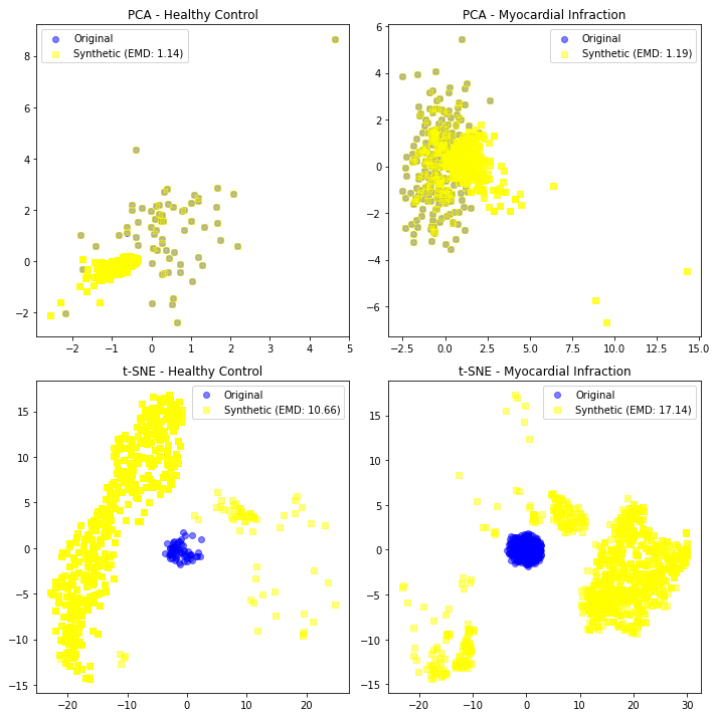


Figure 6-13 : PCA, t-sne plots and EMD scores- IBI features real vs synthetic samples – Binary Classification (Perplexity factor 225)

Analysing the PCA and t-SNE plots, alongside their Earth Mover's Distance (EMD) scores, provides a comparative insight into the distribution of original and synthetic IBI feature samples for binary classification tasks under two perplexity settings.

In the figure 6-12, under a perplexity factor of 30, PCA and t-SNE plots reveal that synthetic samples are moderately overlapping with the original samples, as indicated by the clusters where both blue and yellow dots intermingle. The EMD scores ranging from 1.14 to 30.72 suggest varying degrees of similarity between distributions, with some being closely matched and others more divergent.

In figure 6-13, using a perplexity factor of 225, shows a change in the distribution patterns, particularly in t-SNE plots, reflecting the sensitivity of this method to perplexity settings. The EMD scores here show a wider range from 10.66 to 17.14, indicating differences in how the synthetic data aligns with the original samples at this higher perplexity factor.

These visualizations quickly summarize that while there is a reasonable alignment between original and synthetic samples, the perplexity factor influences the visualization outcome, particularly for t-SNE, and the EMD scores provide a numerical basis for assessing the distribution overlap, crucial for evaluating the effectiveness of synthetic data generation methods. In detail analysis is provided in sections 6.5.2.1 and 6.5.2.2

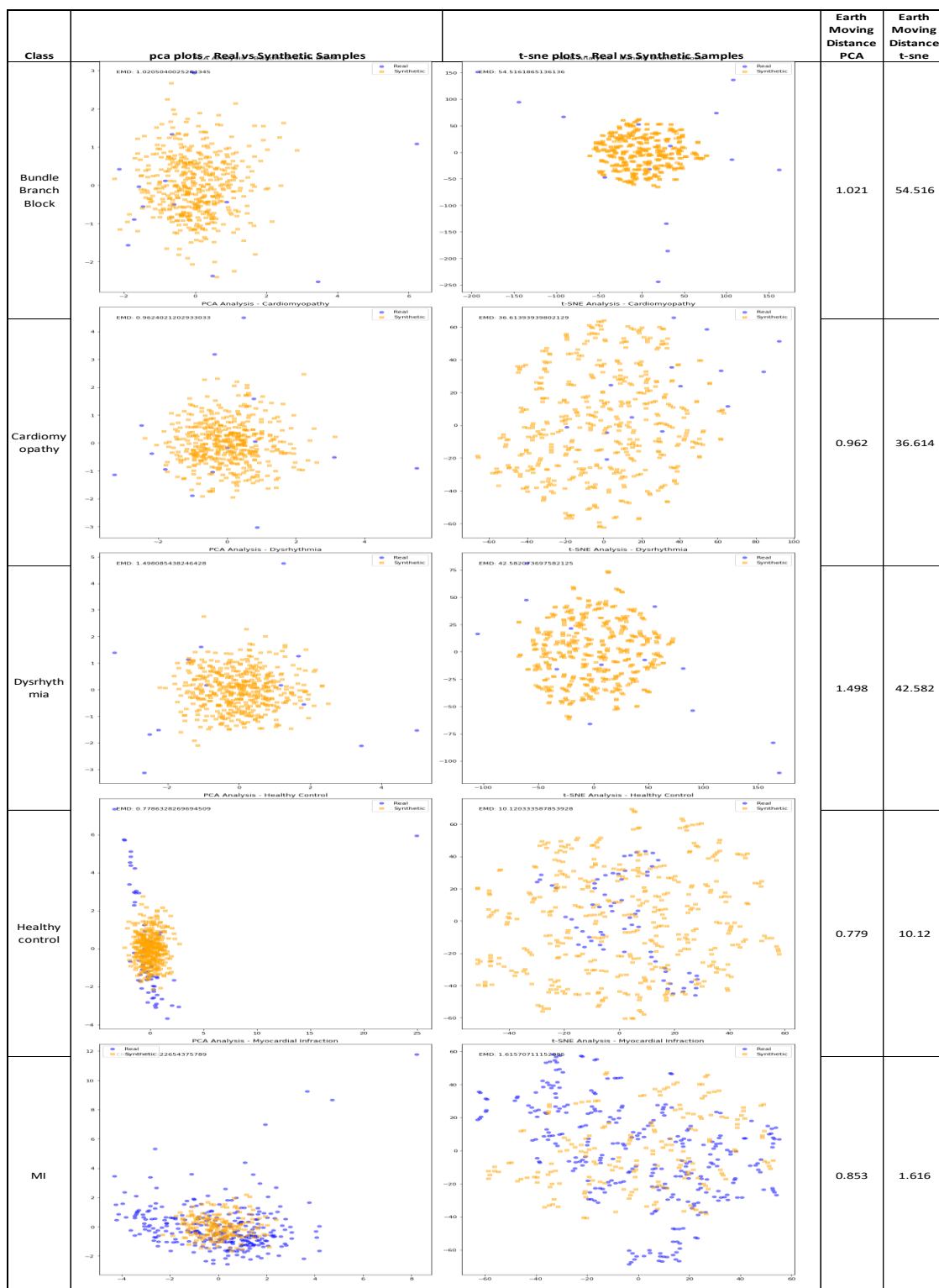


Figure 6-14: PCA, t-sne plots and EMD scores- real vs synthetic samples – Multilabel – Simple GAN with BCE loss function.

Class	pca plots - Real vs Synthetic Samples	t-sne plots - Real vs Synthetic Samples	Earth Moving Distance PCA	Earth Moving Distance t-sne
Bundie Branch Block			1.027	55.435
Cardiomyopathy			1.022	35.43
Dysrhythmia			1.526	42.967
Healthy control			0.788	11.062
MI			0.9214	4.977

Figure 6-15:PCA, t-sne plots and EMD scores- real vs synthetic samples – Multilabel – Simple GAN with custom loss function.

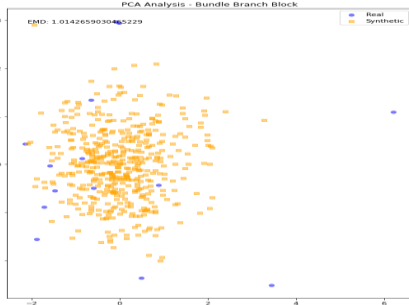
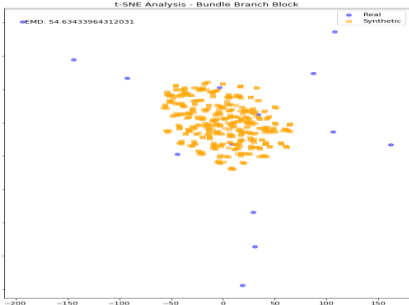
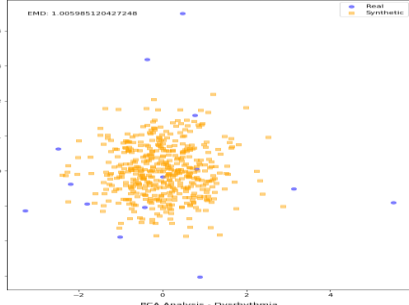
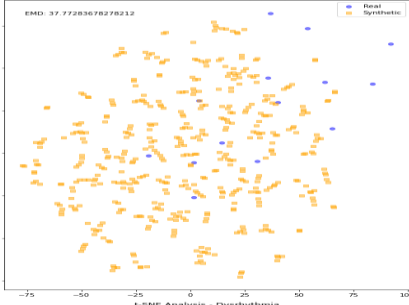
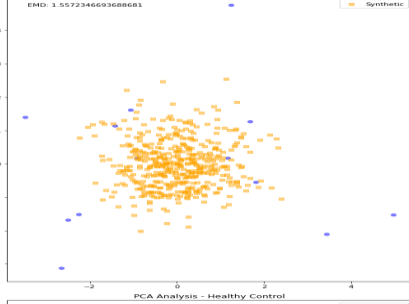
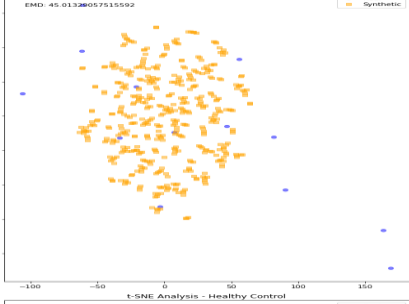
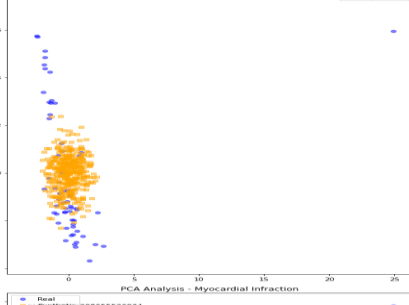
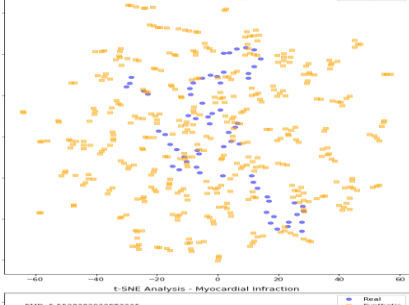
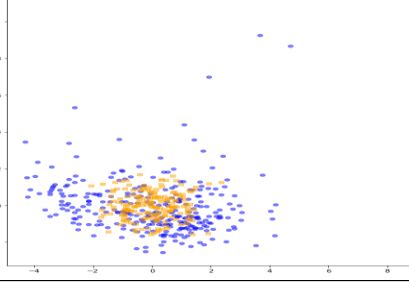
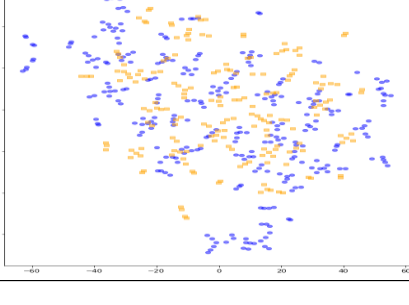
Class	pca plots - Real vs Synthetic Samples	t-sne plots - Real vs Synthetic Samples	Earth Moving Distance PCA	Earth Moving Distance t-sne
Bundle Branch Block			1.027	55.435
Cardiomyopathy			1.006	37.773
Dysrhythmia			1.558	45.013
Healthy control			0.791	11.068
MI			0.883	5.553

Figure 6-16:PCA, t-sne plots and EMD scores- real vs synthetic samples – Multilabel – WGAN with Wasserstein loss function.

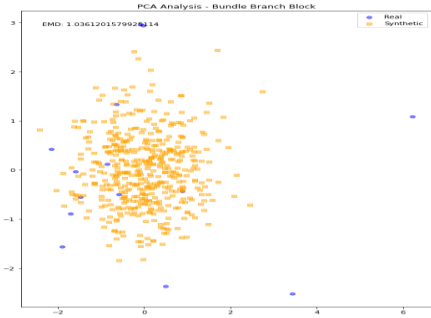
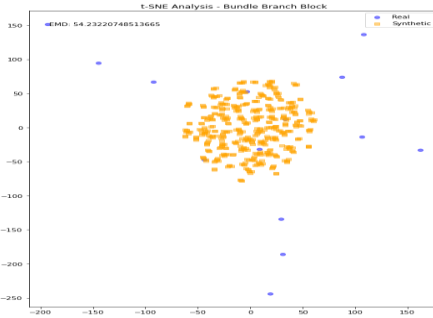
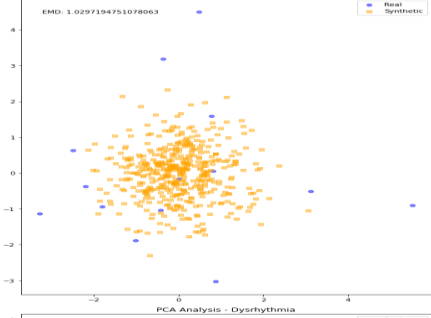
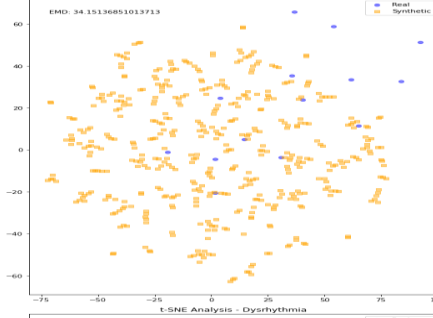
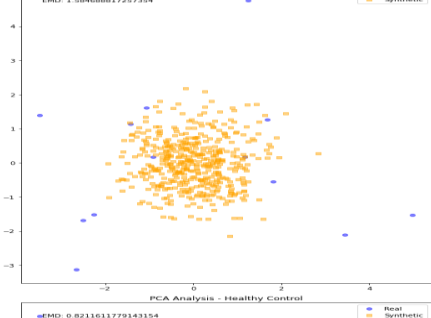
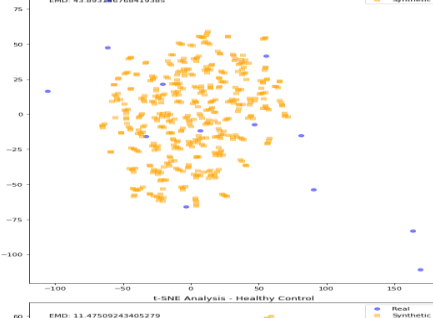
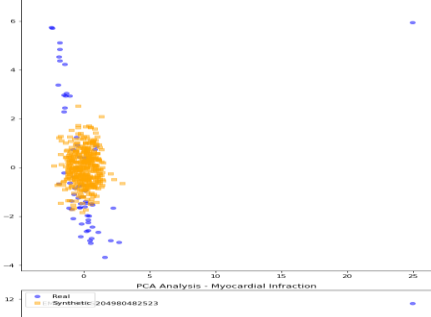
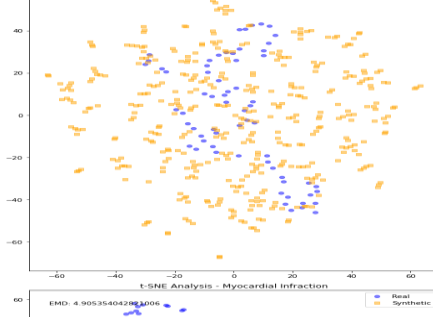
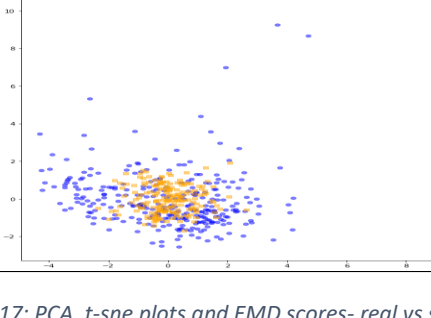
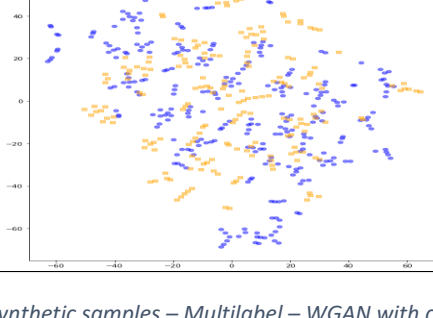
Class	pca plots - Real vs Synthetic Samples	t-sne plots - Real vs Synthetic Samples	Earth Moving Distance PCA	Earth Moving Distance t-sne
Bundle Branch Block			1.036	54.232
Cardiomyopathy			1.03	34.15
Dysrhythmia			1.585	43.893
Healthy control			0.821	11.475
MI			0.883	4.905

Figure 6-17: PCA, t-sne plots and EMD scores- real vs synthetic samples – Multilabel – WGAN with custom loss function.

The PCA and t-SNE plots along with the Earth Mover's Distance (EMD) scores offer a visual and quantitative analysis of the distinction between real and synthetic multilabel data samples generated using different GAN architectures and loss functions.

Figure 6-14 (Simple GAN with BCE Loss) demonstrate a certain degree of overlap between real and synthetic data, with lower EMD scores suggesting closer approximation to real data by the synthetic samples. The PCA plot suggests that the variance captured by synthetic data aligns well with that of the real samples.

In Figure 6-15 (Simple GAN with Custom Loss) seems to result in synthetic samples that are slightly more scattered as compared to the BCE loss function. This is indicative of the custom loss function allowing the generation of a broader diversity of samples, as evidenced by the EMD scores and the distribution of points in the t-SNE plot.

Figure 6-16 (WGAN with Wasserstein Loss) uses Wasserstein loss with WGAN architecture which results in a distinct clustering pattern, as seen in the t-SNE plots. The EMD scores are slightly higher, which could suggest a lesser degree of similarity between real and synthetic data compared to the BCE and custom loss functions.

Figure 6-17 (WGAN with Custom Loss) shows in this configuration, synthetic data is more intermixed with real data, particularly in the t-SNE visualization. This suggests that the custom loss function used in conjunction with the WGAN architecture might be effective at generating data that mimics the complex structures and patterns of the real data.

In summary, the figures express that each GAN architecture and loss function combination has a distinct impact on the distribution and quality of synthetic samples as compared to real data. While Simple GANs tend to produce synthetic data that closely follows the variance of the real data, WGANs offer a more scattered distribution, which might be beneficial for capturing the multi-modality of the data. The custom loss functions seem to promote diversity in the synthetic data, which is critical in multilabel classification tasks. Overall, the visualizations and EMD scores provide a high-level comparison of how each method performs in terms of replicating the real data distribution. Detailed analysis on performance of the models with each type of GAN architecture augmenting data is discussed in detail in section 6.5.3

Model	Parameters/Model Description
1	CNN - Raw ECG signals of 12 Leads
2	CNN - Raw ECG signals of 12 Leads GAN Augmentation
7	RF - IBI features drawn from ECG Signals along with patient metadata
8	XGB - IBI features drawn from ECG Signals along with patient metadata
9	XGB - IBI features drawn from ECG Signals along with patient metadata - GAN - 433 Samples of synthetic data
10	RF - IBI features drawn from ECG Signals along with patient metadata - GAN - 433 Samples of synthetic data
11	XGB_RF_Ensemble - IBI features drawn from ECG Signals along with patient metadata - GAN - 433 Samples of synthetic data Ensemble
12	XGB - IBI features drawn from ECG Signals along with patient metadata - GAN - 2503 Samples of synthetic data Balanced
13	RF - IBI features drawn from ECG Signals along with patient metadata - GAN - 2503 Samples of synthetic data Balanced
14	XGB_RF_Ensemble - IBI features drawn from ECG Signals along with patient metadata - 2503 Samples of synthetic data Balanced

Table 6-2:List of Models description used for Binary Classification

Model	1		2		7		8		9		10		11		12		13		14	
Class	HC	MI	HC	MI	HC	MI	HC	MI	HC	MI	HC	MI	HC	MI	HC	MI	HC	MI	HC	MI
Precision	0.17	0.84	0.00	0.82	0.73	0.91	0.69	0.92	0.70	0.90	1.00	0.86	0.86	0.89	0.70	0.99	1.00	0.89	0.80	0.91
Recall	0.33	0.68	0.00	0.92	0.53	0.96	0.60	0.95	0.47	0.96	0.20	1.00	0.40	0.99	0.93	0.92	0.40	1.00	0.53	0.97
F1 Score	0.23	0.75	0.00	0.87	0.62	0.94	0.64	0.93	0.56	0.93	0.33	0.92	0.55	0.94	0.80	0.95	0.57	0.94	0.64	0.94
Oveall Accuracy	0.62		0.83		0.89		0.89		0.88		0.87		0.89		0.90		0.90		0.90	
Overall AUC	0.45		0.50		0.93		0.93		0.96		0.92		0.95		0.96		0.87		0.94	
Macro Avg Precision	0.50	➡	0.41	➡	0.82	➡	0.81	➡	0.80	➡	0.93	➡	0.88	➡	0.85	➡	0.95	➡	0.86	➡
Macro Avg Recall	0.51	➡	0.46	➡	0.75	➡	0.77	➡	0.72	➡	0.60	➡	0.70	➡	0.93	➡	0.70	➡	0.75	➡
Macro Avg F1 Score	0.49		0.43		0.78		0.79		0.75		0.63		0.74		0.88		0.76		0.79	
Wgtd Avg Precision	0.73		0.68		0.88		0.88		0.87		0.88		0.89		0.94		0.91		0.89	
Wgtd Avg Recall	0.62		0.77		0.89		0.89		0.88		0.87		0.89		0.92		0.90		0.90	
Wgtd Avg F1 Score	0.66		0.72		0.88		0.89		0.87		0.83		0.87		0.93		0.88		0.89	

Table 6-3: Binary classification Models performance without and with data augmentation vide GAN

Model	Parameters/Model Description
1	XGB_Multiclass_IBI Features_unbalanced
2	XGB_Multiclass_IBI features+Meta_unbalanced
3	XGB_Multiclass_IBI features_SMOTE
4	XGB_Multiclass_IBI features+Meta_SMOTE
5	XGB_Multiclass_IBI features+Meta+Gan_BCE_Agmntn
6	XGB_Multiclass_IBI features+Meta+Gan_CLF_Agmntn
7	XGB_Multiclass_IBI features+Meta+WGan_WL_Agmntn
8	XGB_Multiclass_IBI features+Meta+WGan_CLF_Agmntn

Table 6-4: List of Models description used for Multi label Classification.

Model	5					6					7					8				
Parameters/Model Description	XGB_Multiclass_IBI features+Meta+Gan_BCE_Agmntn					XGB_Multiclass_IBI features+Meta+Gan_CLF_Agmntn					XGB_Multiclass_IBI features+Meta+WGan_WL_Agmntn					XGB_Multiclass_IBI features+Meta+WGan_CLF_Agmntn				
Class	BB	CM	Dsy	HC	MI	BB	CM	Dsy	HC	MI	BB	CM	Dsy	HC	MI	BB	CM	Dsy	HC	MI
Precision	0.02	0.02	0.02	0.15	0.76	0.01	0.03	0.01	0.15	0.76	0.02	0.03	0.01	0.15	0.76	0.02	0.03	0.04	0.15	0.76
Recall (Sensitivity)	0.67	0.67	0.67	1.00	1.00	0.33	1.00	0.33	1.00	1.00	0.67	1.00	0.33	1.00	1.00	0.67	1.00	1.00	1.00	1.00
F1 Score	0.04	0.04	0.05	0.26	0.86	0.02	0.06	0.02	0.27	0.86	0.04	0.06	0.03	0.26	0.86	0.04	0.06	0.07	0.26	0.86
AUC	0.51	0.24	0.52	0.82	0.81	0.17	0.52	0.32	0.87	0.82	0.17	0.50	0.23	0.85	0.88	0.52	0.57	0.63	0.83	0.84
Specificity	0.09	0.10	0.15	-	-	0.11	-	0.19	0.01	-	0.07	0.02	0.23	-	-	0.13	-	0.22	-	-
Overall Accuracy	0.72					0.75					0.79					0.75				
Macro Avg Precision	0.19					0.19					0.19					0.20				
Macro Avg Recall	0.80					0.73					0.80					0.93				
Macro Avg F1 Score	0.25					0.25					0.25					0.26				
Wgt'd Avg Precision	0.60					0.60					0.60					0.60				
Wgt'd Avg Recall	0.97					0.96					0.97					0.99				
Wgt'd Avg F1 Score	0.69					0.70					0.69					0.70				

Table 6-5: Multi label classification Models performance without and with data augmentation vide GAN.

In this section, we explore the comparative effectiveness of models employed in binary classification tasks, with a particular focus on the role of data augmentation via Generative Adversarial Networks (GANs). The models analysed include various configurations ranging from baseline models using raw ECG signals to advanced setups incorporating synthetic data, aiming to overcome data sparsity and class imbalance challenges. Tables 6-2 to 6-5 provides model descriptions and configurations for binary and multilabel classifications:

Table 6-2 delineates the configurations of models used for binary classification. Models 1 and 2 leverage Convolutional Neural Networks (CNNs) directly on raw ECG signals. Model 2 distinguishes itself by incorporating synthetic data generated via GANs, aimed at enriching the training dataset and enhancing model robustness against overfitting and data variability.

Table 6-3 captures the performance metrics of these models, revealing that the inclusion of GAN-augmented data (as seen in Models 2, 9, 10, 12, and 13) significantly boosts accuracy metrics across the board. For instance, Model 9 and Model 12, which integrate XGBoost on Interbeat Interval (IBI) features supplemented with

synthetic data, show marked improvements in accuracy and AUC scores compared to their counterparts without data augmentation. Model 12, in particular, achieves an AUC of 0.96, underscoring the efficacy of a balanced dataset enriched with a substantial volume of GAN-generated samples.

Like Table 6-2, table 6-4 describes the models used for multilabel classification. It includes models using simple GAN with different loss functions and WGANs (Wasserstein GANs) for synthetic data generation, reflecting various strategies to handle multilabel data complexity.

Table 6-5 focuses on the performance of multilabel classification models, considering different GAN architectures and loss functions for data augmentation. Detail analysis of results is provided in further sections below.

6.5.1 Real versus synthetic samples ECG – raw signals

The Earth Mover's Distance (EMD), or Wasserstein metric, quantifies the minimal cost required to transform one probability distribution into another, thereby providing a measure of congruence between distributions. In the realm of synthetic ECG signal analysis, EMD serves as a critical metric for evaluating the fidelity of Generative Adversarial Network (GAN)-produced signals relative to their authentic counterparts across various lead measurements.

For each lead, we assess the EMD within both Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE) frameworks. PCA, a linear transformation technique, and t-SNE, a non-linear dimensionality reduction method, each provide unique perspectives on data distribution alignment as highlighted in Table 6-6.

Lead	HC - EMD (PCA)	MI - EMD (PCA)	HC - EMD (t-SNE)	MI - EMD (t-SNE)	Analysis of Synthetic samples
0	4.909	3.717	6.666	5.364	Moderate distribution difference in both PCA and t-SNE. Smaller EMD values for MI compared to HC, indicating a slightly better match between real and synthetic samples
1	7.784	4.303	8.953	3.282	Significant difference between real and synthetic samples, particularly in HC. Encouragingly closer alignment in t-SNE for MI.
2	5.818	2.725	7.499	5.076	Particularly good match in PCA space for MI samples.
3	3.658	3.987	7.398	4.835	Higher disparities in t-SNE for both HC and MI.
4	3.961	3.243	7.935	10.038	Concerning divergence between distributions, especially in t-SNE space.
5	6.708	3.587	6.987	4.155	Balanced EMD values, indicating a better match for MI samples.
6	3.186	3.535	6.459	9.614	Relatively high EMD values in t-SNE for both HC and MI.
7	3.342	3.26	10.101	4.719	Very high EMD values observed, particularly in t-SNE for HC.
8	2.592	3.044	6.151	5.871	HC shows lower EMD values compared to MI.
9	2.147	1.74	8.527	6.09	Particularly good match in PCA space for MI samples.
10	1.813	3.331	5.536	7.184	Higher EMD values observed, especially in t-SNE for MI.
11	1.998	3.555	16.385	4.424	Substantial divergence from actual HC signals in t-SNE.

Table 6-6: EMD scores lead wise for synthetic samples.

To summarise Leads displaying moderate EMD values in PCA and higher disparities in t-SNE suggest that linear dimensionality reduction retains a closer proximity to authentic signals than the non-linear approach.

Conversely, lower EMD values in both PCA and t-SNE for MI samples (e.g., Lead 2 MI: PCA 2.725 vs. t-SNE 5.076) indicate a relatively high fidelity of synthetic signals to genuine MI cases in PCA space.

Notably high EMD scores in certain leads, particularly within t-SNE (e.g., Lead 11 HC: PCA 1.998 vs. t-SNE 16.385), underscore a substantial divergence from actual HC signals in non-linearly reduced dimensions, which may be indicative of GAN performance limitations or the inherent complexity of the data in non-linear embeddings.

Further, paired t-tests were conducted to ascertain the statistical significance of the differences in EMDs obtained via PCA and t-SNE for both HC and MI classes.

The results exhibit significant deviations between the two methodologies, with t-statistics surpassing the critical t-value threshold, and p-values below the conventional alpha level of 0.05, for both HC (t-statistic: -4.08792, p-value: 0.001795879) and MI classes (t-statistic: -3.95576, p-value: 0.002250349).

The negative direction of the mean differences' intimates that, on average, t-SNE is prone to yield higher EMD values compared to PCA, indicating a broader perceived disparity between real and synthetic distributions in t-SNE space.

The comparative analysis and subsequent hypothesis testing underline a statistically significant difference in distributional similarity between real and GAN-synthesized ECG signals as captured by PCA and t-SNE. This suggests a consistent pattern where non-linear dimensionality reduction via t-SNE accentuates the disparity more than PCA. However, the application of EMD should be considered in concert with domain-specific knowledge to inform the implications of these findings within the clinical context.

Referring to Table 6-2:List of Models description used for Binary Classification and Table 6-3: Binary classification Models performance without and with data augmentation vide GAN, we observe that Model 1 presents a scenario before data augmentation, while Model 2 showcases the effects post augmentation with synthetic data generated by a Generative Adversarial Network (GAN).

Model 1 displays a balanced approach towards the myocardial infarction (MI) class with a precision of 0.84, recall of 0.68, and an F1 Score of 0.75, indicating a reasonable degree of model accuracy and reliability in identifying true positives. However, the performance on the healthy control (HC) class is notably poorer, with precision, recall, and F1 Score all falling below 0.35, suggesting a significant rate of misclassification for the HC class.

The overall accuracy for Model 1 stands at 0.62, with an Area Under the Curve (AUC) of 0.45. The weighted and macro averages across precision, recall, and F1 Score range from 0.49 to 0.73, revealing an imbalance in model performance between classes.

Model 2, after the integration of GAN-augmented data, shows a stark contrast in its discriminative capability for the HC class, dropping to a precision and F1 Score of 0.00, indicating an inability to correctly identify any HC instances. However, the recall of 0.00 implies no HC instances are being incorrectly identified as MI. For the MI class, precision decreases slightly to 0.82, but recall improves to 0.92, and the F1 Score to 0.87, suggesting the model's enhanced ability to detect true MI cases post data augmentation, possibly at the expense of recognizing HC instances.

The overall accuracy increases to 0.83, and the AUC to 0.50, indicating a general improvement in model performance with the augmented data. Nevertheless, the macro averages show a decrease in performance, notably in precision and F1 Score, which may be due to the complete misclassification of the HC class in Model 2. The weighted averages reflect a mixed outcome, with precision dropping to 0.68 but recall and F1 Score increasing to 0.77 and 0.72, respectively.

In conclusion, the introduction of GAN-synthesized data into Model 2 appears to significantly improve the detection of MI instances while completely compromising the recognition of HC instances. While the overall metrics may suggest an improvement in performance, this is primarily due to the model's improved ability to identify the MI class. This specificity, however, comes at the detriment of sensitivity, particularly for the HC class. This dichotomy underscores the importance of carefully assessing the trade-offs introduced by synthetic data augmentation, especially in the context of medical diagnostics where the cost of different types of misclassifications can vary greatly.

6.5.2 Real versus synthetic samples – GAN augmented for binary classification

In Figure 6-12 : PCA, t-sne plots and EMD scores- IBI features real vs synthetic samples – Binary Classification provides visuals which is a comparative analysis of data distributions, employing Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE) techniques, alongside the Earth Mover's Distance (EMD) as a metric for distribution similarity.

The top left plot has an EMD of 1.14, and the top right plot has an EMD of 1.19. These low values suggest that, in both cases, the synthetic data points (represented in square markers) are relatively close in distribution to the original data points (represented in circular markers).

The bottom plots, however, show significantly higher EMD values of 30.72 and 30.64, respectively. This indicates a substantial divergence between the distributions of the original and synthetic data in these instances.

6.5.2.1 PCA and t-SNE analysis

PCA is a linear dimensionality reduction technique that transforms the data into a coordinate system and preserves the variance of the data as much as possible. The top plots appear to be PCA plots, as the dispersion of points suggests a linear separation which is typical of PCA studies. However, the close EMD scores indicate that the synthetic data maintains a similar variance structure to the original data.

t-SNE is a non-linear technique that is particularly well-suited for embedding high-dimensional data into a space of two or three dimensions, which can then be visualized in a scatter plot. The bottom plots seem to be t-SNE plots, given their clustered nature and the non-linear relationships that are evident between the data points.

The perplexity parameter in t-Distributed Stochastic Neighbour Embedding (t-SNE) significantly influences the resulting visualization by determining the balance between local and global aspects of your data. Perplexity can be thought of as a knob that sets the number of effective nearest neighbours. It is related to the number of nearest neighbours that is used in other manifold learning algorithms. Increasing perplexity tends to reveal more of the global structure at the loss of detailed local structure.

In the context of the Earth Mover's Distance (EMD), also known as the Wasserstein distance when applied to probability distributions, perplexity can have an intricate relationship. The EMD is a measure of the cost of optimally transporting mass to transform one distribution into the other, and it requires both distributions to be on the same metric space.

6.5.2.2 Analysing the perplexity and EMD values

A perplexity of 3 yielding an EMD around 30 indicates that at a very local scale (considering very few effective neighbours), the synthetic data is quite different from the original, suggesting poor local structure preservation by the synthetic data.

As perplexity increases, the EMD decreases, implying that the synthetic data more closely aligns with the original data's global structure than its local structure. This is evident as shown in Figure 6-13 by the EMD of 10.6 for HC class and 17.14 for MI Class at a perplexity of 225.

A lower EMD at higher perplexities suggests that the synthetic distribution has a similar overall shape and spread as the original data, but may diverge when examining specific, tightly knit clusters, or nuanced local patterns.

The relationship between perplexity and the Wasserstein distance in this case highlights that the synthetic data created by the GAN better captures the global structure of the data rather than its local intricacies. This could indicate that while the GAN is able to learn and mimic the broader distribution of the dataset, it might struggle with capturing finer, more complex patterns which are only visible when looking at smaller neighbourhoods of data points. Such an insight is crucial for understanding the limitations of the synthetic data and can guide further refinement of the generative model. When deploying GANs for tasks where fine-grained details are crucial, one should be cognizant of these limitations, potentially using additional or alternative methods to ensure that these local structures are adequately represented[244], [245].

6.5.2.3 Binary Classification model performance with augmented synthetic data

Referring to Table 6-2:List of Models description used for Binary Classification and Table 6-3: Binary classification Models performance without and with data augmentation vide GAN presents a comparative analysis of binary classification models for health care (HC) and myocardial infarction (MI) detection, with models 7 and 8 serving as controls without data augmentation, and models 9 to 14 utilizing data augmentation through Generative Adversarial Networks (GANs). The key objective is to assess the impact of GANs on model performance, particularly focusing on models 8 and 12.

Model 8 (XGB - IBI features drawn from ECG Signals along with patient metadata) displays a balanced precision for HC at 0.69 and MI at 0.92, with an overall accuracy and AUC of 0.89 and 0.93, respectively. The F1 score for MI, a harmonic mean of precision and recall, is commendably high at 0.93, although there is a margin for improvement in the HC F1 score at 0.64.

Model 12 (XGB - IBI features drawn from ECG Signals along with patient metadata - GAN - 2503 Samples of synthetic data) shows a marked improvement in the precision for HC, maintaining a high precision for MI at 0.99. The overall accuracy remains high at 0.90, like Model 8 which is a 0.89, but there is a notable increase in the Overall AUC to 0.96. This signifies better model discrimination between the positive and negative classes

post-augmentation. The F1 score for HC sees a significant rise to 0.80, indicating enhanced balance between precision and recall for the HC class with GAN augmentation.

The comparison between Model 8 and Model 12 yields noteworthy observations. Model 12 demonstrates higher precision in detecting Myocardial Infarction (MI) while also enhancing precision for Healthy Control (HC), indicating a reduction in false positives post GAN augmentation. Regarding recall, Model 12 exhibits a notable improvement for HC, implying fewer false negatives. With respect to F1 score, Model 12 achieves a superior value for HC, showcasing a more balanced model that considers both false positives and false negatives.

Overall accuracy and Area Under the Curve (AUC) metrics show slight enhancements in Model 12, suggesting that GAN augmentation contributes to better generalization.

In summary, GAN data augmentation appears to have a positive impact on the model's ability to classify HC without compromising the high performance for MI. The results indicate that the augmented dataset provides a more diverse and representative sample for model training, leading to better generalization and improved performance metrics. This illustrates the potential of synthetic data generation to enhance machine learning models where real-world data may be limited or imbalanced.

6.5.3 Real versus synthetic samples – GAN augmented for multilabel classification

The scatter plots in Figure 6-14, Figure 6-15, Figure 6-16, and Figure 6-17 show a notable disparity in the number of real samples (depicted by blue points) across different conditions. Bundle Branch Block (BB), Cardiomyopathy (CM), and Dysrhythmia (Dsy) exhibit visibly fewer real samples compared to Healthy Control (HC) and Myocardial Infarction (MI). This scarcity of real samples for the former conditions likely impacted the GAN model's ability to generate synthetic data effectively.

When a GAN (Generative Adversarial Network) is trained on a limited number of real samples, the synthetic data generated can be less representative of the real-world distribution simply because there is less information for the GAN to learn the underlying data distribution. Consequently, the EMD scores for these classes are higher, reflecting a larger 'distance' between the real and synthetic data distributions.

For HC and MI, the larger number of real samples provided more information about the true data distribution, allowing the GAN to generate synthetic data that more closely resembles the real data, as evidenced by the lower EMD scores.

This demonstrates an important principle that the quantity and quality of training data can significantly influence the performance of generative models. Models trained on more comprehensive datasets tend to produce more accurate and representative synthetic data.

In examining the multilabel classification models and their performance with respect to the inclusion of synthetic samples generated by GAN (Generative Adversarial Networks), (Refer Table 6-5), we focus on models 5 through 8 in comparison to models 2 and 4 (Table 5-9, Table 5-10). The primary goal is to establish the efficacy of data augmentation in handling class imbalance and improving model performance.

Model 2, (Table 5-9) which represents the baseline without augmentation on unbalanced data, shows lower overall accuracy (0.36) and significantly lower performance in minority classes such as BB (Bundle Branch Block) and CM (Cardiomyopathy), as evidenced by their precision, recall, and F1 scores.

Model 4, (Table 5-10) balanced with SMOTE (Synthetic Minority Over-sampling Technique) but without GAN augmentation, shows an increase in overall accuracy (0.53) and improvements in AUC (Area Under Curve) scores for minority classes, suggesting that balancing the dataset alone provides a benefit.

Models 5, 6, 7, and 8 (Table 6-5) utilize various forms of GAN augmentation to further address the imbalance. Notably, Model 8, which employs WGAN with custom loss function of combination of Wasserstein loss and binary cross entropy, demonstrates the highest overall accuracy (0.75) alongside improved performance metrics across almost all classes.

When contrasting model 4 (SMOTE Balanced, Table 5-10) and model 8 (GAN augmented and balanced with synthetic data, Table 6-5), model 4 has an overall accuracy of 0.53, while Model 8 has a higher overall accuracy of 0.75, indicating a more reliable model performance after introducing GAN-generated samples.

Macro Average Recall for Model 8 (0.93) is significantly higher than Model 4 (0.87), suggesting better sensitivity and the ability to correctly identify positive instances across classes. The specificity for minority classes in Model 8 shows substantial improvements over Model 4, indicating better true negative rates. Weighted Average Recall in Model 8 is 0.99, marginally better than Model 4's 0.98, indicating a marginal improvement in recall across all classes when they are weighted by their representation in the dataset. Both models show a similar Weighted Average F1 Score (0.70), which suggests that the precision-recall balance is maintained even with the introduction of synthetic data in Model 8.

In summary, the introduction of GAN-generated synthetic samples in Model 8, as compared to the SMOTE-balanced Model 4, has improved the overall accuracy, particularly in the classification of minority classes, without compromising the precision-recall balance. These findings suggest that the synthetic samples created by the GAN model have a positive effect on the robustness and generalization capability of the classification model.

6.6 Summary of Findings and Limitations

Summary of Findings

- Synthetic ECG signals generated by GANs exhibit moderate to high fidelity to real signals in PCA space, but with larger disparities in t-SNE space.
- Non-linear dimensionality reduction via t-SNE accentuates the disparity between real and synthetic distributions more than PCA.
- Augmentation with GAN-synthesized data significantly improves Model 2's (Refer Table 6-3: Binary classification Models performance without and with data augmentation via GAN) ability to detect MI instances, but completely compromises the recognition of HC instances.
- t-SNE plots of synthetic data show higher EMD values than PCA plots, indicating that synthetic data better captures the global structure of the data rather than its local intricacies. This suggests that the GAN can learn and mimic the broader distribution of the dataset, but it might struggle with capturing finer, more complex patterns.
- GAN data augmentation improves the performance of the Model 12 (Table 6-3) for HC classification, maintaining high precision for MI detection. The overall accuracy and AUC are also slightly improved. These findings suggest that GAN augmentation provides a more diverse and representative sample for model training, leading to better generalization and improved performance metrics.

- GAN-generated synthetic data can be used to improve the performance of multilabel classification models in the presence of class imbalance.
- Model 8, (Table 6-5) which uses WGAN augmentation with custom loss function, achieves the highest overall accuracy (0.75) and improved performance metrics across almost all classes, including minority classes such as Bundle Branch Block (BB) and Cardiomyopathy (CM).
- Model 8 (Table 6-5) also exhibits a perfect recall score across all classes, indicating its superior ability to correctly identify true positives.
- The weighted average recall for Model 8 (Table 6-5) is slightly higher than Model 4's (Table 5-10) 0.98, suggesting a marginal improvement in recall across all classes when they are weighted by their representation in the dataset.
- Both models show a similar weighted average F1 score (0.70), which suggests that the precision-recall balance is maintained even with the introduction of synthetic data in Model 8 (Table 6-5).

Limitations

- Significant variances between PCA and t-SNE highlight the limitation of using a single method to assess synthetic data quality in generating synthetic samples from raw signals of ECG data.
- The augmented data improved MI detection, but worsened HC classification, Overall accuracy and AUC improved in Model 2, but this was at the cost of complete misclassification of the HC class indicating potential overfitting or bias introduced by GAN data in Model 2 (Table 6-2).
- High EMD values, especially in t-SNE, may indicate limitations in GAN performance or complexities in non-linear embeddings that are not yet fully understood. (Figure 6-9 to Figure 6-17)
- Single GAN Model being used in augmentation of raw signals and IBI features in binary classification task.
- Feature engineering could lead to high dimensionality, which might cause the model to perform worse due to the curse of dimensionality, especially when data is sparse.
- The effectiveness of GANs is heavily reliant on the availability of a diverse and representative set of real samples. In cases where real data is extremely limited, the GAN may not capture the full complexity of the underlying data distribution, which could lead to suboptimal synthetic data generation.
- Extensive feature engineering and hyperparameter tuning require significant computational time and resources, which may not be feasible in all operational environments.
- The clinical relevance of the observed discrepancies in EMD values and their impact on diagnostic accuracy is not directly addressed, which is a significant limitation in a medical context.
- The models' improved performance on synthetic data requires validation on external, real-world datasets to confirm that the findings are not artifacts of the specific dataset or synthetic generation process used in this study.

6.7 Conclusion

In conclusion, the objectives set forth in Chapter 6 have been effectively met through the utilization of various GAN architectures for the augmentation of ECG and IBI data. The exploration of different GAN models has significantly contributed to our understanding of synthetic data's role in enhancing model accuracy and handling class imbalance, thereby advancing the capabilities of machine learning in cardiac health diagnostics. This work has also underscored the complexity of integrating synthetic data into clinical models, highlighting both the potential and the limitations of current generative models.

The integration of GAN-augmented data into our machine learning models has not only demonstrated the feasibility of such approaches in academic settings but also suggests a promising avenue for real-world applications, pending further validation. These findings align with the overarching goals of this thesis, which aim to enhance the precision of diagnostic models and support the advancement of personalized medicine through innovative AI techniques. Moving forward, Chapter 7 will delve into the practical challenges and potential clinical implementations of these technologies, ensuring that our research contributions are both scientifically robust and clinically relevant.

7 Discussion: Bridging AI Insights with Clinical Applications

7.1 Research objectives restatement

The primary objectives of this research were to explore and enhance the interpretability and reliability of machine learning models in medical imaging and signal processing. Specifically, the study aimed to:

- provide a thorough literature review on AI algorithms for modelling Cardiovascular adverse outcomes using multi-modal data, establishing a foundational understanding of current methodologies, and identifying areas for innovation.
- to construct an AI model using Cardiac Magnetic Resonance (CMR) for segmentation task, thereby advancing the precision of AI driven segmentation of cardiac imaging analysis.
- sought to expand upon the AI model by integrating demographic, clinical records, and ECG modalities to identify multi-modal biomarkers of AF and associated adverse cardiac outcomes.
- focused on evaluating the extracted biomarkers for their prediction power, approximate costs, and usability, aiming to provide actionable insights for clinical application.

These objectives were pursued with the intent to bridge the gap between machine learning outputs and clinical applicability, ensuring that the developed models are not only accurate but also transparent and useful for healthcare professionals.

In addressing these objectives, the study identified a process gap in the comprehensive pre-processing and feature engineering required to fully harness the clinical insights and HRV data within the PTB dataset. The research highlighted the necessity for a broader spectrum of machine learning models and acknowledged the untapped potential of GANs for dataset augmentation. GANs' ability to generate synthetic data that can improve model performance, diversity, and quality was recognized as a significant opportunity for future exploration.

7.2 Implications of the Findings

7.2.1 MRI Segmentation of model explainability

In the domain of MRI segmentation, we focused on the explainability of automatic segmentation models. Grad-CAM and saliency maps were applied to the U-net architecture to visualize and understand the areas of the MRI scans that the model deemed significant for segmentation tasks.

Advantages of Grad-CAM:

- Provides visual explanations for decisions made by convolutional neural networks.
- Helps in localizing the regions of interest in MRI images that contribute most to the outcomes.
- Can enhance trust in the model by offering insights into its reasoning.

Disadvantages of Grad-CAM:

- Limited to convolutional layers, thus may not provide insights into the decision-making process of fully connected layers.
- The heatmap generated can sometimes be ambiguous and may require expert interpretation to be clinically useful.

Interpretation of Grad-CAM Images:

The Grad-CAM images Figure 4-13 demonstrated regions in the MRI scans that were pivotal in predicting the segmentation mask. Notably, the model also highlighted pixels of the same class that were not part of the segmented mask, indicating a broader recognition of relevant features beyond the immediate area of interest.

Pros of Saliency Maps:

- Highlights the pixels that have the most significant impact on the model's output.
- Can be used to verify that the model's attention aligns with clinically relevant features.

Cons of Saliency Maps:

- Often produces noisy visualizations that can be difficult to interpret.
- May not always align with human intuition, especially in complex medical images.

Interpretation of Saliency Maps:

The Vanilla and Smooth Grad Vanilla saliency results Figure 4-14 and Figure 4-15 identified pixels essential to the segmentation task. However, they lacked the specificity that Grad-CAM provided, offering a less localized and more diffuse interpretation of influential regions.

Implications for Clinical Fraternity:

The interpretability techniques applied in this study, such as Grad-CAM and Saliency Maps, have significant implications for clinical practice by providing insights into the AI's decision-making process. To ensure these techniques are practically applicable in a clinical setting, future iterations of this research will involve closer collaboration with clinicians. This will aid in refining the methods to produce visualizations that are not only clear and actionable but also aligned with clinical knowledge and workflows.

Further Research Directions:

To advance this field, future research could focus on:

- Involving Clinicians in Design and Validation: Future projects will integrate clinical professionals not only in the testing and validation phases but also during the model design process. This collaboration will ensure that the development of AI tools aligns with real-world clinical needs and usability.
- Exploring alternative visualization techniques that could offer more direct clinical insights.

These efforts will enhance the clinical relevance of the AI tools developed, ensuring they contribute effectively to patient care and medical practice.

7.2.2 ECG Classification enhancements

The advancement of Objective 3 in our research has been pivotal in integrating demographic and clinical records with ECG modalities. This has provided a framework for AI models to not only diagnose but also prognosticate, identifying multi-modal biomarkers that are associated with adverse cardiac outcomes and potentially enhancing the management of cardiovascular diseases.

Implications for ECG Classification:

- Provide a thorough literature review on AI algorithms for modeling cardiovascular adverse outcomes using multi-modal data, establishing a foundational understanding of current methodologies and identifying areas for innovation.
- Construct an AI model using Cardiac Magnetic Resonance (CMR) for the segmentation task, thereby advancing the precision of AI-driven segmentation of cardiac imaging analysis.
- Expand upon the AI model by integrating demographic, clinical records, and ECG modalities to identify multi-modal risk markers for atrial fibrillation (AF) and associated adverse cardiac outcomes. These risk markers include clinically significant indicators that are measurable and quantifiable.
- Focus on evaluating the extracted risk markers for their prediction power, approximate costs, and usability, aiming to provide actionable insights for clinical application.

Machine Learning Explainability:

Transparency in AI Decision-Making: In line with the objectives outlined in Section 5.9, the research has emphasized the importance of machine learning explainability. This ensures that healthcare professionals can understand and trust AI decisions, aligning model outputs with clinical reasoning and practice.

Significance of HRV Measures:

- Superior Classification Power: Our findings indicate that HRV measures possess a greater classification prowess compared to raw signal classification. This underscores the potential for these measures to serve as reliable indicators for cardiac health assessment.
- Wearable Technology Integration: The extraction of HRV measures can be effectively carried out using wearable devices, which opens avenues for continuous, non-invasive monitoring and real-time data analysis.

Challenges and Further Research:

- Data Integration: The amalgamation of diverse data types presents significant challenges in ensuring data compatibility and necessitates sophisticated pre-processing techniques.
- Model Complexity: The complexity of models increases with the integration of diverse data types, requiring advanced algorithms capable of handling intricate data without sacrificing performance.
- Clinical Validation: For any multi-modal biomarkers discovered, rigorous clinical validation is crucial to establish their relevance and applicability in real-world settings.

Limitations and Future Scope:

- **Validation with External Datasets:** A limitation of the current study is the validation of findings with other datasets. Future research should aim to corroborate these results across various populations and datasets to ensure the generalizability of the models.
- **Real-World Application:** Moving forward, the implementation of these AI models in real-world clinical settings will require validation studies, particularly focusing on the utility and impact of wearable technology in continuous cardiac health monitoring.

In summary, the research has made significant strides in enhancing ECG classification through the integration of multi-modal data and the exploration of machine learning explainability. The potential of HRV measures as extracted through wearable technology has been highlighted as a significant advancement. The validation of these findings with external datasets remains a crucial next step, ensuring that the developed AI models can withstand the rigors of diverse clinical environments and contribute meaningfully to the field of personalized medicine.

7.2.3 GAN-Based Augmentation

The acknowledgment of the process gap has catalysed the exploration of GANs for dataset augmentation, particularly to ameliorate class imbalances and augment the AI model's robustness [246], [247]. The capability of GANs to synthesize authentic-looking samples could dramatically enhance the quality and volume of data available for training AI models in the context of adverse cardiac outcomes modelling.

Implications for Data Augmentation:

- **Addressing Data Scarcity:** GANs can generate additional data where samples are limited, especially for underrepresented classes in medical datasets.
- **Model Generalization:** With a more varied dataset, models are less likely to overfit and more likely to generalize well to unseen data.
- **Improved Model Robustness:** Synthetic samples can introduce necessary variability into the training process, leading to more robust models capable of handling real-world variations.

Challenges and Further Research:

- **Quality of Synthetic Data:** Ensuring that the synthetic data is of high quality and truly representative of real-world distributions is essential.
- **Model Bias:** There is a risk that GANs might replicate or even exacerbate existing biases in the data, leading to skewed models.
- **Ethical Considerations:** The use of synthetic data raises ethical questions, particularly in terms of patient consent and data privacy.

Future Directions:

- Ethical Framework: Developing a robust ethical framework for the use of synthetic data in medical AI research.
- Hybrid Models: Exploring hybrid models that can effectively integrate real and synthetic data for optimal performance.
- Clinical Partnerships: Collaborating with clinicians to validate the efficacy of GAN-augmented datasets in clinical trials and diagnostic settings to address potential biases and ensure the utility of synthetic data in real-world scenarios.

Together, the implications and future directions outlined in 7.2.2 and 7.2.3 affirm the potential of AI to revolutionize ECG classification and highlight the innovative role of GANs in overcoming current data challenges. The next phase of research must address these challenges head-on, ensuring that the advancements in AI not only push the boundaries of what is technically possible but also align with clinical needs and ethical standards.

8 Concluding Insights – The Future of AI in Cardiac Health

8.1 Summary of Research contributions

This research journey has traversed the intricate pathways of ECG analysis, MRI segmentation, and the innovative realm of GAN-based data augmentation, all through the lens of artificial intelligence. We have delved deep into the literature, unearthed the potential of multi-modal datasets, and harnessed the prowess of AI to unveil biomarkers that lay concealed within complex cardiac signals and imagery.

The significant outcomes of this research are manifold:

- This study has successfully integrated demographic, clinical records, and ECG modalities, paving the way for multi-modal biomarker discovery for adverse cardiac outcomes.
- The study has showcased the enhancement of ECG classification through meticulous feature engineering and model explainability, providing a beacon for future efforts in personalized medicine.
- The application of GANs to mitigate class imbalances has demonstrated the untapped potential of synthetic data in bolstering the diversity and quality of datasets for AI models.

8.2 Impact on the field of ECG Analysis and AI

The methodologies refined and the insights gleaned from this study have made a substantial contribution not only to the field of ECG analysis but also to the broader practice of healthcare where AI integration is concerned. This research has demonstrated how advanced AI algorithms can be employed to enhance the interpretative capabilities of medical professionals, leading to more accurate diagnostics and improved patient outcomes.

The application of AI in ECG analysis, as developed through this research, offers significantly improved predictive accuracy. By integrating multi-modal data, including demographic and clinical records alongside traditional ECG readings, AI models developed in this study can detect subtle nuances in ECG signals that might be overlooked in conventional analysis. This capability could lead to earlier detection of cardiovascular conditions and other rhythm anomalies, potentially guiding more timely and targeted therapeutic interventions.

The integration of AI tools in ECG analysis has the potential to transform clinical decision-making processes. By providing clinicians with deeper insights and more reliable diagnostic tools, AI can support more informed decisions, reducing diagnostic errors and enhancing the personalization of treatment plans. This shift not only improves patient care but also optimizes resource allocation within healthcare systems.

On a system-wide level, the successful integration of AI into ECG analysis could drive significant changes in healthcare delivery. Enhanced diagnostic tools can lead to more efficient patient management, reduce the burden on healthcare facilities, and decrease overall healthcare costs by curtailing the need for repetitive tests and reducing the incidence of misdiagnosis.

8.3 Future Horizons and Practical Applications

The study bridges the gap between theoretical AI models and practical healthcare applications, offering a robust foundation for future advancements in the cardiology domain. The integration of these AI models into clinical workflows has the potential to transform real-time diagnostics and patient monitoring.

The advent of personalized healthcare technologies, bolstered by the power of AI, is on the horizon. Wearable devices, infused with the intelligence from this study's multi-modal data analysis, could soon provide early warnings for cardiac events, improving the scope and efficacy of preventative healthcare.

Validation of these AI models across diverse patient populations and clinical environments is the critical next step. The objective is to scale and adapt these models, ensuring their effectiveness and reliability in detecting a wide range of cardiac health variations.

The role of generative AI in creating high-quality, synthetic data for robust training datasets is particularly promising, especially for underrepresented conditions in medical datasets. These GAN-augmented datasets are poised to enhance the generalizability and resilience of AI models, marking a significant stride in diagnostic, and monitoring applications within clinical settings.

However, the ethical implications of using synthetic data, such as patient privacy, consent, and data integrity, must be carefully navigated. As we harness the advancements in AI technology, developing a robust ethical framework is essential to ensure the responsible use of synthetic data in medical research.

As we look to the future, the key challenge will be to seamlessly blend AI sophistication with the nuanced requirements of cardiology. Ensuring that AI complements and augments clinical expertise is vital for achieving improved patient outcomes and establishing new benchmarks in healthcare delivery.

In conclusion, the convergence of AI with cardiology promises to not only enhance patient outcomes but to redefine the paradigms of healthcare delivery. This study contributes a pivotal chapter to this narrative, setting the stage for the next wave of innovation in AI-driven cardiology solutions.

Through this journey, the study has laid down the markers for ethically responsible, AI-enriched cardiology practices that prioritize patient welfare, marking our research as a catalyst for continued progress in the field.

9 References

- [1] M. Kassa and J. Grace, "The Global Burden and Perspectives on Non-Communicable Diseases (NCDs) and the Prevention, Data Availability and Systems Approach of NCDs in Low-resource Countries," in *Public Health in Developing Countries*, E. E. Anugwom and N. Awofeso, Eds. Rijeka: IntechOpen, 2019.
- [2] H. Singh and J. Bharti, "Non-Communicable Diseases and Their Risk Factors: Review," *EAS J. Parasitol. Infect. Dis.*, vol. 3, no. 6, pp. 83–86, 2021, doi: 10.36349/easjpid.2021.v03i06.001.
- [3] I. Olier, S. Ortega-Martorell, M. Pieroni, and G. Y. H. Lip, "How machine learning is impacting research in atrial fibrillation: Implications for risk prediction and future management," *Cardiovasc. Res.*, vol. 117, no. 7, pp. 1700–1717, 2021, doi: 10.1093/cvr/cvab169.
- [4] K. Strimbu and J. A. Tavel, "What are Biomarkers?," *Curr Opin HIV Aids*, vol. 5, no. 6, pp. 463–466, 2011, doi: 10.1097/COH.0b013e32833ed177.What.
- [5] R. M. Califf, "Minireview Biomarker definitions and their applications," *Exp. Biol. Med.*, pp. 213–221, 2018, doi: 10.1177/1535370217750088.
- [6] T. Turner, M. Misso, C. Harris, and S. Green, "Development of evidence-based clinical practice guidelines (CPGs): Comparing approaches," *Implement. Sci.*, vol. 3, no. 1, pp. 1–8, 2008, doi: 10.1186/1748-5908-3-45.
- [7] G. Y. H. Lip, P. Kakar, and T. Watson, "Atrial fibrillation — the growing epidemic," *Heart*, vol. 93(5), no. May, pp. 542–543, 2007, doi: 10.1136/hrt.2006.110791.
- [8] C. Abbafati *et al.*, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," *Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020, doi: 10.1016/S0140-6736(20)30925-9.
- [9] C. A. Morillo, A. Banerjee, P. Perel, D. Wood, and X. Jouven, "Atrial fibrillation: The current epidemic," *J. Geriatr. Cardiol.*, vol. 14, no. 3, pp. 195–203, 2017, doi: 10.11909/j.issn.1671-5411.2017.03.011.
- [10] G. Lippi, F. Sanchis-Gomar, and G. Cervellin, "Global epidemiology of atrial fibrillation: An increasing epidemic and public health challenge," *Int. J. Stroke*, vol. 16, no. 2, pp. 217–221, 2021, doi: 10.1177/1747493019897870.
- [11] B. J. Hansen, J. Zhao, V. V. Fedorov, and N. Zealand, "Fibrosis and Atrial Fibrillation: Computerized and Optical Mapping; A View into the Human Atria at Submillimeter Resolution LA - eng," *JACC. Clin. Electrophysiol.*, vol. 3, no. 6, pp. 531–546, 2018, doi: 10.1016/j.jacep.2017.05.002.Fibrosis.
- [12] T. A. Csepe, B. J. Hansen, V. V. Fedorov, C. Biology, D. Heart, and T. O. State, "Atrial fibrillation driver mechanisms: insight from the isolated human heart," *Trends Cardiovasc Med*, vol. 27, no. 1, pp. 1–11, 2018, doi: 10.1016/j.tcm.2016.05.008.Atrial.
- [13] O. V. Aslanidi, M. A. Colman, J. Stott, H. Dobrzynski, and R. Mark, "3D virtual human atria: A computational platform for studying clinical atrial fibrillation," *Prog Biophys Mol Biol*, vol. 107, no. 1, pp. 156–168, 2011, doi: 10.1016/j.pbiomolbio.2011.06.011.3D.
- [14] M. A. Colman *et al.*, "Pro-arrhythmogenic effects of atrial fibrillation-induced electrical remodelling : insights from the three-dimensional virtual human atria," *J. Physiol.*, vol. 17, pp. 4249–4272, 2013, doi:

10.1113/jphysiol.2013.254987.

- [15] S. Prabhu, A. J. A. Mclellan, T. E. Walters, and M. Sharma, "ScienceDirect Atrial Structure and Function and its Implications for Current and Emerging Treatments for Atrial Fibrillation," *Prog. Cardiovasc. Dis.*, vol. 58, no. 2, pp. 152–167, 2015, doi: 10.1016/j.pcad.2015.08.004.
- [16] R. Beinart and S. Nazarian, "Role of magnetic resonance imaging in atrial fibrillation ablation," *Curr. Treat. Options Cardiovasc. Med.*, vol. 16, no. 6, 2014, doi: 10.1007/s11936-014-0316-3.
- [17] C. B. Marcu, A. M. Beek, and A. C. Van Rossum, "Clinical applications of cardiovascular magnetic resonance imaging Constantine," *Can. Med. Assoc. J.*, vol. 175(8), no. October, pp. 911–917, 2006.
- [18] C. McGann *et al.*, "Atrial Fibrillation Ablation Outcome Is Predicted by Left Atrial Remodeling on MRI," *Circ. Arrhythmia Electrophysiol.*, pp. 23–30, 2013, doi: 10.1161/CIRCEP.113.000689.
- [19] Neeraj Sharma and Lalit M Aggarwal., "Automated Medical image segmentation techniques," *J. Med. Phys.*, vol. 35, no. 1, 2010, doi: 10.4103/0971-6203.58777.
- [20] D. Kaur and Y. Kaur, "International Journal of Computer Science and Mobile Computing Various Image Segmentation Techniques: A Review," 2014.
- [21] C. A. G. & G. W. Horgan, *Image Analysis for the Biological Sciences*, 1995th ed. 1996.
- [22] "1-s2.0-S001048251930321X-mmc1."
- [23] B. Preim and D. Bartz, *Visualization in Medicine*. Burlington, MA, USA: Morgan Kaufmann Publishers, 2007.
- [24] P. D. S. Report, "Graph Cuts for Image Segmentation," Indian Institute of Technology Mumbai.
- [25] U. of T. Eppel, Sagi, "Tracing liquid level and material boundaries in transparent vessels using the graph cut computer vision approach.le," *Res. Gate*, 2016.
- [26] F. Yi, "Image Segmentation : A Survey of Graph-cut Methods," in *International Conference on Systems and Informatics*, 2012, no. Icsai, pp. 1936–1941.
- [27] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning : A Survey," *Arxiv*, pp. 1–22.
- [28] R. Kaushik and S. Kumar, "Image Segmentation Using Convolutional Neural Network," *Int. J. Sci. Technol. Res.*, vol. 8, no. 11, pp. 667–675, 2019, doi: ISSN 2277-8616.
- [29] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on," *Zeitschrift ffr Medizinische Phys.*, vol. 29, no. 2, pp. 102–127, 2019, doi: 10.1016/j.zemedi.2018.11.002.
- [30] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," 2015, doi: 10.1038/nature14539.
- [31] Ian Goodfellow and Yoshua Bengio and Aaron Courville, *Deep Learning*. MIT Press, 2016.
- [32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015.
- [33] K. He, Z. Xiangyu, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016, doi: 10.1109/CVPR.2016.90.

- [34] R. Girshick, J. Donahue, T. Darrell, J. Malik, and U. C. Berkeley, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.
- [35] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986, doi: 10.1038/323533a0.
- [36] R. Karim *et al.*, "A Method to Standardize Quantification of Left Atrial Scar From Delayed-Enhancement MR Images," *IEEE J. Transl. Eng. Heal. Med.*, vol. 2, no. February, pp. 1–15, 2014, doi: 10.1109/JTEHM.2014.2312191.
- [37] L. M. Positano V, Pingitore A, Giorgetti A, Favilli B, Santarelli MF, Landini L, Marzullo P, "A fast and effective method to assess myocardial necrosis by means of contrast magnetic resonance imaging.," *J Cardiovasc Magn Reson*, vol. 7, no. 2, pp. 487–94, 2005, doi: 10.1081/jcmr-200053630.
- [38] B. E. Faculty, A. Kolipaka, G. P. Chatzimavroudis, R. D. White, T. P. O. Donnell, and M. Setser, "Segmentation of Non-viable Myocardium in Delayed Enhancement Magnetic Resonance Images Publisher's Statement," *Int. J. Cardiovasc. Imaging*, vol. 21, pp. 303–311, 2005, doi: 10.1007/s10554-004-5806-z.
- [39] A. Hennemuth *et al.*, "A Comprehensive Approach to the Analysis of Contrast Enhanced Cardiac MR Images," *IEEE Trans Med Imaging*, vol. 27, no. 11, pp. 1592–1610, 2008, doi: 10.1109/TMI.2008.2006512.
- [40] A. Schmidt *et al.*, "Infarct Tissue Heterogeneity by Magnetic Resonance Imaging Identifies Enhanced Cardiac Arrhythmia Susceptibility in Patients With Left Ventricular Dysfunction," *Circulation*, vol. Apr 17, no. 115(15), pp. 2006–2014, 2007, doi: 10.1161/CIRCULATIONAHA.106.653568.
- [41] A. T. Yan *et al.*, "Characterization of the Peri-Infarct Zone by Contrast-Enhanced Cardiac Magnetic Resonance Imaging Is a Powerful Predictor of Post – Myocardial Infarction Mortality," *Circulation*, pp. 32–39, 2006, doi: 10.1161/CIRCULATIONAHA.106.613414.
- [42] D. Perry *et al.*, "Automatic classification of scar tissue in late gadolinium enhancement cardiac MRI for the assessment of left-atrial wall injury after radiofrequency ablation," in *SPIEDigital Library Org*, 2021, no. February 2012, doi: 10.1117/12.910833.
- [43] Y. Lu, G. Wright, and P. E. Radau, "Automatic myocardium segmentation of LGE MRI by deformable models with prior shape data," *J. Cardiovasc. Magn. Reson.*, vol. 15, no. Suppl 1, pp. 14–15, 2013, doi: 10.1186/1532-429X-15-S1-P14.
- [44] R. S. Oakes *et al.*, "Detection and Quantification of Left Atrial Structural Remodeling With Delayed-Enhancement Magnetic Resonance Imaging in Patients With Atrial Fibrillation," *Circulation*, pp. 1758–1767, 2009, doi: 10.1161/CIRCULATIONAHA.108.811877.
- [45] B. R. Knowles *et al.*, "3-D Visualization of Acute RF Ablation Lesions Using MRI for the Simultaneous Determination of the Patterns of Necrosis and Edema," *IEEE J. Biomed. Eng.*, vol. 57, no. 6, pp. 1467–1475, 2010.
- [46] D. Ravanelli *et al.*, "A Novel Skeleton Based Quantification and 3-D Volumetric Visualization of Left Atrium Fibrosis Using Late Gadolinium Enhancement Magnetic Resonance Imaging," *IEEE Trans. Med. Imaging*, vol. 33, no. 2, pp. 566–576, 2014, doi: 10.1109/TMI.2013.2290324.
- [47] G. Veni, Z. Fu, S. P. Awate, and R. T. Whitaker, "Bayesian Segmentation of Atrium Wall Using Globally-

- Optimal Graph Cuts on 3D Meshes,” *Inf Process Med Imaging*, vol. 23, pp. 656–667, 2013, doi: 10.1007/978-3-642-38868-2_55.
- [48] and R. T. W. Gopalkrishna Veni, Zhisong Fu, Suyash P. Awate, “PROPER ORDERED MESHING OF COMPLEX SHAPES AND OPTIMAL GRAPH CUTS APPLIED TO ATRIAL-WALL SEGMENTATION FROM DE-MRI,” in *IEEE Int Symp Biomed Imaging*, 2014, pp. 1296–1299, doi: 10.1109/ISBI.2013.6556769.PROPER.
 - [49] C. Jin *et al.*, “Left atrial appendage segmentation and quantitative assisted diagnosis of atrial fibrillation based on fusion of temporal-spatial information,” *Comput. Biol. Med.*, vol. 96, no. February, pp. 52–68, 2018, doi: 10.1016/j.combiomed.2018.03.002.
 - [50] J. E. Ortuño *et al.*, “Automatic estimation of aortic and mitral valve displacements in dynamic CTA with 4D graph-cuts,” *Med. Image Anal.*, vol. 65, 2020, doi: 10.1016/j.media.2020.101748.
 - [51] P. Agrawal, R. T. Whitaker, S. Y. Elhabian, and A. Member, “An Optimal , Generative Model for Estimating Multi-Label Probabilistic Maps,” *IEEE Trans. Med. Imaging*, vol. 39, no. 7, pp. 2316–2326, 2020, doi: 10.1109/TMI.2020.2968917.
 - [52] H. C. Kang, J. Lee, and J. Shin, “Automatic Four-Chamber Segmentation Using Level-Set Method and Split Energy Function,” *Healthc. Inform. Res.*, vol. 22, no. 4, pp. 285–292, 2016, doi: 85-292. <https://doi.org/10.4258/hir.2016.22.4.285>.
 - [53] P. G. Christian Wachinger, “Spectral label fusion,” *Med Image Comput Comput Assist Interv*, vol. 15, no. (pt3), pp. 410–417, 2012, doi: 10.1007/978-3-642-33454-2_51.
 - [54] Dp. Orod Razeghi, PhDIain Sim, MBBSCaroline H. Roney, PhDRashed Karim, PhDHenry Chubb, MBBSJohn Whitaker, BM, BCh, PhDLouisa O’Neill, MBRahul Mukherjee, MBBSMatthew Wright, MB, BS, PhDMark O’Neill, MB, BCh, DPhilSteven E. Williams, PhDSteven Niederer, “Fully Automatic Atrial Fibrosis Assessment Using a Multilabel Convolutional Neural Network _ Enhanced Reader.pdf,” *Circ. Cardiovasc. Imaging*, vol. Dec, no. 13(12), 2020, doi: 10.1161/CIRCIMAGING.120.011512.
 - [55] Z. Xiong *et al.*, “A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging,” *Med. Image Anal.*, vol. 67, Jan. 2021, doi: 10.1016/j.media.2020.101832.
 - [56] Y. Lu, X. Fu, X. Li, and Y. Qi, “Cardiac Chamber Segmentation Using Deep Learning on Magnetic Resonance Images from Patients Before and After Atrial Septal Occlusion Surgery,” in *IEEE Eng Med Biol Soc. 2020*, 2020, pp. 1211–1216, doi: 10.1109/EMBC44109.2020.9175618.
 - [57] G. Yang *et al.*, “Simultaneous left atrium anatomy and scar segmentations via deep learning in multiview information with attention,” *Futur. Gener. Comput. Syst.*, vol. 107, pp. 215–228, 2020, doi: 10.1016/j.future.2020.02.005.
 - [58] X. Du *et al.*, “Segmentation and visualization of left atrium through a unified deep learning framework,” *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 4, pp. 589–600, Apr. 2020, doi: 10.1007/s11548-020-02128-9.
 - [59] L. Li *et al.*, “Atrial scar quantification via multi-scale CNN in the graph-cuts framework,” *Med. Image Anal.*, vol. 60, Feb. 2020, doi: 10.1016/j.media.2019.101595.
 - [60] N. F. M. and F. B. S. Chao Huang, Stephen L. Wasmund, Takanori Yamaguchi, Nathan Knighton, Robert W. Hitchcock, Irina A. Polejaeva, Kenneth L. White, “Towards Automated Quantification of Atrial

- Fibrosis in Images from Catheterized Fiber-Optics Confocal Microscopy Using Convolutional Neural Networks,” *Funct Imaging Model Hear.*, vol. June, no. 11504, pp. 168–176, 2019, doi: 10.1007/978-3-030-21949-9.
- [61] Z. Xiong, V. V. Fedorov, X. Fu, E. Cheng, R. Macleod, and J. Zhao, “Fully Automatic Left Atrium Segmentation From Late Gadolinium Enhanced Magnetic Resonance Imaging Using a Dual Fully Convolutional Neural Network,” *IEEE Trans. Med. Imaging*, vol. 38, no. 2, pp. 515–524, Feb. 2019, doi: 10.1109/TMI.2018.2866845.
 - [62] G. Yang *et al.*, “Multiview Sequential Learning and Dilated Residual Learning for a Fully Automatic Delineation of the Left Atrium and Pulmonary Veins from Late Gadolinium-Enhanced Cardiac MRI Images,” in *IEEE Eng Med Biol Soc. 2018*, 2018, pp. 1123–1127, doi: 10.1109/EMBC.2018.8512550.
 - [63] A. Chakravarty and J. Sivaswamy, “RACE-Net: A Recurrent Neural Network for Biomedical Image Segmentation,” *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 3, pp. 1151–1162, May 2019, doi: 10.1109/JBHI.2018.2852635.
 - [64] L. Zhu, Y. Gao, A. Yezzi, and A. Tannenbaum, “Automatic segmentation of the left atrium from MR images via variational region growing with a moments-based shape prior,” *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5111–5122, 2013, doi: 10.1109/TIP.2013.2282049.
 - [65] N. Vu and B. S. Manjunath, “Shape Prior Segmentation of Multiple Objects with Graph Cuts,” 2008, doi: 10.1109/CVPR.2008.4587450.
 - [66] W. J. Kutra D, Saalbach A, Lehmann H, Groth A, Dries SP, Krueger MW, Dössel O, “Automatic Multi-model-Based Segmentation of the Left Atrium in Cardiac MRI Scans _ Enhanced Reader.pdf,” *Med Image Comput Comput Assist Interv*, vol. 15, no. (Pt.2), pp. 1–8, 2012, doi: 10.1007/978-3-642-33418-4_1.
 - [67] G. P. Depa M, Sabuncu MR, Holmvang G, Nezafat R, Schmidt EJ, “Robust Atlas-Based Segmentation of Highly Variable Anatomy: Left Atrium Segmentation,” *Stat Atlases Comput Model. Hear.*, vol. 6364, pp. 85–94, 2010, doi: 10.1007/978-3-642-15835-3.
 - [68] J. Koikkalainen *et al.*, “Methods of Artificial Enlargement of the Training Set for Statistical Shape Models,” *IEEE Trans Med Imaging*, vol. 27, no. 11, pp. 1643–1654, 2008, doi: 10.1109/TMI.2008.929106.
 - [69] M. Firouznia M, Feeny AK, LaBarbera MA, McHale M, Cantlay C, Kalfas N, Schoenhagen P, Saliba W, Tchou P, Barnard J, Chung MK, “A. Machine Learning-Derived Fractal Features of Shape and Texture of the Left Atrium and Pulmonary Veins From Cardiac Computed Tomography Scans Are Associated With Risk of Recurrence of Atrial Fibrillation Postablation,” *Circ Arrhythm Electrophysiol*, vol. Mar, no. 14(3), pp. 329–336, 2021, doi: 10.1161/CIRCEP.120.009265.
 - [70] C. Liu, S. Chang, H. Chen, W. Chen, and Y. Lin, “The Clinical Application of the Deep Learning Technique for Predicting Trigger Origins in Patients With Paroxysmal Atrial Fibrillation With,” *Circ. Arrhythmia Electrophysiol.*, no. November, pp. 1286–1295, 2020, doi: 10.1161/CIRCEP.120.008518.
 - [71] J. Li, H. Chen, F. Zhu, C. Wen, H. Chen, and L. Wang, “Automatic Pulmonary Vein and Left Atrium Segmentation for TAPVC Preoperative Evaluation Using V-Net with Grouped Attention,” in *Annu Int Conf IEEE Eng Med Biol Soc. 2020*, 2020, pp. 1207–1210, doi: 10.1109/EMBC44109.2020.9175907.
 - [72] H. H. Chen *et al.*, “Automated extraction of left atrial volumes from two-dimensional computer tomography images using a deep learning technique,” *Int. J. Cardiol.*, vol. 316, pp. 272–278, Oct. 2020,

doi: 10.1016/j.ijcard.2020.03.075.

- [73] H. H.-S. L. Hung-Hsun Chen, Chih-Min Liu, Shih-Lin Chang, Paul Yu-Chun Chang, Wei-Shiang Chen, Yo-Ming Pan, Ssu-Ting Fang, Shan-Quan Zhan, Chieh-Mao Chuang, Yenn-Jiang Lin, Ling Kuo, Mei-Han Wu, Chun-Ku Chen, Ying-Yueh Chang, Yang-Che Shiu, Shih-Ann Chen, "Automated extraction of left atrial volumes from two-dimensional computer tomography images using a deep learning technique," 2020. doi: <https://doi.org/10.1016/j.ijcard.2020.03.075>.
- [74] L. Baskaran *et al.*, "Automatic segmentation of multiple cardiovascular structures from cardiac computed tomography angiography images using deep learning," *PLoS One*, vol. 15, no. 5, May 2020, doi: 10.1371/journal.pone.0232573.
- [75] L. Baskaran *et al.*, "Identification and Quantification of Cardiovascular Structures From CCTA: An End-to-End, Rapid, Pixel-Wise, Deep-Learning Method," *JACC Cardiovasc. Imaging*, vol. 13, no. 5, pp. 1163–1171, May 2020, doi: 10.1016/j.jcmg.2019.08.025.
- [76] C. Jin, J. Feng, L. Wang, H. Yu, and J. Liu, "Left Atrial Appendage Segmentation Using Fully Convolutional Neural Networks and Modified Three-Dimensional Conditional Random Fields," *IEEE J Biomed Heal. Inf.*, vol. 22, no. 6, pp. 1906–1916, 2018, doi: 10.1109/JBHI.2018.2794552.
- [77] C. S. Perone and J. Cohen-adad, "Promises and limitations of deep learning for medical image segmentation," *J. Med. Artif. Intell.*, vol. 2, pp. 1–2, 2019, doi: 10.21037/jmai.2019.01.01.
- [78] Z. Eaton-rosen and F. Bragman, "Improving Data Augmentation for Medical Image Segmentation," in *1st Conference on Medical Imaging with Deep Learning (MIDL 2018)*, 2018, no. Midl, pp. 1–3.
- [79] A. Committee on Reproducibility and Replicability in Science; Board on Behavioral, Cognitive *et al.*, *Reproducibility and Replicability in Science*. Washington DC: The National Academies Press, 2019.
- [80] F. Renard, S. Guedria, N. De Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Sci. Rep.*, no. 0123456789, pp. 1–16, 2020, doi: 10.1038/s41598-020-69920-0.
- [81] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high - class imbalance in big data," *J. Big Data*, 2018, doi: 10.1186/s40537-018-0151-6.
- [82] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0197-0.
- [83] L. Taylor and G. Nitschke, "Improving Deep Learning using Generic Data Augmentation," *Arxiv*, no. August, 2017.
- [84] S. Chen and J. H. Lee, "A Group-Theoretic Framework for Data Augmentation," *J. Mach. Learn. Res.*, vol. 21, no. 20–163, pp. 1–71, 2020.
- [85] G. Kang, X. Dong, L. Zheng, and Y. Yang, "PatchShuffle Regularization," *Arxiv*, no. Nips, 2017.
- [86] H. Inoue, "Data Augmentation by Pairing Samples for Images Classification," *Arxiv*, no. April, 2018.
- [87] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," 2017.
- [88] R. Zhang, P. Isola, and A. A. Efros, "Colorful Image Colorization," *Arxiv - Comput. Res. Repos.*, vol. abs/1603.0, 2016.
- [89] K. Chatfield, K. Simonyan, and A. Zisserman, "Return of the Devil in the Details : Delving Deep into

- Convolutional Nets,” in *British Machine Vision Conference*, 2014, p. 12, doi: <http://dx.doi.org/10.5244/C.28.6>.
- [90] A. Jurio, M. Pagola, and M. Galar, “A Comparison Study of Different Color Spaces in A Comparison Study of Different Color Spaces,” in *Communications in Computer and Information Science*, 2016, no. January, doi: 10.1007/978-3-642-14058-7.
 - [91] C. Gormond, “Influence of Photometric Data Augmentation Influence of Photometric Data Augmentation,” *Res. Gate*, no. September, pp. 0–5, 2021.
 - [92] A. Fawzi and P. Frossard, “DeepFool : a simple and accurate method to fool deep neural networks,” *Arxiv*, pp. 2574–2582, 2016.
 - [93] S. Li and L. G. Apr, “Learning More Robust Features with Adversarial Training,” *Arxiv*, no. April, pp. 1–7, 2018.
 - [94] I. Goodfellow, “Generative Modeling Generative Modeling,” *Arxiv*, 2016.
 - [95] W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and and I. Goodfellow, “Many Paths to Equilibrium: GANs Do Not Need to Decrease a Divergence At Every Step,” *Arxiv*, pp. 1–21, 2018.
 - [96] L. A. Gatys, A. S. Ecker, M. Bethge, and C. V Sep, “A Neural Algorithm of Artistic Style,” *Arxiv*, pp. 3–7, 2015.
 - [97] J. Ma *et al.*, “Loss odyssey in medical image segmentation,” *Med. Image Anal.*, vol. 71, no. 2021, 2021, doi: 10.1016/j.media.2021.102035.
 - [98] O. Bernard *et al.*, “Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis : Is the Problem Solved ?,” *IEEE Trans. Med. Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018, doi: 10.1109/TMI.2018.2837502.
 - [99] K. Jamart, Z. Xiong, G. D. Maso Talou, M. K. Stiles, and J. Zhao, “Mini Review: Deep Learning for Atrial Segmentation From Late Gadolinium-Enhanced MRIs,” *Front. Cardiovasc. Med.*, vol. 7, May 2020, doi: 10.3389/fcvm.2020.00086.
 - [100] S. Jadon, “A survey of loss functions for semantic segmentation,” 2020, doi: 10.1109/CIBCB48159.2020.9277638.
 - [101] M. A. Vi-de and L. I. U. Qing, “Automated Image Segmentation Using Improved,” in *2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, 2004, pp. 743–746.
 - [102] T. Lin, F. Ai, and P. Doll, “Focal Loss for Dense Object Detection,” *Arxiv*, 2018.
 - [103] F. Milletari, “V-Net : Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation,” in *Fourth International Conference on 3D Vision*, 2016, pp. 567–573, doi: 10.1109/3DV.2016.79.
 - [104] R. El, C. Petitjean, P. Honeine, V. Cheplygina, and F. Abdallah, “High-level Prior-based Loss Functions for Medical Image Segmentation : A,” *Comput. Vis. Image Underst.*, pp. 1–24, 2020.
 - [105] M. Drozdal, E. Vorontsov, G. Chartrand, and C. V Sep, “The Importance of Skip Connections in Biomedical Image Segmentation,” *Arxiv*, 2016.
 - [106] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. J. Cardoso, and T. I. Group, “Generalised Dice

- overlap as a deep learning loss function for highly unbalanced segmentations,” *Arxiv*, pp. 1–8, 2017, doi: 10.1007/978-3-319-67558-9_28.
- [107] A. Rahman and Y. Wang, “Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation,” 2016, doi: 10.1007/978-3-319-50835-1_22.
 - [108] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, “Tversky loss function for image segmentation using 3D fully convolutional deep networks,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10541 LNCS, pp. 379–387, 2017, doi: 10.1007/978-3-319-67389-9_44.
 - [109] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and H. Klaus, “Automated Design of Deep Learning Methods for Biomedical Image Segmentation,” *Arxiv*, pp. 1–55, 2020, doi: 10.1038/s41592-020-01008-z.
 - [110] J. E. Vogt, “Interpretability and Explainability : A Machine Learning Zoo Mini-tour,” *Arxiv - Comput. Res. Repos.*, pp. 1–24, 2020.
 - [111] B. H. M. Van Der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis,” *Arxiv - Comput. Res. Repos.*
 - [112] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” *Arxiv - Comput. Res. Repos.*, 2012.
 - [113] K. Simonyan, “Deep Inside Convolutional Networks : Visualising Image Classification Models and Saliency Maps arXiv : 1312 . 6034v2 [cs . CV] 19 Apr 2014,” *Arxiv - Comput. Res. Repos.*, pp. 1–8, 2013.
 - [114] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization,” *Arxiv - Comput. Res. Repos.*, pp. 2921–2929, 2015.
 - [115] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” *Arxiv*, Oct. 2016, doi: 10.1007/s11263-019-01228-7.
 - [116] I. V Nov, “Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks,” *Arxiv - Comput. Res. Repos.*, pp. 1–20, 2021.
 - [117] S. Mousavi, F. Afghah, and U. R. Acharya, “HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks,” *Comput. Biol. Med.*, vol. 127, 2020, doi: 10.1016/j.compbiomed.2020.104057.
 - [118] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, “Interpretability of machine learning-based prediction models in healthcare,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 5, pp. 1–13, 2020, doi: 10.1002/widm.1379.
 - [119] M. Saiz-viv, C. Fonfr, and L. Mart, “Supervised Domain Adaptation for Automated Semantic Segmentation of the Atrial Cavity,” *Entropy*, vol. 23, no. 7, 2021, doi: 10.3390/e23070898.
 - [120] E. Puybureau, Z. Zhou, Y. Khoudli, E. Carlinet, J. Yongchao Xu, Lacotte, and G. Thierry, “Left Atrial Segmentation In a Few Seconds Using Fully Convolutional Network and Transfer Learning,” in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges --- 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers*, 2019, pp. 339–347, doi: 10.1007/978-3-

- [121] C. G. Fonseca *et al.*, “The Cardiac Atlas Project — an imaging database for computational modeling and statistical atlases of the heart,” *Bioinformatics*, vol. 27, no. 16, pp. 2288–2295, 2011, doi: 10.1093/bioinformatics/btr360.
- [122] R. Karim *et al.*, “Evaluation of current algorithms for segmentation of scar tissue from late Gadolinium enhancement cardiovascular magnetic resonance of the left atrium: An open-access grand challenge,” *J. Cardiovasc. Magn. Reson.*, vol. 15, no. 1, pp. 1–17, 2013, doi: 10.1186/1532-429X-15-105.
- [123] Y. B. Yann LeCun, Patrick Haffner, Léon Bottou, “Object Recognition with Gradient-Based Learning,” in *Shape, Contour and Grouping in Computer Vision*, 1999, no. 0, p. 319.
- [124] S. Niyas, S. J. Pawan, M. Anand Kumar, and J. Rajan, “Medical image segmentation with 3D convolutional neural networks: A survey,” *Neurocomputing*, vol. 493, pp. 397–413, 2022, doi: 10.1016/j.neucom.2022.04.065.
- [125] P. Wagner *et al.*, “PTB-XL , a large publicly available electrocardiography dataset,” *Sci. Data*, pp. 1–15, 2020, doi: 10.1038/s41597-020-0495-6.
- [126] S. Smigiel and D. Ledzi, “Study of the Few-Shot Learning for ECG Classification Based on the PTB-XL Dataset,” pp. 1–25, 2022.
- [127] H. Seo, G. Yoon, S. Joo, and G. Nam, “Computer Methods and Programs in Biomedicine Multiple electrocardiogram generator with single-lead electrocardiogram,” *Comput. Methods Programs Biomed.*, vol. 221, p. 106858, 2022, doi: 10.1016/j.cmpb.2022.106858.
- [128] K. Pałczy and S. Smigiel, “ECG Signal Classification Using Deep Learning Techniques Based on the PTB-XL Dataset,” pp. 1–20, 2021.
- [129] W. Ullah, I. Siddique, R. M. Zulqarnain, M. M. Alam, I. Ahmad, and U. A. Raza, “Classification of Arrhythmia in Heartbeat Detection Using Deep Learning,” vol. 2021, 2021.
- [130] R. Brisk *et al.*, “WaSP-ECG : A Wave Segmentation Pretraining Toolkit for Electrocardiogram Analysis Artificial Intelligence for,” vol. 13, no. March, pp. 1–17, 2022, doi: 10.3389/fphys.2022.760000.
- [131] H. Martin, U. Morar, W. Izquierdo, M. Cabrerizo, A. Cabrera, and M. Adjouadi, “infarction detection ☆,” vol. 121, no. February, 2021.
- [132] X. Chen, W. Guo, L. Zhao, W. Huang, L. Wang, and A. Sun, “Acute Myocardial Infarction Detection Using Deep Electrocardiograms,” vol. 8, no. August, pp. 1–7, 2021, doi: 10.3389/fcvm.2021.654515.
- [133] K. Tseng, J. Li, Y. Tang, C. W. Yang, and F. Lin, “Healthcare knowledge of relationship between time series electrocardiogram and cigarette smoking using clinical records,” vol. 20, no. Suppl 3, pp. 1–11, 2020.
- [134] R. Ecgs, A. S. Rodrigues, R. Augustauskas, M. Lukoševič, and P. Laguna, “Deep-Learning-Based Estimation of the Spatial QRS-T Angle from Reduced-Lead ECGs,” pp. 1–22, 2022.
- [135] H. Chen, S. Das, J. M. Morgan, and K. Maharatna, “Prediction and classification of ventricular arrhythmia based on phase-space reconstruction and fuzzy c-means clustering,” vol. 142, no. December 2021, 2022.
- [136] H. Mohan, K. Chatterjee, and S. Dashkevych, “The prediction of cardiac abnormality and enhancement

- in minority class accuracy from imbalanced ECG signals using modified deep neural network models,” vol. 150, no. March, 2022.
- [137] T. Yang, R. E. Gregg, and S. Babaeizadeh, “Detection of strict left bundle branch block by neural network and a method to test detection Detection of strict left bundle branch block by neural network and a method to test detection consistency.”
 - [138] Q. Sun *et al.*, “A dynamic learning-based ECG feature extraction method for myocardial infarction detection A dynamic learning-based ECG feature extraction method for myocardial infarction detection,” 2022.
 - [139] Y. Jo *et al.*, “Explainable arti fi cial intelligence to detect atrial fi brillation using electrocardiogram,” *Int. J. Cardiol.*, vol. 328, pp. 104–110, 2021, doi: 10.1016/j.ijcard.2020.11.053.
 - [140] M. M. Farag, “A Tiny Matched Filter-Based CNN for Inter-Patient ECG Classification and Arrhythmia Detection at the Edge,” pp. 1–23, 2023.
 - [141] Q. Geng *et al.*, “An ECG Classification Method Based on Multi-Task Learning and CoT Attention Mechanism,” pp. 1–13, 2023.
 - [142] E. Zvuloni, J. Read, A. H. Ribeiro, A. L. P. Ribeiro, and J. A. Behar, “On Merging Feature Engineering and Deep Learning for Diagnosis , Risk Prediction and Age Estimation Based on the 12-Lead ECG,” *IEEE Trans. Biomed. Eng.*, vol. 70, no. 7, pp. 2227–2236, 2023, doi: 10.1109/TBME.2023.3239527.
 - [143] A. Rath, D. Mishra, and G. Panda, “Imbalanced ECG signal-based heart disease classification using ensemble machine learning technique.”
 - [144] H. Zhang *et al.*, “Cardiac Arrhythmia classi fi cation based on 3D recurrence plot analysis and deep learning,” no. July, pp. 1–13, 2022, doi: 10.3389/fphys.2022.956320.
 - [145] S. Biton *et al.*, “Atrial fibrillation risk prediction from the 12-lead electrocardiogram using digital biomarkers and deep representation learning,” pp. 576–585, 2021, doi: 10.1093/ehjdh/ztab071.
 - [146] U. R. Detection, P. Dataset, K. Pałczy, and S. Smigiel, “Deep Learning Techniques in the Classification of ECG Signals,” pp. 1–18, 2021.
 - [147] Q. Sun *et al.*, “Computer Methods and Programs in Biomedicine Early detection of myocardial ischemia in 12-lead ECG using deterministic learning and ensemble learning,” vol. 226, 2022.
 - [148] D. W. Feyisa, T. G. Debelee, Y. M. Ayano, S. R. Kebede, and T. F. Assore, “Lightweight Multireceptive Field CNN for 12-Lead ECG Signal Classification,” vol. 2022, 2022.
 - [149] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, “Deep Learning for ECG Analysis : Benchmarks and Insights from PTB-XL,” vol. 25, no. 5, pp. 1519–1528, 2021, doi: 10.1109/JBHI.2020.3022989.
 - [150] B. T. Pham, P. T. Le, T. C. Tai, Y. C. Hsu, Y. H. Li, and J. C. Wang, “Electrocardiogram Heartbeat Classification for Arrhythmias and Myocardial Infarction,” *Sensors*, vol. 23, no. 6, pp. 1–18, 2023, doi: 10.3390/s23062993.
 - [151] M. F. Safdar, R. M. Nowak, and P. Pałka, “A Denoising and Fourier Transformation-Based Spectrograms in ECG Classification Using Convolutional Neural Network,” 2022.
 - [152] X. Sun, S. S. Nunes, H. Abdolazadeh, N. Khoshdel, R. Xiao, and C. Ding, “Detecting and interpreting myocardial infarction using fully convolutional neural networks Detecting and interpreting myocardial

- infarction using fully convolutional neural networks,” 2019.
- [153] D. Chumachenko, M. Butkevych, D. Lode, M. Frohme, K. J. G. Schmailzl, and A. Nechyporenko, “Machine Learning Methods in Predicting Patients with Suspected Myocardial Infarction Based on Short-Time HRV Data,” 2022.
 - [154] L. Boudnik, J. Graßhoff, F. Vollmer, and P. Rostalski, “Muscle Artifact Removal in Single-Channel Electrocardiograms using Temporal Convolutional Networks,” pp. 1902–1905, 2022, doi: 10.1109/EMBC48229.2022.9871723.
 - [155] J. Lee, K. Oh, B. Kim, and S. K. Yoo, “Synthesis of Electrocardiogram V-Lead Signals From Limb-Lead Measurement Using R-Peak Aligned Generative Adversarial Network,” vol. 24, no. 5, pp. 1265–1275, 2020, doi: 10.1109/JBHI.2019.2936583.
 - [156] J. Liu, C. Zhang, T. Ristaniemi, S. Member, F. Cong, and S. Member, “Detection of Myocardial Infarction from Multi-lead ECG using Dual-Q Tunable Q-Factor Wavelet Transform,” *2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 1496–1499, 2019.
 - [157] H. Zhang, C. Liu, Z. Zhang, Y. Xing, X. Liu, and R. Dong, “Recurrence Plot-Based Approach for Cardiac Arrhythmia Classification Using Inception-ResNet-v2,” vol. 12, no. May, pp. 1–13, 2021, doi: 10.3389/fphys.2021.648950.
 - [158] A. Bousseljot, R. Kreisler, D. Schnabel, “Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über R. Bousseljot, D. Kreisler und A. Schnabel Physikalisch-Technische Bundesanstalt, Institut Berlin Einleitung: Für die Qualitätssicherung von automatisch auswertenden Elektrokardiogrammen, zur Weiterent,” *Biomed. Tech. / Biomed. Eng.*, vol. 40, no. s1, pp. 317–318, doi: 10.1515/bmte.1995.40.s1.317.
 - [159] L. Xu, D. Zhang, K. Wang, N. Li, and X. Wang, “Baseline wander correction in pulse waveforms using wavelet-based cascaded adaptive filter,” vol. 37, pp. 716–731, 2007, doi: 10.1016/j.compbiomed.2006.06.014.
 - [160] E. Technologies, N. Rastogi, R. Mehra, T. T. Training, and T. T. Training, “A ANALYSIS OF S AVITZKY -G OLAY F ILTER FOR B ASELINE W ANDER C ANCELLATION IN ECG USING W AVELETS,” vol. 6, no. 1, pp. 15–23, 2013.
 - [161] N. Rastogi and R. Mehra, “Analysis of Butterworth and Chebyshev Filters for ECG Denoising Using Wavelets,” vol. 6, no. 6, pp. 37–44, 2013.
 - [162] H. Yan, T. Xu, P. Wang, L. Zhang, H. Hu, and Y. Bai, “MEMS Hydrophone Signal Denoising and Baseline Parameter-Optimized Variational Mode,” 2019.
 - [163] L. Zebin, L. Wei, G. Lifu, and Z. Jinsi, “Research on ECG Denoising method Based on Empirical Mode Decomposition and,” pp. 675–679, 2020, doi: 10.1109/ICSIP49896.2020.9339369.
 - [164] V. Gupta, M. Mittal, and V. Mittal, “Performance Evaluation of Various Pre-Processing Techniques for R-Peak Detection in ECG Signal Performance Evaluation of Various Pre-Processing Techniques for R-Peak Detection in ECG Signal,” 2022, doi: 10.1080/03772063.2020.1756473.
 - [165] M. Das, R. Kumar, and B. C. Sahana, “Implementation of Effective Hybrid Window Function for E . C . G Signal Denoising Traitement du Signal Implementation of Effective Hybrid Window Function for E . C . G Signal Denoising,” no. March, 2020, doi: 10.18280/ts.370116.

- [166] R. Sameni *et al.*, "A Nonlinear Bayesian Filtering Framework for ECG Denoising," vol. 54, no. 12, pp. 2172–2185, 2007.
- [167] B. Chen, Y. Li, X. Cao, W. Sun, and W. He, "Removal of Power Line Interference From ECG Signals Using Adaptive Notch Filters of Sharp Resolution," *IEEE Access*, vol. 7, pp. 150667–150676, 2019, doi: 10.1109/ACCESS.2019.2944027.
- [168] Y. Chen, P. Lin, and Y. Lin, "Biomedical Signal Processing and Control A novel PLI suppression method in ECG by notch filtering with a modulation-based detection and frequency estimation scheme," *Biomed. Signal Process. Control*, vol. 62, no. October 2019, p. 102150, 2020, doi: 10.1016/j.bspc.2020.102150.
- [169] Z. Peng and G. Wang, "Study on Optimal Selection of Wavelet Vanishing Moments for ECG Denoising," pp. 1–11, 2017, doi: 10.1038/s41598-017-04837-9.
- [170] P. G. Malghan, "A review on ECG filtering techniques for rhythm analysis," pp. 171–186, 2020.
- [171] S. Goel, P. Tomar, and G. Kaur, "A Fuzzy Based Approach for Denoising of ECG Signal using Wavelet Transform," vol. 8, no. 2, pp. 143–156, 2016.
- [172] S. H. Nawab and J. T. Ludwig, "Approximate Signal Processing," vol. 200, pp. 177–200, 1997.
- [173] J. G. Proakis, *Digital signal processing: principles, algorithms, and applications*. Pearson Education India, 2007.
- [174] A. Srivastava *et al.*, "A deep residual inception network with channel attention modules for multi-label cardiac abnormality detection from reduced-lead ECG A deep residual inception network with channel attention modules for multi-label cardiac abnormality detection from reduce," 2020.
- [175] Z. Xiong, M. K. Stiles, and J. Zhao, "Robust ECG Signal Classification for Detection of Atrial Fibrillation Using a Novel Neural Network Waikato Hospital , Hamilton , New Zealand Spectrograms provided a representation of ECGs in the," vol. 44, no. Figure 1, pp. 1–4, 2017, doi: 10.22489/CinC.2017.066-138.
- [176] N. Fayyazifar, S. Ahderom, D. Suter, A. Maiorana, G. Dwivedi, and F. S. Hospital, "Impact of Neural Architecture Design on Cardiac Abnormality Classification Using 12-lead ECG Signals," vol. 47, pp. 12–15, 2020, doi: 10.22489/CinC.2020.161.
- [177] X. Wang *et al.*, "Conditional Generative Adversarial Net," *IEEE J. Biomed. Heal. Informatics*, vol. 26, no. 7, pp. 2929–2940, 2022, doi: 10.1109/JBHI.2022.3169325.
- [178] P. Xu, H. Liu, X. Xie, S. Zhou, M. Shu, and Y. Wang, "Interpatient ECG Arrhythmia Detection by Residual Attention CNN," vol. 2022, 2022.
- [179] R. Aggarwal and S. Kumar, "Measurement : Sensors HRV based feature selection for congestive heart failure and normal sinus rhythm for meticulous presaging of heart disease using machine learning," *Meas. Sensors*, vol. 24, no. October, p. 100573, 2022, doi: 10.1016/j.measen.2022.100573.
- [180] T. Pham, Z. J. Lau, and S. H. A. Chen, "Heart Rate Variability in Psychology : A Review of HRV Indices and an Analysis Tutorial," pp. 1–20, 2021.
- [181] P. Van Gent, H. Farah, N. Van Nes, and B. Van Arem, "HeartPy : A novel heart rate algorithm for the analysis of noisy signals," *Transp. Res. Part F Psychol. Behav.*, vol. 66, pp. 368–378, 2019, doi: 10.1016/j.trf.2019.09.015.

- [182] H. Mitrova, B. Koteska, A. Madevska, F. Lehoccki, B. Ondrusova, and N. Ackovska, "Evaluation of Python HeartPy Toolkit for Heart Rate extraction from PPG," pp. 1–9.
- [183] T. H. Ma *et al.*, "Fractal Analysis and Time- and Frequency-Domain Measures of Heart Rate Variability as Predictors of Mortality in Patients With Heart Failure," vol. 87, no. 00, 2001.
- [184] M. A. Salo, M. Sc, H. V Huikuri, T. Seppanen, and D. Ph, "Ectopic Beats in Heart Rate Variability Analysis : Effects of Editing on Time and Frequency Domain Measures," 2001.
- [185] A. Helminen *et al.*, "Time domain , geometrical and frequency domain analysis of cardiac vagal out - ow : effects of various respiratory patterns," vol. 1, pp. 365–376, 2001.
- [186] A. Diker, E. Avci, and M. Gedikpinar, "EKG İşaretinde R tepelerinin Hilbert Dönüşümü ve Pan-Tompkins Algoritmaları Kullanılarak Belirlenmesi," *2017 25th Signal Process. Commun. Appl. Conf. SIU 2017*, pp. 20–23, 2017, doi: 10.1109/SIU.2017.7960235.
- [187] M. A. Z. Fariha, R. Ikeura, S. Hayakawa, and S. Tsutsumi, "Analysis of Pan-Tompkins Algorithm Performance with Noisy ECG Signals," *J. Phys. Conf. Ser.*, vol. 1532, no. 1, 2020, doi: 10.1088/1742-6596/1532/1/012022.
- [188] J. B. Copas and P. F. Malley, "A robust," no. April, pp. 4267–4278, 2008, doi: 10.1002/sim.
- [189] R. Fluss, D. Faraggi, and B. Reiser, "Estimation of the Youden Index and its associated cutoff point," *Biometrical J.*, vol. 47, no. 4, pp. 458–472, 2005, doi: 10.1002/bimj.200410135.
- [190] E. M. Molanes-López and E. Letón, "Inference of the Youden index and associated threshold using empirical likelihood for quantiles," *Stat. Med.*, vol. 30, no. 19, pp. 2467–2480, 2011, doi: 10.1002/sim.4303.
- [191] R. Mitchell, E. Frank, and G. Holmes, "GPUShap: Massively parallel exact calculation of SHAP scores for tree ensembles," *PeerJ Comput. Sci.*, vol. 8, 2022, doi: 10.7717/PEERJ-CS.880.
- [192] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.
- [193] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.
- [194] J. Wang *et al.*, "IRGAN: A minimax game for unifying generative and discriminative information retrieval models," *SIGIR 2017 - Proc. 40th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 515–524, 2017, doi: 10.1145/3077136.3080786.
- [195] Z. Gao, L. Li, and T. Xu, "Data Augmentation for Time-Series Classification: An Extensive Empirical Study and Comprehensive Survey," pp. 1–18, 2023.
- [196] D. Kiyasseh *et al.*, "PlethAugment: GAN-Based PPG augmentation for medical diagnosis in low-resource settings," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 11, pp. 3226–3235, 2020, doi: 10.1109/JBHI.2020.2979608.
- [197] A. Dash, J. Ye, and G. Wang, "A review of Generative Adversarial Networks (GANs) and its applications in a wide variety of disciplines -- From Medical to Remote Sensing," *IEEE Access*, vol. PP, p. 1, 2021, doi: 10.1109/ACCESS.2023.3346273.
- [198] A. N. Wu, R. Stouffs, and F. Biljecki, "Generative Adversarial Networks in the built environment: A

- comprehensive review of the application of GANs across data types and scales,” *Build. Environ.*, vol. 223, no. June, p. 109477, 2022, doi: 10.1016/j.buildenv.2022.109477.
- [199] F. N. Hatamian, N. Ravikumar, S. Vesal, F. P. Kemeth, M. Struck, and A. Maier, “The Effect of Data Augmentation on Classification of Atrial Fibrillation in Short Single-Lead ECG Signals Using Deep Neural Networks,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2020-May, pp. 1264–1268, 2020, doi: 10.1109/ICASSP40776.2020.9053800.
 - [200] M. F. Safdar, P. Pałka, R. M. Nowak, and A. Al Faresi, “A novel data augmentation approach for enhancement of ECG signal classification,” *Biomed. Signal Process. Control*, vol. 86, no. July, 2023, doi: 10.1016/j.bspc.2023.105114.
 - [201] S. Haradal, H. Hayashi, and S. Uchida, “Biosignal Data Augmentation Based on Generative Adversarial Networks,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2018-July, pp. 368–371, 2018, doi: 10.1109/EMBC.2018.8512396.
 - [202] T. Lan, Q. Hu, X. Liu, K. He, and C. Yang, “Arrhythmias Classification Using Short-Time Fourier Transform and GAN Based Data Augmentation,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2020-July, pp. 308–311, 2020, doi: 10.1109/EMBC44109.2020.9176733.
 - [203] M. Kuntalp and O. Düzyel, “A new method for GAN-based data augmentation for classes with distinct clusters,” *Expert Syst. Appl.*, vol. 235, no. August 2023, 2024, doi: 10.1016/j.eswa.2023.121199.
 - [204] W. Zhuang, C. Chen, and L. Lyu, “When Foundation Model Meets Federated Learning: Motivations, Challenges, and Future Directions,” vol. 2, no. c, 2023.
 - [205] J. T. Guibas, T. S. Virdi, and P. S. Li, “Synthetic Medical Images from Dual Generative Adversarial Networks,” no. Nips, pp. 1–9, 2017.
 - [206] T. Burt, K. S. Button, H. H. Z. Thom, R. J. Noveck, and M. R. Munafò, “The Burden of the ‘False-Negatives’ in Clinical Development: Analyses of Current and Alternative Scenarios and Corrective Measures,” *Clin. Transl. Sci.*, vol. 10, no. 6, pp. 470–479, 2017, doi: 10.1111/cts.12478.
 - [207] M. Petticrew, A. Sowden, and D. Lister-Sharp, “False-negative results in screening programs: Medical, psychological, and other implications,” *Int. J. Technol. Assess. Health Care*, vol. 17, no. 2, pp. 164–170, 2001, doi: 10.1017/S0266462300105021.
 - [208] D. Marin *et al.*, “Efficient segmentation: Learning downsampling near semantic boundaries,” *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-Octob, pp. 2131–2141, 2019, doi: 10.1109/ICCV.2019.00222.
 - [209] Q. Wen *et al.*, “Time Series Data Augmentation for Deep Learning: A Survey,” *IJCAI Int. Jt. Conf. Artif. Intell.*, pp. 4653–4660, 2021, doi: 10.24963/ijcai.2021/631.
 - [210] S. Steinarsson, “Downsampling Time Series for Visual Representation,” p. 87, 2013.
 - [211] Y. Wang, “A Mathematical Introduction to Generative Adversarial Nets (GAN),” pp. 1–30, 2020.
 - [212] Y. Li, Q. Wang, J. Zhang, L. Hu, and W. Ouyang, “The theoretical research of generative adversarial networks: an overview,” *Neurocomputing*, vol. 435, pp. 26–41, 2021, doi: 10.1016/j.neucom.2020.12.114.
 - [213] S. Mahdizadehaghdam, A. Panahi, and H. Krim, “Sparse generative adversarial network,” *Proc. - 2019 Int. Conf. Comput. Vis. Work. ICCVW 2019*, pp. 3063–3071, 2019, doi: 10.1109/ICCVW.2019.00369.

- [214] L. Gonog and Y. Zhou, "A review: Generative adversarial networks," *Proc. 14th IEEE Conf. Ind. Electron. Appl. ICIEA 2019*, pp. 505–510, 2019, doi: 10.1109/ICIEA.2019.8833686.
- [215] A. Figueira and B. Vaz, "Survey on Synthetic Data Generation, Evaluation Methods and GANs," *Mathematics*, vol. 10, no. 15, pp. 1–41, 2022, doi: 10.3390/math10152733.
- [216] M. Esmailpour, N. Chaalia, A. Abusitta, F. X. Devailly, W. Maazoun, and P. Cardinal, "Bi-discriminator GAN for tabular data synthesis," *Pattern Recognit. Lett.*, vol. 159, pp. 204–210, 2022, doi: 10.1016/j.patrec.2022.05.023.
- [217] J. Moon, S. Jung, S. Park, and E. Hwang, "Conditional tabular GAN-based two-stage data generation scheme for short-term load forecasting," *IEEE Access*, vol. 8, pp. 205327–205339, 2020, doi: 10.1109/ACCESS.2020.3037063.
- [218] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, 2019.
- [219] C. W. Liu, M. C. Su, and S. S. Tsay, "Application of a novel fuzzy neural network to real-time transient stability swings prediction based on synchronized phasor measurements," *IEEE Trans. Power Syst.*, vol. 14, no. 2, pp. 685–692, 1999, doi: 10.1109/59.761898.
- [220] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially Learned One-Class Classifier for Novelty Detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3379–3388, 2018, doi: 10.1109/CVPR.2018.00356.
- [221] S. Tabik *et al.*, "COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 12, pp. 3595–3605, 2020, doi: 10.1109/JBHI.2020.3037127.
- [222] P. Perera, R. Nallapati, and B. Xiang, "OCGAN: One-class novelty detection using gans with constrained latent representations," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 2893–2901, 2019, doi: 10.1109/CVPR.2019.00301.
- [223] C. Lu, C. K. Reddy, P. Wang, D. Nie, and Y. Ning, "Multi-Label Clinical Time-Series Generation via Conditional GAN," *IEEE Trans. Knowl. Data Eng.*, vol. PP, no. X, pp. 1–13, 2023, doi: 10.1109/TKDE.2023.3310909.
- [224] X. Ke, J. Zou, and Y. Niu, "End-to-End Automatic Image Annotation Based on Deep CNN and Multi-Label Data Augmentation," *IEEE Trans. Multimed.*, vol. 21, no. 8, pp. 2093–2106, 2019, doi: 10.1109/TMM.2019.2895511.
- [225] M.-P. Cote, B. Hartman, O. Mercier, J. Meyers, J. Cummings, and E. Harmon, "Synthesizing Property & Casualty Ratemaking Datasets using Generative Adversarial Networks," 2020.
- [226] D. Hazra, W. Shafqat, and Y. C. Byun, "Generating synthetic data to reduce prediction error of energy consumption," *Comput. Mater. Contin.*, vol. 70, no. 2, pp. 3151–3167, 2022, doi: 10.32604/cmc.2022.020143.
- [227] Q. Wang *et al.*, "WGAN-Based Synthetic Minority Over-Sampling Technique: Improving Semantic Fine-Grained Classification for Lung Nodules in CT Images," *IEEE Access*, vol. 7, pp. 18450–18463, 2019, doi: 10.1109/ACCESS.2019.2896409.
- [228] J. D. Katsilometes, "How good is my maintenance program?," *Improv. Optim. Oper. Things That*

Actually Work - Plant Oper. Forum 2004, no. 5302, pp. 27–37, 2005.

- [229] H. A. Khodja and O. Boudjeniba, “Application of WGAN-GP in recommendation and Questioning the relevance of GAN-based approaches,” pp. 1–8.
- [230] J. Zhu, G. Yang, and P. Lio, “How can we make gan perform better in single medical image super-resolution? A lesion focused multi-scale approach,” *Proc. - Int. Symp. Biomed. Imaging*, vol. 2019-April, no. Isbi, pp. 1669–1673, 2019, doi: 10.1109/ISBI.2019.8759517.
- [231] L. Weng, “From GAN to WGAN,” 2019.
- [232] K. Liu and G. Qiu, “Lipschitz constrained GANs via boundedness and continuity,” *Neural Comput. Appl.*, vol. 32, no. 24, pp. 18271–18283, 2020, doi: 10.1007/s00521-020-04954-z.
- [233] S. Cui and Y. Jiang, “Effective lipschitz constraint enforcement for wasserstein GAN training,” *2017 2nd IEEE Int. Conf. Comput. Intell. Appl. ICCIA 2017*, vol. 2017-Janua, pp. 74–78, 2017, doi: 10.1109/CIAPP.2017.8167183.
- [234] D. Terjék, “Adversarial Lipschitz Regularization,” *8th Int. Conf. Learn. Represent. ICLR 2020*, pp. 1–17, 2020.
- [235] Z. Zhou, J. Shen, Y. Song, W. Zhang, and Y. Yu, “Towards Efficient and Unbiased Implementation of Lipschitz Continuity in GANs,” 2019.
- [236] R. Silva and P. Melo-Pinto, “t-SNE: A study on reducing the dimensionality of hyperspectral data for the regression problem of estimating oenological parameters,” *Artif. Intell. Agric.*, vol. 7, pp. 58–68, 2023, doi: 10.1016/j.aiia.2023.02.003.
- [237] F. Anowar, S. Sadaoui, and B. Selim, “Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE),” *Comput. Sci. Rev.*, vol. 40, 2021, doi: 10.1016/j.cosrev.2021.100378.
- [238] K. Pal and M. Sharma, “Performance evaluation of non-linear techniques UMAP and t-SNE for data in higher dimensional topological space,” *Proc. 4th Int. Conf. IoT Soc. Mobile, Anal. Cloud, ISMAC 2020*, pp. 1106–1110, 2020, doi: 10.1109/I-SMAC49090.2020.9243502.
- [239] T. Wang, S. Meng, and J. Bian, “Indexing earth mover’s distance over network metrics,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 6, pp. 1588–1601, 2015, doi: 10.1109/TKDE.2014.2373359.
- [240] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher, “Earth mover’s distances on discrete surfaces,” *ACM Trans. Graph.*, vol. 33, no. 4, 2014, doi: 10.1145/2601097.2601175.
- [241] J. R. Potts, M. Auger-Méthé, K. Mokross, and M. A. Lewis, “A generalized residual technique for analysing complex movement models using earth mover’s distance,” *Methods Ecol. Evol.*, vol. 5, no. 10, pp. 1012–1022, 2014, doi: 10.1111/2041-210X.12253.
- [242] V. M. Panaretos and Y. Zemel, “Statistical aspects of wasserstein distances,” *Annu. Rev. Stat. Its Appl.*, vol. 6, pp. 405–431, 2019, doi: 10.1146/annurev-statistics-030718-104938.
- [243] S. Steinerberger, “Wasserstein distance, Fourier series and applications,” *Monatshefte fur Math.*, vol. 194, no. 2, pp. 305–338, 2021, doi: 10.1007/s00605-020-01497-2.
- [244] R. Mir, B. Felbo, N. Obradovich, and I. Rahwan, “Evaluating style transfer for text,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp.

495–504, 2019, doi: 10.18653/v1/n19-1049.

- [245] S. Van Gassen, B. Gaudilliere, M. S. Angst, Y. Saeys, and N. Aghaeepour, “CytoNorm: A Normalization Algorithm for Cytometry Data,” *Cytom. Part A*, vol. 97, no. 3, pp. 268–278, 2020, doi: 10.1002/cyto.a.23904.
- [246] I. Vaccari, V. Orani, A. Paglialonga, E. Cambiaso, and M. Mongelli, “A generative adversarial network (GAN) technique for internet of medical things data,” *Sensors*, vol. 21, no. 11, pp. 1–14, 2021, doi: 10.3390/s21113726.
- [247] M. Abedi, L. Hempel, S. Sadeghi, and T. Kirsten, “GAN-Based Approaches for Generating Structured Data in the Medical Domain,” *Appl. Sci.*, vol. 12, no. 14, 2022, doi: 10.3390/app12147075.

10 Appendices

10.1 GAN Architectures – Binary Classification – Raw Signals

In our study, we implemented a comprehensive approach involving 24 distinct GAN architectures to train each of the 12 leads for Myocardial Infarction (MI) and Healthy control classes. This extensive array of architectures was essential to accommodate the unique characteristics of each lead.

The core strategy involved the hyper-tuning of various parameters across these architectures. Key aspects such as the number of dense layers, activation functions, loss functions, and optimizers were meticulously adjusted to optimize the performance of each GAN. Such hyper-tuning is crucial in tailoring the model to capture the nuances of ECG signal characteristics accurately.

For most leads, the architecture followed a consistent pattern, with a bi-directional LSTM model employed for the generator and a CNN for the discriminator. This combination was chosen for its effectiveness in capturing temporal dependencies (via LSTM) and spatial features (via CNN) within the ECG data. However, even within this framework, slight variations were necessary.

These modifications included adjustments in dense layers, filters, and learning rates, driven by hyperparameter tuning. The process of tuning was guided by the need to balance model complexity with computational efficiency, ensuring that each model could effectively learn from the data without overfitting or underperforming.

This section will detail the architecture for a few representative leads, illustrating the commonalities and variations across the different GANs used in our study. It's important to note that while the underlying structure remains consistent, the specific configurations were tailored to the peculiarities of each ECG lead, demonstrating the adaptability and versatility of our approach to GAN architecture in handling complex biomedical data.

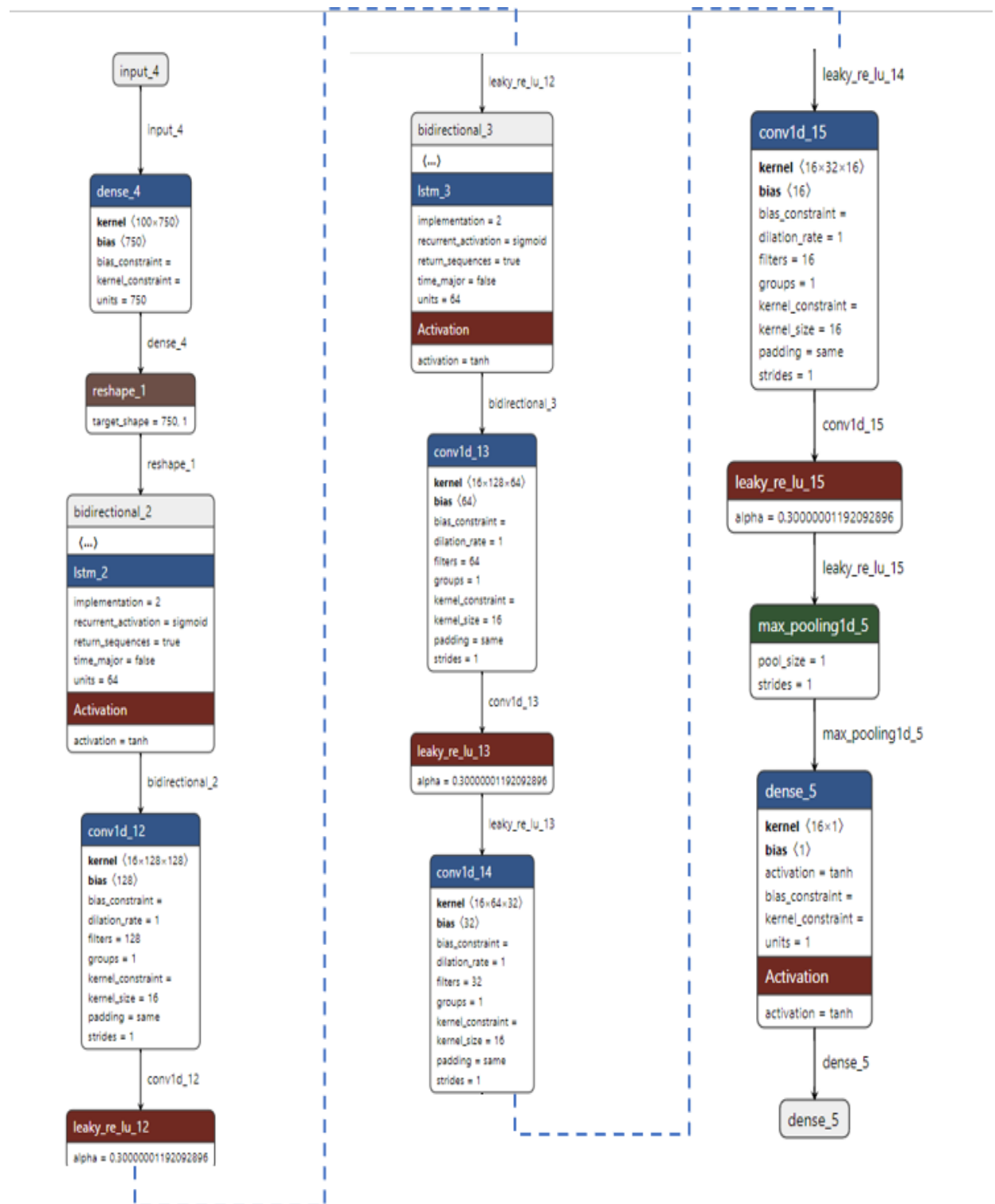


Figure 10-1: Generator Model for synthetic samples of HC Class - Lead V6

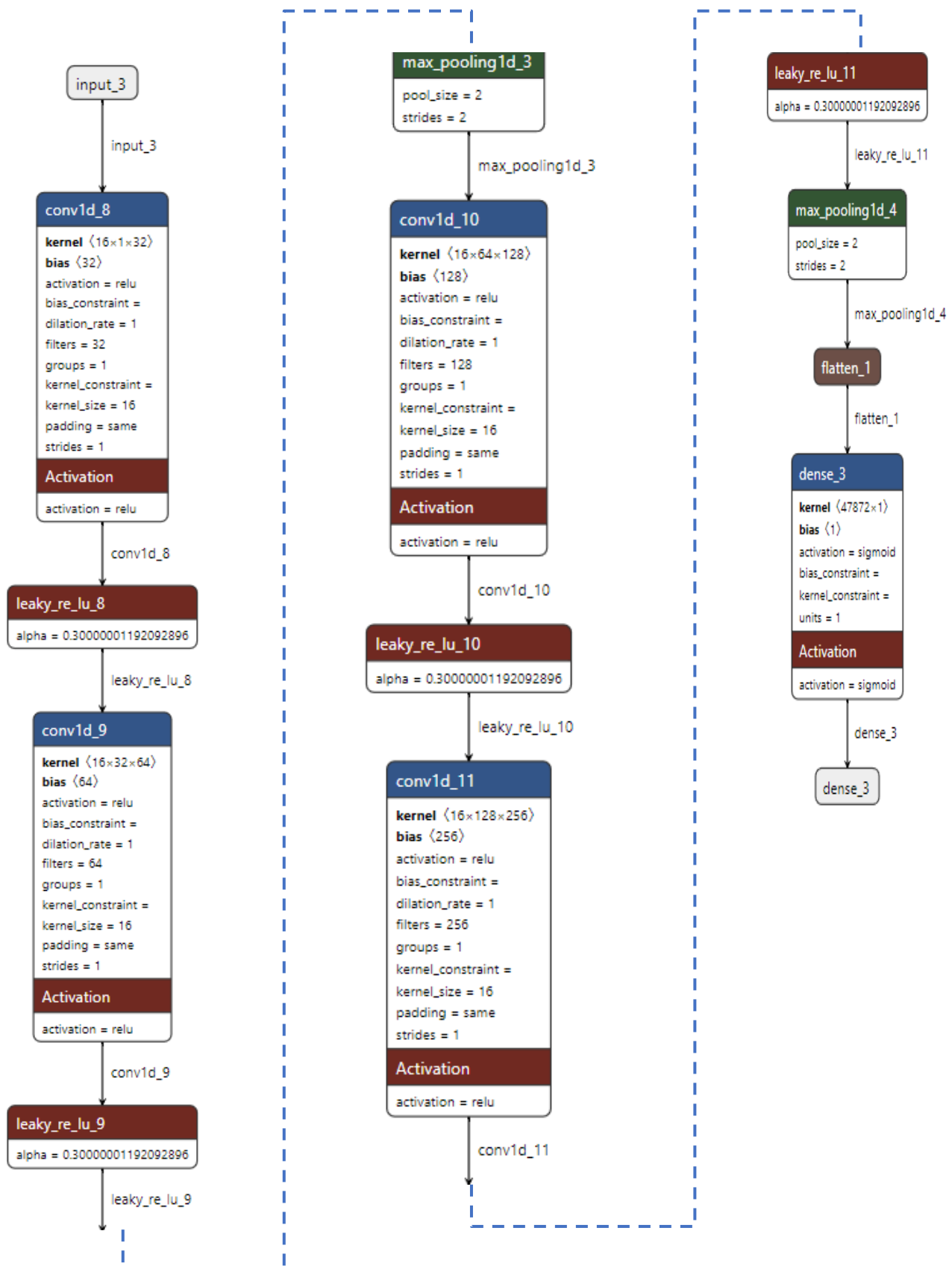


Figure 10-2: Discriminator Model for synthetic samples of HC Class - Lead V6

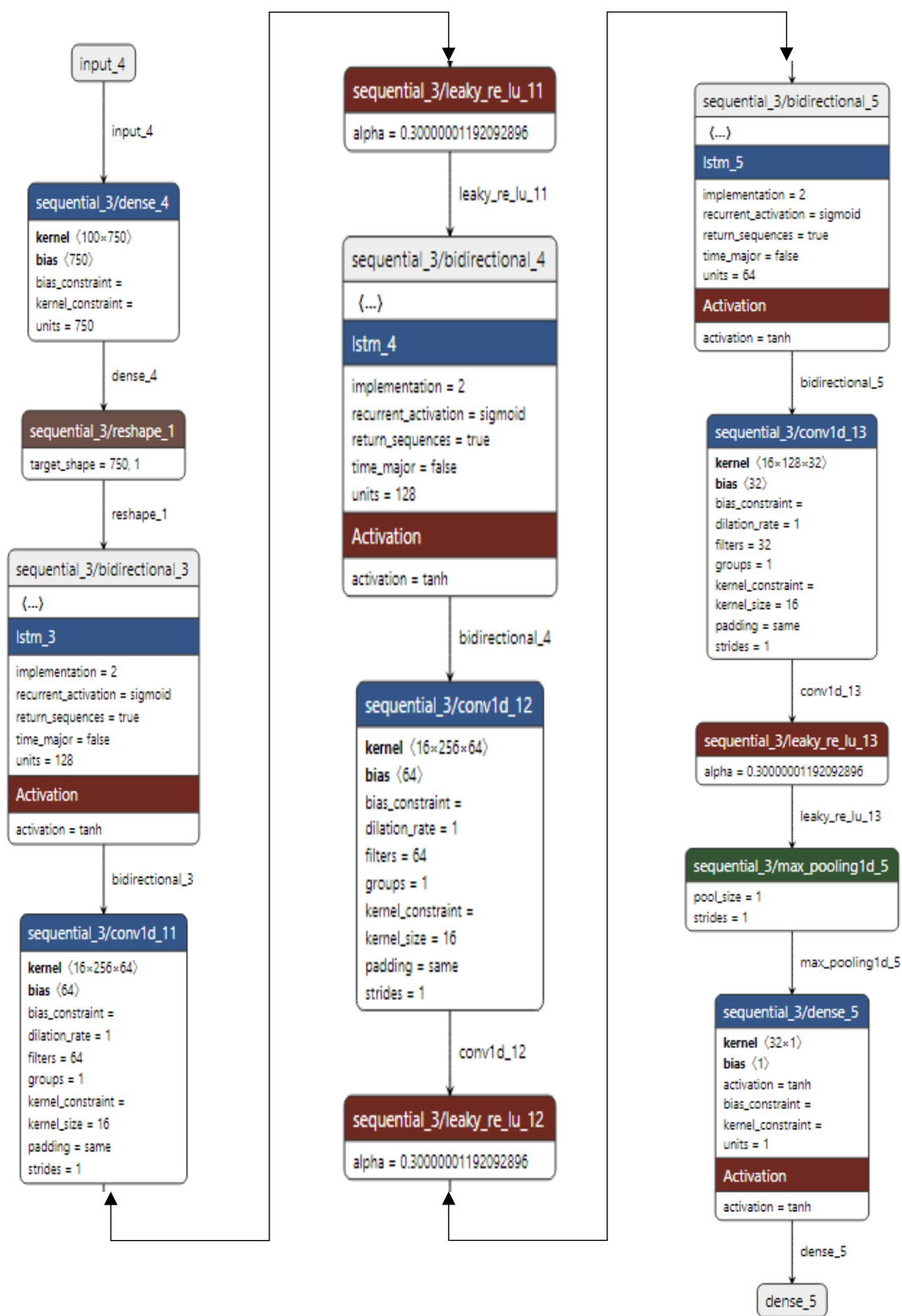


Figure 10-3: Generator Model for synthetic samples of MI Class - Lead V6

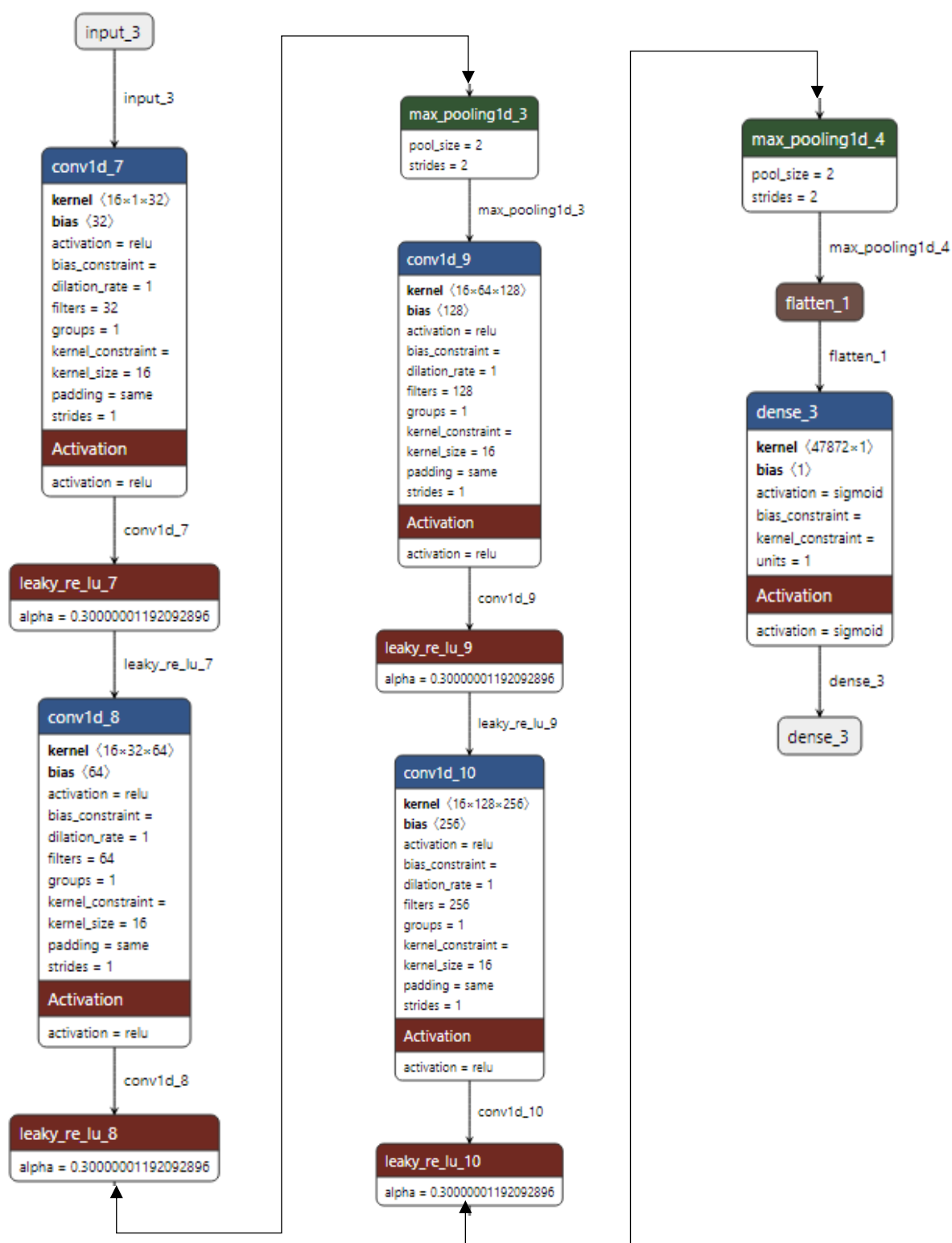


Figure 10-4:Discriminator Model for synthetic samples of MI Class - Lead V6

10.2 Image Segmentation Experiments and results

Table 10-1 : Segmentation Experiments and results performed on Local resource

No.	Experiments	Epochs & Batch Size	Training Accuracy Test Accuracy	Training Mean IOU Test Mean IOU	Train Images Test Images
1	Simple Vanilla model without any image augmentations with image resized to 200*200	Epoch = 15 Batch Size = 8	0.9993 0.9976		1320 440
2	Simple Vanilla model without any image augmentations with image resized to 200*200	Epoch = 15 Batch Size = 8		0.5496 0.5405	1320 440
3	Geometric transformations of images with vertical, flipping, horizon flips, angular rotations, image cropping with image resized to 200*200	Epoch = 15 Batch Size = 8		0.5834 0.5563	1320 440
4	Adaptive histogram equalization for both LA Image and binary segmentation (Ground Truth) with image resized to 200*200	Epoch = 15 Batch Size = 8		1.0 0.4967	1320 440
5	Adaptive histogram equalization for only LA Image with image resized to 200*200	Epoch = 15 Batch Size = 8		0.4967 0.4967	1320 440
6	Simple Vanilla model without any image augmentations, with image resized to 160*160	Epoch = 15 Batch Size = 8		0.5496 0.5405	1320 440
7	contrast stretching equalization for only LA image with image resized to 200*200	Epoch = 50 Batch Size = 8		0.6933 0.4964	1320 440
8	contrast stretching equalization for only LA image with image resized to 200*200	Epoch = 75 Batch Size = 16		0.6965 0.6339	1320 440
9	Simple Vanilla model without any image augmentations, with image resized to 200*200	Epoch = 100 Batch Size = 128		0.8532 0.7993	8800 4752

Table 10-2: Segmentation experiments and results done on Server

No.	Model Hyperparameters	Data	Training Time and Processor	Model Performance			Remarks
				Training Data	Validation Data	Test Data	
1	Unet Levels - 4 Initial Features - 32 Kernel Size = 3 Image Size = 200*200 Batch Size = 128 Total Trainable Parameters - 2140065 Image Preprocessing - CLAHE (*Double Normalisation of images, CLAHE function and DataArgs) Loss Function - Binary Cross Entropy Optimizer Adam	Training Images - 8800 LA MRI Images and 8800 LA Segmented Masks Test Data - 4752 LA MRI Images and 4752 LA Segmented Masks	Intel - CPU 35 Hours	Mean IOU 0.4971 Loss* 0.0021	NA	Mean IOU 0.4964 Loss* 0.2833	Mlg Server
2	Unet Levels - 4 Initial Features - 32 Kernel Size = 3 Image Size = 200*200 Batch Size = 64 Total Trainable Parameters - 2140065 Image Preprocessing - Contrast Stretching only for Train images data and not for segmented masks (*Double Normalisation of images, CLAHE function and DataArgs) Loss Function - Binary Cross Entropy Optimizer Adam	Training Images - 8800 LA MRI Images and 8800 LA Segmented Masks Test Data - 4752 LA MRI Images and 4752 LA Segmented Masks	Intel - CPU 35 Hours	Mean IOU 0.7015 Loss* 4.6219e-04	NA	Mean IOU 0.4965 Loss* 0.1896	Mlg Server
3	Unet Levels - 4 Initial Features - 32 Kernel Size = 3 Image Size = 200*200 Batch Size = 128 Total Trainable Parameters - 2140065 Image Preprocessing - Contrast Stretching for Train images, Test images and Test masks (*Double Normalisation of images, CLAHE function and DataArgs)	Training Images - 8800 LA MRI Images and 8800 LA Segmented Masks Test Data - 4752 LA MRI Images and 4752 LA Segmented Masks	Intel - CPU 35 Hours	Mean IOU 0.8128 Loss* 5.873e-04		Mean IOU 0.4982 Loss* 0.1501	Mlg Server
4	Unet Levels - 4 Initial Features - 32 Kernel Size = 3 Image Size = 200*200 Batch Size = 128 Total Trainable Parameters -	Training Images - 8800 LA MRI Images and 8800 LA Segmented Masks	Intel - CPU 35 Hours	Mean IOU 0.8519 Loss* 0.0002452	Mean IOU 0.8099 Loss* 0.0120	Mean IOU 0.7662 Loss* 0.0178	Mlg Server

	2140065 Image Preprocessing - Contrast Stretching for Train, Test and Validation MRI images only and not for segmented mask images in Train, Test and Validation sets (*one time normalization of images) Loss Function - Binary Cross Entropy Optimizer Adam	Validation Data - 2376 LA MRI Images and 2376 LA Segmented Masks Test Data - 2376 LA MRI Images and 2376 LA Segmented Masks					
5	Unet Levels - 4 Initial Features - 32 Kernel Size = 3 Image Size = 200*200 Batch Size = 128 Total Trainable Parameters - 2140065 Image Preprocessing - CLAHE (Adaptive Equalisation) for Train, Test and Validation MRI images only and not for segmented mask images in Train, Test and Validation sets (*one time normalization of images) Loss Function - Binary Cross Entropy Optimizer Adam	Training Images - 8800 LA MRI Images and 8800 LA Segmented Masks Validation Data - 2376 LA MRI Images and 2376 LA Segmented Masks Test Data - 2376 LA MRI Images and 2376 LA Segmented Masks	Intel - CPU 35 Hours	Mean IOU 0.8109 Loss* 4.5647e ⁻⁰⁴		Mean IOU 0.7955 Loss* 0.0129	Mlg Server
6	Unet Levels - 4 Initial Features - 32 Kernel Size = 4 Image Size = 192*192 Batch Size = 128 Total Trainable Parameters - 3803265 Image Preprocessing - CLAHE (Adaptive Equalisation - Clip limit 0.01) for Train, Test and Validation MRI images only and not for segmented mask images in Train, Test and Validation sets (*one time normalization of images) Loss Function - Binary Cross Entropy Optimizer Adam	Training Images - 8800 LA MRI Images and 8800 LA Segmented Masks Validation Data - 2376 LA MRI Images and 2376 LA Segmented Masks Test Data - 2376 LA MRI Images and 2376 LA Segmented Masks	Intel - CPU 60 Hours	Mean IOU 0.7907 DICE 0.9759 Loss* 0.0005274	Mean IOU 0.7498 DICE 0.8838 Loss* 0.0093	Mean IOU 0.7247 DICE 0.8573 Loss* 0.0123	Mlg Server
7	Unet Levels - 4 Initial Features - 32 Kernel Size = 4 Image Size = 192*192 Batch Size = 16 Total Trainable Parameters - 3803265	Training Images - 8800 LA MRI Images and 8800 LA Segmented Masks Validation Data -	Nivida - GPU 5 Hrs	Mean IOU 0.76 DICE 0.98 Loss* .0003	Mean IOU 0.67 DICE 0.89 Loss* .009	Mean IOU 0.66 DICE 0.87 Loss* .001	Pandora Server

	Image Preprocessing - CLAHE (Adaptive Equalisation - Clip limit 0.02) Loss Function - Binary Cross Entropy Optimizer Adam	2376 LA MRI Images and 2376 LA Segmented Masks Test Data - 2376 LA MRI Images and 2376 LA Segmented Masks					
8	Unet Levels - 5 Initial Features - 32 Kernel Size = 3 Image Size = 192*192 Batch Size = 16 Total Trainable Parameters - 8629921 Image Preprocessing - CLAHE (Adaptive Equalisation - Clip limit 0.03) Loss Function - Binary Cross Entropy Optimizer Adam	Training Images - 8800 LA MRI Images and 8800 LA Segmented Masks Validation Data - 2376 LA MRI Images and 2376 LA Segmented Masks Test Data - 2376 LA MRI Images and 2376 LA Segmented Masks	Nivida - GPU 5 Hrs	Mean IOU 0.85 DICE 0.85 Loss* .0009	Mean IOU 0.79 DICE 0.70 Loss* .01	Mean IOU 0.76 DICE 0.65 Loss* .017	Pandora Server
9	Unet Levels - 4 Initial Features - 32 Kernel Size = 5 Image Size = 192*192 Batch Size = 32 Total Trainable Parameters - 5941665 Image Preprocessing - CLAHE (Adaptive Equalisation - Clip limit 0.05) Loss Function - Binary Cross Entropy Optimizer Adam	Training Images - 8800 LA MRI Images and 8800 LA Segmented Masks Validation Data - 2376 LA MRI Images and 2376 LA Segmented Masks Test Data - 2376 LA MRI Images and 2376 LA Segmented Masks	Nivida - GPU 3 Hrs 20 Mins	Mean IOU 0.69 DICE 0.97 Loss* .0005	Mean IOU 0.64 DICE 0.87 Loss* .009	Mean IOU 0.63 DICE 0.85 Loss* .01	Pandora Server
10	Unet Levels - 5 Initial Features - 32 Kernel Size = 4 Image Size = 192*192 Batch Size = 16 Total Trainable Parameters - 15339393 Image Preprocessing - Contrast Stretching Loss Function - Binary Cross Entropy Optimizer Adam	Training Images - 8800 LA MRI Images and 8800 LA Segmented Masks Validation Data - 2376 LA MRI Images and 2376 LA Segmented Masks Test Data - 2376 LA MRI	Nivida - GPU 3 Hrs	Mean IOU 0.82 DICE 0.99 Loss* .0002	Mean IOU 0.75 DICE 0.90 Loss* .008	Mean IOU 0.73 DICE 0.88 Loss* .012	Pandora Server

		Images and 2376 LA Segmented Masks					
11	Unet Levels - 5 Initial Features - 32 Kernel Size = 5 Image Size = 192*192 Batch Size = 32 Total Trainable Parameters - 23965857 Image Preprocessing - Contrast Stretching Loss Function - Binary Cross Entropy Optimizer Adam	Training Images - 8800 LA MRI Images and 8800 LA Segmented Masks Validation Data - 1760 LA MRI Images and 2376 LA Segmented Masks Test Data - 4752 LA MRI Images and 2376 LA Segmented Masks	Nivida - GPU 3 Hrs 10 Mins	Mean IOU 0.85 DICE 0.98 Loss* .0003	Mean IOU 0.82 DICE 0.91 Loss* .008	Mean IOU 0.77 DICE 0.87 Loss* .013	Pandora Server
12	Unet Levels - 5 Initial Features - 32 Kernel Size = 4 Image Size = 192*192 Batch Size = 16 Total Trainable Parameters - 15339393 Image Preprocessing - CLAHE (Adaptive Equalisation - Clip limit 0.0125) Loss Function - Tversky Loss Function Optimizer Adam	Training Images - 7040 LA MRI Images and 8800 LA Segmented Masks Validation Data - 1760 LA MRI Images and 2376 LA Segmented Masks Test Data - 4752 LA MRI Images and 2376 LA Segmented Masks	Nivida - GPU 3 Hrs	Mean IOU 0.94 DICE 0.94 Tversky 0.94 Loss* .06	Mean IOU 0.88 DICE 0.87 Tversky 0.87 Loss* .13	Mean IOU 0.89 DICE 0.87 Tversky 0.87 Loss* .13	Pandora Server
13	Unet Levels - 4 Initial Features - 32 Kernel Size = 4 Image Size = 192*192 Batch Size = 32 Total Trainable Parameters - 3803265 Image Preprocessing - CLAHE (Adaptive Equalisation - Clip limit 0.01) Loss Function - Tversky Loss Function Optimizer Adam	Training Images - 7040 LA MRI Images and 8800 LA Segmented Masks Validation Data - 1760 LA MRI Images and 2376 LA Segmented Masks Test Data - 4752 LA MRI Images and 2376 LA Segmented Masks	Nivida - GPU 3 Hrs	Mean IOU 0.93 DICE 0.93 Tversky 0.93 Loss* .07	Mean IOU 0.89 DICE 0.87 Tversky 0.87 Loss* .13	Mean IOU 0.89 DICE 0.87 Tversky 0.87 Loss* .13	Pandora Server

14	Unet Levels - 5 Initial Features - 32 Kernel Size = 4 Image Size = 256*256 Batch Size = 16 Total Trainable Parameters - 15339393 Image Preprocessing - CLAHE (Adaptive Equalisation - Clip limit 0.015) Loss Function - Dice Coefficient Loss Optimizer Adam	Training Images - 7040 LA MRI Images and 8800 LA Segmented Masks Validation Data - 1760 LA MRI Images and 2376 LA Segmented Masks Test Data - 4752 LA MRI Images and 2376 LA Segmented Masks	Nivida - GPU 3 Hrs	Mean IOU 0.97 DICE 0.97 Tversky 0.97 Loss* .03	Mean IOU 0.87 DICE 0.85 Tversky 0.92 Loss* .14	Mean IOU 0.89 DICE 0.88 Tversky 0.91 Loss* .12	Pandora Server
15	Unet Levels - 3 Initial Features - 32 Kernel Size = 4 Image Size = 256*256 Batch Size = 16 Total Trainable Parameters - 918785 Image Preprocessing - CLAHE (Adaptive Equalisation - Clip limit 0.01) Loss Function - Dice Coefficient Loss Optimizer Adam	Training Images - 7040 LA MRI Images and 8800 LA Segmented Masks Validation Data - 1760 LA MRI Images and 2376 LA Segmented Masks Test Data - 4752 LA MRI Images and 2376 LA Segmented Masks	Nivida - GPU 3 Hrs	Mean IOU 0.97 DICE 0.97 Tversky 0.97 Loss* .03	Mean IOU 0.88 DICE 0.86 Tversky 0.86 Loss* .14	Mean IOU 0.89 DICE 0.88 Tversky 0.88 Loss* .12	Pandora Server