UNIVERSITÀ DEGLI STUDI DI MILANO

CORSO DI DOTTORATO INDUSTRIALE
IN INTERSECTORAL INNOVATION – DOTTORATO
INTERSETTORIALE PER L'INNOVAZIONE

XXXVIII Cycle
DEPARTMENT OF COMPUTER SCIENCE

Domain Knowledge-Guided Learning for Robust Myocardial Infarction Detection from 12-Lead Electrocardiograms

INF/01 - INFO-01/A - INFORMATICA

TUTOR:
Prof. Roberto SASSI

CO-TUTOR:
Dr. Massimo Walter RIVOLTA

DIRECTOR OF DOCTORAL PROGRAMME:
Prof. Ernesto DAMIANI

DOCTORAL DISSERTATION OF:
Silvia IBRAHIMI
Matr.: R14103
ORCID n.: 0009-0001-2888-4544

Academic Year 2024/2025

Acknowledgment

This doctoral research was supported by Novartis and by the Biomedical Image and Signal Processing Lab at the Department of Computer Science, University of Milan.

Abstract

Background: Myocardial infarction (MI) represents one of the leading causes of morbidity and mortality on a global scale. Diagnosis is primarily based on the interpretation of the 12-lead electrocardiogram (ECG) according to established clinical guidelines that specify alterations in ECG components. Manual ECG interpretation is time-consuming and prone to inter-observer variability, and motivated the development of automated MI diagnosis. In this context, deep learning (DL) has emerged as a promising approach for identifying MI from 12 lead ECG.

Challenges: Despite the encouraging results reported for MI diagnosis, DL models are hindered by several key limitations. First, existing models often overlook the electrocardiographic domain knowledge (DK) codified into clinical decision rules. Without the explicit incorporation of such DK, these models may learn feature representations that are not physiologically grounded, thereby reducing their clinical generalisability. Second, when a DL model is trained on biased datasets, it may rely on spurious correlations associated with age rather than learning MI-relevant features. Third, prior work primarily addresses MI detection, stage classification, and localisation as separate tasks, while rarely integrating all three within a unified framework. As a result, the physiological interdependencies among these diagnostic tasks remain underexploited. Finally, ensuring robustness under dataset shifts remains a critical challenge in DL-based MI diagnosis. Models developed and evaluated primarily on internal datasets may fail to maintain performance when applied to external populations with different demographic distributions or ECG acquisition protocols, thereby limiting their generalisability.

Objectives: The main objectives of this thesis were to improve the robustness of a DL model for MI diagnosis from 12-lead ECGs by: i) mitigating age bias for MI diagnosis; ii) incorporating DK into the DL model (DK-DL) to enhance clinically meaningful representations, and to compare its performance with a DL model trained in a standard way (B-DL) and with an implemented rule-based algorithm (RBA) that followed clinical guideline criteria; iii) introducing a multitask learning framework that simultaneously modelled MI stage (acute *vs.* prior *vs.* normal) classification and localisation by explicitly leveraging interdependencies among related MI diagnostic tasks; and iv) conducting a comprehensive evaluation of the DL models for MI diagnosis using external datasets to assess their robustness across diverse populations and ECG acquisition protocols.

Methods: In this thesis, an adversarial multitask learning framework was proposed to train a DL model using contrastive objectives for MI diagnosis while mitigating age-related spurious correlations. In addition, a DL training framework was designed to incorporate DK to perform both MI detection, staging and localisation. Two strategies were proposed to inject DK into the DL model through two custom regularisation terms in the objective function to control the latent space. Specifically, the strategies were aimed at: i) learning specific ECG components, such as ST-segment, Q and R wave amplitudes, and Q wave durations; and ii) reconstructing a latent space from a set of differentiable approximations of clinical rules. The methods, *i.e.*, DK-DL, B-DL and RBA were developed on PTB-XL+ dataset. In addition, their performance were evaluated on three external datasets, namely CODE, MIMIC-IV and Chapman-Shaoxing.

Results: In this thesis, the proposed AML strategy effectively mitigated age-related bias, decreasing the Pearson correlation coefficient between predictions and age from 0.67 to -0.03, while maintaining an accuracy of 0.85. In addition, incorporating DK into the DL model improved performance in MI staging and localisation tasks. Specifically, the DK-DL model outperformed both the B-DL and RBA in acute MI detection, achieving a higher average recall (0.70 *vs.* 0.53 and 0.65, respectively). When considering overall MI detection (including both acute and prior MI) performance across all four available datasets, the DK-DL model maintained superior performance (0.91), compared with B-DL (0.89) and RBA (0.75). For MI localisation, the DK-DL model achieved mean recall values exceeding 0.84 across all anatomical regions, with major improvements in the lateral territory. These findings demonstrate that AML effectively mitigated age-related bias, whereas DK incorporation enhanced robustness of DL model for MI diagnosis.

Conclusions: This thesis advances MI diagnosis by incorporating DK into DL frameworks, enabling more clinically aligned and robust ECG-based analysis. The proposed methodologies demonstrated improved generalisation and reduced reliance on spurious correlations, thereby enhancing the robustness of automated MI diagnosis across heterogeneous clinical environments. Collectively, the contributions of this thesis provide a foundation for the development of clinically generalisable DL models for 12-lead ECG.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AML	Adversarial Multitask Learning
AV	Atrioventricular
BCE	Binary Cross Entropy
B-DL	Baseline Deep Learning
CNN	Convolutional Neural Network
DK	Domain Knowledge
DL	Deep Learning
ECG	Electrocardiogram
G-mean	Geometric Mean
GRL	Gradient Reversal Layer
LAD	Left Anterior Descending artery
LBBB	Left Bundle Branch Block
LCx	Left Circumflex Artery
LVH	Left Ventricular Hypertrophy
MI	Myocardial Infarction
ML	Machine Learning
MSE	Mean Squared Error
MTL	Multitask Learning
NOMI	Non-Occlusion MI
NORM	Normal Sinus Rhythm
NSTEMI	Non ST-segment Elevation MI
OMI	Occlusion MI
RBA	Rule Based Algorithm
RCA	Right Coronary Artery
ReLU	Rectified Linear Unit
ROC-AUC	Area Under the Receiver Operating Characteristic Curve
ResNet	Residual Networks
SA	Sinoatrial
STEMI	ST-segment Elevation MI
UDMI	Universal Definition of MI
Uni-G	University of Glasgow
VM	Vector Magnitude

Chapter I

Introduction

1.1 Motivation

Cardiovascular diseases represent the leading cause of mortality worldwide. According to the World Health Organization, they account for approximately 32% of all global deaths, corresponding to nearly 20 million individuals, of which 85% were attributable to myocardial infarction (MI) and stroke [2]. Despite significant advances in therapeutic management and improved control of cardiovascular risk factors, MI continues to pose a major global public health challenge.

Over recent decades, epidemiological studies indicate a general decline in mortality rates of acute MI in several countries [3, 4, 5]. For instance, a longitudinal study analysing data collected between 1985 and 2010 in several European countries reported a declining trend in acute MI attack rates over time, with higher rates among individuals aged over 65 years [3]. Similar trends have been observed across 27 EU countries from 2012 to 2020, with a decline in age-adjusted acute MI-related mortality, with a more pronounced decline among individual aged over 65 [4]. In the United States, acute MI-related mortality has also decreased among adults aged 15-44 years between 1999 and 2020 [5].

However, these improvements in relative mortality rates obscure a concerning demographic reality. Due to global population growth and ageing, the absolute number of MI-related deaths continues to rise. Over the past three decades, deaths attributable to MI have increased by 60% [6], and projections estimate that annual cardiovascular deaths could reach 35.6 million by 2050 [7]. This trend is particularly concerning in regions such as Northern Africa,

Western and Central Asia, and Latin America, where the proportion of older adults is expected to double by 2050 [8].

In addition to these global demographic trends, disparities in MI incidence and outcome are observed across sex and age groups. A study showed that the prevalence of acute MI was approximately 10% among individuals over 60 years of age, whereas it was 3.8% in those under 60 [9]. MI remains more prevalent among men [5]. However, women, particularly younger, often experience worse outcomes, in part due to delayed recognition and treatment [5, 10]. While older populations have benefited most from declining mortality trends [4, 5], MI in younger individuals represents a growing clinical and socioeconomic concern. Because MI in this population occurs during the most economically productive years of life, its societal impact is substantial. For instance, in Spain, productivity losses due to acute MI-related premature deaths between 2013 and 2022 have been estimated at over €5.5 billion, with a rising annual trend [11].

Beyond mortality, MI imposes a burden in terms of morbidity, quality of life, and healthcare costs. Post-MI survivors are at elevated risk of arrhythmia, heart failure, valvular dysfunction, and premature death [12]. The economic consequences are equally significant: healthcare resource utilisation in the year following an acute MI has been shown to triple compared to the year preceding the event [13], with similar patterns reported in Italy, where MI is recognised as the leading cause of escalating healthcare expenditure [14].

Therefore, early and accurate diagnosis of MI is critical to improve clinical outcomes. Rapid intervention significantly reduces myocardial damage and subsequent complications. However, timely diagnosis remains challenging due to the heterogeneity of clinical presentations. Many patients present with atypical symptoms or remain asymptomatic, leading to unrecognised or delayed diagnoses [15]. While chest pain is the most common symptom, many patients, particularly women, may present with non–chest pain symptoms, leading to underrecognition and delayed treatment [10]. Such diagnostic delays contribute to worse outcomes and higher mortality, especially among younger women [10].

In current clinical practice, early MI diagnosis relies on the interpretation of the electrocardiogram (ECG), a process which is time-consuming, requires specialised expertise, and is subject to inter-observer variability [16, 17]. Further details on the ECG waveform and MI-related alterations are provided in Chapter II. The growing global burden of MI underscores the need

for scalable and reliable diagnostic support systems to assist clinicians in early identification and management.

Recent advances in artificial intelligence (AI), particularly deep learning (DL), have enabled data-driven approaches for automated ECG analysis. DL-based models have demonstrated promising performance in detecting subtle ECG patterns associated with MI [18]. By enabling rapid and automated ECG interpretation, these approaches have the potential to improve diagnostic accuracy and clinical outcomes.

However, despite encouraging results, many existing DL models remain limited by insufficient external validation, restricted generalisability across populations, and potential performance disparities across age and sex groups. Ensuring robustness is essential before such models can be safely integrated into clinical practice. These challenges highlight the need for the development and rigorous evaluation of reliable DL-based frameworks for MI detection.

1.2 Automatic ECG analysis

Automatic ECG analysis refers to the use of computational methods to extract clinically meaningful information from ECG signals and to support diagnostic decision-making. Automated ECG interpretation typically involves two main stages [19]. The first stage is related to signal processing and includes operations such as filtering, beat detection, waveform delineation, and measurement extraction. The fidelity of these measurements is critical since diagnostic interpretations depend on the accuracy of these extracted features. Different signal processing pipelines across devices may alter the measure of the extracted features, leading to different diagnostic statements [19, 20]. The second stage applies diagnostic algorithms to the extracted features.

Traditionally, cardiologists interpret the standard 12-lead ECG by visual inspection, identifying characteristic patterns associated with cardiac abnormalities. Considering that an estimated 1.5 to 3 million ECGs performed daily worldwide [21], ensuring consistent, rapid, and scalable interpretation has become increasingly important within the healthcare system.

Therefore, automated ECG analysis systems were developed to support clinical decision-making by enhancing diagnostic consistency, reducing interpretation time (by approximately 24%–28%), and potentially decreasing healthcare expenditures [19, 22]. Over the past several

decades, these systems have undergone substantial methodological evolution, progressing from early rule-based computer algorithms to feature-engineered ML approaches and, more recently, to DL models capable of end-to-end representation learning.

1.2.1 Early Computer-Based ECG Interpretation Systems

The development of automated ECG analysis dates back to the second half of the 1950s with the analog-to-digital conversion of the ECG signals, which enabled computer-based signal processing [23]. Early pioneering work was conducted by Pipberger and colleagues, who developed ECG analysis systems based on three orthogonal leads and introduced probabilistic models for diagnostic classification [24]. On the other hand, Caceres and colleagues developed the first 12-lead ECG signal analysis program based on conventional clinical criteria [25].

A significant contribution was provided by the University of Glasgow, which, starting in the late 1960s, systematically investigated automated ECG interpretation techniques [26]. The resulting University of Glasgow ECG interpretation program (Uni-G) evolved from orthogonal lead analysis to full 12-lead interpretation and has been widely adopted in clinical practice. The system integrates patient metadata, such as age and sex, and the interpretation relies on a rule-based diagnostic framework. Moreover, smoothing techniques were introduced to reduce abrupt decision boundaries between normal and abnormal cases, thereby improving interpretative consistency. Among the various clinical criteria implemented, the program incorporates diagnostic criteria for MI in accordance with clinical guidelines.

The success of these academic systems enabled their translation into commercial applications. In the late 1970s, Marquette Electronics introduced the 12SL ECG Analysis Program, the first commercially available program to analyse all 12 leads simultaneously and integrated into ECG carts [27]. Like earlier systems, this software is based on a rule-based framework and has undergone continuous refinement over time. For instance, in 2000, sex-specific threshold values were introduced for MI classification.

An important milestone in the standardisation of ECG interpretation was the introduction of the Minnesota Code Manual of Electrocardiographic Findings in the 1960s by Blackburn, Keys, and colleagues at the University of Minnesota [28]. The Minnesota Code provides a classification of ECG waveforms based on predefined criteria and has been extensively used in epidemiological studies and clinical trials. Rather than producing an ECG interpretation,

it categorises ECG morphologies, including abnormal Q waves, ST-T changes, conduction defects, and arrhythmias. However, the system relies on rigid thresholds and does not account for patient-specific factors such as age or sex, which limits its diagnostic flexibility. Although revisions and software implementations have been proposed [29], threshold-based classification remains inherently constrained.

Rule-Based Approaches

Automated ECG interpretation systems predominantly relied on deterministic, rule-based logic derived from expert consensus. Such algorithms use decision trees or Boolean logic [19]. These systems offered transparency and clinical interpretability, which contributed to their widespread acceptance in clinical practice. However, their reliance on rigid thresholds introduces a fundamental limitation, as borderline cases or subtle morphological variations may lead to misclassification. To improve stability and thus reduce the sensitivity to small measurement variations, statistical diagnostic algorithms were introduced [19]. These approaches, which improved reproducibility, incorporate probability estimates into diagnostic statements, often using Bayesian approaches [19].

However, for individual diagnoses, these systems generally underperfom compared to expert cardiologists [19]. In the context of MI, misinterpretation of ECG findings may lead to false-negative diagnoses, resulting in delays in reperfusion therapy [22]. Automated systems may exhibit false-positive rates of up to 42% and false-negative rates ranging from 22% to 42% [22]. Artifacts and non-ischaemic ST-segment alterations, such as early repolarisation patterns, have been identified as the primary causes of misclassification by automated ECG interpretation systems [22].

Notably, differences between deterministic and probabilistic systems emerge when performance is evaluated using different gold standards. Algorithms based on classical deterministic criteria have been observed more closely aligned with cardiologist interpretation, while probabilistic systems presented greater concordance with diagnosis derived from comprehensive clinical evaluation [30]. This aspect reveals that automated ECG interpretation systems have different performance according to the selected gold standard (cardiologist interpretation or clinical diagnosis).

An additional challenge is the lack of standardisation across automated ECG interpretation programs [22]. Different systems often implement heterogeneous signal processing methods,

criteria and thresholds, leading to inconsistent diagnosis for the same ECG recording [20]. For these reasons, exclusive reliance on automated ECG interpretations must be avoided, and computer-generated reports must always be reviewed by clinicians.

1.2.2 Machine Learning Methods

With increasing computing power and the availability of digitised ECG data, machine learning (ML) approaches emerged as an alternative to purely rule-based ECG interpretation systems. Typically, ML methods treat ECG analysis as a supervised classification problem, in which model parameters are leaned from annotated data to derive predictive decision functions. These models rely on the extraction of handcrafted features derived from well-established electrophysiological knowledge. In the context of electrocardiographic analysis, commonly employed features can be broadly categorised into i) temporal features (*e.g.*, QRS duration, QT interval), ii) amplitude based features (*e.g.*, ST-segment amplitude), iii) morphological features which characterise the shape of specific ECG waveforms, iv) frequency-domain and time-frequency features, obtained through spectral analysis or wavelet decomposition [31, 32].

Following feature extraction, dimensionality reduction techniques are applied to remove redundancy and ensure that the most informative features are selected for the classification task. Since some features can be correlated, selecting a subset of informative features can enhance generalisability and reduce computational cost [31]. The resulting feature vectors are used to train classifiers, including logistic regression, k-nearest neighbours, support vector machines, decision trees, and ensemble methods such as random forests [33].

ML based methods have been extensively investigated for MI detection, and some studies have shown to improve MI diagnosis compared to the commercial ECG interpretation system [17, 34]. ML methods for MI detection differ in the nature of the features. For instance, in [35], time-domain, amplitude and morphological features are derived from 12-lead ECG signals. Random forest feature selection and classification were applied for MI detection and localisation. While the method achieved high performance for MI detection task (accuracy 0.97) on the PTB dataset [36], performance for multi-class localisation decreased substantially (accuracy 0.81), highlighting the increased complexity of MI localisation. Other approaches incorporate time-frequency domain features. Han *et al.* [37] extracted wavelet-based energy entropy features and combined with morphological features extracted from 12-lead ECG

signals. Principal component analysis was applied for dimensionality reduction. MI detection was performed using support vector machines.

However, ML based methods exhibit inherent limitations. First, they are based on handcrafted features, which encode only predefined ECG characteristics. This dependence limits the ability of ML models to capture the temporal sequence of ECG events [33]. Second, morphological and temporal features are typically derived through fiducial point detection, a method that identifies the onset, peak, and offset of ECG waves. The delineation of such points is of crucial importance to ensure the reliability of the features of interest. A noisy ECG signal can lead to inaccurate identification of fiducial points, leading to error [31]. Third, in MI analysis and in analogy to rule-based algorithms, reliance on ST-segment deviation as a primary feature may be insufficient, since similar ST-segment changes can be observed in other cardiac conditions [38].

1.2.3 Deep Learning Methods

In contrast to traditional rule-based and feature-engineered approaches, DL models learn representations directly from raw or minimally preprocessed ECG signals. By leveraging large-scale datasets, these models automatically extract hierarchical spatio-temporal features without explicit manual specification. In addition, this data-driven framework reduces the reliance on handcrafted feature engineering and predefined decision rules. Such capability is particularly relevant for MI diagnosis. The ECG manifestations of MI are spatially distributed across multiple leads and evolve over time.

The application of DL to MI analysis has progressed in response to increasing clinical demands. Early investigations predominantly addressed binary MI detection, whereas subsequent studies expanded to anatomically detailed localisation and stage classification. These tasks differ in both methodological complexity and clinical significance, reflecting a shift from coarse diagnostic discrimination toward a finer characterisation of MI.

MI Detection

Early DL studies predominantly formulated MI diagnosis as a binary classification task, differentiating ECG recordings of MI patients from those of healthy controls. Several investigations have demonstrated high classification performance within this binary framework, highlighting the potential of DL methods for automated MI detection [39, 40, 41].

Most approaches employ standard 12-lead ECG recordings. These signals are analysed in their entirety or segmented into individual beats to increase the number of training samples. Convolutional neural networks (CNNs) were among the first architectures adopted for this purpose. CNNs extract local morphological features through convolutional filters. Multiple convolutional and pooling layers are stacked to learn progressively more abstract signal representations.

For example, Liu *et al.* [40] trained a CNN on 12-lead ECG beats from the PTB dataset and reported a sensitivity of 0.94 and a specificity of 0.86 under an inter-patient evaluation scheme, with both metrics exceeding 0.99 in an intra-patient setting. Similarly, Rai *et al.* [41] combined CNNs with long short-term memory networks to model temporal dependencies through gated memory mechanisms, and achieved high classification accuracy on the same dataset.

Beyond purely discriminative architectures, generative models have also been explored. For instance, variational autoencoders have been employed to reconstruct missing leads and support MI detection using reduced-lead configurations [42]. These findings suggest that discriminative information relevant to binary MI detection can be captured even from a limited subset of leads. Importantly, several studies have demonstrated that reported performance is strongly influenced by the adopted evaluation protocol, particularly with respect to intra- *vs.* inter-patient data partitioning.

Overall, these studies establish the feasibility of DL models for distinguishing MI from normal ECG recordings. However, binary detection represents only the initial step toward a comprehensive characterisation of MI.

MI Localisation

Compared with binary MI detection, infarct localisation constitutes a more complex task. Electrocardiographic alterations associated with MI are not restricted to a single, clearly delineated myocardial territory. Injury may extend beyond a single myocardial region, resulting in overlapping ECG manifestations that complicate precise regional classification.

An additional challenge pertains to dataset composition. Infarct territories are not uniformly represented in publicly available datasets; certain classes, such as posterior MI, occur less frequently and are consequently under-represented. This class imbalance may introduce bias during model training and adversely affect the reliability of predictions for minority categories.

Early DL approaches primarily addressed localisation through coarse-grained classification, typically distinguishing between two broad anatomical regions, most commonly anterior and inferior MI, without accounting for more specific subregional involvement (*e.g.*, anteroseptal MI) [43, 44]. For instance, Strodthoff *et al.* [43] and Wang *et al.* [44] formulated the problem as a binary MI localisation task using the PTB dataset.

Subsequent studies extended this framework to multi-class localisation in order to provide a more detailed characterisation of anatomical involvement [39, 45, 46, 47]. Using the PTB dataset, Jian *et al.* [46] classified five infarct regions, reporting an accuracy that exceeded 0.60. Comparable performance has also been described by Han *et al.* [45] and Jahmunah *et al.* [47], particularly under intra-patient evaluation protocols.

Recently, multitask learning strategies have been proposed to jointly perform MI detection and localisation within a unified framework [48]. By sharing representations across related tasks, such approaches aim to improve feature learning efficiency and enhance performance in anatomically specific classification. Reported findings indicate that certain anatomically adjacent subclasses remain prone to misclassification, underscoring the intrinsic difficulty of fine-grained localisation.

MI Stage Classification

In addition to MI localisation, the stage of MI represents an important complementary diagnostic objective. Electrocardiographic alterations evolve throughout the course of MI and are typically classified into acute and prior (chronic) stages.

Compared with detection and localisation, stage classification remains largely underexplored. The gradual evolution of ECG patterns, together with partial overlap between stages, introduces ambiguity in annotation and increases inter-observer variability. Nevertheless, several studies have addressed this task using DL models. For example, Prabhakararao *et al.* [49] employed an attention-based recurrent neural network trained on the PTB and STAFF-III datasets [50, 51] to distinguish early, acute, and prior MI, achieving a sensitivity exceeding 0.90. Their framework first encoded temporal variations within each lead using recurrent neural networks, followed by intra-lead attention mechanisms to identify discriminative patterns. An inter-lead attention module then aggregated lead-specific representations according to their clinical relevance, producing a high-level feature representation for final

classification. In addition to stage discrimination, the model also differentiated among MI, non-MI, and healthy recordings.

Recently, transfer learning strategies have been proposed for MI stage classification [52]. In this framework, ECG signals are converted into spectrogram representations instead of being analysed as raw time series. Pretrained CNNs are employed to extract discriminative features from these spectrograms. The resulting features derived from each lead are subsequently aggregated, either by concatenation or averaging, to classify recordings into acute, recent, and prior stages. The model was trained on a private dataset and externally evaluated on the PTB dataset, thereby demonstrating the adaptability DL models across heterogeneous signal representations.

Collectively, these studies demonstrate that DL models are capable of capturing progressively finer-grained ECG characteristics, extending from binary MI detection to anatomical localisation and stage classification.

However, these tasks are inherently interrelated in clinical practice. MI detection, localisation, and stage assessment are not independent decisions but complementary components of a unified diagnostic reasoning process. Treating them as isolated classification problems may limit the ability of DL models to exploit shared physiological information. In this context, multitask learning has emerged as a promising strategy, enabling the joint optimisation of related objectives and the learning of shared representations while preserving task-specific discrimination. Such frameworks may enhance feature efficiency, improve consistency across predictions, and better reflect the integrated nature of clinical diagnosis.

Moreover, the predominantly data-driven nature of conventional DL approaches raises important concerns regarding robustness and generalisation. Decades of cardiology research have established well-defined principles regarding ECG morphology, lead–heart region relationships, and diagnostic criteria for MI. Incorporating this domain knowledge (DK) into DL models, through anatomically informed lead grouping, structured constraints, or hybrid learning paradigms, introduces inductive (learning) biases that can guide DL models toward physiologically meaningful representations. This integration is particularly relevant in MI analysis, where subtle ECG morphological variations and heterogeneity across datasets may lead purely data-driven models to exploit spurious correlations instead of clinical relevant features.

In this context, multitask learning and DK–guided modelling can represent a conceptual shift from purely performance-driven classification to clinically grounded, robust, and generalisable DL models for MI analysis. These considerations provide the foundation for examining the methodological challenges and research gaps that remain in the development of reliable DL-based approaches for MI diagnosis.

1.3 Research Challenges

Despite significant advances in both clinical guidelines and AI for MI diagnosis, several critical challenges remain that hinder the development of accurate, robust, and trustworthy DL models. These challenges form the foundation of this thesis and are outlined as follows:

- **Limited robustness and generalisability of automated MI diagnosis system:** Traditional rule-based systems rely on handcrafted features and threshold-based criteria capture only predefined ECG characteristics and often fail to generalise across heterogeneous populations. Although DL models represent a more flexible, data-driven alternative and have achieved promising performance on individual datasets, they exhibit a similar limitation when evaluated on new patient cohorts or clinical environments. This vulnerability suggests that such models may rely on dataset-dependent patterns rather than truly physiologically meaningful representations, ultimately limiting their clinical deployment.
- **Susceptibility to spurious correlations and data biases:** Another major barrier to developing robust AI systems for MI diagnosis is the tendency of models to exploit spurious correlations that are predictive in training data but do not reflect clinical relevance. These correlations may arise from confounding factors such as demographic attributes, acquisition protocols, or skewed data distributions due to the under-representation of MI patients.
- **Insufficient integration of DK into DL models:** Existing DL models typically overlook electrocardiographic DK that has been codified into clinical decision rules. Without the explicit incorporation of such DK, these models may learn feature representations that are not physiologically grounded, thereby limiting their generalisability.

1.4 Research Objectives

The aim of this thesis is to advance MI interpretation using standard 12-lead ECGs. Accordingly, the research is structured with specific objectives that focus on methodological development, empirical evaluation, generalisability, and robustness. The objectives are detailed below:

- **Mitigating age bias via adversarial multitask learning:** Implementing adversarial multitask learning strategies to reduce the influence of confounding factors, with particular emphasis on age-related bias in MI detection.
- **Developing a multitask DL framework for MI diagnosis:** A multitask DL framework is introduced to jointly model MI detection, localisation, and stage classification within a unified architecture. By explicitly leveraging interdependencies among related MI tasks, the framework enhances clinically coherent predictions and generalisation across datasets.
- **Incorporating DK into DL framework for MI diagnosis:** Integrating DK into the DL model through custom regularisation terms that control the latent representation. Two complementary strategies are pursued: i) constraining intermediate representations to capture clinically relevant ECG features, including ST-segment elevation, Q- and R-wave amplitudes, and Q-wave durations; and ii) structuring the latent space using differentiable formulations of clinical decision rules derived from the Fourth Universal Definition of MI (UDMI).
- **Comparing performance evaluation with rule-based algorithm across datasets:** A comprehensive comparison of the DK-guided DL framework with traditional data-driven DL model and rule-based algorithm to assess performance and elucidate their respective strengths and limitations across four datasets.

1.5 Summary of the Aims

Despite significant progress in automated MI diagnosis through DL, important challenges persist with respect to model generalisability, susceptibility to bias, and robustness in heterogeneous clinical environments. These limitations hinder the clinical translation of AI systems. Addressing these challenges requires not only architectural advancements in DL but also the

explicit integration of DK, bias mitigation strategies, and rigorous cross-dataset evaluation to ensure reliable generalisation. Accordingly, this thesis focuses on the development and validation of robust, DK-based multitask learning frameworks designed to promote physiologically grounded representations. Through these contributions, this research aims to advance robust AI systems for MI diagnosis using standard 12-lead ECGs.

1.6 Thesis Structure

The remainder of this thesis is structured as follows:

- Chapter II: Focuses on the interpretation of the ECG and introduces MI, with emphasis on the characteristic ECG patterns that are fundamental to its diagnosis.
- Chapter III: Presents an adversarial strategy aimed at mitigating reliance of deep learning models on spurious correlations arising from age for MI detection.
- Chapter IV: Introduces a training strategy designed to enhance robustness of deep learning model for MI diagnosis by integrating domain knowledge into the learning process.
- Chapter V: Summarises the main findings of this thesis and outlines directions for future research.

Chapter II

Fundamentals of ECG and Myocardial Infarction

2.1 Cardiac Conduction System

The heart is a muscular organ composed of four chambers organised into two atria and two ventricles, whose coordinated contraction ensures blood flow through the pulmonary and systemic circulations.

Cardiac contraction is initiated and regulated by the propagation of electrical impulses throughout the myocardium. Each cardiac cycle begins with spontaneous depolarisation of pacemaker cells in the sinoatrial (SA) node, located in the upper region of the right atrium. From the SA node, the electrical impulse propagates through the atrial myocardium, leading to atrial depolarisation and contraction, which facilitates ventricular filling. Since atria and ventricles are electrically insulated, electrical impulse conduction occurs exclusively through AV node. Within the AV node, conduction velocity decreases to allow atrial contraction to complete before ventricular contraction begins. From the AV node, the electrical impulse is transmitted to the bundle of His, which divides into the right and the left bundle branches along the interventricular septum. These branches further divide into Purkinje fibers that spread the impulse through the ventricular myocardium, provoking depolarisation of ventricular myocytes and the subsequent ventricular contraction (systole). Following contraction, ventricular myocytes undergo repolarisation, a phase of electrical recovery that leads to myocardial relaxation (diastole).

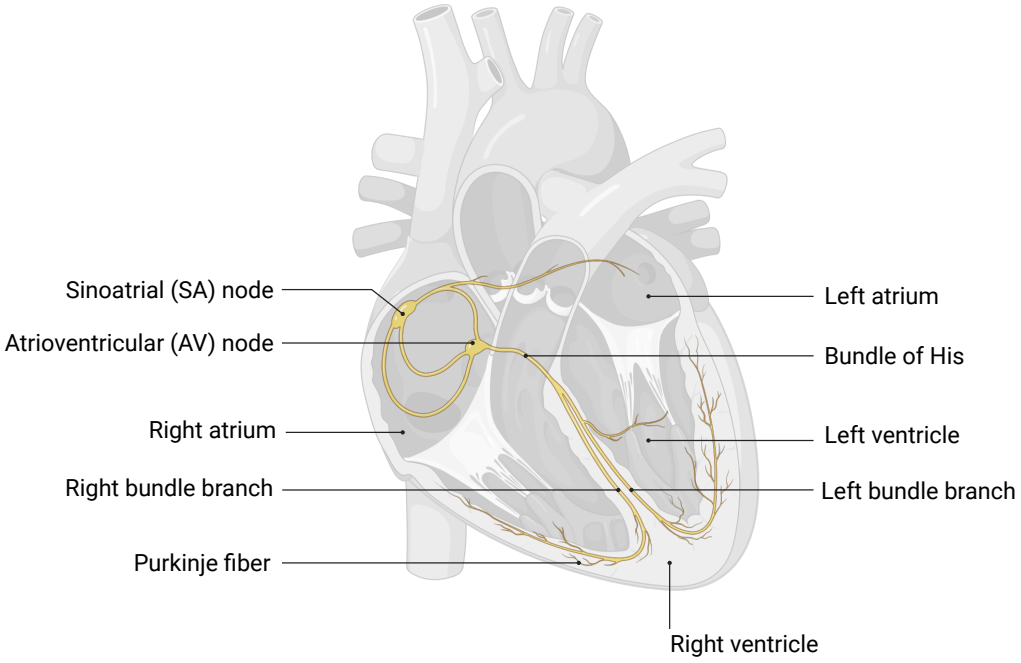


FIGURE 2.1: Schematic representation of the cardiac conduction system. Created in BioRender. Ibrahimi, S. (2026) https://BioRender.com/e118ay0.

The spatial and temporal propagation of depolarisation and repolarisation throughout the myocardium generates electrical potentials that can be measured at the body surface and which constitute the ECG.

2.2 Electrocardiogram

The ECG represents a non-invasive method for evaluating cardiac electrical activity. Owing to its rapid acquisition, cost-effectiveness, and widespread accessibility, the ECG remains a fundamental diagnostic tool for the identification and assessment of cardiac pathologies.

In clinical practice, cardiac electrical activity is typically recorded using the standard 12-lead ECG. This configuration requires the placement of ten surface electrodes: six positioned on the chest (precordial) and four placed on the limbs. From these electrodes, twelve leads are derived, comprising six precordial leads (V1–V6) and six limb leads (I, II, III, aVR, aVL, and aVF). Each lead represents a voltage difference generated by cardiac action potentials of the excitable cardiac cells [53].

The recorded bioelectric signals are transmitted to the ECG device, where they undergo amplification and filtering prior to analysis. Modern ECG systems additionally provide automated measurements, including heart rate, QRS duration as well as a preliminary diagnostic interpretation.

The ECG is traditionally recorded and printed on a standardised graph paper. The grid is composed of small squares of 1 mm and larger squares of 5 mm. Standard calibration settings are a defined by paper speed of 25 mm/s, and a voltage gain of 10 mm/mV. Under these conditions, 1 mm corresponds to 0.1 mV in amplitude and 40 ms in duration. These standardised calibration settings are of essential importance for clinical interpretation. Many diagnostic criteria, including those for myocardial infarction, are defined in terms of millimetres deviation on the ECG tracing (see Section 2.3.1).

2.2.1 ECG Waveform Components

The ECG waveform is composed of characteristic waves, segments and intervals that reflect specific phases of the cardiac cycle. The main components are the P wave, the QRS complex, and the T wave (Figure 2.2). All amplitudes are measured relative to the baseline, commonly referred to as the isoelectric line.

The P wave represents atrial depolarisation. In most leads, it appears as a positive deflection with a normal amplitude below 300 μV and a duration shorter than 120 ms. The P wave is dominated by low-frequency components, generally below 10–15 Hz [53].

The QRS complex corresponds to ventricular depolarisation and typically consists of three deflections: an initial negative Q wave, a prominent positive R wave, and a subsequent negative S wave. The normal QRS duration ranges from 70 to 110 ms, and its amplitude can reach 2–3 mV [53]. The QRS complex contains dominant frequency components in the 10–50 Hz range and is commonly the primary target in automated ECG analysis algorithms, particularly for heartbeat detection due to its high amplitude [53]. Pathological Q waves, characterised by increased depth or duration, may indicate prior myocardial infarction.

The ST-segment extends from the end of the QRS complex (J-point) to the beginning of the T wave. It represents the electrically neutral phase between ventricular depolarisation and repolarisation. Under normal conditions, the ST-segment is isoelectric. Deviations in the form of elevation or depression, particularly when observed in contiguous leads, are clinically significant and may indicate acute myocardial ischaemia or infarction.

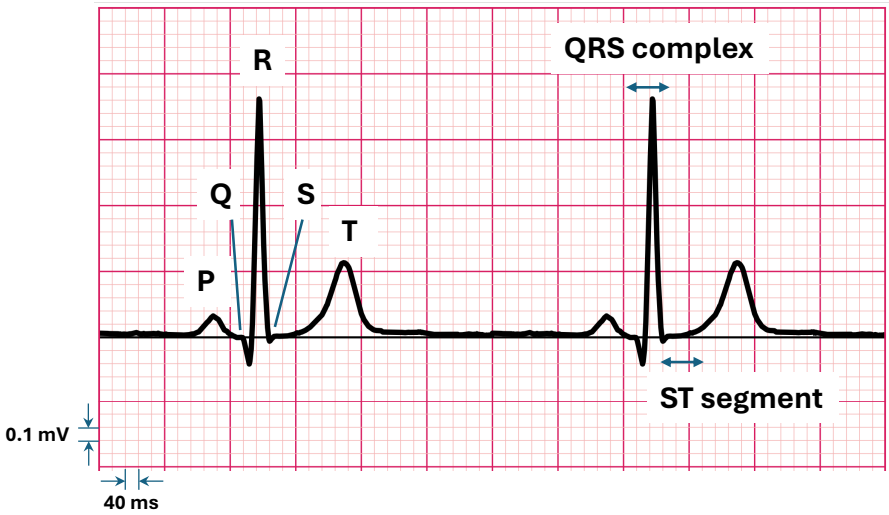


FIGURE 2.2: Representation of an ECG tracing highlighting the main waveform components on a standard ECG paper (calibration: 0.1 mV in amplitude and 40 ms in duration per mm).

The T wave reflects ventricular repolarisation and typically extends approximately 300 ms following the QRS complex [53]. It is generally asymmetric, with a gradual upstroke and a steeper downstroke, and is positive in most leads. Abnormal T wave morphology, including inversion or hyperacute peaking, may be associated with ischaemic processes.

Finally, a small deflection known as the U wave may occasionally follow the T wave. Although its exact origin remains uncertain, it is often attributed to after-repolarisation phenomena [53].

2.2.2 Noises and Interferences

During acquisition, ECG recording are susceptible to various sources of noise and interference which can have a detrimental effect on ECG classification. These unwanted components, both in low and in the high frequency, may deform ECG waveforms limiting the utility of the ECG trace and undermining the clinical interpretation and automated computational analysis. The frequency bandwidth of the ECG signal spans from 0.05 to 100 Hz. However, the majority of the information of the signal is contained within the lower band, *i.e.* below 35 Hz [31], a range that is also susceptible to interference from certain artifacts.

Baseline wander, represents an artifact caused by body movement, respiration or poor

electrode contact. Its spectral content is usually confined to an interval below 1 Hz, but it may contain higher frequencies during strenuous exercise [53]. It is around 15% of the peak-to-peak ECG amplitude [54]. Because this artifact manifests as a slow shift of the isoelectric line, it can distort the ST-segment, thus leading to an incorrect diagnosis of several ST-segment abnormalities including MI [31, 55]. Given that the diagnostic criteria for MI rely on subtle ST-segment deviations, even slight fluctuations in baseline may alter the classification between ST-elevation MI (STEMI) and non–ST-elevation MI (NSTEMI), with significant implications for treatment decisions.

Electrode motion artifacts are caused by a change in the skin-electrode impedance due to skin stretching. Unlike baseline wandering, motion artifacts are harder to tackle since their spectral content overlaps extensively with the PQRST complex, particularly in the 1–10 Hz band [53]. In ECG signals, they appear as large deflections that can sometimes be misinterpreted as QRS complexes.

Powerline interference appears as a 50 or 60 Hz sinusoidal signal, reaching up to 50% of the peak-to-peak ECG amplitude [54]. As a narrowband artifact, it may distort low-amplitude components of the ECG and may introduce spurious waveforms [53], potentially obscuring the P and T waves [56].

Electromyographic noise is caused by electrical activity of skeletal muscles during periods of contraction. This source of noise can be either intermittent or have more stationary noise properties. It reaches an average amplitude of 10% of the peak-to-peak ECG amplitude [54]. The spectral content of electromyographic noise (high frequency, within 0 Hz to 500 Hz but mainly concentrates in the range from 50–150 Hz [31]) overlaps substantially with that of the QRS complex and also spans higher frequency ranges [53]. Because of this overlap, it is quite complex to remove such noise without distorting the ECG morphology.

Respiratory activity has an influence on the ECG in two main ways. On one hand, respiration modulates the heart rate. On the other hand, chest movements and change in the thoracic impedance during respiratory cycle lead to a rotation of the cardiac electrical axis which in turn modifies ECG beat morphology [57].

2.3 Myocardial Infarction

MI represents one of the leading causes of morbidity and mortality worldwide. Its definition and diagnostic criteria have evolved over time, driven by advances in electrocardiography and cardiac biomarkers testing. Contemporary diagnosis integrates biomarker evidence, clinical presentations, and electrocardiographic findings.

Pathophysiologically, MI is defined as the death of cardiac myocytes due to prolonged ischaemia [58]. More in detail, MI occurs when blood flow to a portion of the myocardium is obstructed, most frequently due to atherosclerotic plaque ruptures followed by thrombus formation in the coronary arteries. The resulting ischaemia and necrosis of cardiac tissue can lead to irreversible structural and functional damage if not treated in a timely manner.

Universal Definition and Conceptual Evolution

Before the advent of electrocardiography, MI could only be diagnosed by postmortem examination [59]. The introduction of the ECG in the early 20th century enabled clinical diagnosis [60]. Efforts to standardise diagnosis led to the introduction of an ECG-based definition by the World Health Organization between the 1950s and 1970s [58]. With the introduction of sensitive cardiac biomarkers testing, in 2000, the European Society of Cardiology and the American College of Cardiology redefined MI by integrating biochemical evidence with clinical and electrocardiographic findings [61], laying the foundation for the UDMI.

The first UDMI, published in 2007, established cardiac troponin as the preferred biomarker for myocardial injury and introduced an etiological classification of MI into five distinct types [62]. Subsequent revisions, including the Third UDMI (2012) and the currently adopted Fourth UDMI (2018), refined both electrocardiographic criteria and biochemical thresholds, incorporating age- and sex-specific considerations [58]. A fifth revision is expected to be released in 2026 [63]. These successive revisions underscore the evolving understanding and inherent complexity of MI diagnosis.

2.3.1 Electrocardiographic Manifestations of MI

MI induces alterations in cardiac electrical activity that are reflected in characteristic changes in the ECG morphology. The morphology of ECG waveforms follows a characteristic pattern

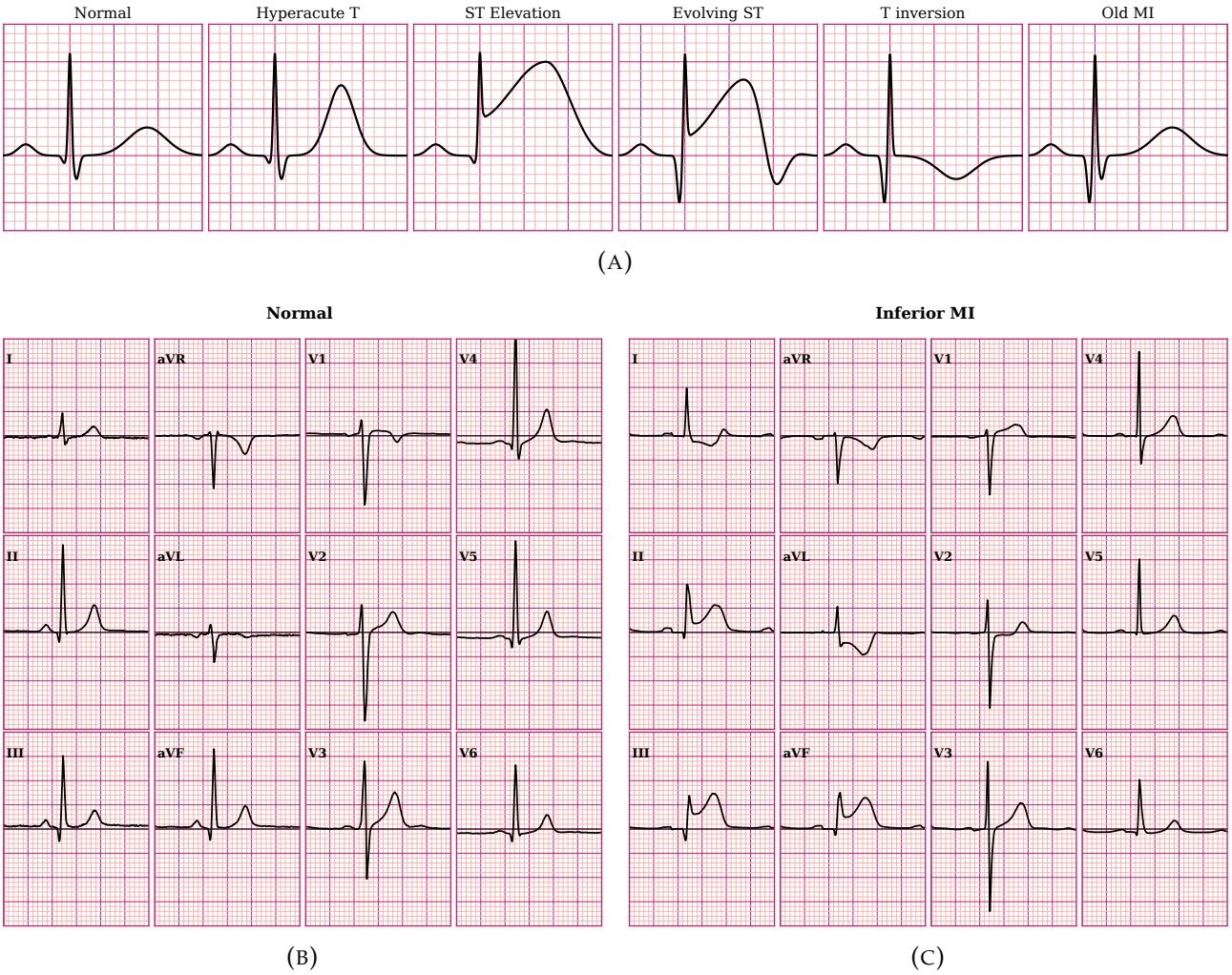


FIGURE 2.3: A) Temporal evolution of ECG alterations during MI (synthetic example, not to scale). B) Representative 12-lead beat extracted from a normal sinus rhythm ECG. C) Representative 12-lead beat extracted from an ECG of a patient with inferior acute MI. In inferior acute MI, ST-segment elevation is expected in the inferior leads (II, III, aVF), with reciprocal ST-segment depression in the lateral leads (I, aVL).

that varies according to the extent and localisation of the infarction. These morphological changes, illustrated in Figure 2.3a, evolve over time in a well-defined sequence:

- **Hyperacute Stage** (minutes to hours): Development of tall, symmetric, and peaked T waves reflecting early ischaemia.
- **Acute Stage** (hours): Development of ST-segment elevation in transmural ischaemia or ST-segment depression in subendocardial ischaemia.

- **Subacute Stage** (hours to days): Appearance of T-wave inversion and pathological Q waves, indicative of evolving necrosis.
- **Chronic Stage** (weeks to months): Resolution of ST-segment abnormalities with persistence of pathological Q waves in cases of transmural MI.

From the Q-Wave Paradigm to ST-Segment for MI Diagnosis

In earlier diagnostic classifications system, MI was retrospectively categorised as Q-wave MI or non–Q-wave MI [64]. Q-wave MI was associated with transmural necrosis, whereas non–Q-wave MI corresponded to subendocardial infarction. However, pathological Q waves typically develop only after irreversible injury and may not be present in all infarctions [65]. Consequently, this classification lacked sufficient sensitivity for early detection.

In 2000, the diagnostic paradigm shifted toward the assessment of ST-segment abnormalities to enable earlier identification of patients requiring urgent reperfusion therapy [61]. For this reason, acute coronary syndromes began to be stratified into STEMI and NSTEMI. The identification of these ECG abnormalities is of critical importance, as therapeutic decisions are often made before cardiac biomarkers results become available [66].

STEMI and NSTEMI Classification and Formal ECG Criteria in the Fourth UDMI

STEMI is typically associated with acute, complete coronary artery occlusion and requires immediate reperfusion therapy. The Fourth UDMI, STEMI is defined by new ST-segment elevation measured at the J-point in two or more anatomically contiguous leads, satisfying the following thresholds:

- ST-segment elevation $\geq$ 1 mm (0.1 mV) in all leads except V_2–V_3.
- In leads V_2–V_3, sex- and age-specific thresholds apply:
  - $\geq$ 1.5 mm (0.15 mV) in women,
  - $\geq$ 2.0 mm (0.2 mV) in men aged $\geq$ 40 years,
  - $\geq$ 2.5 mm (0.25 mV) in men aged $<$ 40 years.

These criteria should be applied in the absence of QRS confounders (*e.g.*, left bundle branch block (LBBB) or left ventricular hypertrophy (LVH)) that may obscure or mimic ischaemic ST-segment deviations.

In an ECG recording, in addition to the elevation of the ST-segment observed in specific groups of leads reflecting the ischaemic territory, depression of the ST-segment may be present in leads opposite to the infarcted region. These are known as reciprocal changes and contribute to differentiating ST-elevation MI from other conditions associated with ST-segment elevation, such as pericarditis and early repolarisation [58]. Figure 2.3 provides an illustrative comparison between a beat extracted from a 12-lead ECG recording of a healthy subject and that from a subject with inferior wall STEMI.

NSTEMI is generally associated with partial coronary occlusion and is managed with a less urgent, though still time-sensitive, strategy. The Fourth UDMI defines relevant ECG findings as:

- New horizontal or downsloping ST-segment depression $\geq$ 0.5 mm (0.05 mV) in two or more contiguous leads, and/or
- T-wave inversion $>$ 1 mm (0.1 mV) in two or more contiguous leads in the presence of a prominent R wave or R/S ratio $>$1.

In addition to the criteria for acute MI, the Fourth UDMI also provides specific criteria to identify a previous or unrecognised MI from an ECG. Among these, the presence of pathological Q waves is regarded as the primary indicator. The criteria are:

- Any Q wave in leads V2–V3 with duration $>$0.02 s or QS complex in leads V2–V3.
- Q wave $\geq$0.03 s and $\geq$1 mm deep or QS complex in leads I, II, aVL, aVF or V4-V6 in any two leads of a contiguous lead grouping (I, aVL; V1-V6; II, III, aVF).

Limitations of STEMI/NSTEMI Paradigm: Despite its widespread adoption, the STEMI/NSTEMI paradigm has important limitations. A subset of patients classified as NSTEMI may have total coronary artery occlusion, resulting in delayed reperfusion and worse clinical outcomes compared with patients presenting with STEMI [67]. Conversely, ST-segment elevation may occur in the absence of acute coronary occlusion.

To address these limitations, the Occlusion MI (OMI) *vs.* Non-Occlusion MI (NOMI) paradigm has been proposed [65]. In contrast to previous classifications, the OMI/NOMI is defined based on underlying pathophysiology rather than ECG waveform changes. Beyond ST-segment changes, the OMI/NOMI framework also takes into account additional ECG features that may indicate coronary occlusion [68].

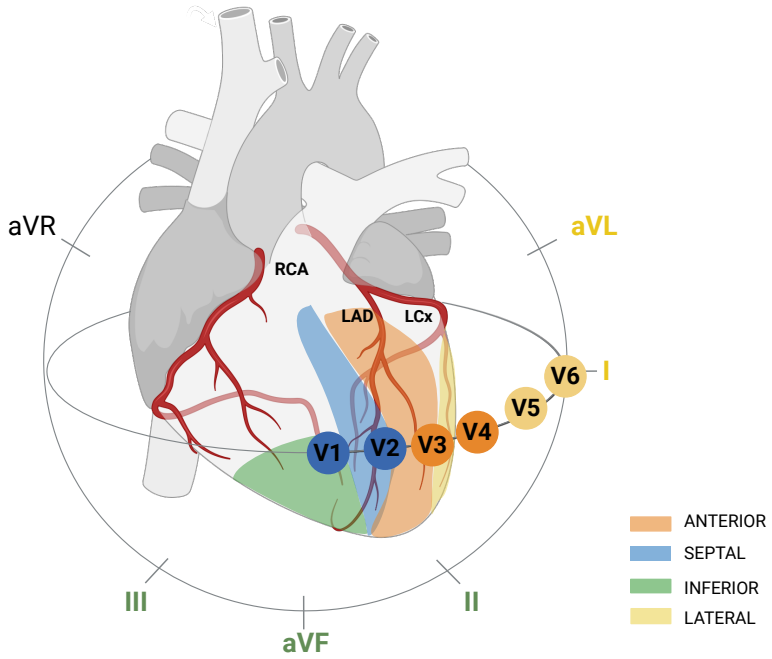


FIGURE 2.4: Schematic representation of coronary arteries (LAD, RCA, LCx) with corresponding MI territories and ECG leads. Created in BioRender. Ibrahimi, S. (2026) https://BioRender.com/jrsbeso.

In terms of diagnostic performance, the STEMI/NSTEMI criteria exhibit a pooled sensitivity for detecting total coronary occlusion of 43.6%, specificity 96.5%. In comparison, the OMI/NOMI paradigm achieves a pooled sensitivity of 78.1%, and a specificity of 94.4% [68]. Earlier, the Q-wave MI/non–Q-wave MI classification exhibited lower sensitivity (38.0%) despite reasonably high specificity (86.9%) [69].

2.3.2 MI Localisation

MI predominantly involves the left ventricle, reflecting fundamental structural and physiological differences between left and right ventricle [70]. The left ventricle is characterised by a greater myocardial mass and oxygen demand. In contrast, isolated right ventricular MI is less common and occurs in association with inferior left ventricular MI [71]. Consequently, electrocardiographic localisation of MI primarily focuses on identifying the affected regions of the left ventricle.

Within the left ventricle, MI does not occur uniformly across anatomical regions. Infarction

TABLE 2.1: ECG lead changes and corresponding coronary arteries for different anatomical regions [1].

MI Localisation	ECG Changes	Affected Coronary Artery
Septal	V_1–V_2	Septal LAD
Anterior	V_3–V_4	LAD
Lateral	I, aVL, V_5, V_6	LCx, diagonals
Inferior	II, III, aVF	LCx (15%), RCA (85%)
Posterior	V_7, V_8, V_9	RCA

most commonly affects the anterior and anteroseptal walls, which are supplied by the left anterior descending artery (LAD), as well as the inferior wall, which is typically associated with the occlusion of right coronary artery (RCA) and, in a subset of cases, the left circumflex artery (LCx). The lateral wall is supplied by LCx artery or diagonal branches of the LAD artery, and MI occurs less frequently. Posterior MI is usually caused by the occlusion of the RCA or LCx and most often occurs in association with inferior or lateral MI.

Each of the 12 ECG leads provides information regarding specific myocardial regions. Abnormalities observed in particular leads enable MI localisation and, consequently, suggest the culprit coronary artery. Leads V1-V2 reflect the involvement of the septal wall, while leads V3-V4 correspond to the anterior wall. The lateral wall is assessed thorough leads I, aVL, and V5–V6. Inferior wall MI manifests in leads II, III, and aVF. Posterior MI is not directly visualised by the standard 12-lead ECG. On the standard ECG, posterior MI is suggested by reciprocal changes (*i.e.* ST-segment depression) in the leads V1-V3. Direct confirmation requires the use of posterior leads V7-V9. Figure 2.4 and Table 2.1 provide a summary of these lead-specific correlations.

Chapter III

Mitigating Age Bias in ECG-Based Myocardial Infarction Diagnosis via Adversarial Multitask Learning

3.1 Introduction

Despite promising performance in MI diagnosis, many current DL models have been developed and evaluated on small datasets or on data biased toward a specific clinical condition. In this context, spurious correlations between demographic factors and DL model predictions represent a critical challenge.

In healthcare applications, spurious correlations are particularly concerning due to the structure of clinical data [72, 73, 74]. Medical datasets often exhibit heterogeneous acquisition protocols, demographic imbalances, and age-dependent differences in disease prevalence. Demographic factors such as age, sex, and ethnicity influence both ECG morphology and prevalence of the disease therefore they can introduce confounding relationships into predictive models.

Among these factors, age is especially relevant in the context of MI. Cardiovascular risk increases with advancing age, and clinical datasets typically show that patients with MI are older than healthy controls. At the same time, ageing influences ECG morphology. Consequently, DL models trained on such data may exploit age-related ECG characteristics rather than learning features that specifically reflect MI.

Several studies have demonstrated that age can be inferred from ECG signals [75, 76, 77]. As a consequence, MI classifiers trained on imbalanced datasets may rely on ECG alterations associated with ageing, such as reduced P wave amplitude, increased QRS amplitude or QT prolongation [78], rather than abnormalities specific to MI. This phenomenon can lead to biased predictions and poor generalisation performance with different age distributions, particularly in young patients for whom MI is less prevalent.

From a machine learning perspective, bias can arise at multiple stages of the modelling pipeline [79]. Age-related effects are considered a form of representation bias, as they arise from the demographic distribution of the available ECG data.

The importance of explicitly addressing demographic bias in ECG-based DL models has been emphasised by Alday *et al.* [80], who analysed the impact of age, sex and race on arrhythmia detection algorithms and demonstrated that demographic biases are overlooked in model evaluation. These findings highlight the need for training strategies that mitigate the influence of demographic confounders.

Motivated by this gap, the present study investigates adversarial multitask learning (AML) as a strategy to mitigate age bias in MI detection from ECG signals. AML enables the learning of task-relevant representations while discouraging the encoding of unwanted information. The objective of this study is to decorrelate MI detection from patient age in ECG-based DL models. Specifically, the objectives are: i) to quantify the extent of age-related bias in baseline MI classifiers; and ii) to demonstrate that AML reduces this correlation without sacrificing diagnostic accuracy.

3.2 Related Works

Several DL models had been investigated for MI classification from ECG signals. In general, the primary objective of these studies was to improve classification performance. However, comparatively little attention has been paid to analysing how demographic factors influence model behaviour and performance.

With the increasing emphasis on fairness in AI, the problem of demographic bias has begun to receive growing attention. Alday *et al.* [80] pointed out that in the context of ECG analysis, very few studies have considered the bias in terms of metrics, algorithms and databases. By analysing 56 open-source algorithms from the 2021 PhysioNet Challenge on more than

130,000 ECGs, the authors demonstrated that demographic biases were overlooked. They also proposed the inclusion of a bias penalisation term in the loss function to reduce disparities among sex, ethnicity, and age groups.

AML has been proposed as a framework for mitigating known sources of bias in DL algorithms. In this paradigm, a model is trained to optimise a primary task, such as classification, while minimising its dependence on a secondary task associated with a known bias. AML has successfully been applied in speech recognition to extract age-independent and speaker-invariant representations [81, 82]. In the ECG domain, the AML paradigm has been adopted to reduce subject dependency with the aim of improving the generalisation capability of arrhythmia classification models [83].

3.3 Adversarial Multitask Learning

Generally, DL models which are trained under empirical risk minimisation, learn statistical associations between inputs and outputs without distinguishing whether the predictive information arises from causal or spurious (non-causal) correlations [84]. Although age is not explicitly provided as an input variable to the DL model, the learned representation may extract features correlated with age, leading to predictions that may implicitly depend on age. However, mitigating such spurious correlation requires learning representations that extract task-relevant information. To address this issue, AML is introduced as a strategy to mitigate age-related information as a spurious correlation in ECG representations, thereby aiming to improve robustness in MI detection.

Multitask learning (MTL), in which multiple tasks are learned jointly by sharing a common representation [85], provides the foundation for the AML. In the standard case, the shared representation is encouraged to be predictive for all tasks. However, shared representations may inadvertently encode spurious correlations present in the training distribution [86]. Learning representations that are invariant to known confounding factors represents a major challenge in building robust and generalisable DL models. AML has been adopted to tackle this challenge. Within the AML framework, a DL model is optimised to perform a primary task while being penalised whenever it encodes information related to confounding factors.

A common implementation employs a gradient reversal layer (GRL), originally introduced by Ganin and Lempitsky in the context of domain adversarial neural networks for unsupervised

domain adaptation [87]. In this case, a model trained on a labelled source dataset generalises to an unlabelled target dataset with a different distribution. GRL acts as an identity transform during forward propagation, while multiplying the gradients by a negative constant value during backpropagation [88]. As a result, the shared layers of the DL model are optimised to minimise the loss associated with the primary task, while maximising the loss of the adversarial task.

Inspired by this principle, we adopted the GRL to mitigate age-related confounding in MI detection from ECGs. As illustrated in Figure 3.1, in the proposed architecture, a shared feature extractor (backbone), processes the 12-lead average beat input and feeds two parallel task-specific branches: the MI classifier (primary task), which performs MI detection, and the age predictor (secondary task), which estimates the age of patients. A GRL was inserted between the backbone and the age predictor. As a consequence, the backbone was optimised in two competing directions. On the one hand, it minimised the binary cross-entropy loss (BCE) associated with the MI classification task. On the other hand, it maximised the mean squared error (MSE) loss of the age prediction task through GRL. During training, gradients from the MI classifier updated the backbone to improve classification accuracy, while gradients from age predictor, penalised the backbone whenever it encoded age-related information. To implement AML, we followed the approach described in [83].

The backbone network is denoted as b_{θ_B}, the MI classifier as m_{θ_M} and the age predictor as a_{θ_A}. Given an input 12-lead average beat ECG x, the probability of detecting MI is defined as

$$\hat{y} = m_{\theta_M}(b_{\theta_B}(x)) \tag{3.1}$$

where $\hat{y}$ denotes the probability of MI, and the functions m_{θ_M} and b_{θ_b} are neural networks parametrised by θ_M and θ_B, respectively. The predicted age is given by

$$\hat{a}(x) = a_{\theta_A}(b_{\theta_B}(x)) \tag{3.2}$$

where a_{θ_A} is a neural network with parameters θ_A.

The training procedure was organised in three sequential steps, each for a specific purpose: i) learning MI-discriminative representations, ii) training an age predictor, and iii) enforcing adversarial learning to reduce age-related information in the shared backbone.

In the first stage, the backbone b_{θ_B} and the MI classifier m_{θ_M} were optimised jointly to

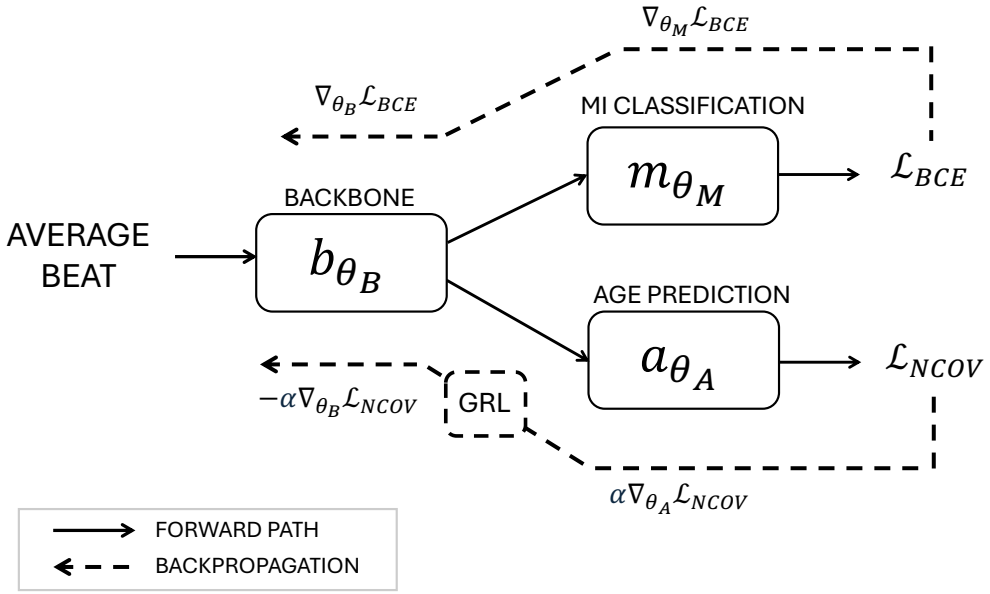


FIGURE 3.1: AML framework for age-bias mitigation. A shared feature extractor (backbone) feeds two task-specific branches: MI classification and age prediction. A GRL is incorporated in the age prediction branch to penalise the encoding of age-related information in the backbone.

minimise the BCE loss associated with the MI classification task. The parameters were updated using stochastic gradient descent according to

$$\begin{aligned} \theta_M &\leftarrow \theta_M - \eta \frac{\partial \mathcal{L}_{\text{BCE}}}{\partial \theta_M} \\ \theta_B &\leftarrow \theta_B - \eta \frac{\partial \mathcal{L}_{\text{BCE}}}{\partial \theta_B} \end{aligned} \tag{3.3}$$

where η denotes the learning rate. This stage ensured that the backbone learned representations that were discriminative for MI detection.

In the second stage, the age predictor a_{θ_A} was trained to minimise the MSE between the predicted and the true age y. The parameter vector θ_A were updated as follows

$$\theta_A \leftarrow \theta_A - \eta \frac{\partial \mathcal{L}_{\text{MSE}}}{\partial \theta_A} \tag{3.4}$$

In this phase, the backbone parameters θ_B were kept fixed to create a model able to predict age of using the same b_{θ_B} model for MI classification.

In the third stage, adversarial training was introduced by inserting a GRL between the backbone and the age predictor. The backbone parameters were updated according to

$$\theta_B \leftarrow \theta_B - \eta \left(\frac{\partial \mathcal{L}_{\text{BCE}}}{\partial \theta_B} - \alpha \frac{\partial \mathcal{L}_{\text{MSE}}}{\partial \theta_B} \right) \tag{3.5}$$

where α is a hyperparameter that controls the strength of GRL.

3.3.1 Compensation of Age-related Information

The age predictor a_{θ_A} was trained using the MSE loss as defined in Eq. (3.4). In the AML formulation instead, MSE loss was driven to worse using GRL, as shown in Eq. (3.5). However, this approach does not necessarily remove age-related information from the backbone. In particular, a strong negative correlation between the true and predicted ages can substantially increase the MSE while still indicating that the representation is heavily dependent on age. In this case, the error is large, but the backbone continues to encode age information, since a simple linear transformation would suffice to recover the true values. This limitation arises because MSE quantifies prediction error rather than statistical dependence. Therefore, maximizing MSE fails to enforce the intended age-invariance.

To address this limitation, we proposed a covariance-based adversarial loss. The covariance between two variables quantifies their linear dependence. Given M training samples with true ages a_i and predicted $\hat{a}_i$, the sample covariance was defined as

$$Cov(a, \hat{a}) = \frac{1}{M} \sum_{i=1}^{M} (a_i - \bar{a}) (\hat{a}_i - \bar{\hat{a}}) \tag{3.6}$$

where $\bar{a}$ and $\bar{\hat{a}}$ denote means of the true and predicted ages, respectively. The adversarial loss was then defined as the negative squared covariance

$$\mathcal{L}_{\text{NCOV}} = -(Cov(a, \hat{a}))^2 \tag{3.7}$$

The squared term ensures that both positive and negative correlations are penalised symmetrically, preventing the model from exploiting an inverted relationship between predicted and true ages. The negative sign is introduced to align the loss with the adversarial optimisation: through the GRL, the age predictor is encouraged to maximise this loss, whereas the backbone

is optimised to minimise it. Under this formulation, the minimum of the objective is attained when the covariance approaches zero, corresponding to the predicted and true ages being linearly uncorrelated.

Replacing the MSE term in Eq. (3.5) with $\mathcal{L}_{\text{NCOV}}$, the update rule for the backbone becomes

$$\theta_B \leftarrow \theta_B - \eta \left(\frac{\partial \mathcal{L}_{\text{BCE}}}{\partial \theta_B} - \alpha \frac{\partial \mathcal{L}_{\text{NCOV}}}{\partial \theta_B} \right). \tag{3.8}$$

The proposed covariance-based loss is related to mutual information applied to the variable age and its estimated version by the network. While mutual information provides a general measure of dependence between two variables, covariance quantifies only the linear relationship with the advantage of an straightforward computation. Nevertheless, in the case of bivariate Gaussian variables, mutual information can be expressed as a function of the squared correlation coefficient, which relates the proposed squared covariance reduction to mutual information minimisation.

3.4 Dataset

In this study, the PTB-XL dataset (version 1.0.3) publicly available on PhysioNet [89], was employed. The dataset consists of 21,799 12-lead ECG recordings collected from 18,869 patients between 1989 and 1996 across multiple German hospitals. Each recording has a duration of 10 seconds and is originally sampled at 500 Hz; a downsampled version at 100 Hz is also provided. The dataset includes metadata such as patient age, sex, and diagnostic annotations encoded according the SCP-ECG standard [90]. Diagnostic labels cover a wide range of cardiac conditions and can be grouped into five principal categories: arrhythmias, conduction blocks, MI, nonspecific abnormalities, and normal ECGs.

For the purposes of this study, a subset of the dataset was selected. Specifically, normal sinus rhythm (NORM) and MI recordings were considered. MI cases were included only in the absence of additional ECG abnormalities, such as bundle branch block, abnormal QRS, high or low QRS voltage, supraventricular tachycardia, sinus tachycardia, paroxysmal supraventricular tachycardia, or LVH. ECG recordings associated with unreliable patient age values (*e.g.*, age set to 300 years) were excluded. All experiments were conducted using the 100 Hz downsampled version of the recordings. After applying the described selection criteria and the preprocessing (see Section 3.5), the final dataset consisted of 7,735 NORM

and 1,695 MI recordings. Notably, the two classes exhibited different age distributions, with NORM subjects having a median age of 54 years [IQR 41.0-65.0], and MI 69 years [IQR 60.0–78.0].

3.5 Data Preprocessing

ECG recordings are commonly affected by multiple sources of noise and interference that can obscure clinically relevant features, distort waveform morphology, and in some cases lead to misinterpretation (see Section 2.2.2). To ensure reliable analysis, a preprocessing pipeline was implemented to attenuate noise while preserving the diagnostic content of the signal.

All selected ECGs were filtered using a zero-phase, third order bandpass Butterworth filter with a pass-band of $0.5 - 40$ Hz to reduce powerline interference, baseline wandering and high frequency noise. After denoising, beats were detected applying the gqrs algorithm [91] on the vector magnitude (VM) of the 12-lead ECG. The VM was computed as the square root of the sum of the squared ECG signals across the 12 leads. The detected beat positions were refined employing the Woody algorithm [92] on the VM. The signal quality of each lead was evaluated by computing the mean Pearson correlation coefficient between each QRS complex (from $Q - 20$ ms to $Q + 100$ ms), and an average QRS template. An ECG trace was considered of good quality if the average cross-correlation was higher than 0.9 in at least 8 leads. ECG recordings failing to meet this quality criterion were excluded from further analysis. For each retained ECG, an average beat was computed for all leads. To reduce the influence of irregular rhythms, only beats with inter-beat time interval, *i.e.*, $QQ_k = Q_k - Q_{k-1}$ with k as the beat index, deviating no more than 50 ms from the median QQ value were included in the averaging process. The resulting average beat had a total duration of 580 ms ($Q - 250$ ms; $Q + 330$ ms).

3.6 DL Model

The backbone b_{θ_B} consists of two one-dimensional convolutional blocks (kernel size=3), with a rectified linear unit (ReLU) activation, batch normalisation, and max-pooling. The convolutional layer were followed by a fully connected layer that produced a shared latent representation. For the primary task, *i.e.* MI classification, the output of the backbone was fed into the m_{θ_M} network, composed by two fully connected layers. The a_{θ_A} network for the

age estimation task consists of a single fully connected layer. The input of the network was 12x58 matrix, representing the 12 lead average beat.

3.7 Experiments

The dataset was partitioned into training (80%) and test sets (20%), ensuring that recordings from the same patient were assigned exclusively to one subset to prevent data leakage. In order to address class imbalance between MI and NORM recordings, the BCE loss was weighted according to the class distribution. The experiments were conducted using a batch size of 32. Different learning rates and number of training epochs were adopted across the training stages.

The AML framework was trained in three sequential stages. In the first stage, the backbone b_{θ_B} and the MI classification network m_{θ_M} were jointly optimised according to Eq. (3.3). Training was performed for 100 epochs using a learning rate of 10^{-4}. This stage also served as a baseline model for the MI classification task only.

In the second stage, the pretrained backbone b_{θ_B} was frozen and paired with the age predictor a_{θ_A}. During this phase, only the parameter θ_A were updated according to Eq. (3.4). The learning rate was increased to 10^{-3} and the number of epochs was set to 200. This stage also served as a baseline model for the age prediction task only.

In the final stage, AML framework was considered. The pretrained backbone b_{θ_B} was updated through GRL according to Eq. (3.5), while the parameters of both the MI classifier and m_{θ_M} and the age predictor a_{θ_A} remained fixed. Training was performed for 100 epochs and the hyperparameter α set to 10^{-3}.

Finally, the AML framework employing the MSE loss for the adversarial task (Eq. (3.5)) was compared with the proposed covariance-based adversarial loss (Eq. (3.8)). The two approaches were evaluated in terms of MI classification accuracy and degree of correlation between predicted and true age. This evaluation framework enabled the assessment of the efficacy of the backbone b_{θ_B} to learn age-invariant representations.

3.8 Results and Discussions

3.8.1 MI Classification Performance

The MI classification baseline model achieved a test accuracy of 0.87, with a sensitivity of 0.88 and a specificity of 0.87. These results indicate discriminative performance across MI and NORM classes and confirm that the backbone architecture was capable of learning clinically relevant representations from beat signals. The baseline performance was therefore used as a reference to assess the effect of the AML framework on MI classification performance and age bias mitigation.

When the AML framework was employed, different patterns were observed depending on the adversarial loss function. Using the MSE loss in the AML framework yielded in a reduced test accuracy of 0.82. In contrast, employing the proposed covariance-based adversarial loss yielded a test accuracy of 0.85, which showed a performance comparable to the baseline model (0.87).

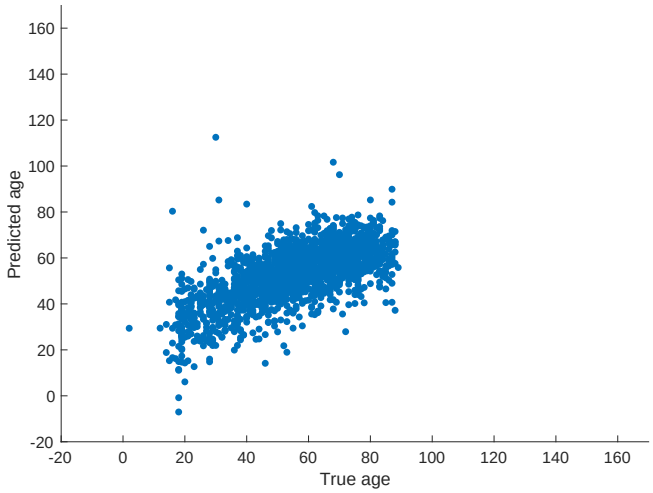
3.8.2 Age Encoding in the Learned Representations

To assess the encoding of age-related information within the learned representations, Pearson correlation coefficient between predicted age and true age was computed on the test set. This prediction was computed with Eq. (3.2) using the three different models.

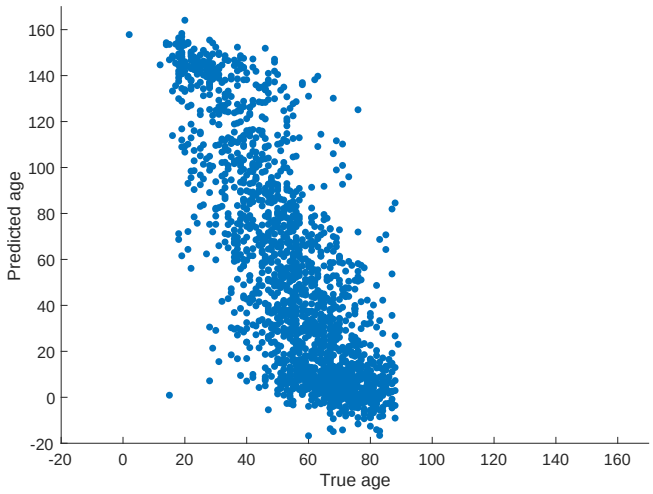
For the age prediction of the baseline model, a correlation coefficient of 0.67 was obtained. Figure 3.2a illustrates a positive linear trend between predicted and true age. This pattern indicated that age-dependent ECG features were encoded in the learning representation and may act as a confounding factor in MI classification. Given the clinical association between age and MI prevalence, the model may partially encode age-related information to support its predictions, potentially leading to biased outputs.

3.8.3 Effect of MSE Loss in AML Framework

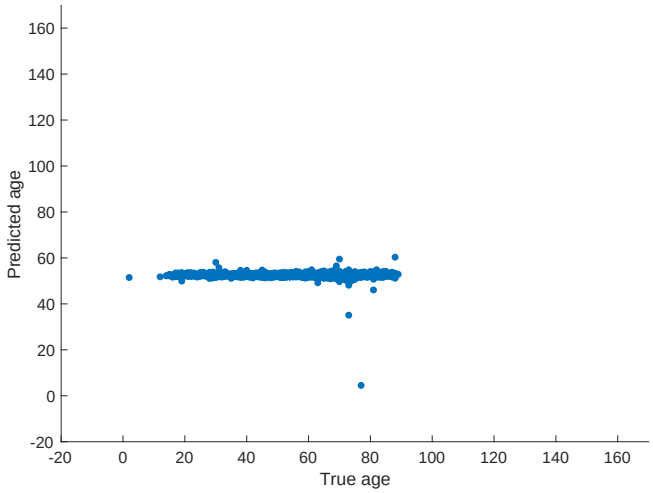
When AML framework was implemented using the MSE loss, a high negative correlation of -0.78 was obtained between predicted and true age. As shown in Figure 3.2b, the scatterplot exhibits a negative linear trend.



(A)



(B)



(C)

FIGURE 3.2: Scatterplots of true *vs.* predicted age (in years) on the test set for different training configurations: A) Baseline age prediction model, B) AML with MSE adversarial loss, C) AML with covariance-based adversarial loss.

Although the direction of the correlation was negative, its large magnitude indicated that substantial age-related information was encoded in the shared representation. In fact, a correlation of -0.78 reflected a strong linear dependence; thus, age could be recovered through a simple affine transformation (*e.g.*, sign inversion and scaling). In other words, the representation retained linearly accessible age information, even though the AML framework increased the regression error. This observation demonstrated that maximizing the regression error of the age predictor did not necessarily eliminate age-related information from the shared representation.

3.8.4 Effect of Negative Covariance Loss Function in AML Framework

In contrast, the adoption of the proposed negative squared covariance loss yielded a Pearson correlation coefficient of -0.03, indicating the absence of linear dependence between predicted and true age. As illustrated in Figure 3.2c, the predicted ages are dispersed across the full range of true ages that indicates the effective mitigation of age-related information.

Unlike the MSE-based AML framework, the covariance-based AML framework explicitly penalised the linear dependence between the learned representation and the age attribute. By directly minimizing statistical dependence rather than maximizing regression error, this framework enhanced decorrelation at the learning representation. Notably, the substantial reduction of age-related information did not compromise MI classification performance.

3.8.5 Comparative Analysis of AML Framework

The obtained results demonstrated that the AML framework can effectively mitigate the age-related information when performing MI classification. While the standard MSE-based AML framework provided high negative correlation, it did not eliminate age information from the shared representation. In contrast, the covariance-based AML framework explicitly reduced statistical dependence and mitigated spurious correlation between MI and age.

Unlike previous approaches, such as the method proposed in [83], where the same loss function, *i.e.*, cross-entropy, was used to optimise the subject-discriminator branch and for the for the AML framework, this study decoupled the learning of age-related information by employing a distinct loss function for the adversarial task. These findings suggest that enforcing statistical decorrelation is more effective than maximizing age-prediction error alone to reduce age-related dependence in MI classification models.

3.9 Conclusion

The experimental results demonstrate that the proposed covariance-based adversarial learning strategy substantially reduced age-related spurious correlations while largely preserving MI classification performance. These findings highlighted the importance of carefully designing adversarial objectives when promoting invariant representation learning in MI diagnosis.

Despite the encouraging results, several aspects require further investigation. Particularly, external validation on independent datasets is required to assess the robustness and generalisability of the learned age-invariant representations. Furthermore, the proposed AML framework is not inherently limited to age-related bias and could be extended to account for multiple demographic confounders. Future work could investigate the inclusion of additional adversarial branches targeting factors such as sex, enabling the simultaneous mitigation of multiple sources of bias.

Chapter IV

Domain Knowledge Injection for MI Diagnosis from ECG and Comparison with Clinical Rule-Based Algorithm

4.1 Introduction

Clinical guidelines such as those codified in the Fourth UDMI [58] provide standardised diagnostic criteria, including ST-segment elevation in anatomically contiguous leads and the development of pathological Q waves. In clinical practice, these criteria are often implemented through automated interpretation systems embedded in commercial ECG machines. These systems are typically rule-based and provide preliminary diagnostic suggestions [26, 27]. For acute MI, sensitivities reported for automated ECG interpretation systems range from approximately 0.62 to 0.78 [93, 94]. While these systems perform reasonably well for certain abnormalities, their accuracy is generally lower than that of cardiologists, mainly due to their limited ability to integrate subtle waveform features [95].

These limitations of rule-based ECG interpretation systems motivate the use of alternative approaches. In this context, DL has emerged as a powerful approach for ECG analysis, capable of automatically extracting representations from ECG signals. Although numerous DL models have been proposed and often achieve encouraging performance, high accuracy does not necessarily guarantee that the learned representations are clinically meaningful. DL models may instead exploit spurious correlations in the training data, raising concerns about their reliability, robustness, and generalisability.

A key limitation of traditional DL training strategies for MI detection is their purely data-driven nature, as they do not explicitly incorporate DK. Injecting DK into the training process represents a promising strategy to mitigate spurious correlations, leading to improved robustness. By guiding the learning process of the DL model and imposing constraints in the latent representations, DK can help extract clinically meaningful features. However, effective incorporation of DK into the DL models remains an open research challenge for MI identification.

In this study, we investigated how the injection of DK during training influenced the latent representations of DL models. Specifically, we proposed two training strategies that constrained an intermediate network layer to estimate clinically meaningful features and evaluate their impact on both classification performance and alignment with established clinical guidelines. In the first strategy (split-wise learning, see Section 4.4), the model parameters were optimised in two consecutive phases: first, the parameters of the ST-segment amplitude estimation module were optimised; subsequently, using the learned representations from the first phase, the parameters of the MI detection module were optimised. In the second strategy (MTL, see Section 4.5), the entire DL model was trained end-to-end to simultaneously optimise parameters to perform two tasks *i.e.*, MI localisation and staging.

In addition, we implement a rule-based algorithm (RBA) grounded in diagnostic criteria that leverages ECG-derived features. By comparing DK–guided DL (DK-DL), purely data-driven DL (B-DL), and the RBA, we aim to quantify the performance gap and better understand the trade-offs among these approaches.

4.2 Domain Knowledge Injection

The incorporation of DK into DL models remains one of the major challenges of AI systems [96]. In the literature, there is no universal definition of DK. Generally, it refers to problem-specific information that is relevant to solving a task [96]. Such information may manifest in several forms, including equations, logical rules, knowledge graphs, or probabilistic relations [97]. DK can represent a powerful strategy to overcome some of the DL limitations. Typically, DL models require a large amount of data to learn meaningful representations. However, in the clinical context, annotated datasets are often limited and DK can restrict the region of the parameters space [98]. In general, DK can be injected into DL models at different levels of the pipeline: i) at the data level, ii) at the model architecture level, iii) at the optimisation level

through the modification of the objective function, and iv) after training through post-hoc refinement [97].

Data Level Integration

At the data level, DK is incorporated by modifying the input data prior to training. One approach involves transforming logical rules into additional input features using propositionalisation techniques [96]. Furthermore, DK can be leveraged to design data augmentation strategies. For example, in contrastive learning for ECG analysis, augmented samples may preserve overall morphological similarity to the original signals while introducing controlled variations in diagnostically relevant regions [99]. In MI samples, this could include masking the QRS complex or scaling the ST-segment. Similarly, prior clinical knowledge about MI can motivate the selection of a reduced subset of leads, thereby changing the input representation [100]. While DK incorporation at the data-level is relatively simple to implement, its impact is restricted to modifying the input representation.

Architecture Level Integration

At the architectural level, DK can be integrated through graph neural networks [97, 101]. In this case, the topology of the graph determines the pattern of information propagation within the network. In the ECG context, for instance, each lead can be processed independently using dedicated feature extraction blocks. The extracted features are aggregated according to defined groups of contiguous leads before being fused to produce a global decision [102]. Attention mechanisms provide another effective strategy for integrating DK at the architectural level, as they allow to emphasise clinically relevant patterns by dynamically weighting important temporal segments or spatial regions [103].

Loss Level Integration

DK is also incorporated by introducing an additional loss component that penalises deviations from domain-specific constraints. In particular, logical rules are translated into differentiable approximations and embedded within the objective function [97, 98, 100]. For example, in ECG-based arrhythmia classification, Sun [100] integrated DK by adding a regularisation term that penalises discrepancies between the class probability distribution predicted by the

DL model and a reference distribution derived from fuzzy logical rules grounded in clinical knowledge.

Post Training Integration

In post-training approaches, DK is incorporated only after the model has completed the learning process, serving as an external validation or correction mechanism [97, 100]. In this case, model predictions are evaluated against predefined domain constraints derived from expert knowledge or clinical guidelines. Predictions that violate these constraints may be rejected, flagged for review, or adjusted to enforce consistency with established rules. In this case, DK does not influence the internal representations during training of the model.

4.3 Related Works

The incorporation of DK into DL models has attracted increasing research interest, particularly for enhancing MI detection and localisation using 12-lead ECG signals [104, 103, 105, 106]. Most existing approaches primarily embedded DK at the architectural level by grouping leads according to clinically defined anatomical regions and processing them through dedicated feature extraction blocks. Among the DK approaches, Prabhakararao *et al.* [103] introduced a method that used a weight-sharing recurrent neural network coupled with intra- and inter-lead attention mechanisms for 12-lead ECG signal. These attention mechanisms allowed to emphasise clinically relevant patterns by dynamically weighting important temporal segments or spatial regions. Evaluated on the PTB dataset, the method achieved an accuracy of 0.98.

Subsequent studies incorporated DK more explicitly through structural modelling of lead–territory relationships. Guo *et al.* [104] developed a DK-guided graph neural network for MI localisation on the PTB-XL dataset, grouping ECG leads into graphs according to correspondences between ECG leads and myocardial regions. A multi-branch dense graph neural network extracted myocardial region-specific features that were refined via a class-level attention mechanism, achieving an accuracy of 0.81 and a compute area under the receiver operating characteristic curve (ROC-AUC) of 0.96. Extending this framework, the authors later introduced an overlapping lead grouping strategy grounded in clinical DK to capture spatial dependencies among leads [106]. The advanced architecture modelled intra- and inter-lead relationships using convolutional layers, squeeze-and-excitation blocks, residual

blocks, and positional transformer modules, followed by a weighted fusion mechanism. This allowed to improve MI localisation performance, achieving an accuracy of 0.85 and a ROC-AUC of 0.95 on the same dataset.

Similarly, Sun *et al.* [102] proposed a DK driven multi-lead group ResNet that integrated both static and dynamic ECG features for MI detection and localisation. The framework employed deterministic learning to capture subtle ECG dynamics and adopted a multi-branch ResNet architecture guided by clinically defined the relationship between leads and MI region. Then, local and global fusion modules aggregated features within and across lead groups to obtain comprehensive representations. Using five-fold cross-validation on the PTB-XL dataset, the method achieved a detection accuracy and localisation accuracy of approximately 0.94 and 0.87, respectively.

Other lead grouping strategies included graph-based approaches which explicitly modelled the relationships between ECG leads and MI regions. Yuan *et al.* [107] constructed a hypergraph in which leads were represented as nodes and MI classes as hyperedges, enabling higher-order modelling of clinically related leads. Their framework combined lead-level and MI classes-level representations and incorporated Wasserstein-distance-based domain alignment strategy to improve cross-domain MI localisation without requiring manual annotations for new patients. Further advancing graph-based modelling strategies, Guo *et al.* [108] proposed a DK-driven graph representation learning framework that integrated ECG signals, morphological features, demographic attributes, and lead–territory associations within a unified graph structure. This framework allowed the model to learn interactions across interconnected entities through a parallel multi-branch embedding network and a relation-aware graph aggregation module. MI localisation was formulated as a link prediction task between patient nodes and MI label nodes.

Notably, MI staging together with MI localisation has been explicitly addressed in only a limited number of studies. Han *et al.* [109] proposed a framework that jointly performed MI staging and localisation by incorporating established clinical diagnostic criteria into a DL pipeline. Their approach employed a densely connected convolutional network architecture to classify beat-level morphological patterns, such as ST-segment elevation. The extracted lead-wise morphological features were then integrated into a knowledge graph structured according to clinical criteria for MI diagnosis. Subsequently, these criteria were applied to determine whether specific morphological patterns occurred across anatomically contiguous

leads, enabling simultaneous identification of MI stage (hyperacute, acute, or prior) and MI localisation.

In summary, existing studies on MI identification from ECGs mainly incorporated DK through architectural design, such as lead grouping. While these strategies enhanced representation by embedding anatomical priors into the model architecture, DK was generally confined to the feature extraction phase. The optimisation process was driven by generic objective functions, such as cross-entropy loss, which did not explicitly encode clinical diagnostic criteria. As a result, learned representations were not directly regularised toward clinically meaningful representation, thereby limiting the full potential of DK to improve robustness and generalisation. Furthermore, early studies did not comprehensively evaluate the performance of DL models using external test sets. Instead, most studies relied primarily on internal validation strategies, such as random train-test splits or cross-validation within the same dataset [102]. Without rigorous external validation, reported performance metrics may overestimate true clinical utility and fail to reflect potential distribution shifts across institutions, devices, or demographic groups. Moreover, the joint modelling of MI detection, localisation, and staging within a unified multi-task framework remains unexplored. Although these tasks were inherently interrelated and relied on shared features, most existing approaches treated them independently or considered MI stage and localisation as secondary objectives. Such task-specific modelling may lead to redundant feature learning and suboptimal use of shared representations. To address these gaps, we propose an unified framework to: i) jointly optimise MI detection, localisation, and staging; ii) incorporate DK into the training; and iii) test the performance on three external datasets whose MI diagnoses were confirmed by cardiologists.

4.4 Domain Knowledge Injection via Split-wise Training

Parts of the study presented in this section are based on previously published work [110].

4.4.1 Motivation

As discussed in Chapter II, deviations of the ST-segment amplitude in the 12-lead ECG represent a cornerstone of acute MI detection and localisation, and are extensively employed in both clinical practice and automated ECG analysis systems. Incorporating DK at the architectural level of a DL model can enhance the learning representations. Given that

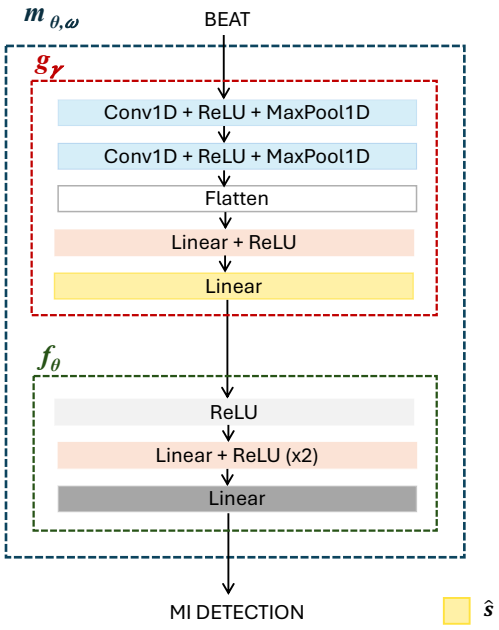


FIGURE 4.1: Architecture of the DL model. The yellow block denotes the linear layer constrained to estimate ST-segment amplitude ($\hat{\mathbf{s}}$) in split-wise model training strategy.

ST-segment amplitude is a clinically established physiological marker and inherently a continuous variable, it was modelled using a regression-based formulation as a form of DK incorporation into the DL model. The underlying rationale was that by enforcing the extraction of clinically significant features at the lower levels of the network (hidden layers), the subsequent classification layers were compelled to rely on these clinically significant features.

4.4.2 Split-wise Model Training

The proposed approach adopted a two-phase training strategy in which an intermediate representation was explicitly constrained to encode clinically meaningful ECG features before being used for the final classification. Formally, the DL model was designed as the composition of two mathematical functions

$$\hat{y} = m_{\gamma,\theta}(\mathbf{x}) = f_\theta(g_\gamma(\mathbf{x})) \tag{4.1}$$

where $\mathbf{x}$ denotes the 12-lead average beat input, $g_\gamma(\mathbf{x})$ maps the input to the constrained hidden representation corresponding to ST-segment amplitudes, and f_θ corresponds to the subsequent layers leading to the final classification output. θ and γ denote the parameters of f_θ and g_γ, respectively.

Figure 4.1 illustrates the architecture of the proposed DL model. The model, denoted as $m_{\gamma,\theta}$, comprised two main components: f_θ and g_γ. The component g_γ consisted of two one-dimensional convolutional blocks. The first block included a 1D convolutional layer with 16 output channels (kernel size = 5), followed by a rectified linear activation unit (ReLU) activation function and a 1D max-pooling layer (kernel size = 2). The second block followed the same structure, employing a 1D convolutional layer with 20 output channels. Subsequently, the resulting feature maps were flattened and passed through two fully connected layers with 256 and 12 neurons, respectively. The final layer of g_γ was explicitly constrained to estimate the ST-segment amplitudes, thereby incorporating clinically relevant DK into the learning process. The f_θ was implemented as a multilayer perceptron composed of three fully connected layers with 32, 64, and 2 neurons, respectively. Each intermediate layer was followed by a ReLU activation function, while the final layer was followed by a Softmax activation function to produce class probabilities corresponding to MI and NORM.

To explicitly supervise the constrained representation, the dataset was augmented with auxiliary targets corresponding to clinically relevant features. Specifically, the dataset was defined as $\mathcal{D} = \{\mathbf{x}_i, \mathbf{s}_i, y_i\}_{i=1}^{M}$ where M is the total number of samples and $\mathbf{s}_i$ is the vector of ST-segment amplitudes of the i-th sample, and y_i is the scalar containing the binary diagnostic label (MI *vs.* NORM). The auxiliary targets $\mathbf{s}_i$ were used exclusively during training.

Model training was carried out in two consecutive phases. In the first phase, the parameters γ were optimised to estimate the ST-segment amplitudes by minimising the MSE between the predicted values ($g_\gamma(\mathbf{x})$) and corresponding ground truth ST-segment amplitudes $\mathbf{s}_i$. In the second phase, the parameters γ were frozen, and the parameters θ were optimised by minimising the cross-entropy loss between the model output $m_{\gamma,\theta}(\mathbf{x}_i)$ and the diagnostic label y_i.

4.4.3 Dataset

The proposed training strategy was evaluated on the PTB Diagnostic ECG Database [36, 111], which contains 549 ECG recordings from 290 subjects with a sampling frequency of 1 kHz.

ECG preprocessing and average beat extraction were carried out as described in Section 3.5. After preprocessing, a total of 52 NORM and 145 MI recordings (patients) were retained for analysis. The dataset was split into training set (0.7) and test set (0.3).

4.4.4 Experiments

To assess the effectiveness of DK incorporation, two DL models were considered. Both models shared the same network architecture, but differed in the training strategy. Specifically, the proposed split-wise model was trained using the two-phase procedure described in Section 4.4.2, whereas the baseline model was trained in a single phase, where all network parameters were jointly optimised using the final classification objective, without intermediate supervision. Both models were trained under the same optimisation conditions. Training was performed for 100 epochs with a batch size of 8, and learning rate of 10^{-4}, using Adam optimiser [112].

4.4.5 Explainability Analysis via Lead Occlusion

The occlusion procedure consisted of iteratively removing the information from a single ECG lead by setting its entire signal to zero, while leaving the other leads unchanged. After each occlusion step, the modified input was fed to the trained model to obtain a new predicted probability of MI. The absolute change between this probability and the baseline probability computed from the original (non-occluded) average beat was then computed. This value quantified the influence of the removed lead on the prediction of the model. For each MI recording, the three leads associated with the largest probability variation were identified as the most influential for MI classification. For each MI region, the leads most frequently identified as relevant were compared with those established in clinical context.

4.4.6 Preliminary Results and Discussions

The end-to-end baseline model achieved a higher test accuracy than the split-wise model (0.85 *vs.* 0.69). However, the two models exhibited distinct lead-importance patterns. In particular, the split-wise model, trained in two separate phases, demonstrated a stronger tendency to rely on clinically relevant leads compared to the baseline model. For instance, in anterior MI cases, the split-wise model prioritised leads V2–V4, consistent with clinical

guidelines. In contrast, the baseline model more frequently attributed importance to leads that are less strongly associated with anterior MI.

These findings suggest that constraining the intermediate representation to estimate ST-segment amplitudes encourages the model to focus on clinically meaningful features. However, this constraint resulted in reduced classification performance. The two-phase training procedure inherently limited the flexibility of the model, as the feature extraction parameters were fixed before the classification task, preventing joint optimisation of all parameters. This limitation along with the promising results of this training approach motivated the development of an end-to-end training strategy.

4.5 DK Injection in End-to-End Multitask Framework

A manuscript authored by Ibrahimi *et al.*, based on the results presented in this section, has been finalised and is ready for submission to the IEEE Transactions on Biomedical Engineering.

4.5.1 Multitask Learning for Comprehensive MI Characterisation

The two-phase training strategy described in Section 4.4.2 presents inherent limitations when multiple tasks, *i.e.*, MI detection, localisation and staging, are performed. Particularly, split-wise optimisation becomes less efficient as the number of objectives (tasks) increases, since each task is optimised independently rather than jointly. This approach limits the ability of the model to learn shared representations that are meaningful across related tasks. To overcome these limitations, the previously proposed framework was extended to an end-to-end MTL framework, in which all tasks are optimised simultaneously within a single training phase.

In the context of MI interpretation, the clinical assessment involves both MI localisation and staging. Therefore, a DL model should not be limited to a single classification objective, but rather designed to provide a comprehensive characterisation of MI. The localisation task was formulated as a multi-label classification problem across four anatomical regions (anterior, septal, inferior, and lateral). For each sample, the model predicted the presence or absence of pathological changes within each region. The MI stage classification task was formulated as a three-class problem, distinguishing acute MI, prior MI, and NORM cases.

In this study, to jointly model MI stage and localisation, a MTL framework was adopted, comprising a shared feature extraction backbone and task-specific output heads. MI localisation was supervised using average binary cross-entropy loss across four classes, which we denoted as $\mathcal{L}_{\text{loc}}$. The prediction of MI stage was supervised using a cross-entropy loss, which we denoted as $\mathcal{L}_{\text{stage}}$. To mitigate the effects of class imbalance, both loss components were weighted according to the class frequencies. The overall training objective was defined as the sum of the two loss terms

$$\mathcal{L}_{\text{baseline}} = \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{stage}} \tag{4.2}$$

which served as the baseline loss function for training the MTL framework.

4.5.2 Regularisation Strategies for DK Injection

We proposed two strategies to inject DK into the DL model. Both strategies employed a custom regularisation term to control the latent space mapped by g_γ. The first strategy followed a bottom-up approach, in the sense that, the network was instructed to learn specific ECG characteristics, such as wave amplitudes and durations, involved in both MI localisation and stage detection. The second strategy was instead a top-down approach, in which the network was instructed to reconstruct a latent space from a set of approximated clinical rules which we made differentiable. The two regularisation terms were then added to the multitask loss and their relative contribution was controlled using two scalar values.

Formally, we built a training dataset $\mathcal{D} = \{e_i\}_{i=1}^{M}$ where M is the total number of training samples and e_i is the i-th tuple defined as $(\mathbf{x}_i, \mathbf{s}_i, \mathbf{q}_{\text{d},i}, \mathbf{q}_{\text{a},i}, \mathbf{r}_{\text{a},i}, a_i, g_i, \mathbf{y}_{\text{loc},i}, \mathbf{y}_{\text{stage},i})$ with $\mathbf{x}_i \in \mathbb{R}^{12\times N}$ is the median beat and N is the beat length, $\mathbf{s}_i \in \mathbb{R}^{12}$ contains the 12 ST-segment amplitudes, $\mathbf{q}_{\text{a},i}$ the 12 Q wave amplitudes, $\mathbf{q}_{\text{d},i}$ the 12 Q wave durations, $\mathbf{r}_{\text{a},i}$ the 12 R-peak amplitudes, a_i is the age in years, g_i is the sex, $\mathbf{y}_{\text{loc},i} \in \{0,1\}^4$ indicates when one or more regions (in the order: anterior, septal, inferior, lateral) are subjected to MI (the region affected is encoded with 1), and $\mathbf{y}_{\text{stage},i} \subset \{0,1\}^3$ is the one-hot encoding of the three possible stages (acute, prior, normal). For convenience, we built the reference matrix $W_i \in \{0,1\}^{4\times 2}$ which encoded the presence of MI across anatomical regions and stages, where $W_i(r,s) = 1$ indicates MI in region r at stage s, 0 otherwise.

The feature vectors $\mathbf{s}_i$, $\mathbf{q}_{\text{d},i}$, $\mathbf{q}_{\text{a},i}$ and $\mathbf{r}_{\text{a},i}$ were z-score normalised in each dimension in the training set, and the 48 averages and standard deviations here computed were also applied to the validation and test sets. Vectors were considered columns. The list of leads was I, II, III,

TABLE 4.1: Differentiable approximations of logical operations used in the DK regulariser.

Logical operation	Differentiable approximation
$\beta > \delta$	$\sigma\left(k(\beta - \delta)\right)$
$\beta \wedge \delta$	$\sigma\left(-\log\left(e^{-k\beta} + e^{-k\delta}\right)\right)$
$\beta \vee \delta$	$\sigma\left(\log\left(e^{k\beta} + e^{k\delta}\right)\right)$

aVR, aVL, aVF, V1, V2, V3, V4, V5 and V6. We kept this order fixed for all our experiments and, for convenience, we associated a progressive number from 1 to 12 to indicate a given lead. To enhance clarity in interpreting the formulas reported in following section, with a little abuse of notation, we refer to the ℓ-th component of a vector with its name rather than the progressive number. For example, the number 8 referred to V2, but also $\mathbf{s}(\ell)$ with $\ell = \text{V2}$ referred to the 8th component of the vector s.

Clinically Relevant Feature Estimation

In the first strategy, we aimed to instruct the network to extract relevant features directly from the ECG, without letting the extraction of features completely handled by the multitask loss. To do so, considering that ST-segment amplitudes, Q wave amplitudes and durations, and R-peak amplitudes across the 12 leads carry fundamental diagnostic information for both localisation and stage detection, we leveraged the additional information provided by the PTB-XL+, in terms of features, to construct a regularisation term which guided the network to learn how to extract such quantities. Specifically, the feature estimation loss was defined as the MSE between the predicted and the true normalised features. Let us define $\mathcal{B}$ as the index set for a mini batch, and the loss as

$$\mathcal{L}_{\text{feat}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \|g_\gamma(x_i) - \mathbf{f}_i\|_2^2 \tag{4.3}$$

where $\mathbf{f}_i = [\mathbf{s}_i, \mathbf{q}_{\text{a},i}, \mathbf{q}_{\text{d},i}, \mathbf{r}_{\text{a},i}]^\intercal$ and $g_\gamma(x_i)$ provides as output $\hat{\mathbf{f}}_i$.

Soft RBA Regularisation

In the second strategy, we encoded a set of diagnostic rules to regularise the latent space. The main aim was to construct an estimated matrix $\hat{W}_i$ from the latent space vectors $\hat{\mathbf{f}}_i$ outputted

by $g_\gamma(x_i)$, and to compare $\hat{W}_i$ with the reference W_i. Unfortunately, the clinical guidelines and the RBA rely on hard, non-differentiable logical rules; therefore, differentiable approximations (Table 4.1) were adopted to enable gradient-based optimisation.

Instrumental for the following calculations, we defined a binary mask $A \in \{0,1\}^{4\times 12}$, mapping contiguous leads to the four anatomical regions. In particular, we set $A(r,\ell) = 1$ when the ℓ-th lead should be considered as part of the contiguous leads for region r.

Let us decompose the vector $\hat{\mathbf{f}}_i$ in $[\hat{\mathbf{f}}_{1,i}, \hat{\mathbf{f}}_{2,i}, \hat{\mathbf{f}}_{3,i}, \hat{\mathbf{f}}_{4,i}]^\intercal$, where each $\hat{\mathbf{f}}_{j,i} \in \mathbb{R}^{12}$ and whose elements are considered connected with the fixed list of leads, *i.e.*, $\hat{\mathbf{f}}_{j,i}(\ell)$ refers to the ℓ-th lead for all j and i. All other vectors in this section follow this convention.

We first described the construction of the first column of W_i which relates to acute MI. We defined the ST-segment elevation for each lead as

$$\mathbf{v}_i = \sigma\left(k(\hat{\mathbf{f}}_{1,i} - \tilde{\mathbf{t}}^{\text{el}})\right) \tag{4.4}$$

where σ is the sigmoid function acting as a differentiable "if" operator, k indicates the steepness of the sigmoid function, and $\mathbf{t}^{\text{el}}$ is the ST elevation threshold, defined as

$$\mathbf{t}^{\text{el}} = \begin{cases} 0.25, & \text{V2-V3, } g = \text{M, } a < 40, \\ 0.20, & \text{V2-V3, } g = \text{M, } a \geq 40, \\ 0.15, & \text{V2-V3, } g = \text{F}, \\ 0.10, & \text{otherwise.} \end{cases} \tag{4.5}$$

and $\tilde{\mathbf{t}}^{\text{el}}$ refers to $\mathbf{t}^{\text{el}}$ after normalisation with the z-score parameters obtained from the training set. Since the latent representations estimated by g_γ correspond to normalised feature values, the clinical thresholds were transformed accordingly so that the differentiable diagnostic rules could be evaluated in the same feature space.

Lead-wise ST-segment elevation scores were aggregated at the anatomical level to enforce lead contiguity. Contiguity was enforced by requiring at least two activated leads within a region. Using the mapping matrix A, the regional abnormality scores were computed as

$$\hat{\mathbf{w}}_{\text{acute},i} = \sigma\left(k(A\mathbf{v}_i - 2)\right) \tag{4.6}$$

where the number 2 was broadcasted for each of the four elements of $A\mathbf{v}_i$.

Regarding the identification of prior MI, the clinical guidelines recommended the verification of any QS complex OR a pathological Q wave. The latter was defined differently depending on the leads. Specifically, it indicated prior MI if, for V2 and V3, Q wave duration was greater than 20 ms, whereas for the remaining leads, Q wave duration greater than 30 ms AND Q wave amplitude lower than -0.1 mV. The OR and AND operators were made differentiable using the approximations reported in Table 4.1.

QS complex detection was based on the R wave amplitude being absent and was modelled as

$$\mathbf{q}_{\mathrm{s},i} = \sigma(-k\hat{\mathbf{f}}_{4,i}) \tag{4.7}$$

while for the pathological Q wave we defined

$$\mathbf{p}_i(\ell) = \begin{cases} \sigma\Big(k\big(\hat{\mathbf{f}}_{3,i}(\ell) - \tilde{\mathbf{t}}^{\mathrm{dur}_1}\big)\Big) & \ell \in \{\mathrm{V2}, \mathrm{V3}\}, \\ \sigma\Big(-\log\Big(e^{-k(\tilde{\mathbf{t}}^{\mathrm{amp}}(\ell) - \hat{\mathbf{f}}_{2,i}(\ell))} + e^{-k(\hat{\mathbf{f}}_{3,i}(\ell) - \tilde{\mathbf{t}}^{\mathrm{dur}_2}(\ell))}\Big)\Big) & \text{otherwise.} \end{cases} \tag{4.8}$$

where $\tilde{\mathbf{t}}^{\mathrm{dur}_1}$, $\tilde{\mathbf{t}}^{\mathrm{dur}_2}$ and $\tilde{\mathbf{t}}^{\mathrm{amp}}$ were the z-score normalised thresholds of 20 ms, 30 ms and -0.1 mV, respectively, as recommended by the guidelines.

The final OR condition was then computed as

$$\mathbf{q}_i = \sigma\left(\log\left(e^{-k\mathbf{p}_i} + e^{-k\mathbf{q}_{\mathrm{s},i}}\right)\right). \tag{4.9}$$

Analogously to the acute MI case, the lead-wise pathological Q wave and QS complex scores were aggregated at the regional level via the mapping matrix A, yielding the final $\hat{\mathbf{w}}_{\mathrm{prior},i}$ using Eq. (4.6) but changing the vector $\mathbf{v}_i$ with $\mathbf{q}_i$. The predicted $\hat{W}$ was then obtained as

$$\hat{W}_i = [\hat{\mathbf{w}}_{\mathrm{acute},i}\ \hat{\mathbf{w}}_{\mathrm{prior},i}]. \tag{4.10}$$

The DK loss was computed with the Frobenius norm

$$\mathcal{L}_{\mathrm{soft}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \|\hat{W}_i - W_i\|_F^2. \tag{4.11}$$

TABLE 4.2: Number of ECGs per dataset and stage. For PTB-XL+ dataset, only ground truth labels are reported (pseudo-labeled cases are excluded). For the Chapman-Shaoxing dataset, MI cases were not differentiated between acute and prior, and thus an overall quantity is reported.

	Train			Val			Test		
Dataset	**Acute**	**Prior**	**Norm**	**Acute**	**Prior**	**Norm**	**Acute**	**Prior**	**Norm**
PTB-XL+	117	826	4843	28	259	1335	13	125	651
CODE	–	–	–	–	–	–	989	–	–
MIMIC-IV	–	–	–	–	–	–	431	2240	–
Chapman-Shaoxing	–	–	–	–	–	–	69		5218

The Final Loss

The final loss function was defined as

$$\mathcal{L} = \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{stage}} + \lambda_1 \mathcal{L}_{\text{feat}} + \lambda_2 \mathcal{L}_{\text{soft}} \tag{4.12}$$

where $\mathcal{L}_{\text{loc}}$ and $\mathcal{L}_{\text{stage}}$ denote the localisation and stage losses of the MTL, respectively. The weighting parameters λ_1 and λ_2 control the relative contributions of the DK-driven terms.

4.5.3 Clinical Rule-Based Algorithm

The RBA for MI detection and localisation encoded the clinical criteria [58] into a set of decision rules, designed to reproduce the ECG interpretation. According to these guidelines (see Section 2.3.1), acute and prior MI are characterised by abnormalities in ST-segment, and Q wave occurring in anatomically contiguous leads. In the RBA, rule satisfaction was determined by comparing ECG features with a threshold defined by the clinical criteria. In addition to ECG features, age and sex were incorporated to apply guideline specific thresholds for the acute MI stage. In accordance with the clinical guidelines and the dataset, rules related to posterior MI and NSTEMI were excluded from the algorithm.

The RBA evaluated rule satisfaction within region-specific lead groups to identify candidate MI regions. When multiple regions met the criteria, localisation was determined by aggregating the identified regions. The MI stage was inferred based on the types of detected abnormalities; in line with the guidelines, we defined: i) ST-segment changes indicate acute

Algorithm 1 RBA for MI Localisation and Staging

Input: a (age), g (sex), $\mathbf{s}$ (array of ST elevations; mV), $\mathbf{q}_\text{a}$ (array of Q wave amplitudes; mV), $\mathbf{q}_\text{d}$ (array of Q wave durations; ms), $\mathbf{r}_\text{a}$ (array of R peak amplitudes; mV)
Output: localisation, stage

Define territories:
$\mathcal{G}_\text{sep} = \{\text{V1}, \text{V2}\},\ \mathcal{G}_\text{ant} = \{\text{V3}, \text{V4}\},$
$\mathcal{G}_\text{lat} = \{\text{I}, \text{aVL}, \text{V5}, \text{V6}\},\ \mathcal{G}_\text{inf} = \{\text{II}, \text{III}, \text{aVF}\}$
$\mathbb{T} = \{\text{sep}, \text{ant}, \text{lat}, \text{inf}\}, \quad \mathcal{A}, \mathcal{P} \leftarrow \varnothing$
Acute MI (ST elevation)

$$\mathbf{t}^\text{el}(\ell) = \begin{cases} 0.25, & \ell \in \{\text{V2}, \text{V3}\}, g = \text{M}, a < 40 \\ 0.20, & \ell \in \{\text{V2}, \text{V3}\}, g = \text{M}, a \geq 40 \\ 0.15, & \ell \in \{\text{V2}, \text{V3}\}, g = \text{F} \\ 0.10, & \text{otherwise} \end{cases}$$

for all $t \in \mathbb{T}$ **do**
  if $|\{\ell \in \mathcal{G}_t : \mathbf{s}(\ell) \geq \mathbf{t}^\text{el}(\ell)\}| \geq 2$ **then**
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{t\}$
  end if
end for
Prior MI (Pathological Q wave OR QS complex)

$$\mathbf{p}(\ell) = \begin{cases} \mathbf{q}_\text{d}(\ell) \geq 20 \text{ ms}, & \ell \in \{\text{V2}, \text{V3}\} \\ \mathbf{q}_\text{d}(\ell) \geq 30 \text{ ms} \wedge \mathbf{q}_\text{a}(\ell) \leq -0.10 \text{ mV}, & \text{otherwise} \end{cases}$$

$\mathbf{q}_s(\ell) := \mathbf{r}_\text{a}(\ell) \leq 0$
for all $t \in \mathbb{T}$ **do**
  if $|\{\ell \in \mathcal{G}_t : \mathbf{p}(\ell) \vee \mathbf{q}_s(\ell)\}| \geq 2$ **then**
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{t\}$
  end if
end for
localisation $\leftarrow \mathcal{A} \cup \mathcal{P}$

$$\text{stage} = \begin{cases} \text{NORM}, & \mathcal{A} \cup \mathcal{P} = \varnothing \\ \text{Acute MI}, & \mathcal{A} \neq \varnothing \\ \text{Prior MI}, & \text{otherwise} \end{cases}$$

return localisation stage

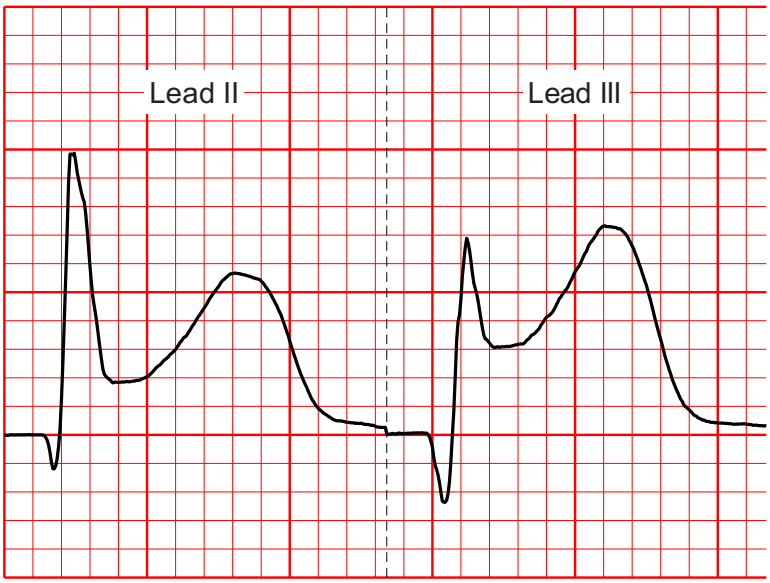


FIGURE 4.2: Representative median beats of lead II and III from an acute MI case. Small horizontal squares represent 40 ms while small vertical squares indicate 0.1 mV.

MI, ii) pathological Q waves indicate prior MI, and iii) the presence of both types of abnormalities indicate acute MI. When none of the guideline-based conditions were satisfied, the ECG was classified as NORM. A description of the RBA is provided in Algorithm 1.

4.5.4 Datasets

In this study, we used the PTB-XL+ dataset (version: 1.0.1) [91, 113], an extension of the PTB-XL dataset, to train the DL models. PTB-XL+ provides 21,799 median beats sampled at 500 Hz, extracted from 10-second 12-lead ECG recordings by means of GE Healthcare's Marquette 12SL [27], ECGDeli [114] and Uni-G software. Each recording is linked to clinical metadata. Also, each beat is associated with a set of features, including wave amplitudes and durations. Here, we used median beats extracted using Uni-G and the ECG features of interest included the amplitudes of the ST-segment at the J-point, Q, and R waves, along with the durations of the Q waves. These features are particularly relevant for detecting ST-segment elevation and Q wave abnormalities.

We applied strict inclusion and exclusion criteria to align with the clinical guidelines. For the MI group, we excluded ECGs that exhibited LVH or LBBB. We also excluded the posterior MI cases due to the low prevalence in the dataset. Diagnostic statements were also associated

to a likelihood. We also excluded the cases for which the likelihood statements was lower than 50% for MI cases and 80% for NORM cases. After this filtering process, we selected 6829 NORM and 2437 MI median beats, derived from 8540 unique patients.

In addition to waveform features, the dataset includes metadata on the MI stage (stadium) and anatomical localisation. Stadium labels were binarised into two classes: "acute" ("Stadium I", "Stadium I-II", or "Stadium II"; 158 beats; this stage categorisation was coherent with [115]); and "chronic/prior" ("Stadium II-III" or "Stadium III"; 2279 beats). However, approximately 40% of the MI ECGs lacked stadium annotations, limiting the use of these metadata in certain analyses.

For MI localisation, codes specifying the MI region were mapped to a binary vector encoding the presence or absence of MI in defined anatomical zones (anterior, septal, lateral, and inferior) to allow the representation of both isolated and concurrent involvement of multiple regions (see Section 4.5.2). In the dataset, 1368 beats showed anterior involvement, 1270 beats septal involvement, 546 beats lateral involvement, and 1528 beats inferior involvement. Counts are overlapping, as a single beat may involve multiple regions.

External validation: We evaluated the proposed methods on three external datasets, each for a different evaluation purpose depending on the available information in a given dataset. The main characteristics of the external datasets, including acquisition settings, and labelling procedures, are summarised in Table 4.3. The same inclusion and exclusion criteria as for the PTB-XL+ dataset were applied, and ECG recordings were preprocessed to extract median beats (see Section 4.5.5 for a description of the preprocessing pipeline).

First, from the Chapman-Shaoxing dataset [116], we selected 69 MI and 5218 NORM median beats. Due to the relatively limited numbers of MI cases, and the lack of detailed annotations, this dataset was primarily used to assess the specificity, *i.e.*, the recognition rate of normal cases.

Second, a subset of the CODE dataset was selected [117]. CODE is a large-scale dataset containing more than two million ECG recordings from over one million patients in Brazil. The ECGs are grouped in 10 clinical categories, encompassing normal recordings, rhythm and conduction disorders, structural abnormalities and ischaemic or infarction changes [118]. Each ECG is associated with an automated diagnostic interpretation by Uni-G and a clinical report by a cardiologist. This subset included 989 acute MI cases, with ECG durations ranging

TABLE 4.3: Characteristics of the external validation datasets.

Dataset	Country	Acquisition Setting	ECG (Duration, Sampling)	Data Labelling	Subset Used
CODE	Brazil	Telehealth Network of Minas Gerais	7–10 s, 300 Hz	Expert annotation	Acute MI
MIMIC-IV-ECG	USA	Beth Israel Deaconess Medical Center (Hospital and emergency department)	10 s, 500 Hz	Automated reports	Acute MI, Prior MI
Chapman-Shaoxing	China	Shaoxing People's Hospital	10 s, 500 Hz	Expert annotation	MI, Healthy

from 7 to 10 s, and served to evaluate acute MI sensitivity. Localisation labels were manually extracted from the clinical reports using a rule-based natural language processing algorithm.

Third, the MIMIC-IV-ECG dataset [119] was used to evaluate prior MI. The dataset comprises more than 800 thousands 10 s 12-lead ECG recordings collected from more than 160 thousands patients between 2008 and 2019. ECG recordings were were acquired at Beth Israel Deaconess Medical Center from the emergency department (25% of the ECGs), hospital (55% of the ECGs) including the intensive care unit and from outpatient clinics. Each ECG signal is associated to a machine-generated report. Although approximately 600 thousands ECGs were reviewed by cardiologists, these expert annotations are not yet publicly available. ECGs collected in hospital and emergency department are linked to a discharge diagnosis with ICD-10-CM codes [120]. Acute MI was defined using ICD-10-CM codes I21 and I22, excluding type 2 MI (I21.A1) and unspecified MI subtypes (I21.9, I21.A9), as these categories do not reliably distinguish between STEMI and NSTEMI. Prior MI was identified using ICD-10-CM code I25.2. MI localisation was extracted from the clinical reports (machine-generated) rather than ICD-10-CM codes, since ICD-10-CM codes do not provide anatomical localisation for prior MI. ECGs were retained only when automated interpretations and ICD-10-CM codes agreed on MI stage. When multiple ECGs were available per emergency department stay or per hospital stay per patient, only the earliest one was included. The resulting set included 431 acute MI and 2240 prior MI cases.

4.5.5 Preprocessing

For each recording in the external datasets, we created the median ECG beats. Specifically, ECG signals were filtered using a Butterworth filter (3rd order, $0.5 - 40$ Hz, zero-phase) to reduce baseline wandering, high-frequency noise, and powerline interference. After denoising, beats were detected using the *gqrs* algorithm [91]. The algorithm was applied on

the VM of the 12-lead ECG. Beat positions were aligned on the R peak using the Woody's algorithm [92] applied to the VM. The signal quality of each lead was evaluated by computing the mean Pearson correlation coefficient between each QRS complex (from $R - 50$ ms to $R + 100$ ms, to include part of the ST-segment), and an average QRS template. An ECG trace was considered of good quality when the average cross-correlation was higher than 0.9 in at least 8 leads. Once the quality check was assessed, we computed the median beat for all leads. The median considered only heart beats whose inter-beat time interval, *i.e.*, $RR_k = R_k - R_{k-1}$ with k as the beat index, did not vary more than 50 ms with respect to the median RR value. The median beat for the external datasets was extracted from $R - 90$ ms and $R + 446$ ms. The median beat for the PTB-XL+ dataset was segmented similarly after detecting the R-peak as the index of the maximum of the vector magnitude. The segmentation boundary of the median beat was selected to exclude the P-wave, which is not involved in the MI and may cause the DL model to leverage spurious correlations during training, and to include T-waves as well. The overall end of the T-wave was computed using Lepeschkin and Surawicz's method [121] on the PTB-XL+ dataset.

ECG recordings of the CODE dataset were originally sampled at 300 Hz. For these recordings, prior to filtering, we resampled the signals at 500 Hz. The sampling rates of the Chapman-Shaoxing and MIMIC-IV-ECG datasets were already in line with the one of the PTB-XL+.

Figure 4.2 reports an example of median beat for acute MI.

4.5.6 DL Model

The DL models architecture consisted of a backbone adapted from the ResNet architecture proposed in [122]. The backbone comprised six residual blocks, each containing two convolutional layers (kernel size = 3), followed by batch normalisation, ReLU, and drop out rate of 0.3. After every second residual block, the temporal resolution was decreased by a factor of two, while the number of feature channels was increased by 64. The number of residual blocks was selected based on the validation set performance. Following the backbone, a fully connected linear layer with 48 neurons was applied. This dimensionality corresponded to the 12-lead features to be estimated. Next, a fully connected linear of 128 neurons was used, followed by an output layer of 7 neurons, comprising 4 neurons for MI localisation and 3 neurons for MI stage detection. Building upon the formulation introduced in Section 4.4.2,

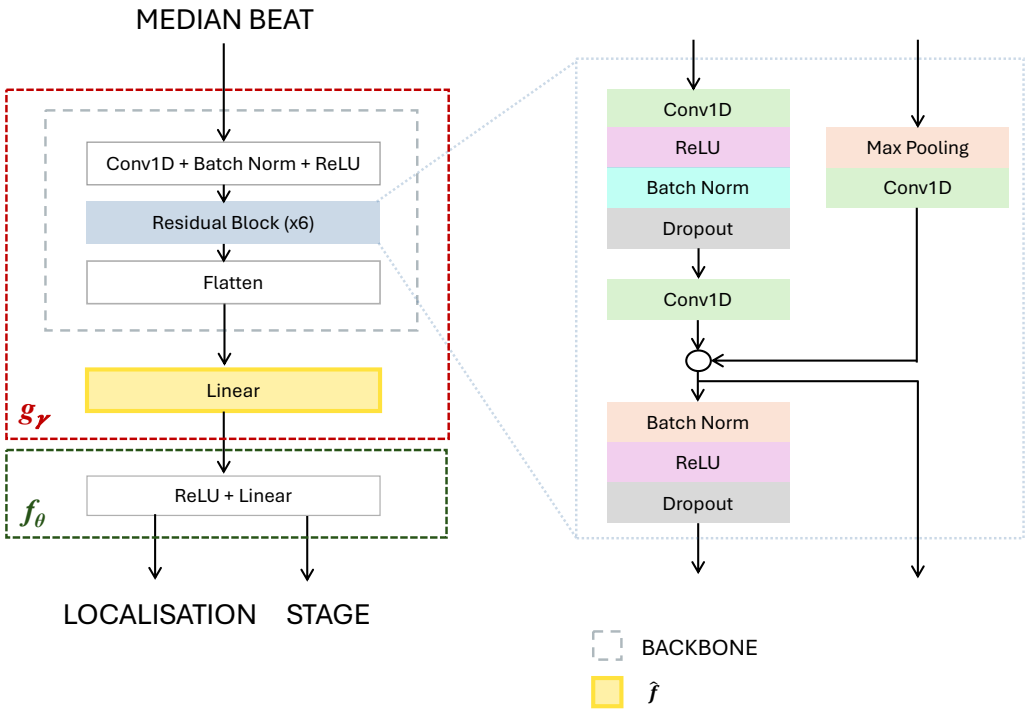


FIGURE 4.3: Diagram of the proposed DL architecture. The yellow block denotes the linear layer that outputs 48 latent features ($\hat{\mathbf{f}}$), corresponding to 12-lead ECG features estimated by the DK-DL model.

model was defined as

$$\hat{\mathbf{y}} = f_\theta(g_\gamma(\mathbf{x})) \tag{4.13}$$

where $\mathbf{x}$ is an ECG, $g_\gamma(\mathbf{x})$ was the neural network related to the backbone and the fully connected layer outputting 48 latent features $\hat{\mathbf{f}} \in \mathbb{R}^{48}$ for each ECG, $f_\theta(\hat{\mathbf{f}})$ was the fully connected layer providing $\hat{\mathbf{y}} \in \mathbb{R}^7$, *i.e.*, the 7 output values, and θ and γ were the learnable parameters of the whole network. Figure 4.3 depicts the structure of the neural network.

4.5.7 DL Experiments

Table 4.2 reports a summary of the number of ECGs in each dataset and stage.

The PTB-XL+ dataset was split into training (70%), validation (20%), and test (10%). All ECG recordings from the same patient were assigned exclusively to the same subset. To promote reproducibility of our study, we used the splits provided by the dataset curators. Specifically, curators split the dataset into ten folds: we considered fold 1 to 7 as training data, fold 8

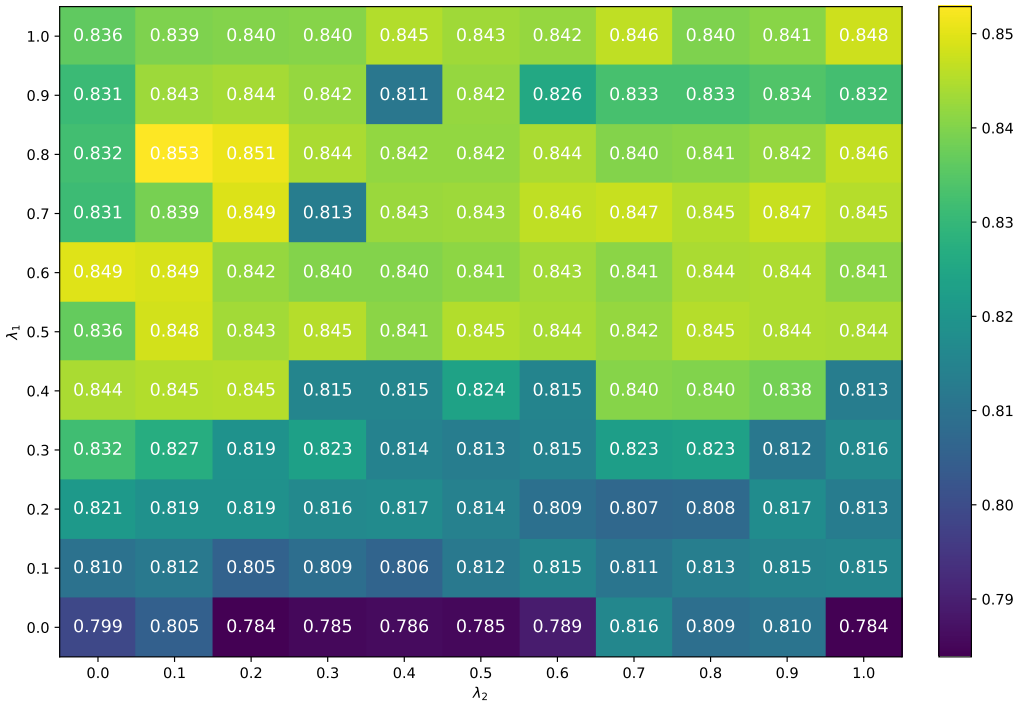


FIGURE 4.4: Heatmap of the G-mean acute MI recall, prior MI recall, and specificity on the validation set for different values of λ_1 and λ_2. The optimal hyperparameter configuration is selected by maximising the G-mean.

and 9 for validation, and fold 10 for testing. The selection of what fold to use for training, validation and test sets was done considering that only fold 9 and 10 contained diagnosis manually confirmed by cardiologists.

For both training and validation sets, when stage labels were unavailable, pseudo-labels were assigned using the RBA (Algorithm 1). Pseudo-labels were not considered for test evaluations.

To assess the impact of DK injection, we compared two DL models sharing the same architecture but trained using different loss functions. The B-DL model was trained exclusively using classification losses for MI stage and localisation, without incorporating DK terms (in other words, $\lambda_1 = \lambda_2 = 0$). The DK-DL model was trained using the proposed loss function in Eq. (4.12). The weighting parameters λ_1 and λ_2 controlling the DK regularisation terms were selected via grid search over the range $[0, 1]$ by maximising the geometric mean (G-mean) of acute MI recall (sensitivity of acute MI), prior MI recall (sensitivity of prior MI), and specificity (true recognition rate of NORM) on the validation set. We selected G-mean

to promote a selection of λ_1 and λ_2 associated with balanced recalls across the three classes. The final values found were $\lambda_1 = 0.8$ and $\lambda_2 = 0.1$. Figure 4.4 reports the G-mean value for the tested ranges. The optimal model architecture described in Section 4.5.6 was found by optimising the same G-mean on the validation set. The steepness parameter of the sigmoid functions was set to $k = 5$. The maximum number of epochs and batch sizes were set to 70 and 64, respectively. We adopted the Adam optimiser, a learning rate of 10^{-3} and early stopping procedure to monitor the loss on the validation set (patience=7). The final B-DL and DK-DL models were trained considering all these hyperparameters associated with the optimal validation loss on the training set.

We compared RBA, B-DL and DK-DL under three different perspectives. The first one aimed to compare the detection rates of MI *vs.* NORM. Here, we considered as MI when either acute or prior MI was detected, while NORM otherwise. Consequently, misclassifications between acute and prior MI did not affect this evaluation, which therefore assessed the ability of the models to detect MI independently of infarction stage. For both B-DL and DK-DL, we considered the output of the staging head for this evaluation. In the second one, we quantified the performance of the three models for MI stage classification for acute MI, prior MI and NORM labels, separately. Finally, in the third one, localisation performance was assessed by computing the per-region recall. Here, the evaluation considered regions separately and independently. Model performances were also evaluated on three external datasets. The main metric used for comparisons was the recall (either sensitivity or specificity depending on the evaluation). Ninety-five percent confidence intervals were computed using the bootstrap method with $10,000$ random resamplings and the percentile method. Evaluations were conducted only on cases for which the labels were available.

To determine whether the observed performance differences between the B-DL and the DK-DL models models were statistically significant, we performed a paired permutation test. The RBA model was included as a reference and was not part of the inferential analysis. Under the null hypothesis that both DL models exhibited equivalent performance, their predictions were assumed to be exchangeable. Permutations were performed at the patient level, such that all predictions associated with a given patient were jointly exchanged between models. For each permutation, the difference in recall between the two DL models was recomputed, generating an empirical null distribution of recall differences. A total of 10,000 permutations were performed for each comparison. Statistical significance was defined as a p-value$<$0.05.

TABLE 4.4: Recalls (sensitivities) with 95% confidence intervals for both MI and stage detection across the three models and four datasets. * denotes a statistically significant difference between DK-DL and B-DL according to a paired patient-level permutation test ($p < 0.05$).

Dataset	Model	Detection		Stage	
		MI	NORM	Acute	Prior
PTB-XL	RBA	0.64 (0.55–0.72)	0.99 (0.98–0.99)	0.62 (0.33–0.90)	0.53 (0.44–0.62)
	B-DL	0.81 (0.74–0.87)	0.96 (0.94–0.97)	0.46 (0.12–0.75)	0.70 (0.61–0.78)
	DK-DL	0.89* (0.83–0.94)	0.97 (0.96–0.98)	0.69 (0.40–0.93)	0.76 (0.68–0.84)
CODE	RBA	0.83 (0.81–0.85)	—	0.65 (0.62–0.68)	—
	B-DL	0.85 (0.83–0.86)	—	0.59 (0.56–0.62)	—
	DK-DL	0.87* (0.85–0.89)	—	0.72* (0.69–0.75)	—
MIMIC-IV-ECG	RBA	0.60 (0.58–0.62)	—	0.69 (0.64–0.73)	0.53 (0.51–0.55)
	B-DL	0.92 (0.91–0.93)	—	0.54 (0.49–0.59)	0.89* (0.88–0.90)
	DK-DL	0.93* (0.92–0.94)	—	0.70 (0.66–0.74)	0.84* (0.83–0.86)
Chapman-Shaoxing	RBA	0.93 (0.86–0.98)	0.96 (0.96–0.97)	—	—
	B-DL	0.96 (0.90–1.00)	0.95 (0.95–0.96)	—	—
	DK-DL	0.96 (0.90–1.00)	0.96* (0.95–0.96)	—	—

4.5.8 Results

We evaluated the performance on the test set of the baseline DL model and the DL model with DK injection, using the RBA as a baseline for comparison.

MI Detection

Table 4.4 reports the performance of the evaluated methods across four different datasets for MI and stage detection. In the PTB-XL+, the RBA showed a strong capability in identifying NORM cases with a recall of 0.99, but lower performance for MI detection which displayed a recall of 0.64. Both DL models improved MI detection while preserving high recall for NORM cases. The B-DL model achieved recall values of 0.81 and 0.96 for MI and NORM cases, respectively. The DK-DL model further improved MI detection, attaining a recall of 0.89, significantly outperforming the B-DL model ($p < 0.05$) for MI cases, while maintaining a comparable recall for NORM cases (0.97 vs 0.96; $p = 0.06$).

To further assess generalisation, model performance was evaluated on external datasets containing only MI cases, namely CODE and MIMIC-IV-ECG. Across these datasets, both DL models outperformed the RBA, with the DK-DL model achieving slightly better performance. On the CODE dataset, DK-DL attained a recall of 0.87, which was significantly higher than

that of B-DL model (0.85; $p < 0.05$), while the RBA achieved a recall of 0.83. A more pronounced performance gap was observed on the MIMIC-IV-ECG dataset, where both DL–based models substantially outperformed the RBA. In particular, a statistically significant increase in recall was observed for DK-DL compared with B-DL (0.93 vs. 0.92; $p < 0.05$). The RBA achieved a recall of 0.60.

Performance was further evaluated on the Chapman-Shaoxing dataset, which included both NORM and MI cases. The RBA achieved a recall of 0.93 for NORM cases and 0.96 for MI cases. The B-DL model achieved a recall of 0.96 for NORM and 0.95 for MI cases, while the DK-DL model showed comparable performance for NORM cases (0.96; p=1.00), and a slightly higher recall (0.96, $p < 0.05$) for MI cases. On this dataset, all methods demonstrated strong performance; however, DL–based models exhibited a more balanced performance across NORM and MI classes.

Recall for NORM cases ranged from 0.96 to 0.99 for RBA, from 0.95 to 0.96 for B-DL and from 0.96 to 0.97 for DK-DL. For MI cases, recall values ranged from 0.60 to 0.93 for RBA, 0.81 to 0.96 for B-DL and from 0.87 to 0.96 for DK-DL. These results for MI detection indicate that DL models provided improved robustness and generalisation compared to the RBA. In particular, the DK-DL model consistently achieved better MI detection performance across both in-distribution and external datasets, without compromising performance on NORM cases.

MI Stage Classification

On the PTB-XL+ dataset, the RBA showed limited capability in correctly classifying both acute MI and prior MI. Recall values were 0.62 for acute MI and 0.53 for prior MI, respectively. The B-DL model showed lower recall for acute MI (0.46) but improved performance for prior MI (0.70), while the DK-DL model achieved the highest performance on this dataset, with recalls of 0.69 for acute MI and 0.76 for prior MI. However, the differences between the DK-DL and the B-DL did not reach statistical significance (acute MI: $p = 0.25$; prior MI: $p = 0.09$).

On the CODE dataset, where only acute MI stages were available, the DK-DL model achieved the highest sensitivity (0.72), outperforming both the B-DL model (0.59; $p < 0.05$) and the RBA (0.65), indicating superior robustness for acute stage classification in an external setting.

Results on the MIMIC-IV-ECG dataset, showed a marked difference in stage detection performance. The RBA achieved limited recall for both acute MI (0.69) and prior MI (0.53).

In contrast, the B-DL model demonstrated the highest performance for prior MI (0.89; $p < 0.05$) but exhibited reduced sensitivity for acute MI (0.54). The DK-DL model achieved instead sensitivities of 0.70 and 0.84 for acute MI and prior MI, respectively, and significantly outperformed the B-DL model in acute MI detection ($p < 0.05$).

For acute MI, recall ranged from 0.62 to 0.69 for RBA, 0.46 to 0.59 for B-DL model and from 0.69 to 0.72 for DK-DL model. In prior MI cases, recall values were comparable across datasets (0.53) for RBA; the ranges for B-DL and DK-DL were 0.70-0.89 and 0.76-0.84, respectively. Across all datasets acute MI was more difficult to detect than prior MI for most methods.

MI Localisation

Table 4.5 summarises the recalls for each of the four involved myocardial regions. On the PTB-XL+ dataset, the RBA exhibited limited capability in localising the affected myocardial region, with both DL models reporting a substantial improvement. For the anterior region, the DK-DL model achieved the highest recall (0.97), significantly outperforming B-DL model (0.91; $p < 0.05$), whereas the RBA showed markedly lower performance (0.35). A similar trend was observed for the septal region, where DK-DL model achieved a significant higher recall than B-DL (0.98 vs. 0.91; $p < 0.05$), while the RBA achieved a recall of 0.66. In the inferior region, both DL models substantially outperformed the RBA, with recall values of 0.85 and 0.91, for B-DL and DK-DL respectively. Statistical analysis confirmed a significant recall advantage of DK-DL model over B-DL ($p < 0.05$). The lateral region proved to be the most challenging for the DL models; B-DL achieved a recall of 0.67, while the DK-DL model achieved a significant higher recall of 0.87 ($p < 0.05$). For RBA, recall was 0.17.

Results on the CODE dataset confirmed the superior performance of the DL models over the RBA across all myocardial regions. The inferior region, yielded the highest recall for all methods (0.86 for DK-DL, 0.82 for B-DL and 0.76 for RBA). Although DK-DL achieved a higher recall than B-DL in this region, the difference did not reach statistical significance ($p = 0.06$). A similar pattern was observed for the septal region, where the difference between the DL models was not statistically significant ($p = 0.09$). The lateral region appeared to be the most difficult to localise. Nevertheless, the DK-DL achieved a significantly higher recall compared to B-DL (0.75 vs. 0.70; $p < 0.05$), while the RBA a value of 0.47. A significant increase in recall was observed for the DK-DL relative to B-DL in the anterior region (0.86 vs. 0.81; $p < 0.05$), whereas the RBA achieved a recall of 0.68.

TABLE 4.5: Recalls with 95% confidence intervals for MI localisation. * denotes a statistically significant difference between DK-DL and B-DL according to a paired patient-level permutation test ($p < 0.05$).

Dataset	Model	Anterior	Septal	Inferior	Lateral
PTB-XL	RBA	0.35 (0.26–0.46)	0.66 (0.56–0.75)	0.45 (0.37–0.53)	0.17 (0.07–0.29)
	B-DL	0.91 (0.85–0.96)	0.91 (0.86–0.96)	0.85 (0.79–0.90)	0.67 (0.52–0.81)
	DK-DL	0.97* (0.93–0.99)	0.98* (0.95–1.00)	0.91* (0.86–0.95)	0.87* (0.76–0.96)
CODE	RBA	0.66 (0.64–0.71)	0.61 (0.57–0.65)	0.76 (0.71–0.81)	0.47 (0.42–0.52)
	B-DL	0.81 (0.78–0.84)	0.80 (0.77–0.84)	0.82 (0.77–0.86)	0.70 (0.65–0.75)
	DK-DL	0.86* (0.83–0.88)	0.82 (0.79–0.86)	0.86 (0.82–0.90)	0.75* (0.70–0.79)
MIMIC-IV-ECG	RBA	0.33 (0.30–0.36)	0.69 (0.65–0.72)	0.54 (0.51–0.56)	0.50 (0.45–0.56)
	B-DL	0.93 (0.92–0.95)	0.90 (0.87–0.92)	0.90 (0.88–0.91)	0.77 (0.73–0.82)
	DK-DL	0.94 (0.93–0.95)	0.96* (0.94–0.97)	0.95* (0.94–0.96)	0.91* (0.88–0.94)

Similarly, result on MIMIC-IV-ECG dataset further confirmed the superior capability of the DL models for MI localisation. The lateral region remained the most difficult to localise for both DL models. Across the evaluated regions, DK-DL consistently achieved the highest recall and significantly outperformed the B-DL in three of the four regions ($p < 0.05$), while no statistically significant difference was found in the anterior region ($p = 0.40$).

For the anterior region, recall ranged from 0.33 to 0.68 in RBA, from 0.81 to 0.93 for B-DL and from 0.86 to 0.97. For the septal region, recall ranged from 0.61 to 0.69 in RBA, from 0.80 to 0.91 for B-DL and from 0.82 to 0.96 for DK-DL. For the inferior region, recall ranged from 0.45 to 0.76 in RBA, from 0.82 to 0.90 for B-DL and from 0.86 to 0.95 for DK-DL. For the lateral region, recall ranged from 0.17 to 0.50 in RBA, from 0.674 to 0.77 for B-DL and from 0.75 to 0.91 for DK-DL. These results across all datasets demonstrate that DL models improved MI localisation compared to the RBA. In general, the incorporation of DK further enhanced the performance, with DK-DL model achieving the most robust results.

4.5.9 Discussions

In this study, we aimed to fill the gaps reported in the introduction for MI detection under several perspectives. First, we designed a DL architecture that could tackle both localisation and staging. This was achieved by considering an established technique called multitask learning. Second, we designed a training strategy to make the DL model leverage the DK strongly consolidated by years of clinical research. Third, we evaluated the proposed approaches using some of the largest ECG databases freely available with confirmed diagnosis

and compared the performance of these DL models with that of the rules reported by the clinical guidelines.

Experimental results demonstrated that DL models achieved an overall superior performance compared to rule-based approaches across all tasks. The comparatively lower performance of the RBA in MI detection, staging and localisation, may be attributed to two main reasons. First, the RBA solely relies on a set of extracted ECG features and does not leverage the full information present in the signal. Second, the rules are designed through fixed decision thresholds, which may inadequately capture borderline cases. However, in clinical practice, such thresholds are often flexibly interpreted by cardiologists. The sensitivity to rigid decision boundaries is a known limitation of rule-based systems and has motivated the adoption of statistical diagnostic algorithms in automated ECG analysis software. In contrast, in the DK-DL model, the clinical rules were not enforced as hard constraints. As a result, the influence on hard clinical thresholds on the DK-DL model was mitigated.

Another important aspect is that annotation quality may partially explain the difference in observed performance across datasets. The PTB-XL dataset is known to contain annotation inconsistencies [113, 123]. When evaluated on CODE and Chapman-Shaoxing datasets, which included cardiologist-validated diagnostic annotations, the RBA achieved superior performance. However, RBA proved to achieve strong capability in identifying NORM cases.

In general, both DL models achieved superior results compared to the RBA. The DK incorporation further enhanced model performance, suggesting that the DK-DL model encouraged the learning of physiologically meaningful representations. This aspect promoted the robustness of the DK-DL model, which was found to be consistent across the datasets. When considering MI stage classification, the DK injection was particularly effective for acute MI detection. In the PTB-XL+ dataset, acute MI cases represented the minority class. Due to the limited number of samples, a DL model may exhibit lower capacity in learning robust pattern for this class (even if a weighted loss was used for mitigation). The incorporation of DK proved effective, as it guided the DL model towards meaningful ECG features. However, on MIMIC-IV-ECG dataset, the B-DL demonstrated a higher performance for prior MI detection. These findings suggest that prior MI patterns may benefit from unconstrained feature learning, or that rules for detecting prior MI are still insufficiently good. In the proposed approach, DK injection emphasised pathological Q-wave patterns. However, other ECG features can be relevant in prior MI detection, such as abnormalities in R-wave progression. As a consequence, complementary features relevant to prior MI may be less effectively captured,

leading to a reduced recall for this stage. Moreover, clinical ECG criteria for prior MI were found less sensitive than those for acute MI, which may further limit the effectiveness. In fact, a study showed that the sensitivity of Q-wave criteria for prior MI detection was around 38% [69]. From a clinical perspective, however, the improved acute MI detection is particularly relevant, as early identification impacts clinical decision making and treatment.

An additional factor may contribute to the lower recall observed for MI stage classification compared to MI detection. Since some ECG features associated with MI may be observed across different stages, a degree of uncertainty in stage labels cannot be excluded. This limitation is less relevant for MI detection, where acute and prior MI are merged into a single class.

The localisation task was formulated as region-wise multi-label classification problem to enable independent prediction of anterior, septal, inferior, and lateral myocardial involvement, allowing multiple regions to be localised within the same ECG recording. DL models achieved higher recall compared to the RBA. DL models achieved the lowest recall for the lateral region, which was the least represented class in the training dataset, suggesting sensitivity to class imbalance. However, even with such factor potentially affecting the training, both DL models outperformed the RBA, which showed the worst performance for the anterior region in the PTB-XL+ and MIMIC-IV-ECG datasets, and lateral region for the CODE dataset.

When considering other studies, MI localisation has been typically formulated as a multiclass classification problem for distinguishing NORM from different MI localisation classes. Only one work was found to adopt a multitask learning approach to jointly detect and localise MI [124]. Differently, our approach avoided the mutual exclusivity among classes mathematically imposed by the softmax layer applied to the output. Moreover, DK knowledge has been mainly incorporated for tackling MI localisation and not MI stage. It was either done by fusing features extracted from individual ECG leads according to the lead-region associations [102, 106], or by modelling such relationships through graph-based approaches such as graph neural networks [101, 108, 125]. Our approach could instead predict localisation and stage concurrently.

MI stage detection remains considerably less explored in the literature. This is also due to the difficulty of assigning reliable stage labels from ECG recordings. Han *et al.* [109] addressed

both MI localisation and stage detection by integrating clinical rules through a knowledge-graph-based framework on a private dataset. In their work, a DL model is used to classify the beat morphology for each lead in one of four classes, which should represent different MI conditions. Once predicted, a list of features is extracted from the beat and, together with the predicted class, they queried the knowledge graph to obtain a possible interpretation for both localisation and stage. In contrast, in our work, we considered an end-to-end training, in which DK was directly incorporated into the model, without relying on external knowledge graphs.

This study has several limitations. First, our analysis of the acute stage was restricted to STEMI, as the lack of publicly available and well annotated datasets currently limits the investigation of NSTEMI. Moreover, we excluded MI cases presenting i) LVH and LBBB, since they require specific clinical criteria, and ii) posterior MI given the low prevalence in the training dataset. Second, label quality also represents a critical challenge when incorporating DK into supervised learning frameworks. While DL generally exhibit greater robustness to label noise compared to rule-base approaches, their performance remains dependent on the quality of the reference annotations. Also, the reliance on automatically generated annotations during training may limit the generalisability of the results to real-world clinical setting, since ECG interpretation systems may produce heterogeneous diagnostic reports for the same signal [20]. Third, despite the proposed method can simultaneously detect MI stage and localisation, the two heads of the DL model may contrast with each other. Indeed, one head may predict that no regions are affected by MI while the other head could predict either acute or prior MI, or vice versa. Fortunately, these two errors were quantified to be around 3% on the PTB-XL+, thus suggesting this condition as sufficiently rare. Future studies should consider strategies to incorporate label uncertainty to improve robustness of the DL models in presence of noisy labels, or to make the model hierarchical where MI localisation is performed after detecting MI.

Despite these limitations, the present study demonstrated that the incorporation of DK can mitigate challenges associated with model robustness and generalisation capability. More broadly, these findings align with [126], which emphasise the incorporation of clinical DK to enhance interpretability, performance, and reduce the reliance to spurious correlations. Moreover, DK incorporation may be beneficial in the scenarios characterised by data scarcity. In these setting, combining DK with open-set detection strategies could further improve robustness by enabling the identification of ECGs that are not represented in the modelled

conditions, thus reducing overconfident predictions.

4.6 Conclusion

In this chapter, we demonstrated that incorporating DK into DL models is an effective strategy for mitigating spurious correlations. Two DK injection strategies were investigated. The first, a split-wise approach, enforced the estimation of clinically relevant features during an intermediate phase of training. This strategy encouraged the model to focus on the leads that were consistent with clinical guidelines, suggesting that DK was effectively integrated into the learning process. The second strategy employed a MTL framework, in which DK was incorporated through regularisation terms in the objective function. This approach enabled MI detection, stage and localisation simultaneously while embedding clinically meaningful constraints into the model. The proposed DK-guided model was compared with a RBA constructed according to clinical guidelines and with a traditional DL model trained without DK incorporation. All three models were evaluated on three external datasets to assess their generalisability. The DL-based models outperformed the RBA, suggesting that clinical criteria may require further adjustments. Overall, these findings from two strategies highlight the importance of incorporating DK into DL models to enhance robustness and generalisation.

Chapter V

Conclusions and Final Remarks

5.1 Conclusions

Accurate diagnosis of MI fundamentally depends on meaningful feature extraction that reflects the underlying myocardial injury. In recent years, the rapid advancement of artificial intelligence, particularly DL, has led to improvement in automated MI diagnosis, often exceeding the performance of traditional systems. However, these improvements in predictive accuracy have not been accompanied by rigorous assessment of robustness and generalisability. In heterogeneous real-world clinical practice, variations in patient demographics, acquisition protocols, and disease prevalence introduce distributional shifts that may significantly affect model performance and reliability. Addressing these challenges is essential to ensure that robust algorithms translate into clinically reliable decision-support systems.

Motivated by these challenges, the main objective of this thesis was to enhance the robustness of DL models for MI diagnosis by incorporating DK. To achieve this, several complementary strategies were investigated. The possible involvement of age as a known confounding factor was tackled through adversarial learning in order to enhance physiologically meaningful representations. In addition, a MTL framework was employed to jointly perform MI detection, localisation, and staging. These tasks are clinically interdependent and share common physiological characteristics. By leveraging shared representations across related tasks, the MTL framework aimed to reinforce physiologically coherent feature extraction and improve generalisation.

5.1.1 Mitigating Age Bias in ECG-Based Myocardial Infarction Diagnosis via Adversarial Multitask Learning

Age constitutes a clinically relevant and potentially confounding factor in the diagnosis of MI, since ECG morphology varies with age. In the presence of demographic imbalance within the training data, DL models may exploit age-related features that correlate with MI prevalence, rather than learning features directly indicative of myocardial injury. Such reliance on demographic confounding factors can introduce bias and compromise generalisation.

To address this issue, an AML framework was proposed to disentangle MI-related representations from age-related features within the learned embedding space. The loss function was designed within the AML framework, in which MI detection was the primary task, while age estimation was considered a secondary task. To minimise the correlation between predicted and true age, a negative squared covariance loss was incorporated into the loss function for the secondary task. This constraint allowed the model to retain relevant information for MI detection while suppressing age-related information in the learned representation.

Experimental results demonstrated that the proposed AML framework reduced the Pearson correlation coefficient between true and predicted age to approximately zero, indicating that age-related information was effectively removed from the learned representation. Notably, this strategy did not degrade the performance of MI detection compared to a single-task based MI detection model. Overall, these findings demonstrate that achieving high predictive performance alone is insufficient for developing robust models. Instead, robustness requires the mitigation of confounding factors.

5.1.2 Domain Knowledge Injection for MI Diagnosis from ECG and Comparison with Clinical Rule-Based Algorithm

DK plays a crucial role in the development of reliable and clinically meaningful MI diagnosis systems, where purely data-driven approaches may capture spurious correlations rather than physiologically relevant features. In this chapter, we investigated the impact of incorporating DK into DL models for MI detection by comparing traditional end-to-end DL models, trained without DK incorporation, to DK-guided approaches.

Two DK injection strategies were explored. The first was a split-wise model, in which the model was trained to estimate clinically relevant features during an intermediate phase. This

constraint encouraged the model to focus on clinically appropriate leads that were more consistent with established guidelines. Explainability analysis using lead occlusion provided insight into the internal decision-making processes of the model. Although the baseline end-to-end model achieved higher accuracy, the split-wise model demonstrated stronger alignment with clinical guidelines. In particular, for anterior MI cases, it more consistently prioritised leads V2–V4, which are clinically recognised as the most relevant leads for this condition. In contrast, the baseline model more frequently attributed importance to leads less strongly associated with the affected MI region.

However, MI diagnosis in clinical practice extends beyond confirming infarct presence; it also requires characterisation of temporal stage and anatomical localisation for making therapeutic decisions. To meet this requirement, DK was further incorporated within a MTL framework to enhance robustness and generalisation. In this second strategy, DK was integrated through two complementary ways: i) softly constraining an intermediate layer to estimate ECG features associated with MI, thereby encouraging the model to learn physiologically meaningful representations, and ii) introducing a regularisation term derived from clinical guidelines into the loss function. The proposed framework was evaluated through comprehensive comparisons with a traditional DL model trained without DK-guided constraints as well as an implemented RBA following the clinical criteria. Furthermore, performance was evaluated across four independent datasets to assess robustness under heterogeneous clinical conditions and potential distributional shifts. The DK-guided MTL framework consistently outperformed both baseline approaches, demonstrating improved robustness and generalisation across datasets.

Overall, this work highlights the value of incorporating DK into DL frameworks to enhance physiologically meaningful representations and mitigate reliance on spurious correlations. Such incorporation represents a promising direction toward more reliable and clinically deployable AI systems for automated MI diagnosis.

5.2 Final Remarks

The results presented in this thesis showed that robustness can be enhanced by incorporating DK into the learning process and addressing confounding factors that may bias the learned representations.

Despite the promising results of this thesis, several important limitations need to be acknowledged.

Data-related challenges remain a key concern, particularly with respect to class imbalance and annotation consistency. Certain MI subtypes, such as acute and posterior MI, are underrepresented. This class imbalance restricts the capacity of the DL model to learn discriminative representations. A further limitation concerns annotation consistency. Although a subset of the annotations of the dataset has undergone manual validation, not all diagnostic categories have been systematically reviewed by clinical experts, potentially leading to variability in annotation reliability across diagnostic categories. Such annotation variability can affect the learning process and ultimately hinder the robustness of DL models. Although this thesis focused on improving robustness with respect to a confounding factor and multi-centre datasets, it did not explicitly address annotation uncertainty or annotation noise. Future research should explore approaches that account for annotation uncertainty, such as noise-robust loss functions and probabilistic annotation frameworks for MI diagnosis.

In addition to dataset-related limitations, methodological constraints also need to be considered. Although age was explicitly considered as a demographic confounding factor, ECG morphology can be influenced by multiple additional variables, including sex, medication use and cardiac abnormalities. However, the present thesis does not explicitly model the joint influence of multiple confounders. Therefore, future research should explore strategies that are capable of handling multiple confounding factors. In this context, causal learning approaches may represent a promising direction for mitigating spurious correlations.

Another consideration concerns the representation of clinical ECG criteria within the proposed framework. The framework was developed and evaluated in patients for whom standard ECG criteria for MI interpretation are applicable. Consequently, conditions such as LBBB or LVH, which require different diagnostic criteria, were beyond the scope of the present study. Extending this approach to broader patient populations would require the incorporation of additional clinical criteria. Future research should therefore investigate how larger sets of ECG criteria can be incorporated while preserving robustness.

Finally, important considerations arise regarding clinical translation. Before deployment, substantial improvements in generalisation performance, cross-dataset validation, and robustness across diverse patient populations are required. Moreover, it is important to recognise that MI diagnosis is inherently multimodal. While ECG provides critical and often immediate

diagnostic information, it may not be sufficient in many borderline cases. Clinical diagnosis typically requires integration of cardiac biomarkers, patient symptoms, medical history and other clinical variables. Particularly, MI presentations may not exhibit definitive ECG patterns, making sole reliance on ECG signals inadequate. Therefore, future research should explore multimodal learning frameworks that combine ECG data with additional diagnostic modalities.

Publications

Journal Articles

- S. Ibrahimi, L. D'Andrea, D. Gastaldi, M. W. Rivolta, and P. Vena, "Machine Learning approaches for the design of biomechanically compatible bone tissue engineering scaffolds". Computer Methods in Applied Mechanics and Engineering, vol. 423, p. 116842, Apr. 2024, doi: 10.1016/j.cma.2024.116842.
- S. Ibrahimi *et al.*, "A Domain Knowledge-Guided Deep Learning Framework for Myocardial Infarction Detection, Staging and Localization". Manuscript under preparation for submission.

Conference Proceedings

- S. Ibrahimi, M. W Rivolta, and R. Sassi, "Injecting Domain Knowledge in Deep Learning Models for Automatic Identification of Myocardial Infarction from Electrocardiograms". Computing in Cardiology Conference, Nov. 26, 2023. doi: 10.22489/cinc.2023.388.
- S. Ibrahimi, M. W Rivolta, and R. Sassi, "Adversarial Multitask Learning Reduces the Correlation between Age and Deep Learning Predictions of Myocardial Infarction from Electrocardiograms". Computing in Cardiology Conference, Dec. 01, 2024. doi: 10.22489/cinc.2024.451.
- S. Ibrahimi, M. W Rivolta, and R. Sassi, "A Comparative Study of Clinical Rule-Based and Deep Learning-Based Diagnosis of Myocardial Infarction Using Electrocardiograms". Computing in Cardiology Conference, doi:10.22489/CinC.2025.440
- A. Bhensdadia, S. Ibrahimi, F. Maffezzoli and M. W Rivolta, "Inverse Design of Scaffolds for Bone Tissue Engineering using Artificial Neural Networks and Generative Additive Models". Engineering in Medicine and Biology Society, 2025.

Bibliography

[1] ECGWaves. ECG localization of myocardial infarction / ischemia and coronary artery occlusion (culprit). https://ecgwaves.com/topic/localization-localize-myocardial-infarction-ischemia-coronary-artery-occlusion-culprit-stemi/, 2016. [Accessed 06 September 2025].

[2] World Health Organization. The top 10 causes of death, 2024. Accessed: 2025-08-30.

[3] Irene R Dégano, Veikko Salomaa, Giovanni Veronesi, Jean Ferriéres, Inge Kirchberger, Toivo Laks, Aki S Havulinna, Jean-Bernard Ruidavets, Marco M Ferrario, Christa Meisinger, et al. Twenty-five-year trends in myocardial infarction attack and mortality rates, and case-fatality, in six European populations. *Heart*, 101(17):1413–1421, 2015.

[4] Marco Zuin, Gianluca Rigatelli, Pierluigi Temporelli, Stefania Angela Di Fusco, Furio Colivicchi, Giampaolo Pasquetto, and Claudio Bilato. Trends in acute myocardial infarction mortality in the European Union, 2012–2020. *European journal of preventive cardiology*, 30(16):1758–1771, 2023.

[5] Haibin Li, Jin Zheng, Frank Qian, Xinye Zou, Siyu Zou, Zhiyuan Wu, Xiuhua Guo, and Pixiong Su. Demographic and regional trends of acute myocardial infarction-related mortality among young adults in the US, 1999–2020. *npj Cardiovascular Health*, 2(1):9, 2025.

[6] Nick S Nurmohamed, Quyen Ngo-Metzger, Pam R Taub, Kausik K Ray, Gemma A Figtree, Marc P Bonaca, Judith A Hsia, Santosh Angadageri, James P Earls, Fatima Rodriguez, et al. First myocardial infarction: risk factors, symptoms, and medical therapy. *European heart journal*, 46(38):3762–3772, 2025.

[7] Bryan Chong, Jayanth Jayabaskaran, Silingga Metta Jauhari, Siew Pang Chan, Rachel Goh, Martin Tze Wah Kueh, Henry Li, Yip Han Chin, Gwyneth Kong, Vickram Vijay

Anand, et al. Global burden of cardiovascular diseases: projections from 2025 to 2050. *European journal of preventive cardiology*, 32(11):1001–1015, 2025.

[8] Gregory A Roth, George A Mensah, Catherine O Johnson, Giovanni Addolorato, Enrico Ammirati, Larry M Baddour, Noël C Barengo, Andrea Z Beaton, Emelia J Benjamin, Catherine P Benziger, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *Journal of the American college of cardiology*, 76(25):2982–3021, 2020.

[9] Nader Salari, Fatemeh Morddarvanjoghi, Amir Abdolmaleki, Shabnam Rasoulpoor, Ali Asghar Khaleghi, Leila Afshar Hezarkhani, Shamarina Shohaimi, and Masoud Mohammadi. The global prevalence of myocardial infarction: a systematic review and meta-analysis. *BMC cardiovascular disorders*, 23(1):206, 2023.

[10] Judith H Lichtman, Erica C Leifheit, Basmah Safdar, Haikun Bao, Harlan M Krumholz, Nancy P Lorenze, Mitra Daneshvar, John A Spertus, and Gail D'Onofrio. Sex differences in the presentation and perception of symptoms among young patients with myocardial infarction: evidence from the virgo study (variation in recovery: role of gender on outcomes of young ami patients). *Circulation*, 137(8):781–790, 2018.

[11] Josep Darbà, Meritxell Ascanio, and Antonio Rodríguez. Productivity costs associated with premature deaths due to acute myocardial infarction in Spain: analysis from 2013 to 2022. *The European Journal of Health Economics*, pages 1–9, 2025.

[12] Kristoffer Jarlov Jensen, Jedidiah I Morton, Marius Mølsted Flege, Janne Petersen, and Zanfina Ademi. Healthcare costs of myocardial infarction in Denmark: A nation-wide registry-based cohort study. *Value in health regional issues*, 48:101125, 2025.

[13] Hassan Serrier, Hugo Rabier, Violaine Fernandez, Anne-Marie Schott, Nathan Mewton, Michel Ovize, Norbert Nighoghossian, Antoine Duclos, and Cyrille Colin. Comparison of healthcare costs before and after a STEMI: A French national health data system-based cohort study. *Journal of Cardiology*, 83(1):44–48, 2024.

[14] Prisco Piscitelli, Giovanni Iolascon, Marco Greco, Alessandra Marinelli, Francesca Gimigliano, Raffaele Gimigliano, Pietro Gisonni, Antonio Giordano, Alberto Migliore, Mauro Granata, et al. The occurrence of acute myocardial infarction in Italy: a five-year analysis of hospital discharge records. *Aging Clinical and Experimental Research*, 23(1):49–54, 2011.

[15] Kai Nogami, Masahiro Hoshino, Yoshihisa Kanaji, Tomoyo Sugiyama, Toru Misawa, Masahiro Hada, Masao Yamaguchi, Tatsuhiro Nagamine, Yun Teng, Hiroki Ueno, et al. Prognostic implications of unrecognized myocardial infarction before elective percutaneous coronary intervention. *Scientific Reports*, 12(1):21579, 2022.

[16] Gary J Balady, Vincent J Bufalino, Martha Gulati, Jeffrey T Kuvin, Lisa A Mendes, and Joseph L Schuller. COCATS 4 task force 3: training in electrocardiography, ambulatory electrocardiography, and exercise testing. *Journal of the American College of Cardiology*, 65(17):1763–1777, 2015.

[17] Salah S Al-Zaiti, Christian Martin-Gill, Jessica K Zègre-Hemsey, Zeineb Bouzid, Ziad Faramand, Mohammad O Alrawashdeh, Richard E Gregg, Stephanie Helman, Nathan T Riek, Karina Kraevsky-Phillips, et al. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. *Nature Medicine*, 29(7):1804–1813, 2023.

[18] Ping Xiong, Simon Ming-Yuen Lee, and Ging Chan. Deep learning for detecting and locating myocardial infarction by electrocardiogram: A literature review. *Frontiers in cardiovascular medicine*, 9:860032, 2022.

[19] Paul Kligfield, Leonard S Gettes, James J Bailey, Rory Childers, Barbara J Deal, E William Hancock, Gerard Van Herpen, Jan A Kors, Peter Macfarlane, David M Mirvis, et al. Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society endorsed by the International Society for Computerized Electrocardiology. *Journal of the American College of Cardiology*, 49(10):1109–1127, 2007.

[20] J De Bie, C Martignani, G Massaro, and I Diemberger. Performance of seven ECG interpretation programs in identifying arrhythmia and acute cardiovascular syndrome. *Journal of Electrocardiology*, 58:143–149, 2020.

[21] Annemarijn SM Steijlen, Kaspar MB Jansen, Armagan Albayrak, Derk O Verschure, and Diederik F Van Wijk. A novel 12-lead electrocardiographic system for home use: Development and usability testing. *JMIR mHealth and uHealth*, 6(7):e10126, 2018.

[22] Jürg Schläpfer and Hein J Wellens. Computer-interpreted electrocardiograms: benefits and limitations. *Journal of the American College of Cardiology*, 70(9):1183–1192, 2017.

[23] Hubert V Pipberger, Edward D Freis, Leonard Taback, and Henry L Mason. Preparation of electrocardiographic data for analysis by digital electronic computer. *Circulation*, 21(3):413–418, 1960.

[24] Pentti M Rautaharju. Eyewitness to history: Landmarks in the development of computerized electrocardiography. *Journal of electrocardiology*, 49(1):1–6, 2016.

[25] Cesar A Caceres, Charles A Steinberg, Sidney Abraham, William J Carbery, Joseph M McBride, Walter E Tolles, and Arthur E Rikli. Computer extraction of electrocardiographic parameters. *Circulation*, 25(2):356–362, 1962.

[26] PW Macfarlane, B Devine, and E Clark. The university of Glasgow (Uni-G) ECG analysis program. In *Computers in Cardiology, 2005*, pages 451–454, 2005.

[27] GE Healthcare. *Marquette 12SL ECG Analysis Program: Physician's Guide*, 2019. Version 2056246-002c.

[28] PW Macfarlane. Minnesota coding and the prevalence of ECG abnormalities, 2000.

[29] JS Duisterhout, JF May, G Van Herpen, CA Distelbrink, JL Talmon, and HWM Plokker. A computer program for ECG classification according to the Minnesota code. *Journal of Electrocardiology*, 10(4):379–386, 1977.

[30] Peter W Macfarlane and Julie Kennedy. Automated ECG interpretation—a brief history from high expectations to deepest networks. *Hearts*, 2(4):433–448, 2021.

[31] Liping Xie, Zilong Li, Yihan Zhou, Yiliu He, and Jiaxin Zhu. Computational diagnostic techniques for electrocardiogram signal analysis. *Sensors*, 20(21):6318, 2020.

[32] Roshan Joy Martis, U Rajendra Acharya, and Hojjat Adeli. Current methods in electrocardiogram characterization. *Computers in biology and medicine*, 48:133–149, 2014.

[33] Mohammed A Chowdhury, Rodrigue Rizk, Conroy Chiu, Jing J Zhang, Jamie L Scholl, Taylor J Bosch, Arun Singh, Lee A Baugh, Jeffrey S McGough, KC Santosh, et al. The heart of transformation: exploring artificial intelligence in cardiovascular disease. *Biomedicines*, 13(2):427, 2025.

[34] Javad Hassannataj Joloudari, Sanaz Mojrian, Issa Nodehi, Amir Mashmool, Zeynab Kiani Zadegan, Sahar Khanjani Shirkharkolaie, Roohallah Alizadehsani, Tahereh Tamadon, Samiyeh Khosravi, Mitra Akbari Kohnehshari, et al. Application of

artificial intelligence techniques for automated detection of myocardial infarction: a review. *Physiological Measurement*, 43(8):08TR01, 2022.

[35] Sahar Ramezani Moghadam and Babak Mohammadzadeh Asl. Automatic diagnosis and localization of myocardial infarction using morphological features of ECG signal. *Biomedical Signal Processing and Control*, 83:104671, 2023.

[36] Ralf-Dieter Bousseljot, D Kreiseler, and A Schnabel. The PTB diagnostic ECG database, 2004.

[37] Chuang Han and Li Shi. Automated interpretable detection of myocardial infarction fusing energy entropy and morphological features. *Computer methods and programs in biomedicine*, 175:9–23, 2019.

[38] Revathi Jothiramalingam, Anitha Jude, and Duraisamy Jude Hemanth. Review of computational techniques for the analysis of abnormal patterns of ECG signal provoked by cardiac disease. *Computer Modeling in Engineering and Sciences*, 128(3), 2021.

[39] Ulas Baran Baloglu, Muhammed Talo, Ozal Yildirim, Ru San Tan, and U Rajendra Acharya. Classification of myocardial infarction with multi-lead ECG signals and deep CNN. *Pattern recognition letters*, 122:23–30, 2019.

[40] Wenhan Liu, Fei Wang, Qijun Huang, Sheng Chang, Hao Wang, and Jin He. MFB-CBRNN: A hybrid network for mi detection using 12-lead ECGs. *IEEE Journal of Biomedical and Health Informatics*, 24(2):503–514, 2020.

[41] Hari Mohan Rai, Kalyan Chatterjee, Alok Dubey, and Praween Srivastava. Myocardial infarction detection using deep learning and ensemble technique from ECG signals. In *Proceedings of International Conference on Computing, Communications, and Cyber-Security: IC4S 2020*, pages 717–730, 2021.

[42] Younghoon Cho, Joon-myoung Kwon, Kyung-Hee Kim, Jose R Medina-Inojosa, Ki-Hyun Jeon, Soohyun Cho, Soo Youn Lee, Jinsik Park, and Byung-Hee Oh. Artificial intelligence algorithm for detecting myocardial infarction using six-lead electrocardiography. *Scientific reports*, 10(1):20495, 2020.

[43] Nils Strodthoff and Claas Strodthoff. Detecting and interpreting myocardial infarction using fully convolutional neural networks. *Physiological measurement*, 40(1):015001, 2019.

[44] HM Wang, Wei Zhao, DY Jia, Jing Hu, ZQ Li, Cong Yan, and TY You. Myocardial infarction detection based on multi-lead ensemble neural network. In *International conference of the engineering in medicine and biology society*, pages 2614–2617, 2019.

[45] Chuang Han and Li Shi. ML-ResNet: A novel network to detect and locate myocardial infarction using 12 leads ECG. *Computer methods and programs in biomedicine*, 185:105138, 2020.

[46] Jia-Zheng Jian, Tzong-Rong Ger, Han-Hua Lai, Chi-Ming Ku, Chiung-An Chen, Patricia Angela R Abu, and Shih-Lun Chen. Detection of myocardial infarction using ECG and multi-scale feature concatenate. *Sensors*, 21(5):1906, 2021.

[47] Vicneswary Jahmunah, Eddie YK Ng, Ru-San Tan, Shu Lih Oh, and U Rajendra Acharya. Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals. *Computers in Biology and Medicine*, 146:105550, 2022.

[48] Weibai Pan, Ying An, Yuxia Guan, and Jianxin Wang. MCA-net: A multi-task channel attention network for myocardial infarction detection and location using 12-lead ECGs. *Computers in Biology and Medicine*, 150:106199, 2022.

[49] Eedara Prabhakararao and Samarendra Dandapat. Myocardial infarction severity stages classification from ECG signals using attentional recurrent neural network. *IEEE Sensors Journal*, 20(15):8711–8720, 2020.

[50] Juan Pablo Martínez, Olle Pahlm, Michael Ringborn, Stafford Warren, Pablo Laguna, and Leif Sörnmo. The STAFF III Database: ECGs recorded during acutely induced myocardial ischemia. *Computing in Cardiology*, 44:266–133, 2017.

[51] J. Pettersson, E. Carro, L. Edenbrandt, O. Pahlm, M. Ringborn, L. Sörnmo, S. Warren, and G. Wagner. Spatial, individual and temporal variation of the high frequency qrs amplitudes in the 12 standard electrocardiographic leads. *American Heart Journal*, 139(2):352–358, 2000.

[52] Girmaw Abebe Tadesse, Hamza Javed, Komminist Weldemariam, Yong Liu, Jin Liu, Jiyan Chen, and Tingting Zhu. DeepMI: Deep multi-lead ECG fusion for identifying myocardial infarction and its occurrence-time. *Artificial Intelligence in Medicine*, 121:102192, 2021.

[53] Leif Sörnmo and Pablo Laguna. *Bioelectrical signal processing in cardiac and neurological applications*. 2005.

[54] Gari D Clifford, Francisco Azuaje, Patrick Mcsharry, et al. ECG statistics, noise, artifacts, and missing data. *Advanced methods and tools for ECG data analysis*, 6(1):18, 2006.

[55] Gustavo Lenis, Nicolas Pilia, Axel Loewe, Walther HW Schulze, and Olaf Dössel. Comparison of baseline wander removal techniques considering the preservation of st changes in the ischemic ECG: a simulation study. *Computational and mathematical methods in medicine*, 2017(1):9295029, 2017.

[56] Amit Singhal, Pushpendra Singh, Binish Fatimah, and Ram Bilas Pachori. An efficient removal of power-line interference and baseline wander from ECG signals by employing fourier decomposition technique. *Biomedical Signal Processing and Control*, 57:101741, 2020.

[57] Raquel Bailón, Leif Sornmo, and Pablo Laguna. A robust method for ECG-based estimation of the respiratory frequency during stress testing. *IEEE transactions on biomedical engineering*, 53(7):1273–1285, 2006.

[58] Kristian Thygesen, Joseph S Alpert, Allan S Jaffe, Bernard R Chaitman, Jeroen J Bax, David A Morrow, Harvey D White, and Executive Group on behalf of the Joint European Society of Cardiology (ESC)/American College of Cardiology (ACC)/American Heart Association (AHA)/World Heart Federation (WHF) Task Force for the Universal Definition of Myocardial Infarction. Fourth universal definition of myocardial infarction (2018). *Circulation*, 138(20):e618–e651, 2018.

[59] W Bruce Fye. The delayed diagnosis of myocardial infarction: it took half a century! *Circulation*, 72(2):262–271, 1985.

[60] Gianfranco Cervellin and Giuseppe Lippi. Of MIs and men—a historical perspective on the diagnostics of acute myocardial infarction. In *Seminars in thrombosis and hemostasis*, volume 40, pages 535–543, 2014.

[61] Elliott Antman, Jean-Pierre Bassand, Werner Klein, Magnus Ohman, Jose Luis Lopez Sendon, Lars Rydén, Maarten Simoons, and Michal Tendera. Myocardial infarction redefined—a consensus document of the joint european society of cardiology/american college of cardiology committee for the redefinition of myocardial

infarction: the joint european society of cardiology/american college of cardiology committee. *Journal of the American College of Cardiology*, 36(3):959–969, 2000.

[62] Kristian Thygesen, Joseph S Alpert, Harvey D White, Allan S Jaffe, Fred S Apple, Marcello Galvani, Hugo A Katus, L Kristin Newby, Jan Ravkilde, Bernard Chaitman, et al. Universal definition of myocardial infarction: Kristian thygesen, joseph s. alpert and harvey d. white on behalf of the joint ESC/ACCF/AHA/WHF task force for the redefinition of myocardial infarction. *European heart journal*, 28(20):2525–2538, 2007.

[63] European Society of Cardiology. Esc congress scientific programme. https://www.escardio.org/events/congresses/esc-congress/scientific-programme/, 2026. [Accessed 02 February 2026].

[64] James CC Moon, Diego Perez De Arenaza, Andrew G Elkington, Anil K Taneja, Anna S John, Duolao Wang, Rajesh Janardhanan, Roxy Senior, Avijit Lahiri, Philip A Poole-Wilson, et al. The pathologic basis of Q-wave and non-Q-wave myocardial infarction: a cardiovascular magnetic resonance study. *Journal of the American College of Cardiology*, 44(3):554–560, 2004.

[65] Jesse McLaren, José Nunes de Alencar, Emre K Aslanger, H Pendell Meyers, and Stephen W Smith. From ST-segment elevation MI to occlusion MI: the new paradigm shift in acute myocardial infarction. *JACC: Advances*, 3(11):101314, 2024.

[66] Thomas Lindow, Henrik Engblom, Olle Pahlm, Marcus Carlsson, Annmarie Touborg Lassen, Mikkel Brabrand, Jakob Lundager Forberg, Pyotr G Platonov, and Ulf Ekelund. Low diagnostic yield of ST elevation myocardial infarction amplitude criteria in chest pain patients at the emergency department. *Scandinavian Cardiovascular Journal*, 55(3):145–152, 2021.

[67] Abdur R Khan, Harsh Golwala, Avnish Tripathi, Aref A Bin Abdulhak, Chirag Bavishi, Haris Riaz, Vishnu Mallipedi, Ambarish Pandey, and Deepak L Bhatt. Impact of total occlusion of culprit artery in acute non-ST elevation myocardial infarction: a systematic review and meta-analysis. *European heart journal*, 38(41):3082–3089, 2017.

[68] José Nunes de Alencar Neto, Matheus Kiszka Scheffer, Bruno Pinotti Correia, Kleber Gomes Franchini, Sandro Pinelli Felicioni, and Mariana Fuziy Nogueira De Marchi.

Systematic review and meta-analysis of diagnostic test accuracy of ST-segment elevation for acute coronary occlusion. *International journal of cardiology*, 402:131889, 2024.

[69] Michelle Lobeek, E Badings, M Lenssen, R Uijlings, K Koster, E van't Riet, and FMAC Martens. Diagnostic value of the electrocardiogram in the assessment of prior myocardial infarction. *Netherlands heart journal*, 29(3):142–150, 2021.

[70] Giuseppe Femia, John K French, Craig Juergens, Dominic Leung, and Sidney Lo. Right ventricular myocardial infarction: pathophysiology, clinical implications and management. *Reviews in Cardiovascular Medicine*, 22(4):1229–1240, 2021.

[71] James A Goldstein, Stamatios Lerakis, and Pedro R Moreno. Right ventricular myocardial infarction—a tale of two ventricles: JACC focus seminar 1/5. *Journal of the American College of Cardiology*, 83(18):1779–1798, 2024.

[72] Myrte Barthels, Elisa Verhofstadt, Inigo Bermejo Delgado, Henri Gruwez, Laurent Pison, Noëlla Pierlet, and Pieter Vandervoort. Artificial intelligence-predicted ECG age gap as a biomarker: bias-adjusted correlation with mortality and cardiovascular risk factors. *European Heart Journal-Digital Health*, 7(2):ztaf137, 2026.

[73] Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. Detecting shortcut learning for fair medical ai using shortcut testing. *Nature communications*, 14(1):4314, 2023.

[74] Cathy Ong Ly, Balagopal Unnikrishnan, Tony Tadic, Tirth Patel, Joe Duhamel, Sonja Kandel, Yasbanoo Moayedi, Michael Brudno, Andrew Hope, Heather Ross, et al. Shortcut learning in medical ai hinders generalization: method for estimating ai model generalization without external data. *NPJ digital medicine*, 7(1):124, 2024.

[75] Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE journal of biomedical and health informatics*, 25(5):1519–1528, 2020.

[76] Zachi I. Attia, Paul A. Friedman, Peter A. Noseworthy, Francisco Lopez-Jimenez, Dorothy J. Ladewig, and et.al. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circulation: Arrhythmia and Electrophysiology*, 12(9):e007284, 2019.

[77] Mohammed Yusuf Ansari, Marwa Qaraqe, Fatme Charafeddine, Erchin Serpedin, Raffaella Righetti, and Khalid Qaraqe. Estimating age and gender from electrocardiogram signals: a comprehensive review of the past decade. *Artificial Intelligence in Medicine*, 146:102690, 2023.

[78] Lourdes Vicent and Manuel Martínez-Sellés. Electrocardiogeriatrics: ECG in advanced age. *Journal of electrocardiology*, 50(5):698–700, 2017.

[79] Charlene Chu, Simon Donato-Woodger, Shehroz S Khan, Tianyu Shi, Kathleen Leslie, Samira Abbasgholizadeh-Rahimi, Rune Nyrup, and Amanda Grenier. Strategies to mitigate age-related bias in machine learning: Scoping review. *JMIR aging*, 7:e53564, 2024.

[80] Erick A Perez Alday, Ali B Rad, Matthew A Reyna, Nadi Sadr, Annie Gu, and et al. Age, sex and race bias in automated arrhythmia detectors. *Journal of Electrocardiology*, 74:5–9, 2022.

[81] Lars Rumberg, Hanna Ehlert, Ulrike Lüdtke, and Jörn Ostermann. Age-invariant training for end-to-end child speech recognition using adversarial multi-task learning. In *Proc. Interspeech 2021*, pages 3850–3854, 2021.

[82] Mostafa Ali Shahin, Beena Ahmed, and Julien Epps. Speaker- and age-invariant training for child acoustic modeling using adversarial multi-task learning. *ArXiv*, abs/2210.10231, 2022.

[83] Mostafa Shahin, Ethan Oo, and Beena Ahmed. Adversarial multi-task learning for robust end-to-end ECG-based heartbeat classification. In *International Conference of the Engineering in Medicine & Biology Society*, pages 341–344, 2020.

[84] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[85] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[86] Ziniu Hu, Zhe Zhao, Xinyang Yi, Tiansheng Yao, Lichan Hong, Yizhou Sun, and Ed Chi. Improving multi-task generalization via regularizing spurious correlation. *Advances in Neural Information Processing Systems*, 35:11450–11466, 2022.

[87] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189, 2015.

[88] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

[89] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset, 2022.

[90] Jos L Willems, Christoph Zywietz, Paul Rubel, Rosanna Degani, Peter W Macfarlane, and Jan H van Bemmel. A standard communications protocol for computerized electrocardiography. *Journal of Electrocardiology*, 24:173–178, 1991.

[91] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch Ivanov, and et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101:e215–e220, 2000.

[92] Charles D Woody. Characterization of an adaptive filter for the analysis of variable latency neuroelectric signals. *Medical and biological engineering*, 5(6):539–554, 1967.

[93] Ziad Faramand, Stephanie Helman, Abdullah Ahmad, Christian Martin-Gill, Clifton Callaway, Samir Saba, Richard E Gregg, John Wang, and Salah Al-Zaiti. Performance and limitations of automated ECG interpretation statements in patients with suspected acute coronary syndrome. *Journal of Electrocardiology*, 69:45–50, 2021.

[94] Yuan-Hui Wu, Ai-Hsien Li, Tsan-Chi Chen, Jen-Kuei Liu, Kuang-Chau Tsai, and Min-Po Ho. Compared with physician overread, computer is less accurate but helpful in interpretation of electrocardiography for ST-segment elevation myocardial infarction. *Journal of Electrocardiology*, 81:60–65, 2023.

[95] I Min Chiu, Yuki Sahashi, Sam S Torbati, Sumeet S Chugh, and David Ouyang. Factors associated with physician modifications to automated ECG interpretations. *European Heart Journal - Digital Health*, 7(1):ztaf119, 2026.

[96] Tirtharaj Dash, Sharad Chitlangia, Aditya Ahuja, and Ashwin Srinivasan. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, 12(1):1040, 2022.

[97] Laura Von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, et al. Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):614–633, 2021.

[98] Soumali Roychowdhury, Michelangelo Diligenti, and Marco Gori. Regularizing deep networks with prior knowledge: A constraint-based approach. *Knowledge-Based Systems*, 222:106989, 2021.

[99] Shuang Zhou, Xiao Huang, Ninghao Liu, Wen Zhang, Yuan-Ting Zhang, and Fu-Lai Chung. Open-world electrocardiogram classification via domain knowledge-driven contrastive learning. *Neural Networks*, 179:106551, 2024.

[100] Jie Sun. Domain knowledge enhanced deep learning for electrocardiogram arrhythmia classification. *Frontiers of Information Technology & Electronic Engineering*, 24(1):59–72, 2023.

[101] Zhaoyang Ge, Huiqing Cheng, Zhuang Tong, Ziyang He, Adi Alhudhaif, Kemal Polat, and Mingliang Xu. A knowledge-driven graph convolutional network for abnormal electrocardiogram diagnosis. *Knowledge-Based Systems*, 296:111906, 2024.

[102] Qinghua Sun, Jiali Li, Chunmiao Liang, Rugang Liu, Jiaojiao Pang, Yuguo Chen, and Cong Wang. A multi-lead group network for myocardial infarction detection and localization based on clinical knowledge-driven and dynamic-static feature fusion. *Expert Systems with Applications*, 274:126901, 2025.

[103] Eedara Prabhakararao and Samarendra Dandapat. Attentive RNN-based network to fuse 12-lead ECG and clinical features for improved myocardial infarction diagnosis. *IEEE Signal Processing Letters*, 27:2029–2033, 2020.

[104] Lin Guo, Yingqi Wu, Nan Ma, and Ying An. KGD-GNN: A knowledge-guided graph neural network for myocardial infarction localization via 12-lead ECG. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025.

[105] Jun Long, Jichao Yang, Lin Guo, and Ying An. KUMA-MI: A 12-lead knowledge-guided multi-branch attention networks for myocardial infarction localization. In *International Symposium on Bioinformatics Research and Applications*, pages 360–372, 2024.

[106] Lin Guo, Qianyun Zhan, Jichao Yang, Ying An, Jun Long, and Nan Ma. Lead-grouped multi-stage learning for myocardial infarction localization. *Methods*, 234:315–323, 2025.

[107] Shuaiying Yuan, Ziyang He, Jianhui Zhao, Zhiyong Yuan, Adi Alhudhaif, and Fayadh Alenezi. Hypergraph and cross-attention-based unsupervised domain adaptation framework for cross-domain myocardial infarction localization. *Information Sciences*, 633:245–263, 2023.

[108] Fengyi Guo, Ying An, Hulin Kuang, and Jianxin Wang. Knowledge-driven graph representation learning for myocardial infarction localization. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[109] Chuang Han, Shihao Pan, Wenge Que, Zhizhong Wang, Yunkai Zhai, and Li Shi. Automated localization and severity period prediction of myocardial infarction with clinical interpretability based on deep learning and knowledge graph. *Expert Systems with Applications*, 209:118398, 2022.

[110] Silvia Ibrahimi, Massimo W Rivolta, and Roberto Sassi. Injecting domain knowledge in deep learning models for automatic identification of myocardial infarction from electrocardiograms. In *Computing in Cardiology*, volume 50, pages 1–4, 2023.

[111] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[112] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[113] Nils Strodthoff, Temesgen Mehari, Claudia Nagel, Philip J Aston, Ashish Sundar, Claus Graff, Jørgen K Kanters, Wilhelm Haverkamp, Olaf Dössel, Axel Loewe, et al. PTB-XL+, a comprehensive electrocardiographic feature dataset. *Scientific data*, 10(1):279, 2023.

[114] Nicolas Pilia, Claudia Nagel, Gustavo Lenis, Silvia Becker, Olaf Dössel, and Axel Loewe. ECGdeli-an open source ECG delineation toolbox for matlab. *SoftwareX*, 13:100639, 2021.

[115] Karolina Janciulevičiūte, Daivaras Sokas, Justinas Bacevičius, Leif Sornmo, and Andrius Petrenas. ECG-based detection of acute myocardial infarction using a wrist-worn device. *IEEE Transactions on Biomedical Engineering*, 2025.

[116] Jianwei Zheng, Hangyuan Guo, and Huimin Chu. A large scale 12-lead electrocardiogram database for arrhythmia study. *PhysioNet*, August 2022. Version 1.0.0.

[117] Antonio Luiz P Ribeiro, Gabriela MM Paixao, Paulo R Gomes, Manoel Horta Ribeiro, Antonio H Ribeiro, Jessica A Canazart, Derick M Oliveira, Milton P Ferreira, Emilly M Lima, Jermana Lopes de Moraes, et al. Tele-electrocardiography and bigdata: the CODE (Clinical Outcomes in Digital Electrocardiography) study. *Journal of electrocardiology*, 57:S75–S78, 2019.

[118] Petrus EOGB Abreu, Gabriela MM Paixão, Jiawei Li, Paulo R Gomes, Peter W Macfarlane, Ana Oliveira, Vinicius T Carvalho, Thomas B Schön, Antonio Luiz P Ribeiro, and Antônio H Ribeiro. CODE-II: A large-scale dataset for artificial intelligence in ECG analysis. *arXiv:2511.15632*, 2025.

[119] Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset. *PhysioNet*, September 2023. Version 1.0.

[120] Nils Strodthoff, Juan Miguel Lopez Alcaraz, and Wilhelm Haverkamp. MIMIC-IV-ECG-Ext-ICD: Diagnostic labels for MIMIC-IV-ECG. *PhysioNet*, August 2024. Version 1.0.1.

[121] Eugene Lepeschkin and Borys Surawicz. The measurement of the Q-T interval of the electrocardiogram. *Circulation*, 6(3):378–388, 1952.

[122] Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature communications*, 11(1):1760, 2020.

[123] Peter Doggart, Alan Kennedy, Emily Foreman, Dewar Finlay, and Raymond Bond. Automated identification of label errors in large electrocardiogram datasets. In *Computing in Cardiology*, volume 498, pages 1–4, 2022.

[124] Yupeng Qiang, Xunde Dong, Xiuling Liu, and Yang Yang. MT-MV-KDF: A novel multi-task multi-view knowledge distillation framework for myocardial infarction detection and localization. *Biomedical Signal Processing and Control*, 95:106382, 2024.

[125] Lin Guo, Yingqi Wu, Nan Ma, and Ying An. KGD-GNN: A knowledge-guided graph neural network for myocardial infarction localization via 12-lead ECG. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025.

[126] Christel Sirocchi, Alessandro Bogliolo, and Sara Montagna. Medical-informed machine learning: integrating prior knowledge into medical decision systems. *BMC Medical Informatics and Decision Making*, 24(Suppl 4):186, 2024.