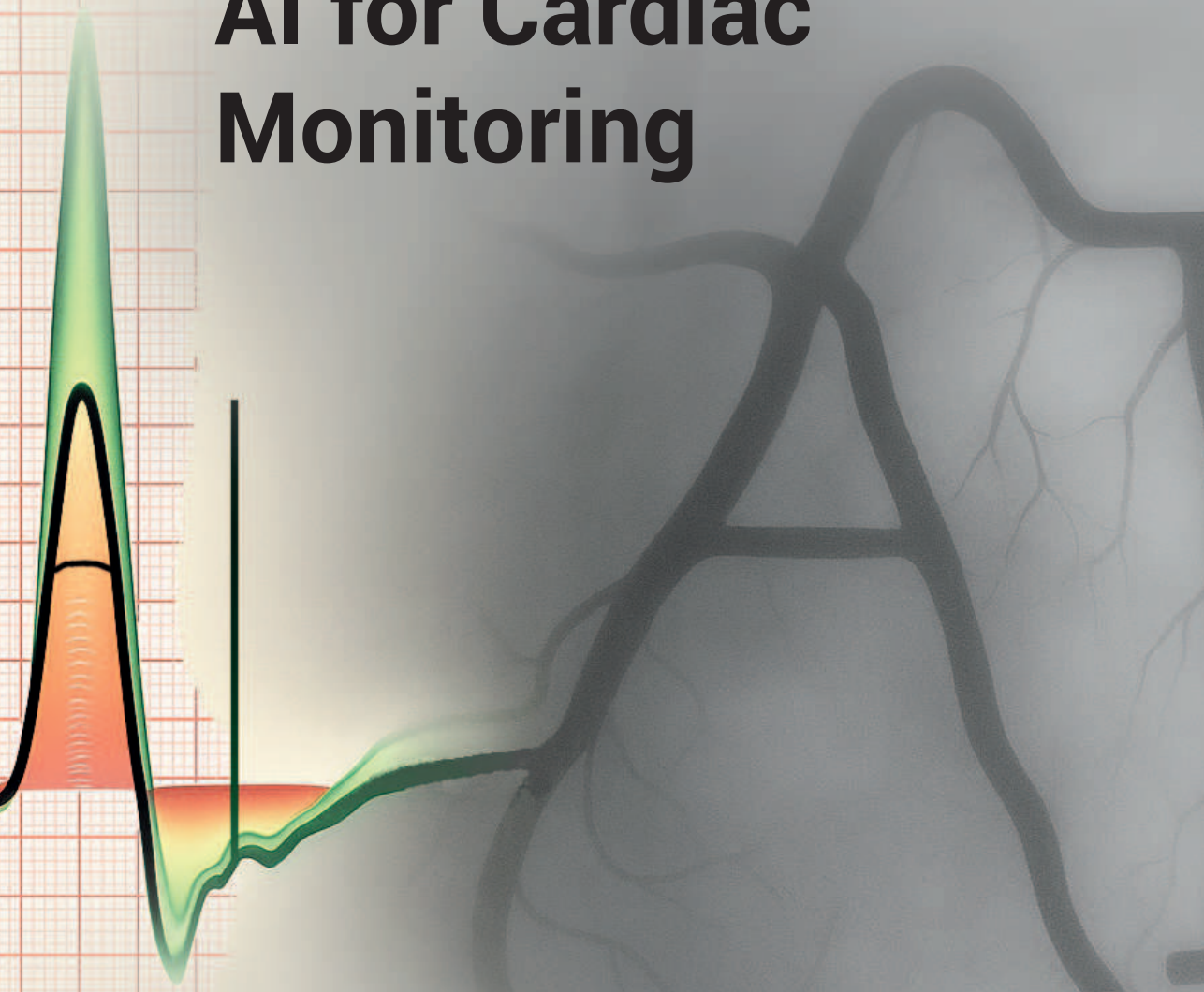
Explainable AI for Cardiac Monitoring



Viktor van der Valk

2026

Leiden University Medical Center

Stellingen behorend bij het proefschrift getiteld

Explainable AI for Cardiac Monitoring

Viktor van der Valk

1. Model comparison is complicated or near impossible if there is no uniformity in benchmarks, thresholds, metrics and datasets, which is often the case in the ICA deep learning field – This thesis (Chapter 2)
2. Smaller LLMs can be finetuned for specific tasks, but then require similar computational hardware as inference on larger generic LLMs – This thesis (Chapter 3)
3. Joint optimization is complex, especially when task difficulties are very different – This thesis (Chapter 4 and 5)
4. Promoting structure in a latent space can help prediction from this (sub)space, but promoting prediction from a (sub) latent space also indirectly structures it – This thesis (Chapter 4)
5. Explainable AI can provide a population context for individual ECG signals – This thesis (Chapter 5)
6. Non-explainable AI prediction models should never be the final goal in medical AI - Research field
7. Cherry picking prediction metrics can make many models seem good – Research field
8. Proper statistical evaluation is not common practice in the field of AI, but very much needed – Research field
9. The quality of the training data is more important than architecture search – Research field
10. It is unwise to be too sure of one's own wisdom (Mahatma Gandhi, 1910), especially if you have multiple academic titles (Viktor van der Valk)
11. Consciousness distinguishes humans from AI, not from animals – Viktor van der Valk
12. Everyone is just figuring it out (Iedereen doet maar wat) – Viktor van der Valk

Explainable AI for Cardiac Monitoring

Viktor van der Valk

2026

Colophon

About the cover:
The cover page was designed by Viktor and Esther van der Valk. On the front cover, you see a combination of two artificially modified medical cardiac images. These two modalities are both subject of the research covered in the thesis. On the left, you see an electrocardiogram slightly modified in shape so it shows the letter A and I. Around the black electrocardiogram graph, a red to green shade is shown, this is inspired by the explainable AI method employed in the thesis. On the right, you see a coronary angiogram in which the coronary arteries are shaped as well in the letter A and I. Claude, a visual language model (VLM) was used to merge and smoothen the final image.

Explainable AI for Cardiac Monitoring
Viktor van der Valk

Thesis layout & cover designed by Viktor and Esther van der Valk
Printed by ProefschriftenMaken

© 2026 Viktor van der Valk, Leiden, the Netherlands
All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the copyright holder.

Explainable AI for Cardiac Monitoring

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr. S. de Rijcke,
volgens besluit van het college voor promoties,
te verdedigen op donderdag 11 juni 2026
klokke 13.00 uur
door
Viktor Olaf van der Valk
geboren te 's-Gravenhage
in 1993

Promotoren: Prof.dr.ir. M. Staring
Prof.dr. D.E. Atsma
Copromotor: dr. R.W.C. Scherptong

Leden promotiecommissie:

Secretaris: Prof.dr.ir. B.P.F. Lelieveldt
Leden: Prof.dr.ir. M.S. Kleinsmann
Technische Universiteit Delft
dr. R.M.A. van der Boon
Erasmus Universiteit Rotterdam
dr.ir. N. Bruining
Erasmus Universiteit Rotterdam

The research in this thesis was performed at the Division of Image Processing (LKEB), Department of Radiology of Leiden University Medical Center, The Netherlands.
The research in this thesis was funded by the EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 860173 and by the European Institute of Innovation and Technology (EIT) Health under grant agreement 230193.

The printing of this thesis was funded by:
LKEB
Center for Human Drug Research (CHDR)
Universitaire Bibliotheken Leiden

Contents

1 General Introduction **1**

1.1 Acute coronary syndrome . . . 1
1.2 The acute coronary syndrome workflow . . . 1
1.2.1 Electrocardiography . . . 3
1.2.2 Invasive coronary angiography . . . 3
1.2.3 Echocardiography . . . 3
1.3 Problem statement . . . 3
1.3.1 Artificial Intelligence . . . 5
1.3.2 ICA . . . 5
1.3.3 ECG and echocardiography . . . 5
1.3.4 Clinical Reports . . . 6
1.4 Thesis overview . . . 7
1.4.1 ICA . . . 7
1.4.2 LLMs . . . 7
1.4.3 ECG . . . 7
1.5 Chapter overview . . . 8

2 The current state of deep learning based invasive coronary angiography image analysis: a systematic review **9**

2.1 Introduction . . . 11
2.2 Literature Search . . . 12
2.3 Results . . . 13
2.3.1 Frame selection . . . 13
2.3.2 Vessel segmentation . . . 13
2.3.3 Lesion localization . . . 17
2.3.4 Lesion assessment . . . 18
2.3.5 Prognosis prediction . . . 20
2.3.6 End-to-end pipelines . . . 20
2.3.7 Other tasks . . . 20
2.3.8 Commercially available software . . . 21
2.4 Discussion . . . 21
2.5 Acknowledgments . . . 22
2.6 Declaration of Generative AI and AI-assisted technologies in the writing process . . . 22
2.7 Appendix A . . . 23

2.8 Appendix B . . . 25

3 Large language models for cardiac electronic health record data classification 27
3.1 Introduction . . . 29
3.2 Methods . . . 30
3.3 Results . . . 35
3.4 Discussion . . . 42
3.5 Sources of Funding . . . 46
3.6 Disclosures . . . 46

4 Joint optimization of a β-VAE for ECG task-specific feature extraction 55
4.1 Introduction . . . 57
4.2 Methods . . . 58
4.2.1 Data . . . 58
4.2.2 Data preprocessing . . . 58
4.2.3 Model overview . . . 59
4.2.4 Model training . . . 60
4.2.5 Feature evaluation . . . 60
4.2.6 Baseline methods . . . 61
4.3 Experiments and Results . . . 61
4.3.1 Experiments . . . 61
4.3.2 Hyperparameter optimization . . . 61
4.3.3 Results on the test set . . . 63
4.4 Discussion . . . 64

5 Explainable ECG analysis by explicit information disentanglement with VAEs 69
5.1 Introduction . . . 70
5.1.1 Related work . . . 72
5.2 Methods . . . 73
5.2.1 Data . . . 73
5.2.2 Data preprocessing . . . 74
5.2.3 Model overview . . . 76
5.2.4 Training of the Partial β-VAE . . . 77
5.2.5 Feature visualization . . . 78
5.2.6 Survival analysis . . . 79
5.3 Experiments and Results . . . 79
5.3.1 Experiments . . . 79
5.3.2 Feature evaluation . . . 80
5.3.3 Results . . . 82
5.4 Discussion . . . 88
5.5 Conclusion . . . 93

6 Leveraging the wearable 1-lead ECG signal: From cardiac rhythm to cardiac function assessment **95**
6.1 Introduction . . . 97
6.2 Methods . . . 98
6.2.1 Data . . . 98
6.2.2 Data preprocessing . . . 99
6.2.3 Left ventricular function . . . 99
6.2.4 Model overview and training . . . 101
6.2.5 Feature visualization . . . 101
6.3 Experiments and Results . . . 102
6.3.1 Experiments . . . 102
6.3.2 Feature evaluation . . . 102
6.3.3 Results . . . 103
6.4 Discussion . . . 104

7 Summary and future work **113**

8 Appendices **125**

1

General Introduction

1.1 Acute coronary syndrome

Cardiovascular disease remains the leading cause of mortality worldwide, placing a large burden on healthcare systems [1]. Among these, acute coronary syndrome (ACS) poses a particularly high risk to patient survival and long-term cardiac function. ACS results from an (abrupt) reduction in coronary blood flow, most commonly due to atherosclerotic plaque rupture and subsequent thrombosis, leading to ischemia and necrosis of myocardial tissue, see Figure 1.1. The resulting ischemia is potentially life threatening and can lead to cardiac complications such as cardiac arrest if not promptly diagnosed and treated.

Even after hospital discharge, ACS survivors face a substantial risk of adverse outcomes, see Figure 1.1. Community-level studies report a one-year case-fatality rate of approximately 12% following ACS, including myocardial infarction (MI) and unstable angina [2]. In addition to mortality, patients remain at heightened risk for complications such as heart failure due to impaired ventricular function, recurrent ischemic events, and arrhythmias. These complications not only impact quality of life but also contribute to frequent hospital readmissions and long-term healthcare dependency. This clinical burden underscores the importance of early risk stratification, targeted intervention, and longitudinal monitoring in ACS care pathways.

1.2 The acute coronary syndrome workflow

Standard clinical practice emphasizes rapid risk stratification, using among others biomarkers, coronary angiography, and electrocardiography (ECG) data, to guide revascularization strategies and pharmacotherapy, thereby reducing infarct size, preserving cardiac function, and improving long-term prognosis [3, 4].

Following an acute coronary event, patients typically undergo longitudinal monitoring for up to a year, aimed at preventing unexpected complications and hospital readmissions.

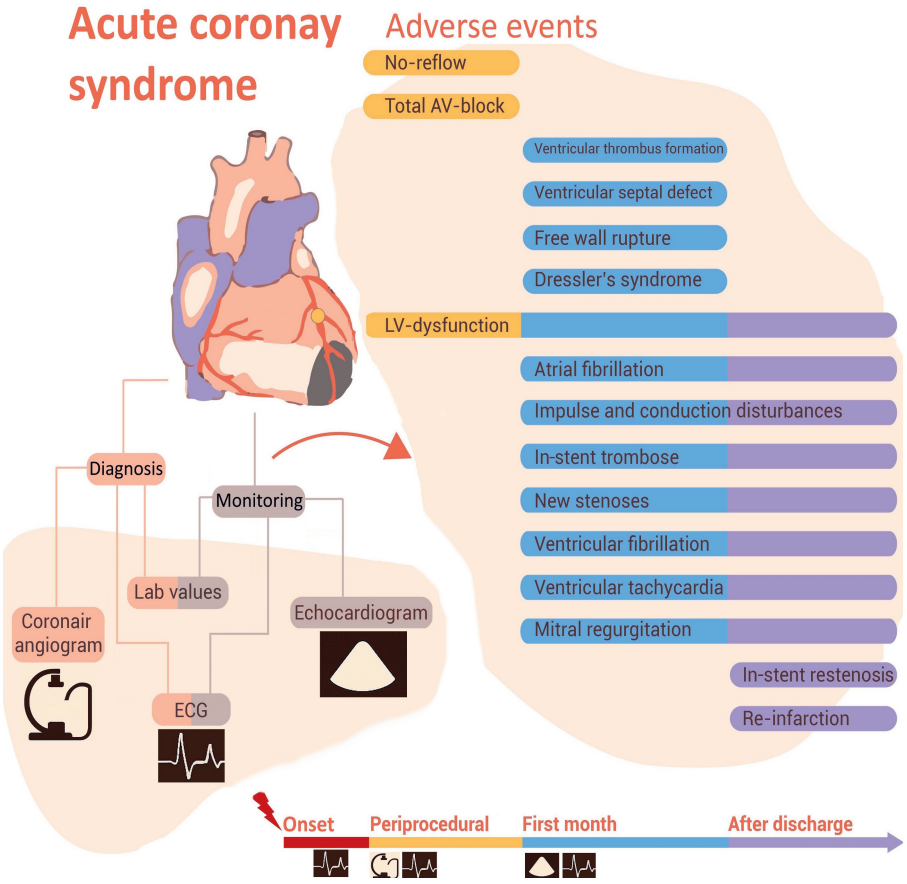


Figure 1.1: Schematic overview of the cardiac assessments and adverse outcomes in acute coronary syndrome. The image of the heart (left) depicts an ischemic lesion which causes a reduced coronary blood flow to the heart muscle downstream. Cardiac assessments for diagnosis (coronary angiography, ECG and lab-values) are used to determine lesion location and severity; cardiac assessments for monitoring (echocardiogram, ECG and lab-values) are used to determine recovery and potential adverse events. Various adverse events, shown on the right, can occur on different timescales from the trauma.

At each follow-up visit, cardiac function is assessed using among other echocardiography and ECG, with particular attention to left ventricular function (LVF), the occurrence of arrhythmias, and other indicators of adverse cardiac remodeling, see Figure 1.1. These evaluations are critical for timely intervention and for guiding long-term management strategies.

1.2.1 Electrocardiography

ECG is an important cardiac assessment technique and provides a rapid, non-invasive recording of the heart's electrical activity through voltage measurements across multiple electrodes placed on the body, see Figure 1.2. Standard 12-lead ECGs are normally interpreted by their spatiotemporal waveforms, P waves, QRS complexes, T waves, which reflect atrial depolarization, ventricular depolarization, and repolarization, respectively. Deviations in waveform morphology or timing can indicate ischemia, hypertrophy, or arrhythmias, but also decline or improvement after cardiac interventions. A Holter ECG is an ambulant version of the ECG which is used mainly for rhythm monitoring during an extended period of time (24-48 hours) with 3-8 electrodes. The 1-lead wearable ECG, as a proxy of lead 1 of the 12-lead ECG, was shown to accurately predict several arrhythmias and has the potential to estimate other important markers for cardiac disease [5, 6, 7].

1.2.2 Invasive coronary angiography

Invasive coronary angiography (ICA) is the gold-standard imaging modality for delineating coronary arteries and quantifying stenotic lesions, see Figure 1.2 [8]. By introducing a radiopaque contrast agent via catheterization, ICA yields X-ray sequences revealing lumen narrowing, vessel tortuosity, and collateral circulation. In the acute setting expert cardiologists visually inspect these sequences to locate the ischemic lesion and guide percutaneous intervention (PCI). In the non-acute setting (multiple of) these sequences are use to determine lesion significance and plan PCI or coronary artery bypass grafting (CABG).

1.2.3 Echocardiography

Echocardiography is another cornerstone of cardiac imaging, offering non-invasive assessment of heart structure and function using ultrasound waves, see Figure 1.2. Transthoracic echocardiography (TTE), the most common modality, provides dynamic two-dimensional views of cardiac chambers, valves, wall motion, and pericardial space. Doppler imaging further enables evaluation of blood flow velocities and pressure gradients, aiding in the detection of valvular stenosis or regurgitation, intracardiac shunts, and ventricular dysfunction. Echocardiography is safe and readily available at the bedside, making it ideal for longitudinal monitoring of cardiac function, particularly LVF, a critical predictor of prognosis following ACS [9].

1.3 Problem statement

Despite advances in diagnostics and treatment, the clinical workflow following ACS remains complex, resource intensive, and not always patient tailored. Patients typically require frequent follow-up visits that involve repeated imaging, electrocardiography, and biomarker analysis to monitor cardiac function and detect adverse events. This multidisciplinary process demands significant time from clinicians and places a high administrative and financial

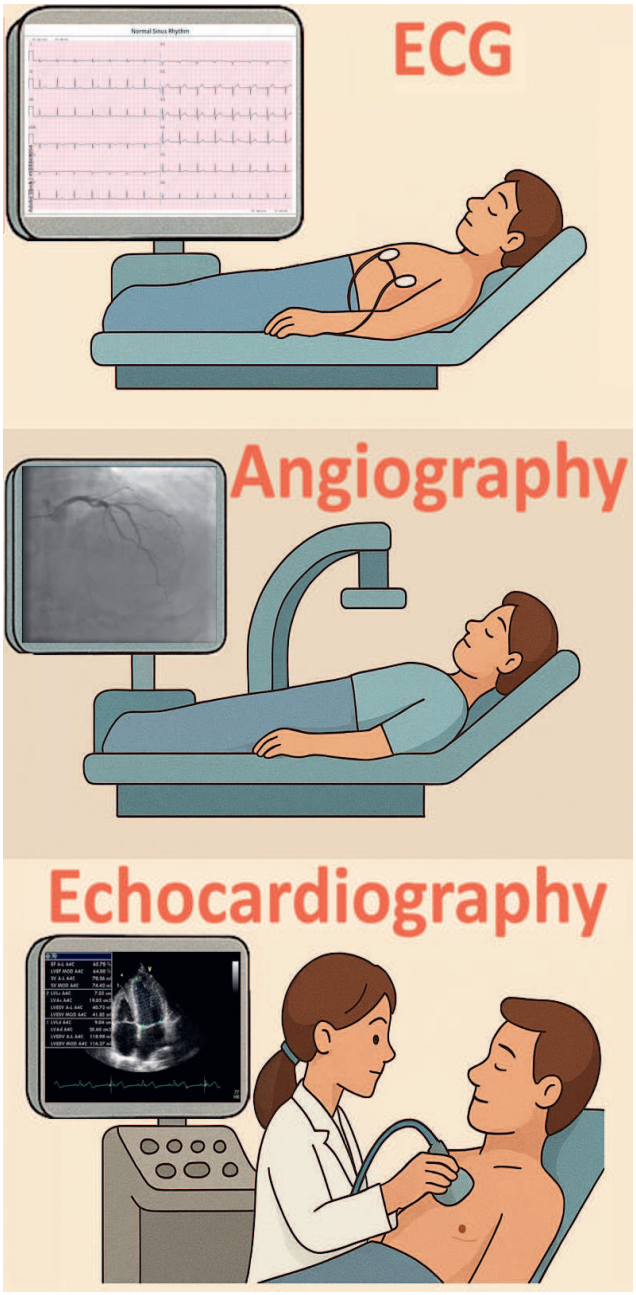


Figure 1.2: The three main imaging modalities used in ACS diagnosis and monitoring: electrocardiography (ECG), coronary angiography, and echocardiography.

burden on healthcare systems. With an aging population and an increasing prevalence of cardiovascular disease, the scalability of this traditional workflow is under pressure. On the other hand, improved patient tailoring could reduce both unnecessary and unplanned clinical visits by adjusting monitoring to patient needs.

The ACS care pathway, as it exists today, is laborious and, without innovation and improved patient tailoring, will become increasingly difficult and costly to sustain in the coming years.

Deep learning has the potential to improve the workflow on several aspects and help tailor it to individual patient needs.

1.3.1 Artificial Intelligence

AI is increasingly transforming medical data analysis by enabling automated interpretation of complex datasets such as medical images, physiological signals, and clinical text. Deep learning methods, particularly convolutional neural networks (CNNs), have shown remarkable success in image-based tasks like lesion detection and segmentation. More recently, transformer architectures and large language models (LLMs) have expanded AI's capabilities to sequential and textual data, supporting applications such as text translation, analysis, and summarization. However, many of these models operate as black boxes, making their decision processes difficult to interpret, especially for image analysis. This has driven the development of explainability techniques to improve transparency and trust in clinical settings. Despite these advances, challenges remain: models can suffer from bias due to imbalanced datasets and overfitting, when a model learns patterns specific to the training data rather than generalizable features, due to small datasets or incorrect training. This makes proper statistical testing of vital importance for these models.

1.3.2 ICA

In ICA, manual sequence review leads to heterogeneous results and can require additional time-consuming measurements. Software tools such as quantitative coronary analysis (QCA) have been developed to aid this process, and more recent deep learning-based models show promising results in terms of automated, accurate ICA analysis.

However, the ICA analysis pipeline is complex and consists of multiple interdependent steps that all require automation to enable efficient, standardized, and ideally real-time interpretation, see Figure 1.3. These steps include: i. the selection of diagnostically relevant, contrast-filled frames from the ICA sequence, ii. segmentation, and ideally also anatomical identification of the coronary vessels, iii. (multiple) lesion localization and iv.) quantitative assessment of lesion severity. Real-time integration of these steps into visualization tools could support clinicians during catheterization procedures and improve diagnostic consistency.

1.3.3 ECG and echocardiography

ECG offers a low-cost, and widely accessible cardiac assessment that can be used in both clinical and remote settings. The ECG contains information that may be predictive of key clinical markers typically derived from echocardiography, such as LVF. Using the ECG

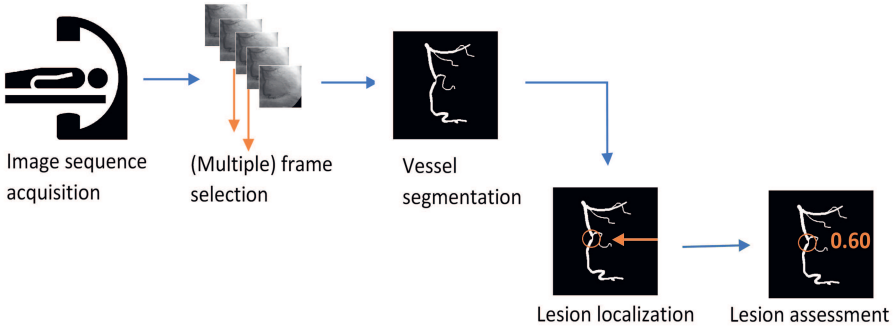


Figure 1.3: Steps in a computational diagnostic ICA workflow.

instead of the echocardiogram for cardiac monitoring where possible, has clinical advantages. Interpretation of echocardiographic findings requires extensive training and remains labor intensive, which can limit scalability in routine care. Furthermore, echocardiography must be performed in a clinical setting, often requiring specialized equipment and trained personnel.

With the increasing availability of portable and even 1-lead ECG devices, there is growing interest in leveraging ECG data for at-home monitoring. However, deep learning models applied to ECG interpretation often lack transparency, especially when used for estimating complex cardiac parameters such as LVF. Explainability remains a major challenge, as clinical adoption of AI-driven ECG analysis depends not only on predictive accuracy but also on the interpretability of model outputs.

1.3.4 Clinical Reports

In order to create deep learning models that can aid in the cardiac workflow, proper training datasets have to be created. Such datasets are ideally large and accurately labeled cardiac data. Labeling of cardiac data is non-trivial and often time consuming. Thereby, for most cardiac imaging and ECG data descriptions are available in accompanying clinical reports. Manual extraction remains the common approach but is labor-intensive and prone to error, underscoring the need for robust, automated solutions that can facilitate efficient, accurate, and scalable clinical data structuring [10]. Echocardiography and ICA reports often describe critical diagnostic and procedural information, such as the presence of coronary lesions, previous cardiac events, ejection fraction, types of cardiomyopathy, and treatment strategies. These reports hold significant clinical and research value but are typically unstructured or semi-structured, limiting their usability for statistical reporting, cohort selection, quantitative analysis, or predictive modeling.

Unlocking this valuable information at scale requires automated methods for extracting structured clinical variables from free-text reports. Large language models (LLMs) have recently emerged as promising tools to enable such extraction, offering flexibility and adaptability across diverse report styles and languages. However, the comparative performance, scalability, and practical feasibility of different LLM approaches, ranging from large commercial models to smaller locally deployed and fine-tuned models, remain underexplored

in cardiology.

1.4 Thesis overview

1.4.1 ICA

Deep learning methods, particularly convolutional neural networks (CNNs), have demonstrated promising performance across various steps in the ICA analysis pipeline, including frame selection, vessel segmentation, and lesion localization and assessment. These developments highlight the potential for automating and standardizing ICA interpretation.

However, research in this domain is currently fragmented, not standardized, and lacks a clear, comprehensive overview. In Chapter 2, we address this gap by presenting a systematic review of deep learning–based ICA analysis. We categorize over sixty studies into four core tasks: frame selection, vessel segmentation, lesion localization, and lesion assessment. Thereby, we provide a critical overview of methodological choices, common challenges, and opportunities for future research to support better alignment and comparability in the field.

1.4.2 LLMs

To enable large-scale reuse of clinical data for research and model development, we explored the use of LLMs to automatically extract structured clinical variables from free-text ICA and echocardiography reports. In Chapter 3, we systematically evaluate both prompt-based querying and fine-tuning approaches across a range of LLMs, spanning large commercial models and smaller locally hosted alternatives. We benchmark these methods on key diagnostic descriptors such as LVF, culprit vessel, and occlusion status, demonstrating that high extraction performance is achievable without manual annotation.

Additionally, we compare the feasibility of deploying these models across different computational infrastructures, including local machines, high-performance computing clusters, and cloud-based platforms. This analysis highlights trade-offs between accuracy, scalability, data governance, and accessibility, factors particularly relevant for medical research groups with limited compute or strict data-sharing constraints.

1.4.3 ECG

To address the challenge of deriving echocardiographic markers such as LVF from more accessible and scalable data sources, we investigated explainable deep learning techniques for automated ECG interpretation. In Chapter 4, we propose a joint training strategy for β-variational autoencoders (VAEs) that combines ECG signal reconstruction with supervised prediction of LVF. This approach encourages the model to encode clinically meaningful features in the latent space while maintaining interpretability through structured disentanglement.

Building on this, Chapter 5 introduces a per-signal supervised disentanglement method that partitions the VAE latent space into clinically relevant and irrelevant components. By explicitly promoting separation between LVF-associated and unrelated latent factors, this technique allows for intuitive visualization of latent manipulations and supports clinician understanding of the model's reasoning. These contributions aim to advance the transparency and trustworthiness of ECG-based diagnostic tools, which is essential for safe

integration into clinical decision-making.

Finally, Chapter 6 explores the adaptation of the disentangled, explainable VAE framework for LVF prediction from 1-lead smart-watch ECGs, a format more suited to out-of-hospital, patient-collected data. While 1-lead ECGs present a more scalable and accessible diagnostic modality, they introduce unique challenges, such as signal orientation errors due to inconsistent watch placement. This chapter extends the applicability of explainable deep learning techniques to ambulatory settings, promoting broader adoption of trustworthy AI tools in longitudinal cardiac monitoring.

1.5 Chapter overview

In summary, this thesis comprises four interconnected studies that advance AI-driven cardiovascular diagnostics:

- Chapter 2 systematically reviews deep learning methods for ICA, categorizing approaches to frame selection, vessel segmentation, lesion localization, and assessment, and proposing guidelines for methodological consistency.
- Chapter 3 evaluates LLM-based classification of cardiac semi-structured reports, contrasting prompt engineering and fine-tuning on echocardiography and ICA text, across different computational infrastructures, to extract structured clinical descriptors.
- Chapter 4 proposes a β-VAE framework jointly trained for high-fidelity ECG reconstruction and task-specific feature extraction, improving predictive performance for cardiac dysfunction.
- Chapter 5 extents and improves the VAE proposed in Chapter 4. It introduces a supervised disentanglement technique within the VAE latent space, isolating dimensions linked to LVF for interpretable per signal ECG feature visualization and robust classification.
- Chapter 6 explores the challenges of using patient-collected ECG data for meaningful predictions with the explainable AI pipeline introduced in Chapter 5.

Together, these contributions demonstrate how imaging, textual, and signal modalities can be analyzed with a consistent focus on explainability, paving the way for standardized, transparent, and clinically adaptable AI solutions in cardiology. By targeting key steps across the MI workflow, from diagnosis to follow-up monitoring and retrospective data curation, this thesis contributes toward partially automating labor-intensive tasks. The ultimate goal is to reduce clinical workload and associated costs while enhancing decision support and model interpretability in cardiac care.

2

The current state of deep learning based invasive coronary angiography image analysis: a systematic review

Van der Valk, V.O., Van der Loo, W., Atsma, D.E., Hjortnaes, J., Staring, M., Scherptong, R.W.C., (2025). The current state of deep learning based invasive coronary angiography image analysis: a systematic review. (in progress)

Abstract

AI, and deep learning in particular, have been making their way into invasive coronary angiography (ICA). Analysis of the ICA sequence poses its own set of domain specific problems, when compared with other imaging modalities, but also in ICA, AI has the potential to partially automate and improve analysis. The papers published in the field, however, have a large variety in objectives and evaluations, which hampers equal comparison, and an overview is currently lacking. This chapter therefore critically reviews the AI applications in the ICA analysis field and provides an overview of the choices made in tasks, ground truths, metrics and datasets as well as on how other domain specific problems are addressed. The studies are categorized into the 4 tasks that were found in ICA analysis: frame selection, vessel segmentation, lesion localization and lesion assessment. Future directions and challenges in the research field are also discussed.

2.1 Introduction

Coronary artery disease (CAD) is one of the leading causes of death worldwide, and the incidence is still increasing [1]. Invasive coronary angiography (ICA) is the gold standard for the diagnosis of CAD as well as for intra-procedural guidance of percutaneous coronary interventions (PCI) [8]. Visual assessment plays a central role both in ICA and PCI. This has several important drawbacks: i) a significant inter- and intra-observer variability, ii) a tendency to over- or underestimate stenosis severity and iii) the inability to diagnose syndromes that are characterized by signs of myocardial ischaemia but without significant focal or culprit lesions [11, 12]. These syndromes include coronary microvascular disease (CMD), myocardial infarction or ischemia with non-obstructed coronary arteries (MINOCA/INOCA) and generalized atherosclerosis without focal lesions [13, 12].

If visual assessment on ICA images is uncertain, quantitative coronary angiography (QCA) could give more clarity. Most of the clinical software packages for coronary angiography have implemented a QCA method, which is used in the clinical setting and in research as a method for structured stenosis assessment. However, QCA still relies on manual initialization and prolongs procedural time. Other intracoronary assessments, such as fractional flow reserve (FFR), instant wave-free ratio (iFR), optical coherence tomography (OCT) or intravascular ultrasound (IVUS) are also used to optimize invasive coronary diagnosis. However, these types of assessment prolong procedural time, increase complication risks and radiation exposure. In addition, these tools are not able to diagnose all of the above-mentioned syndromes [14, 15, 11, 16].

Deep learning algorithms have the potential to facilitate the ICA workflow by improving diagnosis and reducing bias and variability. Moreover, these algorithms can be trained to perform image analysis tasks normally done by cardiologists, which could lead to automation and a decrease in procedure times. Deep learning has been used in medical image analysis in various modalities and for various tasks [17]. MRI and CT image analysis constitute the bulk of the deep learning research performed in medical image analysis, but deep learning is slowly making its way into ICA analysis as well. Segmentation, the process of assigning image pixels to a class (e.g. vessel, background), and classification, the process of assigning a class label (e.g. stenosis or not) to a whole image or a part of it, represent the majority of the research performed in all modalities including ICA [18]. However, the variability in study designs, evaluation methods and clinical relevance is substantial. Therefore, this chapter critically reviews the available literature, thereby providing a framework for future studies. Several reviews exist about deep learning in a wide variety of imaging modalities [19, 18, 20, 17]. Two of these [21, 22] cover deep learning in ICA image analysis, however both reviews only cover a small subset (< 20) of the papers in the field.

The review will be structured according to the clinical workflow, as shown in Figure 2.1. This workflow can be divided into several steps that can all be automated with computational tools. These steps include i) the selection of the contrast-filled frame(s) at the end of the diastolic phase from the ICA sequence, followed by ii) segmentation of the vessels and iii) the localization and iv) assessment of the lesion on these frames. In Appendix 2.7 a brief introduction to Deep Learning and the algorithms that are discussed, can be found. A more extensive introduction to Deep Learning in medical image analysis can be found in

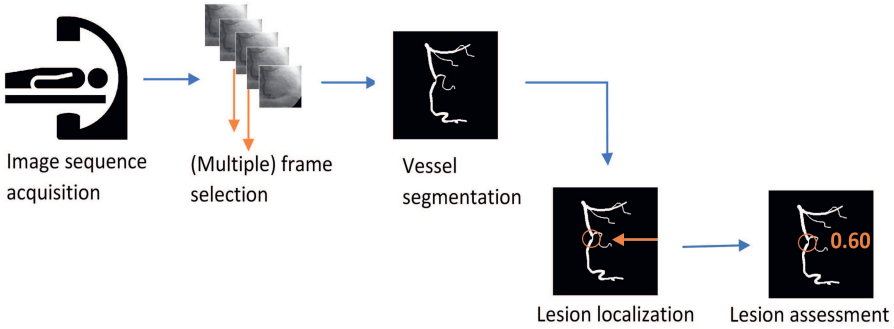


Figure 2.1: Steps in a computational diagnostic ICA workflow.

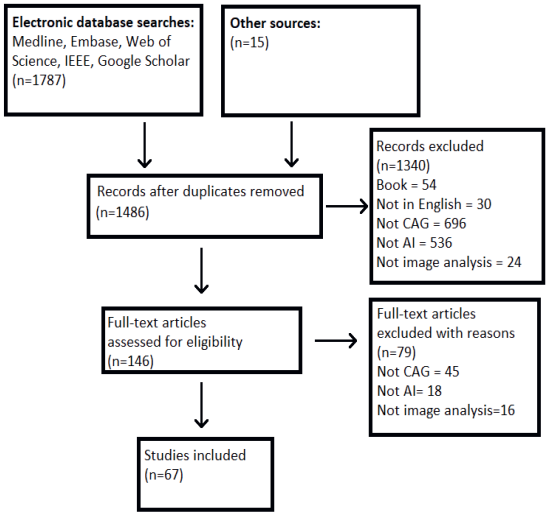
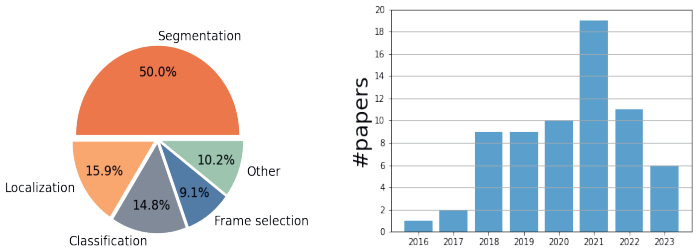


Figure 2.2: The paper inclusion process

a review by Litjens *et al.* [17].

2.2 Literature Search

Eligible studies were identified by searching the following electronic databases: Medline, EMBASE, Web of Science, IEEE, and Google Scholar. Additional literature was identified by checking references of included literature and by a search on ArXiv. The final search was performed on June 1st, 2023. The details of the search process are shown in Figure 2.2 and 2.3, search strategies for each database can be found in 2.8.



(a) The distribution of selected deep learning papers over the tasks.

(b) The number of selected deep learning papers per year.

Figure 2.3: Statistics about the papers selected for this review.

2.3 Results

2.3.1 Frame selection

ICA typically produces multiple sequences comprising an arbitrary number of frames, a subset of which depict vessels filled with contrast. However, for the optimal assessment of stenosis, frames with a sufficient amount of contrast medium within the vessel are crucial. Therefore, (pre)selection of frames from the ICA sequences is necessary, reducing the input data for further processing, unless an algorithm is designed to directly analyze a sequence of arbitrary length. Two main methods are used for automatic frame selection: i) *individual frame scoring*, whereby the best scoring frame is selected, and ii) *whole sequence analysis*, where the entire sequence is analyzed simultaneously.

Individual frame scoring can be done either by selecting the frame with the largest vascular area via segmentation [23, 24] or by employing a CNN to directly predict a score [21, 25].

Three studies have employed the *whole sequence analysis* approach with combinations of an LSTM and a CNN [26, 27] or a Transformer [28].

2.3.2 Vessel segmentation

Segmentation can be seen as a (binary) classification problem, see Figure 2.4. Therefore, segmentation outcomes are often evaluated with classification metrics. Accuracy is one of the most used metrics and reflects the total fraction of correctly classified inputs. Precision and recall reflect respectively the fraction of positive predictions that is correct and the fraction of positive inputs that is classified as positive. The Dice score reflects the harmonic mean between the precision and recall.

Deep learning methods in ICA vessel segmentation can roughly be divided into two categories, namely, those that *predict labels for every patch of an image*, and those that use an *encoder-decoder* design. In Table 2.1 the results of all included segmentation studies are summarized, including the amount of data used, the method category, the task (major or full vessel tree segmentation), the metrics used and, when available from the paper, a comparison with the U-Net.Table 2.2 presents the results of vessel segmentation studies that were performed on the same dataset, enabling a fair comparison. This dataset

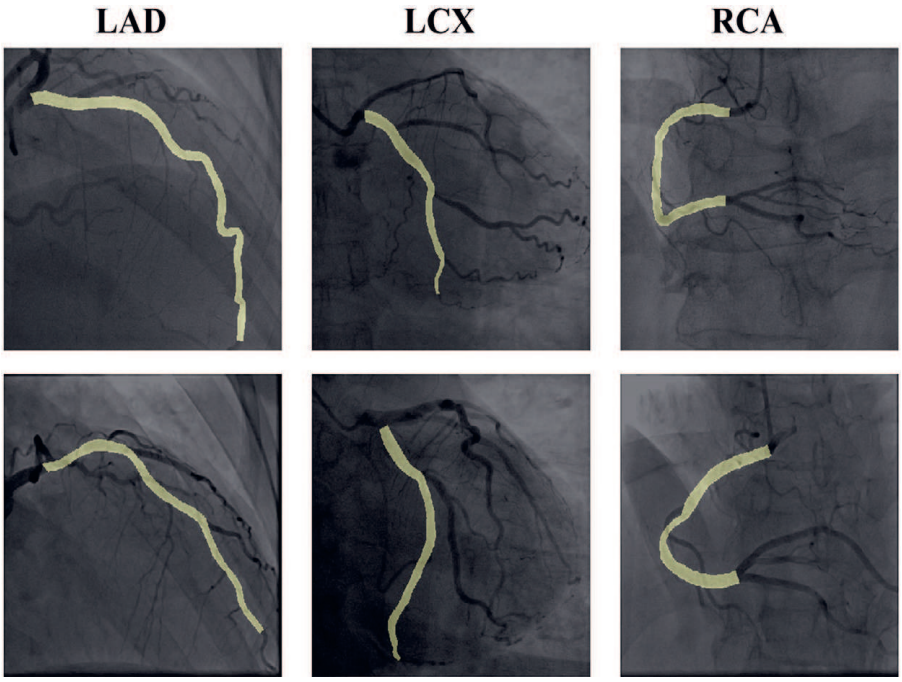


Figure 2.4: Example of major vessel segmentation [29].

(DCA1) is the only open-source data set (`http://personal.cimat.mx:8181/~ivan.cruz/DB_Angiograms.html`), and contains 134 X-ray coronary angiograms segmented by an expert cardiologist.

In *patch-based segmentation*, an ICA image is split into smaller patches, of which the center pixel is classified with a CNN. The patch or window is then slid over the image to obtain a dense segmentation. Patch-based segmentation has been performed as a standalone method [30, 31, 32], but also combined with post processing steps such as the regional parameter expansion and optimal cover tree algorithms [33]. As a standalone method, improvements have been shown by using different patch scales to combine coarse and fine segmentation [31, 32] as well as by using multichannel CNNs with contrast and non-contrast filled frames as input [32], see Table 2.2.

U-Net and its variations are popular *encoder-decoder segmentation* networks used in medical imaging studies. Researchers have implemented the U-Net or U-Net++ on single frames while experimenting with different backbones such as ResNet, ResAttNet [34] or other residual networks [35], DenseNets [36], EfficientNets [37], EfficientUNet++ [38], Inception networks [39], and other multiscale networks [40]. Results for each of these methods are shown in Table 2.1. Silva *et al.* compared U-Nets with combinations of encoders and decoders and concluded that an EfficientUNet++ outperformed combinations of amongst others RegNet, ResNet, ResNeXt, DeeplabV3+ [41], PSPNet [42], UNet and UNet++ in a vessel and catheter segmentation task.

Several studies have explored the use of additional data processing steps (see Ap-

pendix 2.7 for explanation of AI terms) such as pre-processing images with a CNN to enhance image contrast and sharpen boundaries [41]. Multi-frame input has also been investigated, with researchers demonstrating improvements in performance when using multiple frames as input for an accordingly adjusted network [43, 44, 45, 46].

Other architectures and modifications to the U-Net resulted in performance improvements, such as sharing feature maps between multiple encoder and decoder layers [29, 47], combining multiple output feature maps of the encoder [48] or decoder layers [49, 50, 51], deep supervision [47], spatial pyramid pooling [42], atrous spatial pyramid pooling [41, 52], deformable [53] and vessel-specific convolutions [54], squeeze and excitation blocks [55, 38], the use of patch, channel and spatial attention modules [56, 38, 55, 53] and other smart filter and feature map selection mechanisms [50, 57]. Improvements where also found for other major architecture extensions such as creating a generative adversarial network (GAN) by attaching a discriminator to a residual U-Net [35] or by attaching another encoder-decoder network to a U-Net [58], as well as by embedding 2 U-Net++ models in a Mean teacher model [59] and by combining 6 encoder-decoder networks in a bigger U-shaped network [60]. Most of these improvements were achieved by preventing segmentation breakage, suppression of background noise and better segmentation of small vessels.

Transfer learning has been employed to address the limited availability of ICA images. Most studies used a pre-trained network as an encoder backbone, while others experimented with alternative forms of transfer learning. One such approach is to train a GAN using a combination of labeled fundus images and ICA images segmented with the Frangi algorithm [61]. This method has shown improvements in Dice scores over both the Frangi algorithm and a U-Net trained with the fused dataset. Others have considered the use of optical flow to reduce the amount of annotated data needed for vessel segmentation, showing that it can aid in extracting the necessary information for accurate segmentation [62].

Table 2.1: **Vessel segmentation studies**. Columns show the variety in tasks (full or major vessel tree segmentation), metrics (IoU, accuracy or Dice) and data availability. The last column indicates the performance of a U-net as implemented by the authors. Acc = Accuracy, ED = encoder-decoder, ims = images, seq = sequences,- = information not available.

Paper	Data	Type	Vessel tree	Metric	Proposed	U-Net
Wang *et al.* [44]	8835 ims; 170 seq	ED	Full	IoU	0.818	-
Jun *et al.* [29]	4700 ims; - seq	ED	Major	Dice	0.890	0.731
Yang *et al.* [63]	3302 ims; - seq	ED	Major	Dice	0.917	0.871
Xian *et al.* [34]	3200 ims; - seq	ED	Major	Dice	0.921	0.865
Jo *et al.* [57]	1987 ims; - seq	ED	Major	Dice	0.676	0.565
Zhang *et al.* [60]	1893 ims; - seq	ED	Full	Dice	0.916	-
Au *et al.* [52]	1024 patients	ED	Major	Dice	0.704	0.704
Meng *et al.* [47]	616 patients; 210 seq	Full	Dice	0.894	0.880	
Li *et al.* [50]	538 ims; - seq	ED	Full	Dice	0.904	0.882
Iyer *et al.* [41]	462 ims; - seq	ED	Full	Dice	0.864	0.717
Silva *et al.* [38]	270 ims; - seq	ED	Full	Dice	0.8904	-
Hao *et al.* [45]	323 ims; 120 seq	ED	Full	Dice	0.843	-
Zhao *et al.* [51]	314 ims; - seq	ED	Full	Dice	0.890	0.865
Song *et al.* [48]	289 ims; 289 seq	ED	Full	Dice	0.773	0.759
Wan *et al.* [46]	186 patients	ED	Full	Dice	0.90	0.88
Yin *et al.* [33]	173 ims; - seq	Patch	Full	Dice	0.83	-
Mulay *et al.* [64]	164 ims; - seq	ED	Full	Dice	0.717	0.757
Hamdi *et al.* [56]	150 ims; - seq	ED	Full	Dice	0.829	0.743
Hamdi *et al.* [35]	150 ims; - seq	ED (GAN)	Full	Dice	0.8118	0.8012
Alom *et al.* [65]	150 ims; - seq	ED	Full	Dice	0.754	0.743
Mou *et al.* [66]	150 ims; - seq	ED	Full	Dice	0.807	0.743
Fan *et al.*[43]	148 ims; - seq	ED	Full	Dice	0.866	0.760
Gao *et al.* [53]	138 ims; - seq	ED	Full	Dice	0.8074	0.7759
Jiang *et al.* [40]	134 ims; - seq	ED	Full	Dice	0.791	0.757
Cervantes-Sanchez *et al.* [67]	130 ims; - seq	Pre-learned	Full	Dice	0.686	0.552
Zhou *et al.* [21]	120 ims; - seq	ED	Full	Dice	0.876	0.857
Menezes *et al.* [37]	117 ims; - seq	ED	Full	Dice	0.948	-
Zhao *et al.* [49]	- ims; 60 seq	ED	Full	Dice	0.88	0.87
Yang *et al.* [32]	60 ims; 60 seq	Patch	Full	Dice	0.801	-
He *et al.* [59]	58+60 ims; - seq	ED	Full	Dice	0.795	0.773
Nasr-Esfahani *et al.* [31]	44 ims; - seq	Patch	Full	Dice	0.815	-
Yu *et al.* [61]	40+1092 ims; - seq	ED (GAN)	Full	Dice	0.824	0.742
Samuel *et al.* [54]	160 ims; - seq	ED	Full	Acc	0.970	0.966
Zhu *et al.* [42]	109 patients	ED	Full	Acc	0.957	0.915
Cruz-Aceves *et al.* [68]	100 ims; - seq	Pre-learned	Full	Acc	0.957	-
Nasr-Esfahani *et al.* [30]	44 ims; - seq	Patch	Full	Acc	0.935	-
Busto *et al.* [69]	30 ims; - seq	ED	Full	Dice	-	0.94

Table 2.2: **Segmentation studies comparison**: All results are obtained on the DCA1 data set. Multiple scores indicate results of reimplementations of the same method done in different studies.
* = Jiang *et al.*, ** = Tao *et al.*, † = Samuel *et al.*, ‡ = Cervantes-Sanchez *et al.* [40, 54, 67]

Method	Accuracy	Dice
Jiang *et al.* [40]	**0.977***	**0.791***
Samuel *et al.* [54]	0.975*/0.970†	0.774*
Hao *et al.* [45]	0.974*	0.769*
original U-Net [70]	0.976*/0.9707***/0.966†/0.965‡	0.757*/0.552‡
Mulay *et al.* [64]	0.966*	0.717*
FCN [71]	0.9640**	-
SegNet [72]	0.9702**	-
PSPNet [73]	0.9666**	-
Tao *et al* [55]	0.9729**	-

2.3.3 Lesion localization

Lesion localization is a crucial aspect of medical imaging, as it helps determine the appropriate treatment for patients. Various algorithms from the broader field of object detection and localization have been adapted to address this challenge. The intersection over union (IoU) metric is a common method for evaluating lesion localization. It measures the overlap between the predicted and actual locations of the lesion. Localization can also be assessed together with lesion evaluation by setting a minimum IoU value for a correct classification.

Luque-Baena[74] showed that a ResNet50 network outperformed EfficientNetB4 and MobileNet on a 32x32 patch based stenosis detection task. Several studies have employed the Mask R-CNN and Faster R-CNN techniques, of which the latter also creates a segmentation mask as part of the pipeline, for lesion localization, utilizing different backbone networks to improve performance [75, 76, 77, 78, 79]. Combining feature maps from multiple layers and using ensembles of various networks, including the Transformer networks, have been shown to further enhance the accuracy of localization tasks [76, 77, 80, 81]. Joint segmentation and quantification of coronary vessels was also successfully done by Zhang *et al.* [60].

Table 2.3: **Comparison of generic image and video object detection algorithms with tailored stenosis detection methods below the line**. Table from [81]

Method	Multi-frame	Precision	Sensitivity	Dice
Cascade R-CNN [82]	no	0.897	0.794	0.843
Faster R-CNN [83]	no	0.874	0.751	0.808
RetinaNet [84]	no	0.783	0.698	0.738
FoveaBox [85]	no	0.821	0.806	0.816
DETR [86]	no	0.168	0.103	0.127
Deformable DETR [87]	no	0.779	0.855	0.815
FGFA [88]	yes	0.898	0.740	0.811
DFF [89]	yes	0.864	0.613	0.716
SELSA [90]	yes	0.861	0.826	0.842
TRA [91]	yes	0.858	0.797	0.825
TransVOD [80]	yes	0.840	0.846	0.843
Zhao *et al.* [51]	no	0.229	0.799	0.356
Ovalle-Magallanes *et al.* [92]	no	0.407	0.770	0.532
Han *et al.* [81]	yes	0.923	0.896	**0.909**
Wu *et al.* [23]	yes	0.783	0.764	0.773
Pang *et al.* [93]	yes	0.870	0.879	0.875

Some researchers have leveraged multiple frames to extract additional location and classification information, processing these frames in parallel and examining them for temporal consistency to reduce false positives [23, 93, 81]. This approach has led to improvements in localization accuracy. Table 2.3 shows a comparison of several of these networks with generic image and video object detection algorithms in a lesion detection task.

Visualization techniques such as GradCam have been explored for stenosis localization in various studies, showing the potential of combining different aspects of analysis, classification and localization, to achieve more accurate results [27, 24].

2.3.4 Lesion assessment

Determining whether a stenosis is severe enough for invasive treatment is necessary in the clinical workflow. This binary outcome transforms stenosis assessment into a classification problem, with a clinically relevant cut-off value between 60-80% stenosis percentage depending on the clinical context. Intravascular imaging or invasive intracoronary pressure measurement is another type of lesion assessment. Several techniques are available for this purpose that all have specific cut-offs indicating significance of a lesion. For the evaluation of these classification problems, the metrics mentioned in the segmentation section are often used. However, in literature there is a great variety not only in these metrics, but also in tasks chosen to optimize. These tasks range from 25% to 100% stenosis (also known as chronic total occlusion (CTO)) classification tasks to stenoses significance regression tasks, both on either manually annotated or with a QCA or FFR ground truth, and on single or multiple frames, see Table 2.4. The challenge for studies that predict functional assessment (FFR) or stenosis (QCA) lies in the full automation of this assessment, while

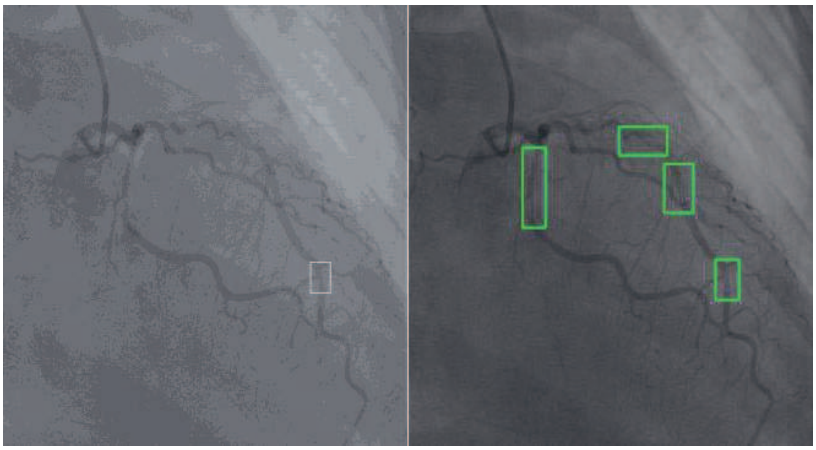


Figure 2.5: Stenosis localization ground truth (left) and example of corresponding Faster-RCNN output (right) [75]

maintaining the quality or at least usability of the ground truth. Different approaches, such as binary classification and regression, have been investigated to address this problem. Some studies used relatively small CNNs, achieving higher AUC values [52] or Dice scores for image patch classification [94], then with larger CNNs. Cong et al. [27] employed an Inception V3 network to differentiate between various degrees of stenoses and chronic total occlusion (CTO), outperforming a strategy involving U-Net segmentation combined with CNN classification. Other studies improved single-frame networks by using multiple frames as input [21, 24], showing that multiple frames do contain additional information over the key frame. However, Yabushita *et al.* [95] used the whole ICA sequence as input, but their performance was significantly lower than the Dice found in the studies mentioned above. The whole ICA sequence likely contains too much irrelevant information to use as input without any form of filtering or attention. In addition to stenosis percentage, some studies explored classifying lesions based on other criteria, such as calcification, thrombosis, and CTO [96, 77]. These studies showed improvement over conventional networks with the addition of FPN inspired network elements. The assessment of CTO is another subject of interest and is done to guide treatment. Liu *et al.* [76] successfully classified CTO based on its shape (blunt vs tapered). A few studies attempted to estimate the exact QCA ground truth value through regression [41, 11]. Iyer *et al.,* [41] used their segmentation results to calculate vessel diameter at the stenosis location. Zhang *et al.* [11] showed that using multiple frames is beneficial in regression by succesfully enriching the key frame with additional frames from the same and different angles. Cho *et al.* [97] examined the use of AI to replace functional assessments. They used an ensemble of boosted decision trees to predict FFR based on computed features from a preselected ICA frame and clinical variables.

Table 2.4: **Classification study comparison:** The different inputs, networks and the way of algorithm evaluation used in classification studies.
acc=accuracy, avg=average, corr=correlation, ex=extension, seq=whole ICA sequence

Study	Input	Network	Metric	Annotation	Score
Au *et al.* [52]	single	CNN (7 layers)	AUC	QCA$\leq$0.7	0.825
Thanou *et al.* [94]	single	ResNet18	Dice	QCA$\leq$0.7	0.571
Cong *et al.* [27]	single	InceptionV3	Dice	QCA$\leq$0.25	0.80
Liu *et al.* [76]	single	ResNet ensemble + FPN + FasterRCNN	Acc.	Expert	0.912
Chen *et al.* [96]	single	ResNet101(ex.)	avg. Dice	Expert	0.946
Du *et al.* [77]	single	DeepDiscern	avg. Dice	Expert	0.824
Iyer *et al.* [41]	single	Deeplabv3+	R^2	QCA	0.973
Zhang *et al.* [98]	21	3DConvNets + self-attention nets	corr.	QCA	0.891
Zhou *et al.* [21]	seq.	ResNet18 + U-Net	Acc.	Expert	0.792
Moon *et al.* [24]	seq.	Inception-v3 + self-attention	Acc.	QCA$\leq$0.5	0.939
Yabushita *et al.* [95]	seq.	3D CNN	Dice	QCA$\leq$0.75	0.60

2.3.5 Prognosis prediction

Future lesion predictions based on ICA images is another use case of AI. Deep learning based on either patched or whole frames was shown to outperform and/or improve future MI predictions based on clinical data [99], with logistic regression [100] and as done by interventional cardiologists who were shown the images or patches [100, 99, 94].

Automatic assessment of collateral physiology before and after revascularization was shown to be another possible use case for vessel segmentation algorithms [101].

2.3.6 End-to-end pipelines

Au *et al.* [52] created a deep learning pipeline that involved stenosis classification after localization and segmentation, all trained separately. Only one study, by Zhou *et al.* [21], attempted to create an end-to-end pipeline, including frame selection, in order to replace the clinical workflow from frame selection until stenosis localization by a single method, see Figure 2.1.

2.3.7 Other tasks

Several studies did not fall into any of the categories discussed above, but instead aim to facilitate the analysis of the steps in the clinical workflow. Fang *et al.* [102] estimated respiratory signals with a ResNet18 and an LSTM with a correlation coefficient of 0.82 between predicted and ground truth signal. Zhang *et al.* [98] improved motion noise reduction of a phantom ICA sequence with a 3D multiframe CNN. Ciusdel *et al.* [25] and Royer-Rivard *et al.* [103] both estimated cardiac phase with multiframe CNN pipelines and achieved an

accuracy of 0.988 and 0.960 respectively. Ma *et al.* [26] implemented a multiframe CNN to detect the inflow of contrast, with a mean 3.6 frame error.

2.3.8 Commercially available software

QCA, 3D-QCA, vessel FFR (vFFR) and quantitative flow ratio or QFR are commercial, FDA/CE approved, software packages to do lesion assessment. Several companies have incorporated an implementation of these in their software. These packages are all integrated via the hospital imaging archive PACS. AI-QCA is another software package and is an ensemble of U-Net++, U^2-Net and DeepLabV3+. It was shown to have a reasonable agreement with manual QCA (for different measures r^2 between 0.65-0.98, percent diameter stenosis 0.73) and to a lesser extend with IVUS (r^2 between 0.21-1.0, percent area stenosis 0.21) [104].

QCA is the oldest of the three and is used in the clinical workflow and often as ground truth in research [105, 16]. QCA still needs manual input and was found to be less useful in moderate stenosis assessment [106]. QCA can also be used with multiple frames as input, from which a 3D image is reconstructed. However, this 3D-QCA (CAAS, PIE Medical Imaging) was outperformed by vFFR in the identification of patients with FFR values of ≤ 0.90 (AUC 0.62 vs 0.98) [107].

Medis Medical Imaging's QFR (at least partly deep learing based) was shown to achieve similar performance (AUC 0.98) on an FFR ≤ 0.80 classification task [108]. Both QFR (Medis Medical Imaging) and vFFR use 2 angiograms, taken from different angles, from which a 3D image is reconstructed. Another QFR implementation by Pulse Medical is able to estimate stenosis severity based on a single frame [109], however, by the time of writing this review this implementation was not FDA/CE approved yet.

2.4 Discussion

All steps in the diagnostic ICA workflow have been addressed by deep learning tools in the literature. In general, recent deep learning architectures or extensions show the best results, and the use of multiple frames typically improves performance. However, a significant variation exists in ground truth, classification thresholds, metrics, and datasets, which makes comparing studies and drawing more general conclusion difficult. The majority of datasets used in these studies are not publicly available and are often used by one research group only. The segmentation studies depicted in Table 2.2 are the only studies that can be compared to a reasonable extent. Nonetheless, even though the data set was the same, the output for the original U-Net was different in the three studies. Moreover, some studies report that for example DeepLabV3+ and PSPNet are outperformed by a U-Net [47], whereas their original papers [41, 42] conclude otherwise. Alterations in data augmentation, preprocessing, cross-validation schemes and implementation could all explain this, however interpretation of the results leads to different conclusions.

From the other studies, only within study comparisons can be made. In general, it can be observed that SOTA algorithms like U-Net, U-Net++ and Faster-RCNN, benefit from various additional pre- or post-processing steps and/or from improvements to their backbones. The use of multiple frames was demonstrated to be beneficial in various tasks in the clinical workflow. A combination of contrast and non-contrast frames, subsequent frames,

frames a few frames apart, frames from different sequences all showed improvements over their single frame counterpart. This indicates that the studied algorithms cannot extract all information needed for segmentation, lesion localization and assessment from a single ICA frame. A comparison of the various ways multiple frames are used, lacks in literature and can not be made based on the studies addressed here since data sets differ between studies.

Competitions or Grand Challenges in classification and segmentation of the same ICA dataset could improve the comparability of algorithms, providing more detailed insights into the networks or adjustments that can improve the desired outcome.

Segmentation algorithms that can process the full cardiac cycle could be a valuable addition to visualization software, as part of a user interface, or as a preprocessing step in lesion assessment software. Lesion significance prediction would be relevant for clinical practice, specifically in intermediate lesions [11]. E.g. algorithms like QFR and vFFR that can reliably predict FFR from the ICA image strongly reduce the need for FFR measurement. This reduces costs, saves time and prevents complications associated with FFR.

In some cases there is a discrepancy between coronary anatomy, the results of non-invasive testing and the presence of anginal symptoms [110]. There is an increasing number of syndromes without significant stenosis on ICA imaging, but with anginal complaints. Recent studies suggest that an important number of patients have anginal complaints and non-obstructed coronaries, with or without abnormal results on ischaemia tests [111]. The prevalence of this syndrome, referred to as INOCA, can be as high as 50% of patients undergoing ICA, depending on the studied population. It is a frequently described phenomenon in women with angina and is associated with increased risk of adverse events [12] A possible explanation for this discrepancy would be that specific patterns within the ICA images cannot be recognized by the human eye. AI-based algorithms can be of help here by uncovering, yet unknown, signs of disease in images of INOCA patients and by discovering relations between patient parameters which can possibly predict the presence of disease in this group. Predicting future lesions might be an interesting next step and could help to explore the specific patterns mentioned above. In this way, novel associations between different patient characteristics and new (sub)populations might be uncovered, which may aid treatment differentiation.

2.5 Acknowledgments

This project received funding from the EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 860173.

2.6 Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used GPT-4 in order to adjust the style of multiple sentences. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

2.7 Appendix A

Deep learning is a fast developing subfield of machine learning in which patterns in images are learned by a deep neural network. Convolutional neural networks (CNN) are currently the most popular deep neural networks in (medical) image analysis [18]. Figure 2.6 shows a schematic overview of a CNN [112]. In CNNs image patches are processed in subsequent layers into multiple usually smaller representations called feature maps. In this so-called down-sampling or encoding path, more relevant contextual information is tried to be captured with every subsequent layer. Accurate location information is lost in this process, since the size of the image is reduced with every layer. The learnable parameters in every layer are the filters. The images/feature maps are convolved with these filters to create subsequent feature maps. This eventually results in the recognition of features that increase in complexity with every layer. After the convolutional layers, a CNN usually has a fully connected layer that weights and combines all feature map pixel values into a final regression or classification outcome. Many CNNs exist with different layer and filter numbers, sizes, and architectures. The term channels is used for the different representations of data used in a Deep Learning algorithm, in that sense it shares some similarity with the term feature maps. Attention is used to refer to selection of for example channels (and thus feature maps).

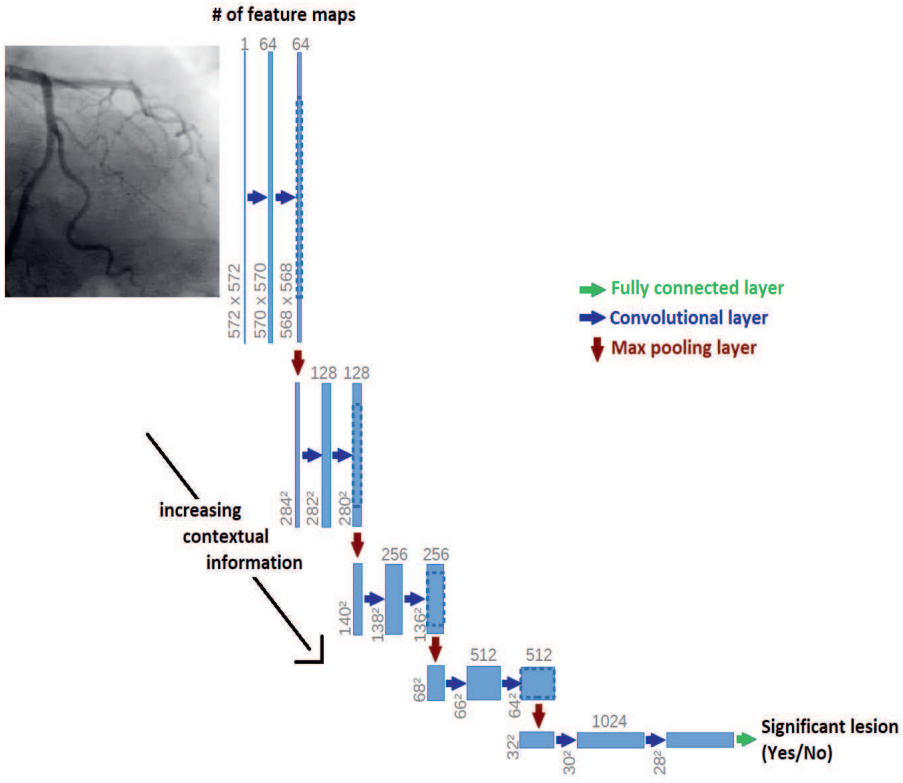


Figure 2.6: A schematic overview of a CNN used in binary classification.

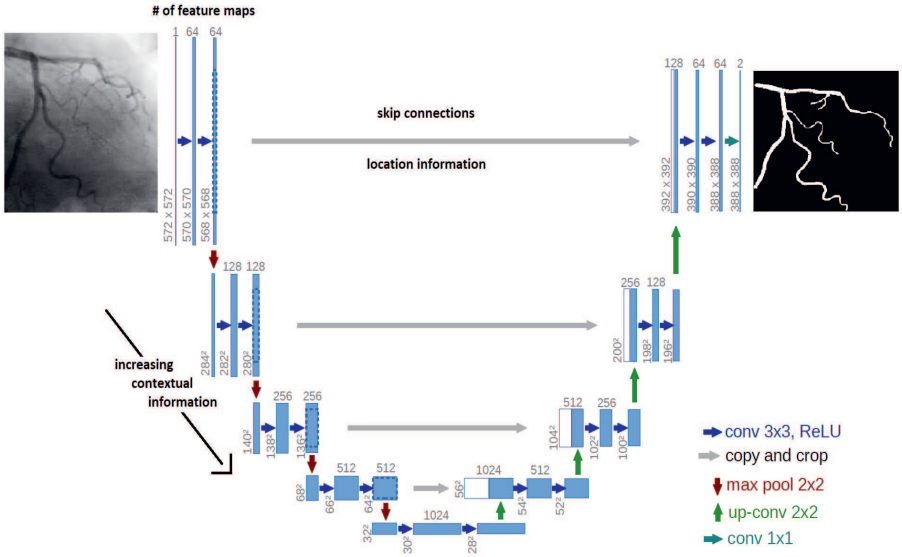


Figure 2.7: A schematic overview of a U-Net used in segmentation.

U-Nets are types of CNNs that are the current state-of-the-art in image segmentation. U-Nets aim to circumvent a loss of information due to the aforementioned downsampling with an up-sampling path, which learns to reconstruct or decode the desired output in subsequent up-sampling layers, see Figure 2.7. To enrich the up-sampling path with the lost location information of the input image, the feature maps created in the down-sampling path are shared in skip connections [70]. The U-Net is an example of an encoder-decoder segmentation network in which an image is encoded to a different representation and then decoded to the desired segmentation output. The CNN that is used in the encoder path of a U-Net is often referred to as its backbone, and is normally mirrored in the decoder path. Many variants on the original U-Net and other encoder-decoder segmentation networks exist. These and various types of CNNs used in medical image analysis will be discussed here.

The neural networks discussed above all follow a sequential way of processing, from layer to layer, and are therefore categorized as feedforward networks. However, networks exist that have connections from deeper layers back to previous layers, the so-called feedback or recurrent connections. These networks, the recurrent neural networks (RNN), are popular for analysis of sequential or time-series data. A long short-term memory network (LSTM) a widely used example of such an RNN which uses recurrent connections to update an internal state kept by the network [113].

One important recent improvement to neural networks in general, including CNNs, is the residual (or sometimes also called the skip) connection. This is essentially a way to let the filters learn the residual signal of feature maps from subsequent layers, by adding the output feature maps of a layer to the feature maps of a previous layer. These residual neural networks (ResNets) do not show the eventual performance decline normal networks show when more layers are added [39]. A DenseNet is a form of a ResNet in which the feature

maps of all previous layers are used as layer input [36]. Both these networks have versions with varying amount of layers, usually denoted by the number right after the network name (eg. ResNet50).

A generative adversarial network (GAN) is in essence a combination of two networks in which one learns to generate new data and the other learns to classify data as being real or generated [114]. GANs in the context of ICA image analysis can be used to create more ICA images to train a network on.

Object localization is another relevant medical image analysis task that can be done by deep learning algorithms. Faster region-based CNN (Faster R-CNN) [83] and gradient-weighted class activation mapping (GradCam) [115] are two popular methods in medical image processing. Faster R-CNN is a two-step algorithm in which a first CNN, the backbone, is used to extract feature maps from an image. A Region Proposal Network (RPN) uses these to propose regions in the image that are likely to contain an object, these regions are subsequently assigned to one of the classes in the training set. GradCam is a form of explainable AI in which the parts of the input image that contributed most to the output prediction are calculated by backwards calculation through the network. Visualization of the GradCam output is often done by overlaying a heatmap, that represents pixel importance scores, with an input image.

2.8 Appendix B

PubMed

("Coronary Angiography"[Majr] OR "Coronary Angiography"[ti] OR "Coronary Angiographies"[ti] OR "coronary angiogram*"[ti]) AND ("Artificial Intelligence"[Majr] OR "Artificial Intelligence"[ti] OR "AI"[tiab] OR "Image Processing, Computer-Assisted"[Majr] OR "automatic image analysis"[ti] OR "machine learning"[ti] OR "deep learning"[ti]) NOT ("CT"[ti] OR "compute*"[ti] OR "tomogr"[ti] OR "SPECT"[ti] OR "MR"[ti] OR "MRI"[ti] OR "magnetic"[ti])

Web of Science

TS=("Coronary Angiography" OR "Coronary Angiographies" OR "coronary angiogram*") AND TI=("Artificial Intelligence" OR "AI" OR "computer assisted Image Processing" OR "automatic image analysis" OR "machine learning" OR "deep learning") NOT TI=("CT" OR "compute*" OR "tomogr" OR "SPECT" OR "MR" OR "MRI" OR "magnetic")

IEEE

"All Metadata":coronary angiography AND ("All Metadata":artificial intelligence OR "All Metadata":deep learning OR "All Metadata":machine learning) NOT ("Document Title":"CT" OR "Document Title":"Computed tomography*")

EMBASE

('artificial intelligence':ti,ab,kw OR 'artificial intelligence'/exp OR 'deep learning':ti,ab,kw OR 'deep learning'/exp OR 'machine learning':ti,ab,kw OR 'machine learning'/exp) AND ('coronary angiography'/exp OR 'coronary angiography':ti,ab,kw) NOT ('computer assisted tomography'/exp OR 'nuclear magnetic resonance imaging'/exp)

Google Scolar "coronary angiography*" "artificial intelligence"—"deep learning" -"computed tomography" -"SPECT" -"MRI" -"kidney"

Ch. 2

3

Large language models for cardiac electronic health record data classification

Van der Loo, W.P.M., Van der Valk, V.O., Atsma, D.E., Staring, M., Scherptong, R.W.C. (2025). Large Language Models for Structured Cardiovascular Data Extraction: A Foundation for Scalable Research and Clinical Applications. European Heart Journal-Digital Health.

Abstract

Background: Automated extraction of information from cardiac reports would benefit both clinical reporting and research. Large language models (LLMs) hold promise for such automation, but their clinical performance and practical implementation across various computational environments remain unclear.

Objectives: To evaluate the feasibility and performance of LLM-based classification of echocardiogram and invasive coronary angiography (ICA) reports, using real-world clinical data across local, high-performance computing and cloud-based platforms.

Methods: The angiography and echocardiography reports of 1000 patients, admitted with acute coronary syndrome, were labeled for multiple key diagnostic elements, including left ventricular function (LVF), culprit vessel and acute occlusions. Report classification models were developed using LLMs via i. prompt-based and ii. fine-tuning approaches. Performance was assessed across different model types and compute infrastructures, with attention to class imbalance, ambiguous label annotations and implementation costs.

Results: LLMs demonstrated strong performance in extracting structured diagnostic information from cardiac reports. Cloud-based models (such as GPT-4o) achieved the highest accuracy (0.87 for culprit vessel and 1.0 for LVF) and generalizability, but also smaller models run on a local high performance cluster (HPC) achieved reasonable accuracy, especially for less complex tasks (0.634 for culprit vessel and 0.984 for LVF). Classification was feasible with minimal preprocessing, enabling potential integration into electronic health record systems or research pipelines. Class imbalance, reflective of real-world prevalence, had a greater impact on fine-tuning approaches.

Conclusions: LLMs can reliably classify structured cardiology reports across diverse compute infrastructures. Their accuracy and adaptability support their use in clinical and research settings, particularly for scalable report structuring and dataset generation.

3.1 Introduction

Invasive coronary angiography (ICA) and transthoracic echocardiography (TTE) are cornerstone imaging modalities in cardiology, essential for diagnosis, treatment and follow-up in patients with coronary artery disease (CAD)[3, 4]. Both procedures result in detailed reports, typically recorded in a semi-structured free text format within electronic health records (EHRs). While this format enables flexible, context-rich documentation, standardization is limited, restricting its usability for data-driven research, clinical decision support systems and artificial intelligence (AI) applications[116, 117, 118].

The absence of structured labels presents a critical bottleneck in AI-driven research, where large, accurately annotated datasets are essential for the development of robust models[119]. Manual annotation remains the predominant method for dataset creation, yet it is labor-intensive and prone to human error, making it impractical for large-scale supervised learning[10]. As a result, much of the valuable data generated in clinical practice remains inaccessible[120].

Recent advances in natural language processing (NLP), particularly the emergence of large language models (LLMs), offer new opportunities for automating the extraction of structured information from free-text reports.[121]. These models, built on transformer-based architectures with attention mechanisms, have demonstrated superior contextual understanding and scalability compared to earlier NLP techniques.[122]. LLMs have already shown promise in automating structured data extraction from free-text radiology reports[119, 123].

We hypothesize that LLM-based methods can automate the structured classification of ICA and TTE reports, enabling scalable dataset creation for AI applications and secondary data use. These methods also offer a robust foundation for extracting standardized data elements required for clinical registries, quality improvement programs and mandatory health system reporting, thereby reducing administrative burden, enhancing data completeness and supporting data-driven oversight across diverse healthcare settings. In this study, two distinct LLM-based approaches, prompt engineering and fine-tuning, were developed and evaluated using free-text cardiology reports obtained from routine clinical practice. Their classification performance and implementation complexity were systematically compared.

3.2 Methods

Datasets

A comprehensive pseudo-anonymized database was constructed from the EHRs of patients treated for acute coronary syndrome (ACS) at Leiden University Medical Center (LUMC), The Netherlands, between 2010 and 2024. This database encompassed all clinical data generated during initial treatment and subsequent follow-up of these patients. The study protocol (nWMODIV2_2022006) was approved by the institutional review board of LUMC, which waived the requirement for obtaining informed consent for the use of data from individual patients. All procedures were conducted in accordance with institutional guidelines and regulations. From the source database, two study-specific datasets were created by random sampling: one consisting of 1000 ICA reports and the other comprising 1000 TTE reports. All reports were written in Dutch, adhered to a semi-structured format and were composed by board-certified cardiologists during routine clinical care. None of the reports contained personally identifiable information, apart from clinical details specific to the modality. Only the textual content of the reports, without any accompanying metadata, was processed through the LLMs. Data processing and model inference were performed within secure, encrypted environments compliant with European data protection regulations, using only anonymized reports that underwent manual review for the presence of personally identifiable data prior to processing. Textual data was pre-processed to enhance quality and consistency prior to model training. Automated cleaning scripts based on regular expressions, a technique for identifying and modifying text patterns, were used to remove unnecessary spaces, redundant line breaks and formatting artifacts introduced during data extraction. This process reduces noise, corrects structural inconsistencies and standardizes the text format across all reports.

Data labeling

A structured annotation process was employed, involving two independent expert reviewers who annotated the reports in two iterative rounds. Annotation guidelines were defined prior to the first round and refined after each iteration to enhance consistency and accuracy. For the final labels, disagreements between annotators were resolved through discussion. Average manual annotation scores are calculated based on the comparison between the labels of the final round of each annotator and the final joint labels. These scores provide a reference for the difficulty of the labeling task and establish a baseline for comparison with

LLM performance.

ICA reports were labeled for: (1) presence of acute occlusion, (2) presence of bypass grafts, (3) presence of significant epicardial CAD, (4) treatment strategy (PCI, referral for coronary artery bypass grafting (CABG) or medical treatment only) and (5) identification of culprit vessel(s), with labels assigned for both the main coronary artery (Main) and specific branch (Sub). For the culprit vessel category, multiple labels could be assigned simultaneously (e.g., RCA/RDA), whereas for all other categories, only a single label was allowed. TTE reports were labeled for left ventricular (LV) systolic function and valvular dysfunction, including type and severity for each cardiac valve. LV function was categorized as normal, mildly, moderately or severely reduced, when multiple gradings were present within a report, the most severe category was selected. Each valve (aortic (AV), mitral (MV), tricuspid (TV) and pulmonary (PV)) was independently assessed for dysfunction type (none, regurgitation, stenosis or both) and severity (none, mild, moderate or severe). A ”no data” label was assigned when relevant findings were absent from the report. A complete overview of all label categories is provided in Table 3.1.

Prompt engineering

In prompt engineering, a set of explicit instructions is provided to a pre-trained LLM, guiding its response generation. The internal parameters of the model are not modified based on these instructions; rather, it utilizes its pre-existing knowledge to interpret the prompt and generate an appropriate response. The model retains the same weights and representations that were established during its initial training phase[124, 125]. Within prompt engineering, a distinction can be made based on whether or not labeled examples are included within the prompt itself. In a zero-shot approach, the model receives only the task description without any examples, relying entirely on its internal knowledge to produce an answer. In contrast, in few-shot prompting, a small number of labeled examples are embedded within the prompt to illustrate the task, helping the model better understand the desired format and reasoning. A comprehensive overview on prompting-techniques is provided by Sivarajkumar et al[125].

Fine-tuning

Fine-tuning entails an additional training phase in which a pre-trained LLM is adapted using a task-specific, labeled dataset. This process adjusts the model's internal parameters

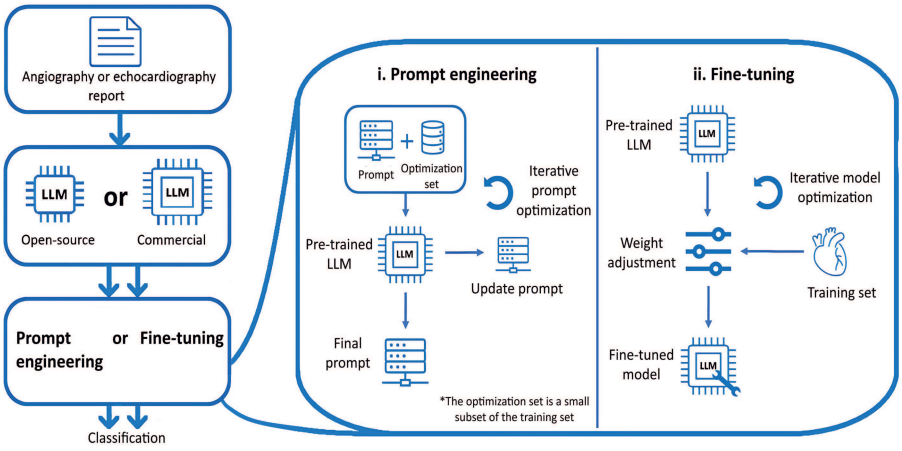


Figure 3.1: Experiment overview: Both angiography and echocardiography reports are tested on open-source and commercial models optimized with either i. the prompt engineering or ii. the fine-tuning method. i. Prompt engineering method, resulting in an optimized prompt, where no model training takes place. ii. Fine-tuning method, a pre-trained LLM is trained on a domain specific dataset, where adjusting the weights of the network results in a fine-tuned model.

to better align with domain-specific objectives[126]. Although fine-tuning typically requires more computational resources and annotated data, it may yield superior performance for specialized tasks, such as the structured labeling of cardiology reports. A visual comparison between both techniques is shown in Figure 3.1.

LLM assessment

Both fine-tuning and prompt engineering (see Figure 3.1) were tested on a commercially available LLM (GPT-4o via Azure OpenAI) and several open-source state-of-the-art (SOTA) models available via Huggingface.co and Ollama.com that can be used on-site. Open-source SOTA models were selected by identifying the two most popular models for text classification or feature extraction across three categories: general-purpose, medical domain-specific and multilingual LLMs. For on-site prompt engineering two different hardware constraints are tested. A smaller local GPU of 16Gb, which would be an average laptop GPU and a bigger high performance cluster (HPC) GPU of 48Gb. For on-site fine-tuning only the HPC hardware was tested, since hardware requirements are too large for fine-tuning LLMs on small GPUs. Selection was done with the important constraint that the model could either be run i) locally on a 16Gb GPU for local prompt engineering or ii) on a 48Gb GPU for HPC prompt engineering or iii) be trained on a 48Gb GPU for fine-tuning.

Label	Values
ICA dataset	
Occlusion	Yes, no
No CAD	Yes, no
Graft	Yes, no
Treatment strategy	PCI, CABG, medical
Main	RDA, RCA, RCx, LM, Graft, IM, no
Sub	RDA, RCA, RCx, LM, Graft, IM, AL, D, MO, PL, RDP, RV, S, no
TTE dataset	
LV function	Normal, mildly, moderately or severly reduced, no data
Valve dysfunction type	None, stenosis, regurgitation, both, no data
Valve dysfunction grade	None, mild, moderate, severe, no data

Table 3.1: Overview of the possible ICA and TTE labels

Fine-tuning

For on-site fine-tuning each model was extended with a classification layer, such that the model directly outputs class indexes. This classification layer was trained without pre-training, while the underlying model was fine-tuned on the manually labeled ICA and TTE datasets. To mitigate overfitting, we implemented early stopping based on the performance on the validation set, along with weight decay as a regularization technique. Additionally, the learning rate, a crucial hyperparameter, is fine-tuned to optimize performance while maintaining stability. For on-site finetuning the following pre-trained models were selected from Huggingface.co on March 2, 2025: multilingual-e5-large from Intfloat [127] (multilingual 1), bert-base-multilingual-uncased-sentiment from NLPTown (multilingual 2), BiomedVLP-BioVil-T from Microsoft [128] (medical 1), MedCPT-Cross-Encoder from NCBI [129] (medical 2), bart-base from Facebook [130] (general 1) and ms-marco-MiniLM-L-6-v2 from Cross-encoder (general 2).

Fine-tuning of the commercially available GPT-4o model was performed using the Azure OpenAI fine-tuning API. The model was adapted to the classification tasks using manually labeled reports. The fine-tuning process is subjected to predefined limitations inherent to the commercial API, which permits adjustment of only a single relevant hyperparameter; the learning rate multiplier. Modification of the model architecture or implementation of training optimizations such as early stopping or weight regularization are not possible. The optimal learning rate was selected based on validation performance and training stability, ensuring generalization to unseen data.

Prompt Engineering

Prompt-based inference was conducted using a few-shot prompting strategy. Each clinical report was combined with standardized labeling instructions to create a structured input format. Structured output templates were used to ensure consistency and facilitate accurate parsing. Prompts were iteratively refined on a subset of the training data, with adjustments to wording and structure aimed at minimizing misclassification and improving labeling accuracy. Model settings were configured to promote consistent outputs, with parameter settings to minimize variability during output generation. For local prompt engineering, the following small pre-trained models were selected from Ollama.com on March 2, 2025: Aya 23 from Cohere [131] (multilingual 1), Llama3.2 from Meta (multilingual 2), MedLlama2 by Sourcell (medical 1), Phi-4 from Microsoft (general 1) and Gemma3 from Google (general 2/multilingual). No second medical model was selected because a decent medical runner up model was not available at this moment. For HPC prompt engineering the following larger pre-trained models were selected from Ollama.com on March 2, 2025: Gemma3 from Google (multilingual 1) and Wizardlm2 from Microsoft (multilingual 2), Meditron (medical 1) [132], Medllama2 by Sourcell (medical 2), Deepseek-R1-Distill-Qwen from DeepSeek (general 1) and Phi-4 from Microsoft (general 2). For some models larger versions exist, the largest version that can fit on either the local GPU (16GB) or the HPC cluster GPU (48GB) was chosen.

Analysis

Both data sets were randomly divided into a training and a test set with a ratio of 70:30. For the prompt engineering method, the training set was used for prompt optimization and example selection. For the fine-tuning method, the training set was again split in a training and validation set with a ratio of 85:15. The training set was used to fine-tune the models, while the validation set is used to monitor overfitting. For all on-site model comparisons a 5-fold cross-validation scheme was used. Model evaluation was done by calculating metrics on the combined predictions on all validation sets, which aggregate to the whole training set. The test set, which remains completely unseen during prompt optimization and fine-tuning, was reserved for final performance evaluation. Models were evaluated using the following performance metrics: accuracy, average recall and macro-averaged F1-score. For the multi-class multi-label tasks (main and sub culprit vessel) a strict evaluation criterion was adopted: classification outputs were only considered correct if the entire set

of predicted labels per report exactly matched the reference labels. Consequently, partially correct classifications, such as predicting “RDA/D” when only “D” was correct, were seen as incorrect. To quantify variability, 95% confidence intervals were estimated using 1000 bootstrap iterations for each metric. To assess statistically significant differences in performance, the Bonferroni corrected p-value (p_b) is calculated per label category, using the best-performing model as the reference for all pairwise comparisons. Finally, a subset of misclassified cases was manually reviewed to identify recurring patterns and potential sources of model error.

3.3 Results

Study population

Both datasets comprised 1000 unique reports. In the ICA dataset, each report represented a unique patient (n = 1000), while in the TTE dataset, the reports corresponded to 736 unique patients, with 264 patients contributing two reports each. No reports were duplicated and no patient appeared more than twice within the TTE dataset. The mean age of the patients was 65.7 ± 11.9 years in the ICA cohort and 62.2 ± 11.2 years in the TTE cohort. The sex distribution in the ICA dataset was 70.7% male (n = 707) and 29.3% female (n = 293). In the TTE dataset, 74.6% were male (n = 746) and 25.4% were female (n = 254). The initial clinical presentation among patients in the ICA dataset was ST-segment elevation myocardial infarction (STEMI) in 33.1% (n = 331) , non-ST-segment elevation myocardial infarction (NSTEMI) in 30.7% (n = 307) and unstable angina in 36.2% (n = 362). In the TTE dataset, 66.2% (n = 487) of cases presented with STEMI, 25.7% (n= 189) with NSTEMI, and 8.2% (n = 60) with unstable angina.

Data labeling

The ground truth label distributions across all variables in the ICA dataset were skewed, with varying degrees of class imbalance, ranging from 88% vs 12% for the no CAD label to 33% vs 67% for the occlusion label. As expected, the proportion of patients without CAD was low (12%), as most patients had an occlusion (67%), underwent PCI (75%) and did not have grafts (85%). In the culprit vessel labels, the ramus descendens anterior (RDA), right coronary artery (RCA) and ramus circumflexus (RCx) were the most frequently identified vessels in both the main (361, 211 and 318 times respectively) and subsegment categories (321, 291 and 147 respectively). In contrast, the anterolateral branch (AL) was not

annotated in any case and other vessels such as the septal (S) and right ventricular (RV) branches were only rarely identified (both 2 times). A full overview of the label distributions is shown in Appendix 3.6

Label imbalance in the TTE dataset was even more pronounced. The most frequent class for LV function was “mildly impaired” (52%) whereas “severely impaired” LV function was observed in only 2% of cases. Complete data on LV function were available for all patients. Across all valves, “no dysfunction” was the most common label (83%), followed by “regurgitation” (overall 11%) in all but the pulmonary valve (PV). The PV had the highest proportion of missing data (9%). Notably, no cases of isolated stenosis were recorded for the mitral, tricuspid or pulmonary valves. For valve dysfunction grading, the most common non-normal category was “mild” (5%) while “moderate” (<1%, 59 cases) and “severe” (<1%, 12 cases) grades were uncommon and for mitral stenosis, pulmonary stenosis and pulmonary regurgitation, and entirely absent for tricuspid stenosis. A full overview of the TTE label distributions is provided in Appendix 3.6

Annotator agreement

The average agreement between both annotators was high, as shown in Table 3.2 and 3.3. In the ICA dataset, accuracy ranged from 0.917 to 0.987, average recall from 0.787 to 0.964 and F1 scores from 0.818 to 0.963. The highest scores were observed for graft presence, while the lowest were found in subsegment-level vessel classification. In the TTE dataset, manual annotation scores were consistently high across all labels, with accuracy ranging from 0.993 to 1.00, average recall from 0.975 to 1.00 and F1 scores from 0.920 to 1.00. Human performance was particularly strong for binary labels in the ICA dataset and for all TTE labels, while greater variability was observed in more complex and ambiguous tasks such as culprit vessel annotation.

Prompt engineering model assessment

Figure 3.2 shows the comparison of the 6 selected LLMs for the classification of ICA and TTE reports. The metrics reflect the average classification performance across all labels per report. Results represent the aggregated outcomes from five-fold cross-validation conducted on the training/validation dataset. Original values are 5-fold cross-validation means on training/validation set, see Figures 3.8-3.13 in Appendix 3.6. For each metric, the mean value and the corresponding 95% confidence interval (CI), derived from 1000 bootstrap samples, are shown. Models run on a local smaller GPU demonstrated slightly inferior

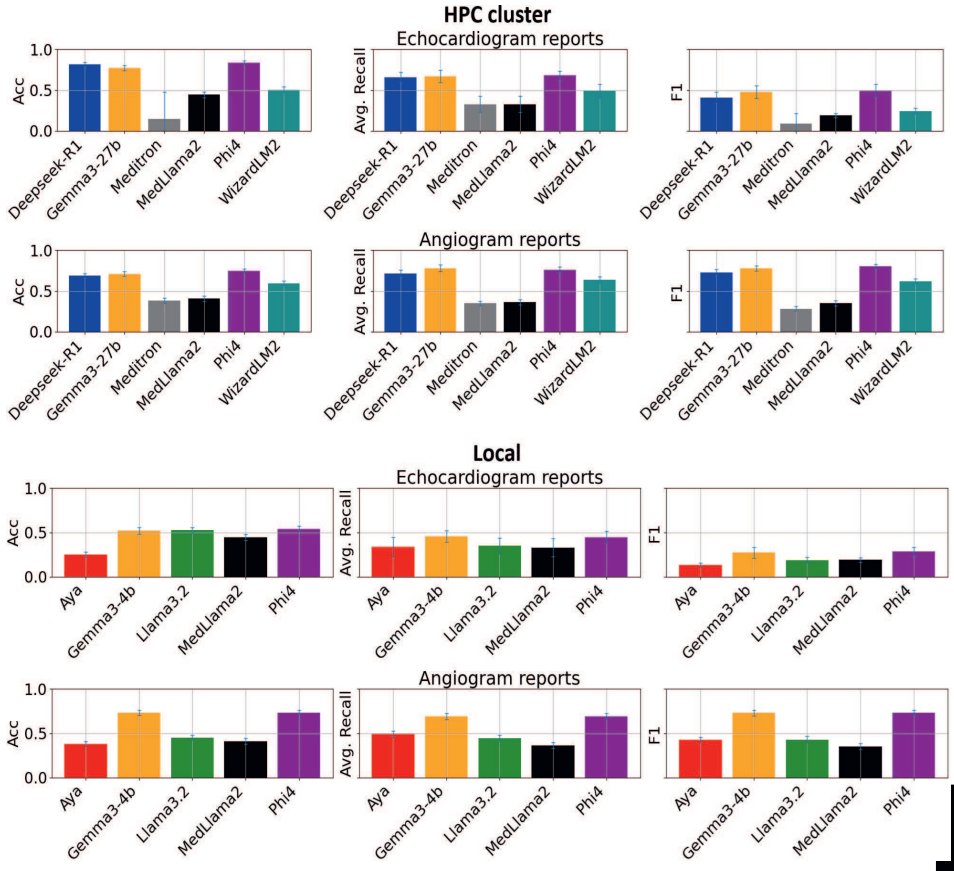


Figure 3.2: Model comparison for prompt engineering on a local machine (16Gb) and on a HPC cluster (48Gb) for ICA and TTE report classification. For each model the average performance (of all labels) is shown. 95% CI of metrics calculated with 1000 bootstrapped samples is indicated with error bars.

performance compared to those trained on a HPC cluster, although the difference was modest. Notably, both the Gemma3 and the Phi4 models showed robust performance in both GPU environments, with the Phi4 model run on the HPC cluster showing best overall performance.

Finetune model assessment

Figure 3.3 shows the comparison of the 6 selected LLMs for the classification of ICA and TTE reports. For each model the average performance (of all labels) is shown. 95% CI of metrics calculated with 1000 bootstrapped samples is indicated with error bars.. The multilingual model from Intfloat showed best overall performance for classification of both

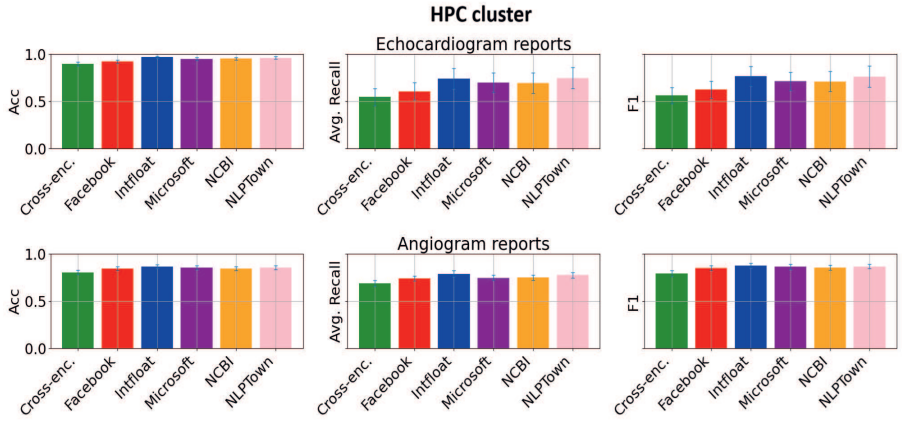


Figure 3.3: Model comparison for finetuning on a HPC cluster on ICA and TTE reports. For each model the average performance (of all labels) is shown. 95% CI of metrics calculated with 1000 bootstrapped samples is indicated with error bars.

report types.

Classification results and model comparison

Table 3.2 and 3.3 show the performance of the best local prompt engineering and fine-tuning models compared with prompt engineering and fine-tuning with the commercial model (GPT-4o) on the held-out test set. The average human annotator performance per task is shown as an indicator of task difficulty. Especially for the more difficult tasks, GPT-4o showed superior performance for both prompt engineering and fine-tuning methods and often approaches or surpasses the human annotator benchmark. For most labels, no statistically significant difference was observed between prompt engineering and fine-tuning when applied to GPT-4o. In contrast, performance differences were more variable across tasks when using the open-source SOTA models, while prompt engineering and fine-tuning yielded comparable average performance, outcomes varied depending on the task. Prompt engineering tended to under perform when the test set lacked representation of categories included in the prompt. Fine-tuning on the other hand exhibited diminished performance in severely imbalanced datasets. Complex tasks such as culprit vessel detection, which are a multi-class and multi-label tasks are challenging for for all on-site models, regardless of training strategy.

Misclassification analysis

All analyses were performed on the held-out test set of 300 ICA and 300 TTE reports.

Label	Accuracy	Avg. recall	F1
Main *(manual)*	*0.943*	*0.920*	*0.963*
FT (Intfloat)	0.634 [0.580-0.683]	0.392 [0.362-0.42]	0.718 [0.672-0.764]
FT (GPT-4o)	**0.870 [0.830-0.907]**	**0.830 [0.774-0.888]**	**0.937 [0.914-0.958]**
PE (Phi-4)	0.522 [0.467-0.580]	0.510 [0.449-0.579]	0.715 [0.671-0.757]
PE (GPT-4o)	0.761 [0.710-0.803]	0.768 [0.693-0.838]	0.881 [0.849-0.91]
Sub *(manual)*	*0.917*	*0.787*	*0.941*
FT (Intfloat)	0.533 [0.477-0.587]	0.212 [0.185-0.252]	0.607 [0.554-0.658]
FT (GPT-4o)	**0.746 [0.693-0.793]**	0.668 [0.576-0.764]	**0.819 [0.776-0.857]**
PE (Phi-4)	0.460 [0.403-0.520]	0.611 [0.528-0.697]	0.797 [0.761-0.833]
PE (GPT-4o)	0.634 [0.580-0.687]	**0.795* [0.719-0.872]**	**0.830 [0.799-0.862]**
No CAD *(manual)*	*0.974*	*0.964*	*0.945*
FT (Intfloat)	**0.960* [0.937-0.980]**	0.883 [0.820-0.939]	**0.918* [0.867-0.959]**
FT (GPT-4o)	**0.964* [0.940-0.983]**	**0.969* [0.940-0.990]**	**0.936* [0.897-0.971]**
Few Shot (Phi-4)	**0.974* [0.953-0.990]**	0.899 [0.865-0.928]	0.782 [0.727-0.838]
Few Shot (GPT-4o)	**0.973* [0.953-0.990]**	**0.950* [0.904-0.985]**	**0.949* [0.913-0.979]**
Occlusion *(manual)*	*0.926*	*0.933*	*0.917*
FT (Intfloat)	0.793 [0.750-0.837]	0.752 [0.701-0.801]	0.761 [0.710-0.810]
FT (GPT-4o)	**0.920* [0.89-0.947]**	**0.920* [0.887-0.951]**	**0.912* [0.878-0.943]**
PE (Phi-4)	0.761 [0.713-0.81]	0.697 [0.646-0.752]	0.709 [0.654-0.767]
PE (GPT-4o)	**0.927* [0.897-0.953]**	**0.917* [0.881-0.949]**	**0.919* [0.884-0.949]**
Treatment strategy *(manual)*	*0.976*	*0.924*	*0.818*
FT (Intfloat)	0.953* [0.930-0.977]	**0.895* [0.825-0.955]**	0.882* [0.810-0.940]
FT (GPT-4o)	**0.987* [0.973-0.997]**	**0.970* [0.926-0.998]**	**0.97* [0.931-0.997]**
PE (Phi-4)	0.940 [0.910-0.967]	0.835 [0.748-0.917]	0.847* [0.762-0.919]
PE (GPT-4o)	**0.977* [0.957-0.993]**	**0.915* [0.841-0.976]**	**0.930* [0.868-0.977]**
Graft *(manual)*	*0.987*	*0.940*	*0.957*
FT (Intfloat)	0.970* [0.950-0.987]	0.901* [0.839-0.959]	0.930* [0.881-0.970]
FT (GPT-4o)	**0.996* [0.990-1.00]**	**0.998* [0.994-1.00]**	**0.992* [0.975-1.00]**
PE (Phi-4)	**0.993* [0.983-1.00]**	0.924* [0.864-0.974]	0.947* [0.905-0.981]
PE (GPT-4o)	**0.993* [0.983-1.00]**	**0.996* [0.990-1.00]**	**0.986* [0.965-1.00]**

Table 3.2: Model comparison for ICA report classification on the test set. Four LLM models and training methods are compared on 6 labels extracted from ICA reports. Comparison is done with 3 metrics: accuracy, average recall and F1. The 95% CI is given in brackets. The best metric score and any score that is not significantly different from this score (p_b=0.0167), is indicated with bold font. Manual annotation scores are averaged human annotator scores and are shown in *italic*. Significant improvement or equality to the manual score is indicated with an asterisk (*). FT = fine-tuning, PE = prompt engineering.

Label	Accuracy	Avg. recall	F1
LV function *(manual)*	*0.997*	*0.998*	*0.998*
FT (Intfloat)	**0.984* [0.967-0.997]**	**0.918 [0.809-0.988]**	**0.947 [0.869-0.993]**
FT (GPT-4o)	**1.00* [1.00-1.00]**	**1.00* [1.00-1.00]**	**1.00* [1.00-1.00]**
PE (Phi-4)	0.727 [0.667-0.777]	0.805 [0.753-0.849]	0.518 [0.429-0.647]
PE (GPT-4o)	**0.997* [0.990-1.00]**	**0.998* [0.995-1.00]**	**0.993* [0.976-1.00]**
Mitral Sten. Grade *(manual)*	*1.00*	*1.00*	*1.00*
FT (Intfloat)	0.969 [0.950-0.987]	0.555 [0.495-0.667]	0.578 [0.488-0.745]
FT (GPT-4o)	**0.997* [0.990-1.00]**	**0.998* [0.995-1.00]**	**0.972* [0.897-1.00]**
PE (Phi-4)	0.796 [0.750-0.837]	0.410 [0.388-0.431]	0.223 [0.215-0.229]
PE (GPT-4o)	**1.00* [1.00-1.00]**	**1.00* [1.00-1.00]**	**1.00* [1.00-1.00]**
Mitral Reg. grade *(manual)*	*0.997*	*0.975*	*0.920*
FT (Intfloat)	0.866 [0.827-0.903]	0.372 [0.298-0.475]	0.379 [0.295-0.504]
FT (GPT-4o)	**0.990* [0.977-1.00]**	**0.963* [0.900-1.00]**	**0.896* [0.738-1.00]**
PE (Phi-4)	0.893 [0.860-0.927]	0.783 [0.688-0.869]	0.707 [0.532-0.853]
PE (GPT-4o)	**0.980 [0.963-0.993]**	**0.946* [0.878-0.989]**	**0.897* [0.743-0.993]**
Aortic Sten. grade *(manual)*	*1.00*	*1.00*	*1.00*
FT (Intfloat)	0.907 [0.873-0.940]	0.219 [0.200-0.250]	0.208 [0.187-0.242]
FT (GPT-4o)	**0.997* [0.987-1.00]**	**0.999* [0.997-1.00]**	**0.991* [0.968-1.00]**
PE (Phi-4)	0.866 [0.823-0.903]	0.791 [0.721-0.850]	0.516 [0.377-0.665]
PE (GPT-4o)	**0.990* [0.977-1.00]**	0.856* [0.753-1.00]	0.836* [0.712-1.00]
Aortic Reg. grade *(manual)*	*0.993*	*0.984*	*0.991*
FT (Intfloat)	0.843 [0.800-0.883]	0.201 [0.200-0.201]	0.184 [0.178-0.19]
FT (GPT-4o)	**0.987* [0.973-0.997]**	**0.975* [0.948-0.999]**	**0.978* [0.952-0.996]**
PE (Phi-4)	0.876 [0.837-0.913]	0.795 [0.772-0.834]	0.661 [0.553-0.744]
PE (GPT-4o)	**0.973 [0.953-0.990]**	**0.885 [0.751-0.967]**	**0.904 [0.742-0.980]**
Tricuspid Sten. grade *(manual)*	*0.993*	*0.997*	*0.957*
FT (Intfloat)	0.963 [0.94-0.983]	0.500 [0.500-0.500]	0.491 [0.485-0.496]
FT (GPT-4o)	0.967 [0.947-0.987]	0.983 [0.972-0.993]	0.831 [0.727-0.925]
PE (Phi-4)	0.838 [0.793-0.877]	0.435 [0.415-0.454]	0.185 [0.177-0.229]
PE (GPT-4o)	**0.997* [0.990-1.00]**	**0.998* [0.995-1.00]**	**0.977* [0.913-1.00]**
Tricuspid Reg.grade *(manual)*	*0.993*	*0.998*	*0.978*
FT (Intfloat)	0.887 [0.850-0.923]	0.260 [0.250-0.333]	0.245 [0.230-0.317]
FT (GPT-4o)	**0.963 [0.943-0.983]**	**0.978 [0.947-0.994]**	**0.905 [0.841-0.955]**
PE (Phi-4)	0.853 [0.81-0.893]	0.797 [0.702-0.878]	0.491 [0.360-0.657]
PE (GPT-4o)	**0.980* [0.963-0.993]**	**0.961 [0.912-0.997]**	**0.901* [0.701-0.985]**
Pulmonary Sten. grade *(manual)*	*0.997*	*0.998*	*0.991*
FT (Intfloat)	**0.987* [0.973-0.997]**	**0.950* [0.896-0.998]**	**0.963* [0.926-0.992]**
FT (GPT-4o)	**0.993* [0.983-1.00]**	**0.982* [0.945-1.00]**	**0.982* [0.950-1.00]**
PE (Phi-4)	0.834 [0.790-0.877]	0.508 [0.459-0.568]	0.260 [0.191-0.384]
PE (GPT-4o)	**0.996* [0.987-1.00]**	**0.998* [0.993-1.00]**	**0.990* [0.967-1.00]**

Continued on next page

Label	Accuracy	Avg. recall	F1
Pulmonary Reg. grade *(manual)*	*0.997*	*0.999*	*0.994*
FT (Intfloat)	0.963 [0.940-0.983]	0.634 [0.597-0.665]	0.638 [0.612-0.659]
FT (GPT-4o)	**0.990* [0.977-1.0]**	**0.939* [0.820-1.00]**	**0.958* [0.873-1.00]**
PE (Phi-4)	0.816 [0.770-0.860]	0.740 [0.684-0.800]	0.427 [0.349-0.507]
PE (GPT-4o)	**0.993* [0.983-1.00]**	**0.950* [0.831-1.00]**	**0.964* [0.876-1.00]**

Table 3.3: Model comparison for TTE report classification on the test set. Four LLM models and training methods are compared on 9 labels extracted from TTE reports. Comparison is done with 3 metrics: accuracy, average recall and F1. The 95% CI is given in brackets. The best metric score and any score that is not significantly different from this score (p_b=0.0167), is indicated with bold font. Manual annotation scores are averaged human observer scores and are shown in *italic*. Significant improvement or equality to this manual score is indicated with an asterisk (*). FT = fine-tuning, PE = prompt engineering.

Open source models

For the ICA report classification (Table 3.2), the fine-tuned model most frequently produced false negatives for "no CAD" (11/300, 4%), while prompt-engineering yielded a higher false-positive rate (46/300, 15%). Both models over predicted occlusions (13–18%). Graft detection was generally reliable. However, the fine-tuned model systematically missed isolated venous grafts (9/300, 3%), while prompt-engineering errors were more evenly distributed. For "Treatment strategy" classification, the fine-tuned model tended to over predict referral for CABG (5/300, 2%), whereas the prompt-engineered model more often missed true PCI cases (8/300, 3%). Culprit vessel classification showed notable limitations. The fine-tuned model only correctly identified 2/24 multi-vessel culprit cases, this pattern was not observed with prompt-engineering, though its overall accuracy was inferior in this category. Among main culprit vessel, 16 (5%) errors of fine-tuned model and 26 (9%) errors with prompt engineering were associated to graft-related misclassification. Notably, prompt engineering failed to assign the "no culprit" label in any case, despite its presence in 52 test samples. In the sub-vessel category, the fine-tuned model only assigned the labels RCA, RCx, RDA and no.

In TTE reports (Table 3.3), the fine-tuned model most frequently misclassified mild regurgitation as none (AV: 25/300, 8%; MV: 14/300, 5%). In contrast, prompt-engineering occasionally overestimated pathology, predicting dysfunction where none was present. Stenosis grading errors followed similar patterns, with the fine-tuned model under-calling and prompt-engineering over-calling severity. Notably, the fine-tuned model assigned the "no dysfunction" label to the AV in all cases.

Commercial models

For commercial models (Table 3.2), both prompt-engineering and fine-tuning showed the lowest performance for culprit vessel classification. Prompt-engineering produced 24% and the fine-tuned model 13% main vessel labeling errors. Common issues included labeling both main and sub-branches (15% and 12%, respectively), and misclassifying graft versus native vessels. Both models consistently failed to recognize the IM branch correctly (3/9 cases each). Errors in classifying the presence of acute occlusions were often related to the presence of CTOs and grafts. Misclassifications also occurred in differentiating pre-existing CTOs from new obstructive disease. In both approaches, the most common misclassifications for the labels graft, no CAD and occlusion were false positives. With the prompt engineering method, these occurred in 2/300 (1%), 6/300 (2%) and 19/300 (6%) cases, respectively, whereas with the fine-tuned model, the corresponding rates were 1/300 ($<$1%), 10/300 (3%), and 6/300 (2%). For ”treatment strategy” classification, the most common error in both models was predicting medical therapy only while a referral for CABG was mentioned, particularly in reports indicating pending diagnostics (prompt engineering: n=5/300, 2%; fine-tuned model: 2/300, 1%).

In TTE data (Table 3.3), valvular assessment was the main challenge, particularly differentiating between ”none” and ”no data” due to inconsistencies between structured data and narrative conclusions. For example, cases where mild MR was quantitatively reported but summarized as “no dysfunction” in the narrative conclusion. Additional errors were linked to ambiguous or non-diagnostic phrases, such as “TV: TR gradient 30 mmHg” or “AV opens well visually, no reliable gradient”. Both models under predicted mild regurgitation, with the fine-tuned model showing fewer errors overall. LV function classification was accurate, with only isolated errors.

3.4 Discussion

LLM-based methods are a very valuable tool in the automation of ICA and TTE report classification. Depending on the accuracy required, the labeling difficulty and the available budget and computational power, different LLM-based methods can be used for the automation of report classification.

In this study, the performance of LLMs using both prompt engineering and fine-tuning was evaluated for the classification of TTE and ICA reports across different computational environments: i. a local server, ii. a HPC cluster and iii. a commercial cloud-based API

(e.g., GPT-4o). The goal was to explore the practical feasibility, performance and trade-offs for different LLM approaches in processing real-world cardiology data.

Model performance and comparative analysis Across platforms, LLMs demonstrated robust performance in classifying key clinical findings with minimal pre-processing, underscoring their potential integration into cardiology workflows such as automated EHR annotation, registry data generation and retrospective data structuring. Compared with manual annotation by a single experienced reviewer, GPT-4o-based approaches demonstrated similar performance for most tasks, except for culprit vessel identification, the most complex task, and tricuspid regurgitation grading with the finetuning approach, where severe class imbalance impaired model performance. Open-source models exhibited similar task-specific performance trends but with consistently lower accuracy. For simpler tasks with balanced classes, both modeling approaches performed comparable to the human annotator. The lowest performance was observed in tasks involving both main and subculprit vessel identification, consistent with the inherent complexity of the task. Additionally, the strict evaluation criterion for these tasks contribute to lower performance metrics. In the context of TTE reports, classification errors often resulted from inconsistencies between structured quantitative measurements and the narrative conclusions. For example, cases describing mild mitral regurgitation in the measurements section were occasionally summarized in the conclusion as “no dysfunction”. Furthermore, the presence of clinically benign phrases, such as “calcified annulus with normal function”, occasionally triggered false-positive classifications. The various approaches revealed different levels of task comprehension, likely influenced by model architecture and scale [133]. For instance, the prompt-engineered open-source model frequently labeled any vessel with atherosclerotic plaque as a culprit lesion, without adequately considering the degree of luminal narrowing, indicating limitations in nuanced clinical reasoning. Importantly, GPT-4o achieved high and consistent performance across both prompt-based and fine-tuned implementations. However, fine-tuning required significantly greater computational resources, a larger volume of annotated training data and incurred substantially higher monetary costs. These trade-offs suggest that for many clinical applications, particularly in resource-constrained environments, prompt engineering with commercial models offers a more pragmatic and scalable solution. Conversely, in complex tasks such as culprit vessel labeling, fine-tuning appeared to meaningfully enhance performance, supporting its use where feasible. While

locally deployed open-source models did not match GPT-4o's performance in complex classification tasks, they remain valuable in settings with simpler classification needs, stricter data privacy constraints, or limited budgets. These models provide a viable alternative for institutions that prefer not to rely on third-party cloud-based services.

Prompt engineering vs. fine-tuning A notable observation in our analysis was that smaller fine-tuned models achieved performance comparable to larger prompt-engineered models on several classification tasks. This aligns with findings from prior computational research, which suggest that while fine-tuning allows for task-specific adaptation, its benefits may be constrained by the underlying model capacity[134]. Conversely, larger prompt-engineered models can leverage broad pre-trained knowledge and perform well without full retraining. These results support a balanced interpretation: a more targeted learning strategy (fine-tuning) applied to a smaller model may yield comparable results to a less customized strategy (prompt engineering) implemented with a larger model. Additionally, it was shown that fine-tuning is more susceptible to class imbalance, which can lead to overfitting if not properly mitigated. This was particularly evident in the fine-tuned open-source model, which learned to classify all aortic valves as non-dysfunctional; a result that maximized performance on the imbalanced training set but failed to generalize. However, this holds true for all use cases with highly imbalanced data. In comparison, prompt-engineered approaches demonstrated greater resilience to label sparsity, suggesting an advantage in scenarios with limited or unevenly distributed annotations. From a practical perspective, this trade-off highlights the importance of aligning methodological choices with both task complexity and available infrastructure. While fine-tuning offers the potential for precise task optimization, it demands considerable computational resources, annotated training data and technical expertise. In contrast, prompt engineering, particularly when applied to large-scale foundation models, offers a more accessible and scalable alternative, especially in structured domains such as diagnostic and procedural cardiology reporting.

The dataset used for training and evaluation reflects a typical ACS population, representative both in patient characteristics and disease distribution.[135]. This real-world alignment enhances the external validity of our findings. However, as a consequence the dataset exhibits substantial class imbalance, mirroring the natural prevalence of clinical findings in ACS, which poses challenges for model training, particularly for fine-tuning strategies. While prompt-based approaches can partially mitigate these effects through tai-

lored task formulations, fine-tuned models remain more susceptible to under performance on underrepresented classes.

Limitations and generalizability Furthermore, overall human annotator scores were high, but lower for more complex labels such as culprit vessel and occlusion, likely due to the ambiguous and complex nature of these annotations. During ICA procedures, the culprit lesion may not be unequivocally identified, particularly in cases of NSTEMI, multivessel disease or diffuse atherosclerosis [136]. Labeling sub-branches of the culprit vessel was even more ambiguous, due to complex anatomical relations and lesions affecting multiple branches at the same time. These ambiguities, exacerbated by the multi-class and multi-label nature of vessel-related classifications, make for highly complex classification tasks. Additionally, a notable proportion of the dataset included patients with prior CABG or CTOs, introducing further anatomical variation and interpretation challenges.

This study primarily focused on moderately structured clinical text, due to the procedural nature of analyzed reports, which typically contain more templated or semi-standardized phrasing. Extending these methods to unstructured clinical narratives, such as clinical rounding notes or discharge summaries, will likely pose greater challenges due to their variability and contextual complexity. Nonetheless, reports used in this analysis were created by numerous different cardiologists, introducing heterogeneity in structure and phrasing.

We anticipate that similar LLM-based strategies can effectively be extended to less structured clinical text, provided that more extensive prompt engineering and carefully defined output schemas are employed to handle the greater variability. Moreover, while our datasets were sourced from a single academic institution, the underlying LLM-based approaches are inherently portable. They can be adapted to local data environments, allowing for institution-specific customization while leveraging generalizable architectures and workflows.

Future Directions Future research should prioritize the development of advanced prompt optimization strategies, including automated prompt generation and hybrid frameworks that integrate prompt engineering with minimal fine-tuning. These approaches may offer a more efficient balance between performance and resource demands, particularly in settings with limited annotated data. In addition, performance could be further enhanced by incorporating post-processing techniques, such as rule-based corrections or ensemble methods, that aggregate outputs from multiple models to enhance reliability.The applica-

tion of LLM-based data extraction to less structured clinical texts, such as progress notes and discharge summaries, warrants further investigation, as these sources contain rich contextual information but pose greater linguistic and structural challenges.

Given the rapid evolution of LLM architectures and hardware, with new models showing both similar performance with a lower parameter count and superior performance with similar or higher parameter counts, we anticipate continuous improvement in performance, accessibility and applicability in clinical cardiology settings[137].

Conclusion In conclusion, both fine-tuning and prompt engineering approaches to LLMs offer valuable tools for the structured classification of cardiology reports. Prompt engineering provides a lightweight, adaptable and cost-efficient strategy, particularly suitable for low-resource settings. Fine-tuning, while resource-intensive, enables more targeted optimization. While commercial LLMs generally outperform open-source models in challenging classification tasks, locally deployed models achieved good performance for more structured or narrowly defined applications. The choice between approaches should be guided by the clinical use case, available infrastructure, local expertise and regulatory or privacy considerations. Critically, the traditional bottleneck of manual data labeling is rapidly becoming obsolete. With LLMs now capable of accurate, scalable information extraction from raw clinical text, the development of machine learning pipelines no longer hinges on large annotated datasets. This shift unlocks new opportunities for rapid deployment of AI in cardiology, from real-time decision support to large-scale data curation. As the technology continues to evolve, LLMs are set to become foundational infrastructure for cardiovascular research and clinical practice alike.

3.5 Sources of Funding

This project has received funding from the Innovative Health Initiative Joint Undertaking under grant agreement 101112022 (iCARE4CVD).

3.6 Disclosures

None

Appendix

Label distributions

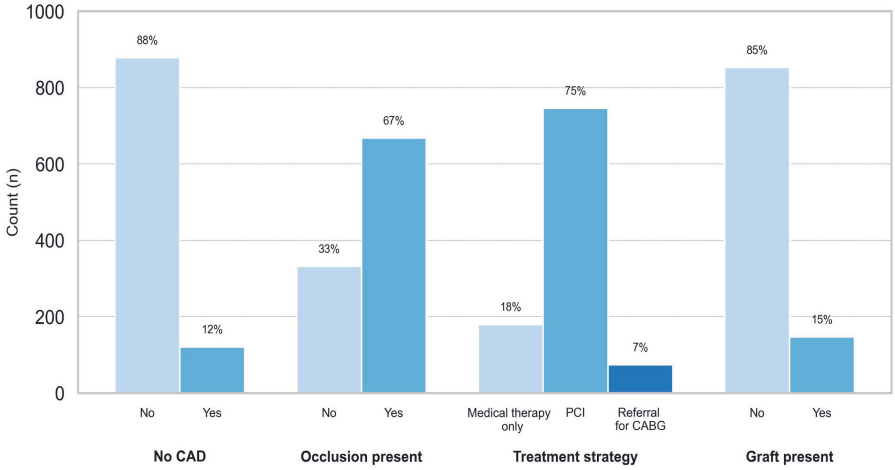


Figure 3.4: Distribution of the manually assigned single-label labels for the ICA dataset

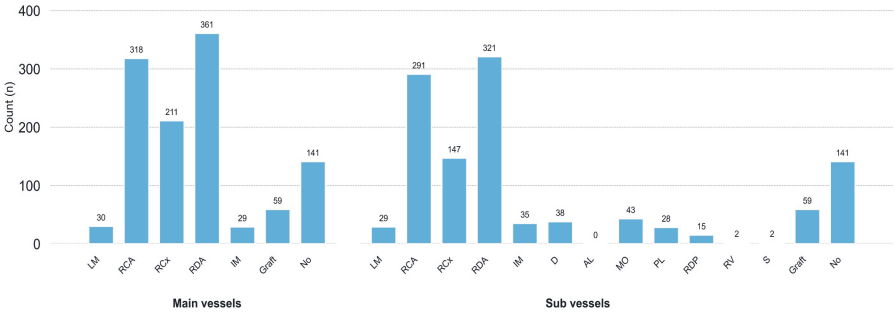


Figure 3.5: Distribution of the manually assigned vessel labels

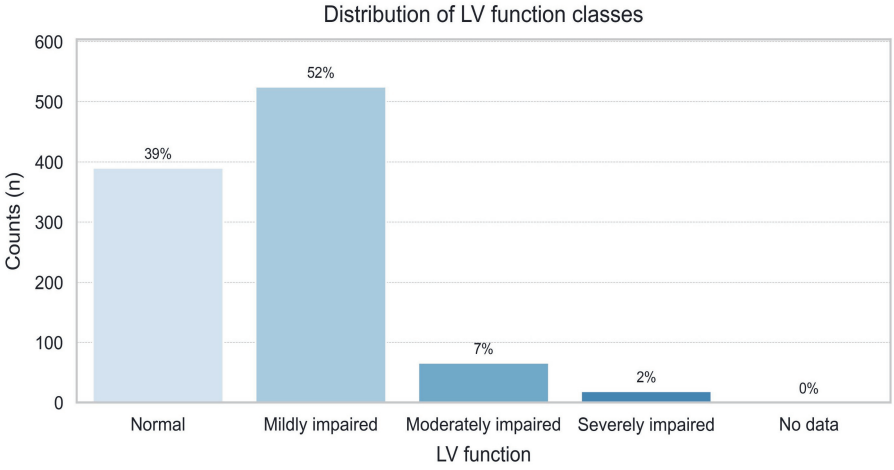


Figure 3.6: Distribution of the manually assigned categorical LV function labels

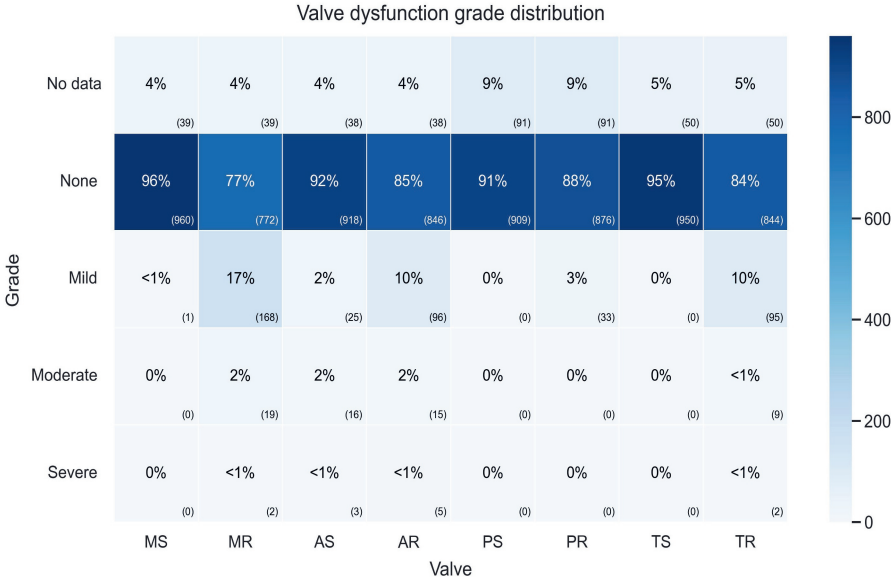


Figure 3.7: Distribution of the manually assigned categorical valve dysfunction labels



Figure 3.8: Model comparison for the finetuning method on the TTE reports. Mean and 95% CI of 1000 bootstrap samples are shown. Results are aggregated metrics of 5 fold cross-validation on the training set.

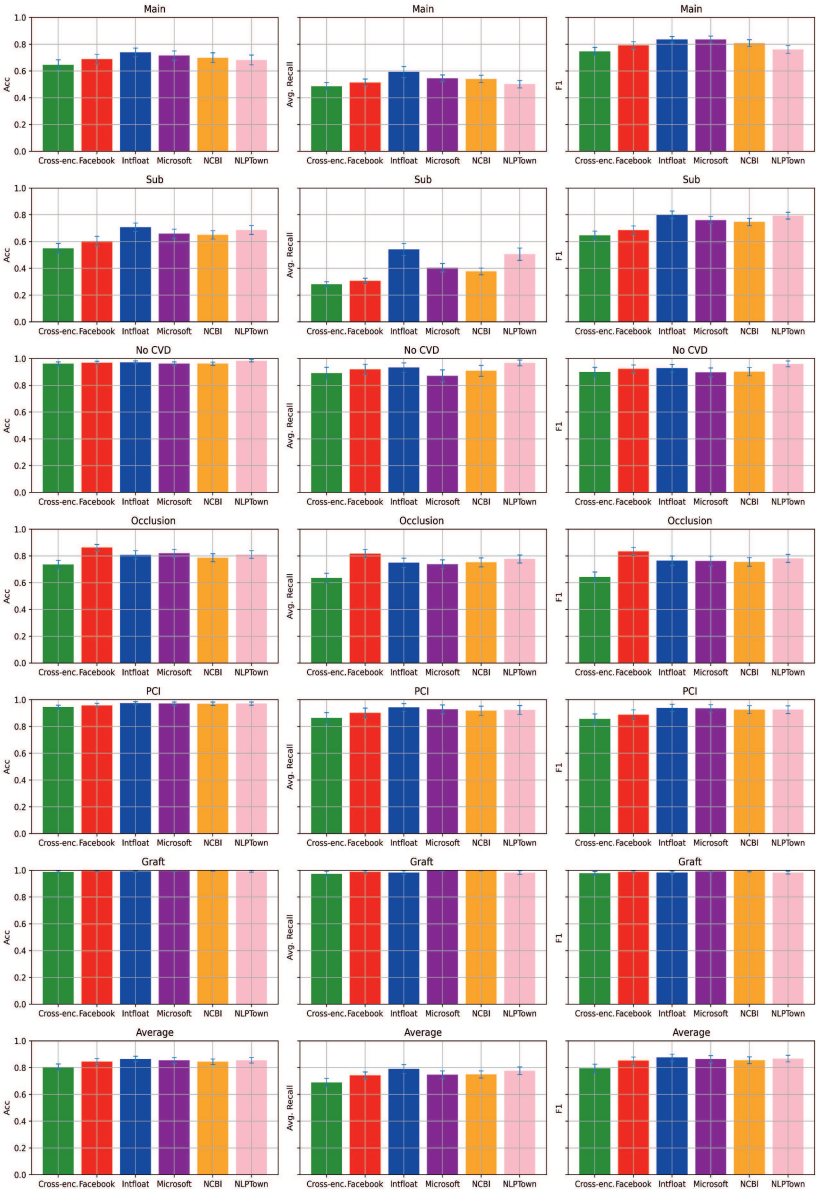


Figure 3.9: Model comparison for the finetuning method on the ICA reports. Mean and 95% CI of 1000 bootstrap samples are shown. Results are aggregated metrics of 5 fold cross-validation on the training set.



Figure 3.10: Model comparison for local prompt engineering on TTE reports. Metrics show aggregated values of 5-fold cross-validation on trainings set. 95% CI of metrics calculated with 1000 bootstrapped samples is indicated with error bars. Sten. = stenosis, reg. = regurgitation

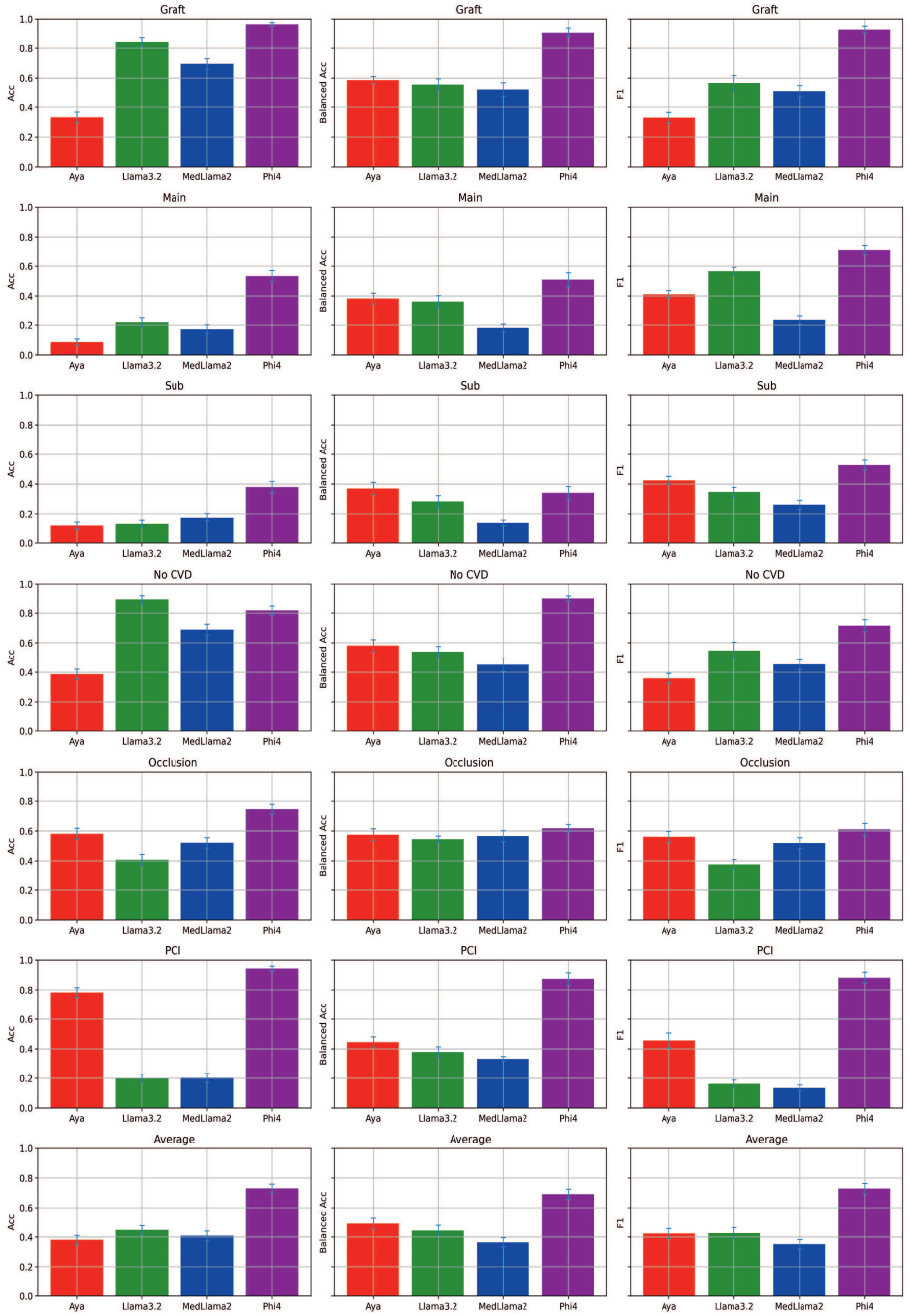


Figure 3.11: Model comparison for local prompt engineering on ICA reports. Metrics show aggregated values of 5-fold cross-validation on trainings set. 95% CI of metrics calculated with 1000 bootstrapped samples is indicated with error bars.

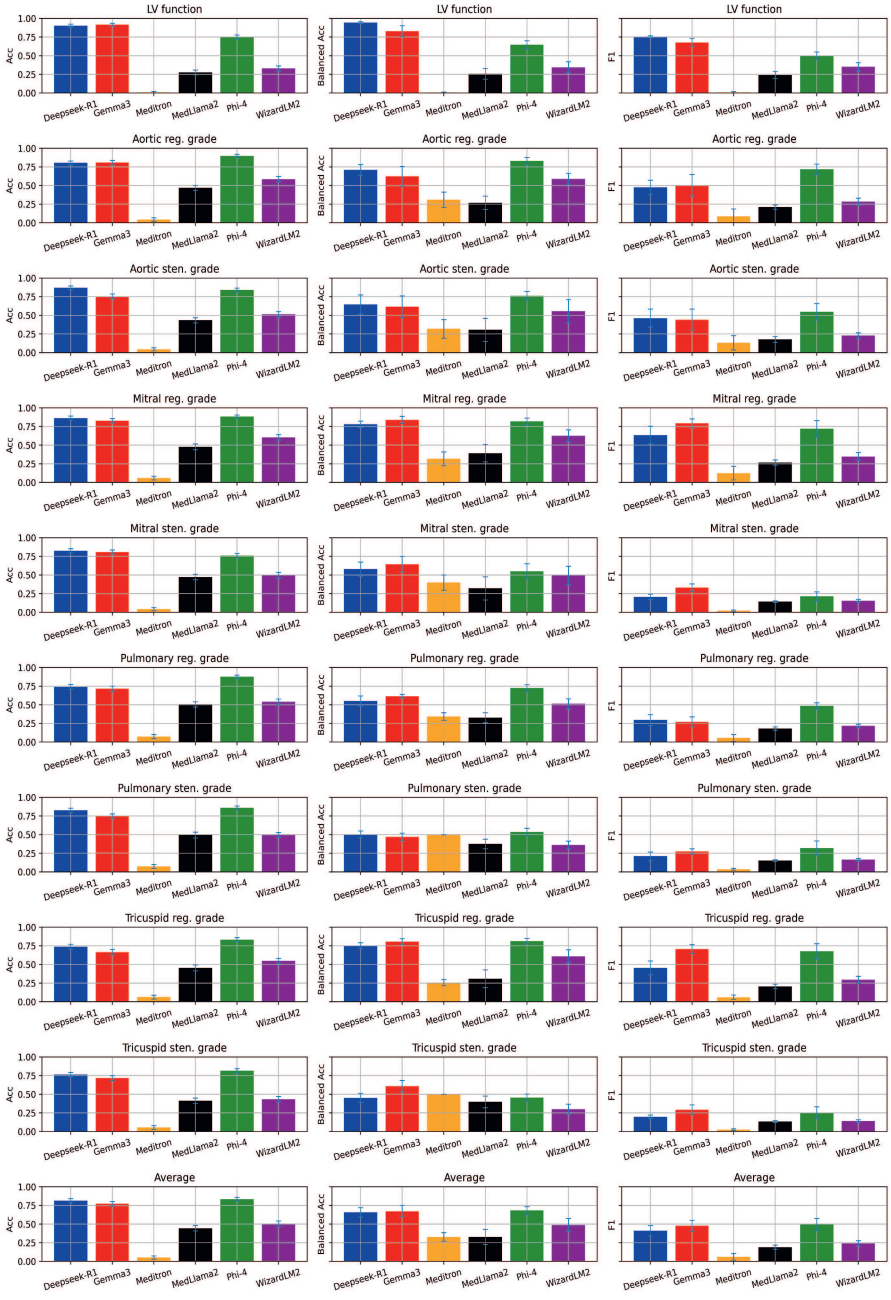


Figure 3.12: Model comparison for HPC prompt engineering on TTE reports. Metrics show aggregated values of 5-fold cross-validation on trainings set. 95% CI of metrics calculated with 1000 bootstrapped samples is indicated with error bars. Sten. = stenosis, reg. = regurgitation

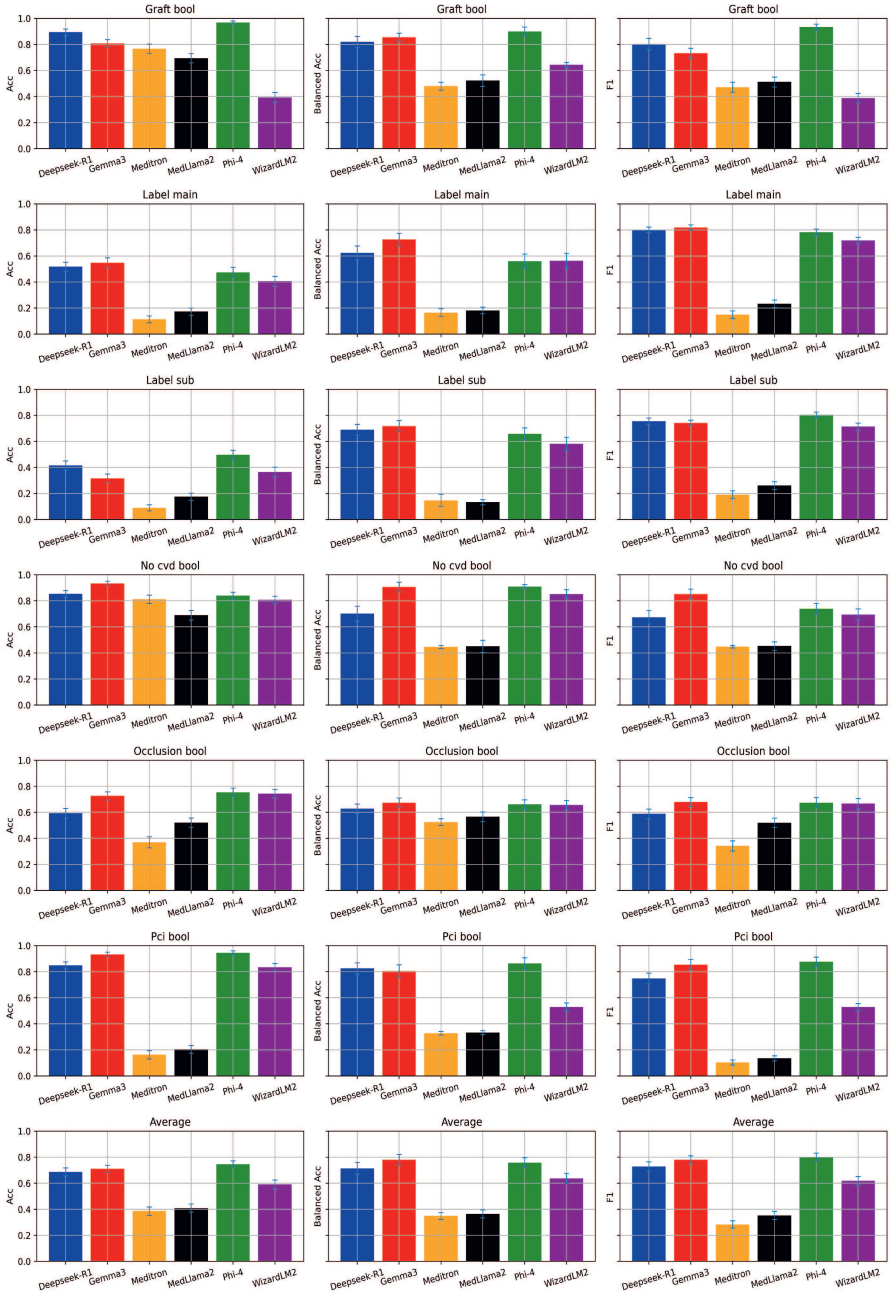


Figure 3.13: Model comparison for HPC prompt engineering on ICA reports. Metrics show aggregated values of 5-fold cross-validation on trainings set. 95% CI of metrics calculated with 1000 bootstrapped samples is indicated with error bars.

4

Joint optimization of a β-VAE for ECG task-specific feature extraction

Van der Valk, V.O., Atsma, D.E., Scherptong, R.W.C., Staring, M. (2023). Joint Optimization of a β-VAE for ECG Task-Specific Feature Extraction in Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, pages 554–563, Cham, 2023, Springer Nature Switzerland.

Abstract

Electrocardiography is the most common method to investigate the condition of the heart through the observation of cardiac rhythm and electrical activity, for both diagnosis and monitoring purposes. Analysis of electrocardiograms (ECGs) is commonly performed through the investigation of specific patterns, which are visually recognizable by trained physicians and are known to reflect cardiac (dis)function. In this work we study the use of β-variational autoencoders (VAEs) as an explainable feature extractor, and improve on its predictive capacities by jointly optimizing signal reconstruction and cardiac function prediction. The extracted features are then used for cardiac function prediction using logistic regression. The method is trained and tested on data from 7255 patients, who were treated for acute coronary syndrome at the Leiden University Medical Center between 2010 and 2021. The results show that our method significantly improved prediction and explainability compared to a vanilla β-VAE, while still yielding similar reconstruction performance.

4.1 Introduction

The electrocardiogram (ECG), is one of the most widely used methods to analyze cardiac morphology and function, by measuring the electrical signal from the heart with multiple electrodes. ECG data is used by clinicians for both diagnostic and monitoring purposes in various cardiac syndromes. A 12-lead ECG is routinely obtained in patients to diagnose and monitor disease development. However, for the interpretation of the ECG signal, the knowledge of an expert is required. Physicians usually analyze the ECG through the recognition of specific patterns, known to be associated with disease. This however requires substantial expertise, and potentially additional relevant information exists in a 12-lead ECG missed by human interpretation. Deep learning has already proven its usefulness in the interpretation of the ECG signal in multiple classification challenges [138, 139] and more recently also in feature discovery by means of explainable AI algorithms [140, 141, 142, 143, 144]. The explainablity of AI algorithms is especially valued in medical settings, where trusting a black box AI algorithm is undesirable[139].

VAEs and in particular β-VAEs have been used as unsupervised explainable ECG feature generators in the explainable AI algorithms mentioned above [145]. It was shown that a β-VAE, trained on reconstruction of the ECG signal, is able to extract features from the ECG signal that can be made more interpretable by visualization of reconstructed latent space samples with the decoder of the β-VAE [140]. This is a first step towards an explainable deep learning pipeline for ECG analysis. However, the features generated by a β-VAE when only trained to minimized reconstruction loss, are likely not optimal for task specific predictions.

The aim of this chapter is to explore further improvement of the latent features by improving their explainability and prediction performance. This is clinically relevant but unexplored for the post myocardial infarction setting. We propose to improve explainability by reducing the dimension of the latent space to a level more manageable for human assessment, while encouraging outcome specific information to be captured in a small part of the latent space, and while maintaining ECG reconstruction performance for visual assessment. To achieve this, we propose a novel method to jointly optimize the β-VAE with a combination of a task specific prediction loss for a subset of the latent space, and KL-divergence and reconstruction loss for the entire latent space. The task chosen to optimize here is left ventricular function (LVF), one of the most important determinants of prognosis

in patients with cardiac disease. Current assessment of LVF requires advanced imaging methods and interpretation by a trained professional. The ECG, on the other hand, can be obtained by a patient at home. In combination with automated analysis this would facilitate remote monitoring of LVF in patients.

4.2 Methods

4.2.1 Data

To train the models for both reconstruction and LVF prediction, two datasets were used: i) A non-labeled dataset consisting of 119,886 raw 10s 12-lead ECG signals taken at 500Hz from 7255 patients diagnosed with acute coronary syndrome between 2010 and 2021 at the Leiden University Medical Center, the Netherlands; ii) A labeled dataset of 33,610 ECGs from 2736 patients of the same cohort. This dataset was labeled by visual assessment of an echocardiogram performed within 3 days before or after the ECG. The label categories, normal, mild, moderate and severe impairment were binarized for model training. When the ECG was taken within two weeks after cardiac intervention a 1-day margin was used. If a cardiac intervention was performed between ECG and echocardiography, the case was excluded. 11.5% of the ECGs were labeled with a moderate to severe impaired LVF. The institutional review board approved the study protocol (nWMODIV2_2022006) and waived the obligation to obtain informed consent.

4.2.2 Data preprocessing

The raw ECG signals were first split into separate heartbeats (400ms before and after the R-peak, the largest peak in the ECG, that represents depolarization of the ventricles) with a peak detection method inspired by RPNet, a U-Net structured CNN with inception blocks, that was trained on manually labeled peak locations [146]. The heartbeats were then filtered with a magnitude and an autocorrelation filter. The magnitude filter removed heartbeats with an average magnitude below a set threshold. The autocorrelation filter removed signals where both the mean and maximum autocorrelation between the heartbeats were below a set threshold. These two criteria were used, since ECG signals showing multiple rhythms can result in low mean autocorrelation, but, if not noisy, will not result in low maximum autocorrelation. The remaining heartbeats were then averaged per ECG lead. The μ and σ of the intervals of the subsequent R peaks were used as an additional feature for LVF prediction.

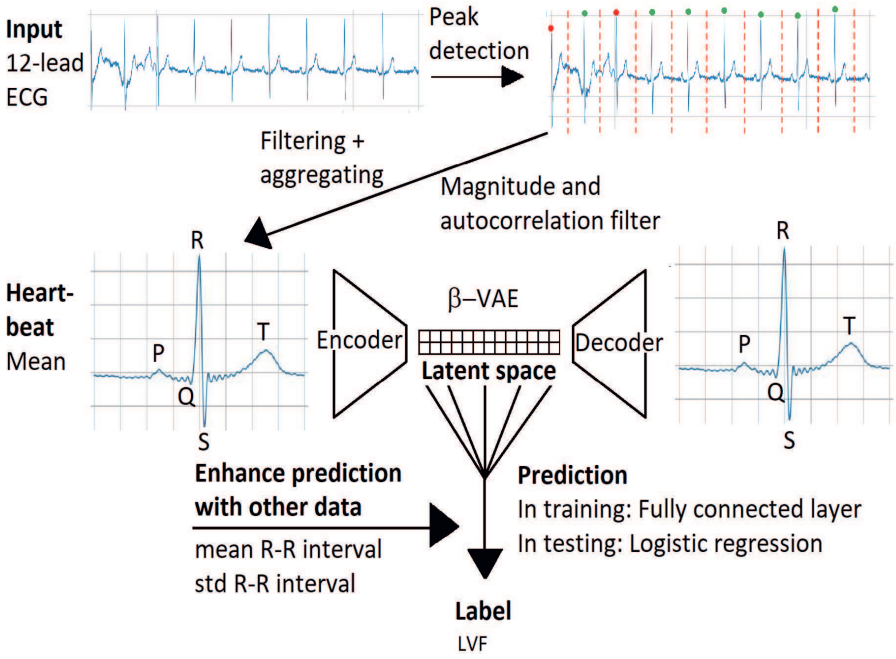


Figure 4.1: Preprocessing, feature extraction and prediction pipeline.

4.2.3 Model overview

To investigate a general improvement to the VAE feature extraction pipeline [140, 141, 142, 143], the proposed method was tested with two architectures: i) A small VAE with 300k parameters consisting of an encoder and a mirrored decoder. Both parts contained 7 2D convolutional layers, of which 3 were residual layers, with respective channel sizes of [8,16,32,64,64,64,64] and a kernel size of 5; ii) A second larger VAE from the FactorECG pipeline as proposed by Van de Leur *et al.* (2022) [140] with 50M parameters. The VAEs were both extended at the bottleneck (the latent space, of size L), with a single fully connected layer for output prediction, in this case the LVF label, see Fig. 6.2. The μ and σ of the RR intervals (time between two subsequent R peaks), were added to the input of the prediction layer, since the information represented by these features is lost in averaging the heartbeats. To maintain explainability of the extracted features, only one fully connected layer is used, as otherwise the features will become weighted combinations of the latent space values, which makes visualization with the decoder and subsequent interpretation complex. However, for pure prediction performance, additional fully connected layers may have been helpful. The extracted features, again with the μ and σ of the RR interval, were

subsequently analyzed with logistic regression using regularized l1 and l2 penalties on the LVF prediction task, ignoring the output of the prediction layer in the VAE. The VAEs were build and trained in the PyTorch 1.12 framework and trained on a Quadro RTX 6000 GPU with CUDA 11.4 [147, 148], while for logistic regression we used the Scikit-learn toolbox [149]. The implementation of our models will be made publicly available via GitHub at `https://github.com/ViktorvdValk/Task-Specific-VAE`.

4.2.4 Model training

The β-VAE was first pretrained in a self-supervised manner with the mean heartbeats of all filtered ECG signals, minimizing i) the mean squared reconstruction error (MSE) between the input and output ECG, and ii) the KL-divergence between the output of the encoder and the standard normal distribution. The KL-divergence loss was weighted with a β factor, like in the original paper [145]. This pretrained VAE was then fine-tuned in two-steps, first the encoder and decoder were fixed, and only the prediction layer was trained, then all layers were trained end-to-end. This training scheme was used to ensure more stable training. For these fine-tuning steps, the loss function was complemented with a binary cross-entropy loss, which was weighted with a γ factor. The *task naive* VAE resulting from pretraining was compared to the *task specific* VAE resulting from both fine-tuning steps. For pretraining, both datasets were combined and split in a training (85%) and a test set (15%). 5-fold cross validation was done with the training set with again an 85%:15% ratio between training and validation set. For fine-tuning, the same procedure was used on just the labeled dataset, making sure labeled ECGs were in the same set in both cases. All data splits were grouped by patient and stratified by label in case of the labeled data splits. Both pretraining and fine-tuning were done until convergence, i.e. until the loss on the validation set stopped improving for 25 epochs. This was done to prevent the advantage of additional training of the *task specific* network. To prevent overfitting, balanced sampling and regularization by means of drop out layers and the Adam optimizer with weight decay were used, this was especially necessary in the fine-tuning phase. To prevent gradient explosion, gradient clipping and He initialization were used [150].

4.2.5 Feature evaluation

The differences between the features from the *task naive* and *task specific* VAEs, were compared w.r.t.reconstruction and prediction. For reconstruction, both MSE and correlation between input and output ECG, and for prediction the Area Under the Receiver Operator

Characteristic Curve (AUROC) and the macro-averaged F1 score were used. Significant difference between AUROC scores was calculated as proposed in Hanley & McNeil (1983) [151]. For visualization of the representation of a latent space feature f in a so called factor traversal, all features except f were sampled at their mean, while f was sampled in a range between $\mu - 3\sigma$ and $\mu + 3\sigma$. Using these samples as input for the decoder, creates a representation of that feature, which can give insight in ECG features that are important for LVF prediction.

4.2.6 Baseline methods

As a baseline method, a principal component analysis (PCA) was performed on the pre-processed ECGs, to extract features. PCA can be considered an ordered task naive linear feature extractor that focuses on the axis of the largest variance, in contrast to the VAEs which are non-ordered non-linear feature extractors, that are optimized for reconstruction. A logistic regression predictor with just sex and age as input was used as an additional baseline.

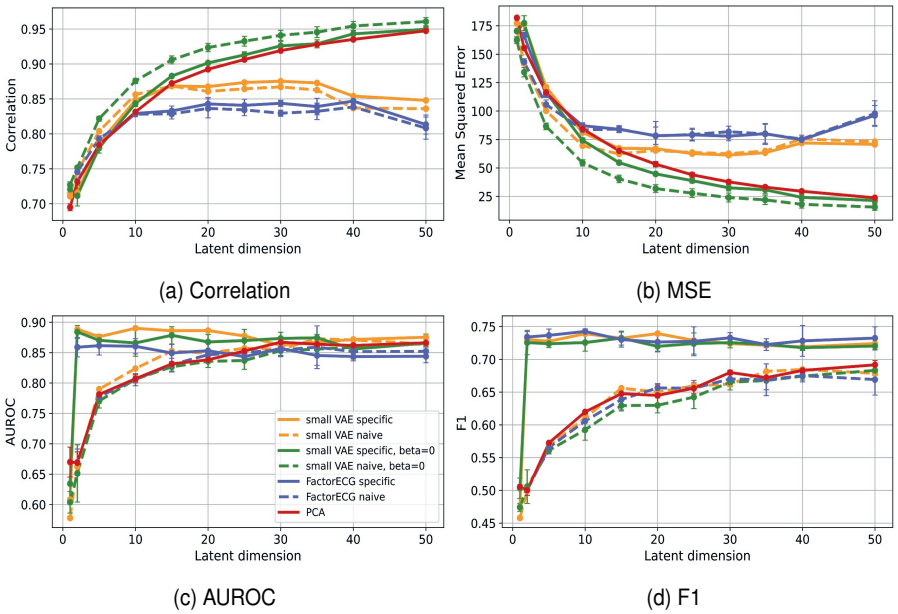
4.3 Experiments and Results

4.3.1 Experiments

The proposed pipeline contains several hyper-parameters, of which the latent space size L was optimized in this study. The influence of the β parameter was also briefly addressed. L was optimized for its importance in the explainability and the reconstruction and prediction quality of the model. A higher L increases the complexity of the model, and consequently decreases its explainability. An L that is too low, on the other hand, restricts the capacity of the model for reconstruction and prediction. The PCA baseline method was considered to give an upper bound of L, since the number of principal components, the PCA analog for L, indicates how many values would be needed to capture sufficient information.

4.3.2 Hyperparameter optimization

The influences of γ on prediction and reconstruction performance was small and was therefore fixed to 500. The influence of L on prediction quality can be seen in Fig. 4.2. The PCA baseline performs more or less equal to the *task naive* networks for all L. For the *task specific* networks, the F1 scores are higher than their *task naive* counterparts and the PCA baseline, especially for lower L. The *task specific* VAEs already reached their best prediction performance starting at L=2, as compared to the *task naive* VAEs and the



(a) Correlation (b) MSE

(c) AUROC (d) F1

Figure 4.2: Influence of the latent dimension L on reconstruction quality: (a) correlation and (b) MSE, and on prediction quality: (c) AUROC and (d) F1-score, for various models. Plotted are the mean and standard deviation of 5-fold cross-validation on the validation set.

PCA baseline that reach their best prediction performance from $L = 30$. The influence of L on reconstruction can be seen in Fig. 4.2a and 4.2b. All networks perform equal to the PCA baseline for low latent dimensions. The reconstruction for the small VAE and the FactorECG VAE does not seems to improve any further for respectively $L >$20 and $L >$ 15, where the PCA baseline reconstruction keeps improving with L. However, setting β to 0 and thereby ablating the variational nature of the VAEs prevents this stagnation of reconstruction performance. The *task specific* networks perform equally well as their *task naive* counterparts, which suggests that the additional joint optimization does not have a major negative impact on reconstruction. The optimization shows that the relevant information for LVF prediction in the ECG signal can be captured in just two features by both VAEs. Reconstruction, on the other hand, requires at least 10/15 features for the VAEs to reach maximum performance. Therefore, in another experiment, a *split task* VAE was trained, in which 8 of the latent space features where only optimized for reconstruction and only 2 also for prediction.

4.3.3 Results on the test set

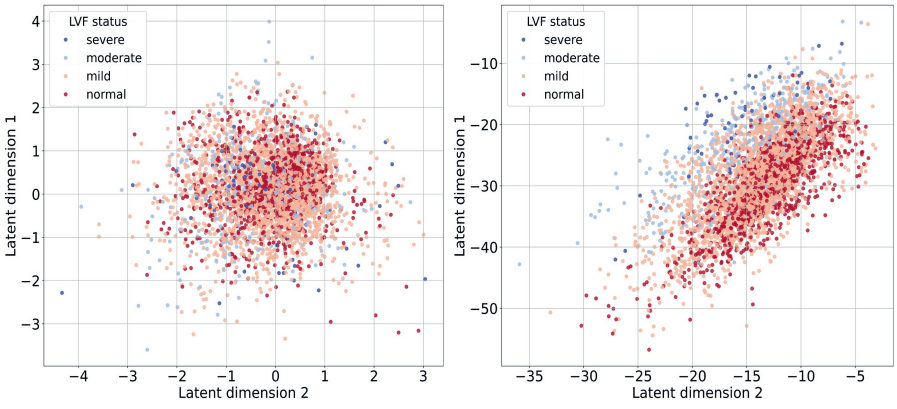
Table 4.1 shows the results on the test set for L=2 and L=10. The *(split) task specific* networks significantly outperform their *task naive* counterparts, the PCA baseline, and the sex and age benchmark w.r.t. LVF prediction. Fig. 4.3a, 4.3b and 4.3c show that the *split task*, in contrast to the *task naive* small VAE with β=0 and β=4 can be used to encode the ECG signals to a landscape that visually separates the signals based on LVF status reasonably well. The factor traversals in Fig. 4.4b and 4.4a show an example of the interpretation of the latent features. Setting β to 0, creates features that appear visually less informative.

Architecture	L	MSE	Correlation	AUROC	F1
Sex and age	2	-	-	0.556 (0.520-592)	0.474
PCA	2	147	0.724	0.656 (0.624-0.688)	0.496
Small VAE task naive	2	133	0.739	0.686 (0.655-0.716)	0.503
Small VAE task specific	2	164	0.711	0.842* (0.822-0.861)	0.682
Small VAE split task	2 (10)	76.5	0.838	0.839 (0.819-0.859)	0.695
Small VAE split task $\beta = 0$	2 (10)	73.6	0.838	0.846 (0.823 -0.862)	0.674
FactorECG [140] task naive	2	139	0.735	0.685 (0.654-0.715)	0.507
FactorECG [140] task specific	2	161	0.724	0.770* (0.745-0.796)	0.695
PCA	10	77.2	0.826	0.761 (0.735-0.787)	0.580
Small VAE task naive	10	63.1	0.854	0.803 (0.781-0.826)	0.586
Small VAE task specific	10	70.6	0.847	0.853* (0.834-0.871)	0.679
FactorECG [140] task naive	10	84.6	0.820	0.770 (0.745-0.796)	0.579
FactorECG [140] task specific	10	87.2	0.822	0.833* (0.813-0.854)	0.707

Table 4.1: Reconstruction and LVF prediction comparison on the test set for the *task naive* and *task specific* architectures. The results show the μ of 5-fold cross-validation. AUROC is shown with its 95% confidence interval.* p-value < 0.01 between AUROC of task naive and specific method for all folds.

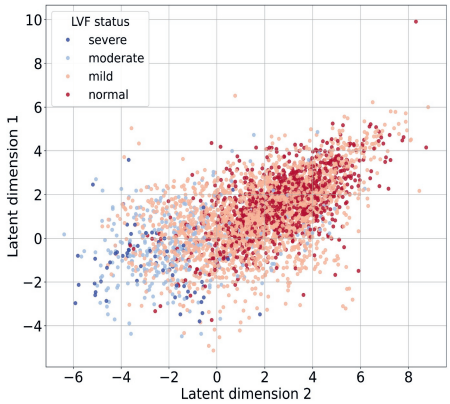
4.4 Discussion

Joint optimization of a β-VAE successfully generated features that contain more information about LVF, without hampering reconstruction of the ECG signal. We hypothesize that the β-VAEs have multiple optima for ECG reconstruction of which only some generate features that are relevant for LVF prediction. This study shows that joint optimization will favor this desired subset of optima, and that this is true for different architectures. In addition, we showed that jointly optimizing only a subset of the latent space features for prediction,



(a) *task naive* with β=4

(b) *split task* β=0



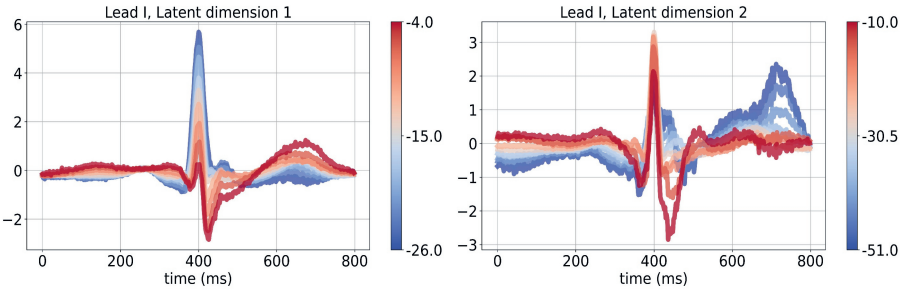
(c) *split task* β=4

Figure 4.3: Comparison of the latent space for different values of β, for the small VAE. For *task specific* methods, the scatter plots show the two dimensions of the latent space that are optimized for prediction: (a) *task naive* (β=4); (b) *split task* (β=0); (c) *split task* ($\beta = 4$).

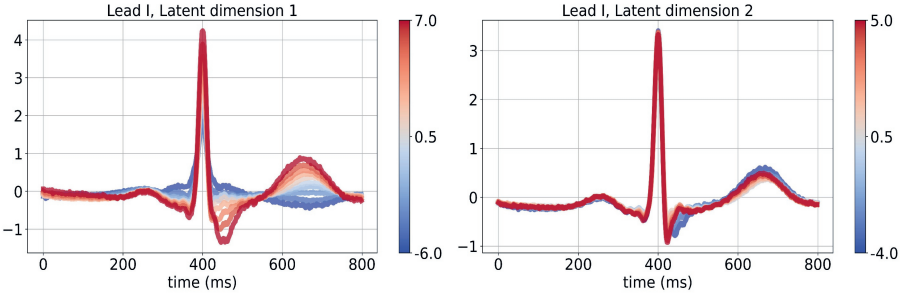
results in aggregation of the predictive information, thereby improving explainability.

The AUROC score of the FactorECG VAE prediction is similar when compared to van der Leur *et al.* (2022) [140] (AUROC≈0.9 for L=36). However, the proposed small VAE achieved equal if not better reconstruction and prediction performance with less than 1% of the parameters as shown in Fig. 4.2.

The F1 score is considered more robust than the AUROC score with data imbalance, which is the case here [152]. From Fig 4.2d we can therefore conclude that the *task specific* networks outperform the *task naive* networks for any L. The differences between the *task specific* networks and their *task naive* versions in prediction, at similar reconstruction,



(a) Factor traversals for β=0



(b) Factor traversals for β=4

Figure 4.4: Comparison of the latent space for different values of β, for the small VAE. The latent space factor traversals (a) and (b) show the visual representation of the features for Lead I of the 12-lead ECG signal: (a) β=0; (b) $\beta = 4$

indicate that the ECG signal can be summarized with a set of latent features of which only a subset is important for LVF prediction. The joint optimization promotes the extraction of this subset especially when L is small. Figs. 4.2b and 4.2a show that the PCA baseline outperforms both VAEs in reconstruction for $L > 20$ when β=4, but not for β=0. This indicates that the VAEs are restricted in reconstruction by the KL-divergence loss. This loss was shown to promote feature disentanglement and a gradient in the latent space [153]. Figs 4.4a and 4.4b show that without this loss (β=0) the latent features are more complex to interpret. This could be explained as a reduction of the disentanglement of the features resulting from the absence of the KL-divergence loss. However, Figs 4.3c and 4.3b both show a gradient in the latent space, which suggests that the prediction loss on its own also promotes a gradient in the latent space. Moreover, Fig 4.3c shows dependence, and thus a lack of disentanglement, between the latent features even when β=4. This complex interplay between the three losses used in the joint optimization, is very relevant for the explainability aspect of this method, but beyond the scope of the current study. We aim

to examine the complex interplay in future work. In conclusion, the proposed joint optimization improves both explainability and prediction performance of VAEs by extraction of a smaller set of LVF specific features from the ECG signal. This could reduce the need of more advanced imaging methods, currently needed to measure the LVF. This opens the way for remote monitoring of left ventricular function in patients.

Acknowledgments

This project received funding from the EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 860173.

5

Explainable ECG analysis by explicit information disentanglement with VAEs

Van der Valk, V.O., Atsma, D.E., Scherptong, R.W.C., Staring, M. (2025) Explainable ECG analysis by explicit information disentanglement with VAEs. IEEE Transactions on Bio-Medical Engineering.

Abstract

Objective: The interpretation of electrocardiogram (ECG) signals is vital for diagnosis of cardiac conditions. Traditional methods rely on expert knowledge, which is time consuming, costly and potentially misses subtle features. AI has shown promise in ECG interpretation, but clinically desired model explainability is often lacking in literature.

Methods: We introduce an explainable AI method for ECG classification by partitioning the variational autoencoder (VAE) latent space into a label-specific and a non-label-specific subset. By optimizing both subsets for signal reconstruction and one subset also for prediction while constraining the other from learning label-specific information with an adversarial network, the latent space is disentangled in a supervised manner. This latent space is leveraged to create enhanced visualizations for ECG feature interpretation by means of attribute manipulation. As a proof of concept, we predict the left ventricular function (LVF), a critical prognostic determinant in cardiac disease, from the ECG.

Results: Our study demonstrates the effective segregation of LVF-specific information within a single dimension of the VAE latent space, without compromising classification performance. We show that the proposed model improves state-of-the-art VAE methods (AUC 0.832 vs. 0.790, F1 0.688 vs. 0.640) in prediction and performs comparable to ground truth LVF (concordance 0.72 vs.0.72) in predicting survival.

Conclusion: The model facilitates the interpretation of LVF predictions by providing visual context to ECG signals, offering a general explainable and predictive AI method.

Significance: Our explainable AI model can potentially reduce time and expertise required for ECG analysis.

5.1 Introduction

The electrocardiogram (ECG) stands as a widely employed method for evaluating cardiac morphology and function by capturing the heart's electrical signals with multiple electrodes. Clinicians rely on ECG data for diagnostic and monitoring purposes in various cardiac syndromes, often obtaining a 12-lead ECG as standard practice for disease diagnosis and progression tracking. However, the interpretation of ECG signals traditionally requires expert knowledge, where physicians identify specific patterns associated with disease.

Despite the proficiency of expert analysis, certain crucial information within a 12-lead ECG may elude human interpretation, prompting the exploration of alternative methodologies. Deep learning has demonstrated its efficacy in interpreting ECG signals for various

classification tasks, among others, atrial fibrillation, tachycardia, and bradycardia detection [138, 139]. Notably, recent advances in explainable artificial intelligence (AI) algorithms have increased the ability to reveal complex features within ECG signals [140, 141, 142, 143]. In a medical setting, the transparency and interpretability of AI algorithms are of paramount importance, given the necessity to understand and trust decision-making processes [139].

In this context, β-VAEs [145] have emerged as unsupervised and explainable feature generators for ECG analysis. Studies have shown that a β-VAE trained to reconstruct ECG signals can extract features that become more interpretable by visualizing the reconstructed latent space with the decoder of the β-VAE [140], a process called attribute manipulation. Analysis of these features resulted in a subset that was shown to be predictive of cardiac function, several of which were similar to known indicators of cardiac health. However, features generated for reconstruction purposes only may not be optimally suited for label-specific predictions, such as cardiac function estimation.

This paper aims to enhance the specificity and predictive performance of latent features derived from ECG signals. The task chosen for this optimization is the assessment of left ventricular function (LVF), a critical prognostic determinant in cardiac disease [9]. Traditionally, LVF assessment necessitates advanced imaging techniques and expert interpretation, however it was shown that ECG signals contain information on LVF as well [154]. A clear understanding of the relationship between LVF and ECG characteristics could facilitate patient monitoring. Thereby, derivatives of the ECG signal show promise in remote monitoring with smart devices. Leveraging these two could enable remote monitoring of LVF in patients, especially in home-based settings.

We propose a method to partition the latent space into label-specific and non-label-specific parts, aiming to enhance the specificity of the latent features. By optimizing both part for signal reconstruction and only one for label (LVF) prediction, while constraining the other part from learning label-specific information through an adversarial network, we aim to enhance the specificity of these features. By means of attribute manipulation, selective sampling from the label-specific latent space only, this latent space can be translated into ECG features. These features are tailored to each individual ECG and demonstrate the algorithm's classification rationale and put the heartbeat in a population context of heartbeats with varying LVF status, thereby enhancing interpretability.

In summary, this study improves the subdivision of the latent space, resulting in su-

pervised disentanglement of label-specific features, which allow improved visualizations by means of attribute manipulation. These visualisations improve interpretability of label-specific features, by illustrating the classification rationale specific to each case. The features are additionally validated with survival analysis. This research extends previous research presented at the MICCAI 2023 conference [155].

5.1.1 Related work

Representation learning, which involves learning subspaces of data that compress or disentangle input data, is used in both generative and explainable AI (XAI). This dynamic field has produced several prominent algorithms and methods, including but not limited to StyleGANs, VAEs, and contrastive learning [156, 157, 158, 159]. These techniques have also been successfully applied in the medical domain, particularly in ECG analysis [160, 161, 141, 142, 162, 163, 140].

In addition to these, other XAI methods have proven effective in ECG analysis. For a comprehensive review of these methods, we refer to Ayano *et al.* (2022) [164]. Wagner *et al.* (2024) [165] also provide an extensive analysis of various XAI techniques for ECG feature extraction. These studies underscore the importance of deriving clinically relevant insights from AI models applied to ECGs.

This paper builds on the principles of XAI and disentanglement research to enhance the interpretability and usability of representations learned by VAEs. In recent years, VAEs have gained significant attention for their ability to generate and learn meaningful data representations. This section reviews the application of these methods with labeled data for classification tasks, with a particular focus on explainable and ECG classification.

Several studies showed the use of classic VAEs as an unsupervised ECGs feature extractor [154, 141, 142, 162, 163]. These features showed to be predictive for cardiac function, myocardial infarction, mortality and several arrhythmias. Moreover, creating artificial data by sampling from the VAE latent space was shown to be a successful ECG data augmentation strategy [143]. The more recent vector-quantized VAE (VQ-VAE) was also successfully used as an unsupervised feature extractor for arrhythmia classification and data augmentation [166].

However, all the studies mentioned above assume that optimizing features for reconstruction implicitly leads to the optimal aggregation and isolation of label-specific information in these features. This is not the case as was shown in Van der Valk *et al.* (2023) [155].

Moreover, the use of a single latent space from which relevant dimensions are selected after training, a method used in some of the studies mentioned above, has its limitations w.r.t. interpretability. It might allow visualization of ECG features important for label prediction by means of attribute manipulation, but label-specific information will most likely be spread over several dimensions, which possibly show complex interaction. This hampers interpretability of these features.

The use of supervision in VAE training could potentially address these issues. Conditional VAEs (CVAEs) [167], classification autoencoders [168], the Attribute-based Regularized VAE [169], the Task-Specific VAE [155] and especially the conditional subspace VAE [170], all use a form of supervision and share some similarities with the model proposed here. These include the use of labeled data, the supervised structuring of the latent space, and an adversarial component that prevents information from being captured in a non label-specific latent space. Several studies have explored the use of these models in ECG analysis. However, the majority of these studies focused solely on the generation of ECG data, rather than the classification or interpretation of the ECG signal [171, 172, 173, 143, 174]. To our knowledge only one study explored the use of supervision in ECG classification with a VAE. Gyawali *et al.* (2018) [175] classified the origin of ventricular activation in the 12-lead ECG with two siamese CVAEs. They extended the CVAE pipeline with a separate deterministic encoder that learned to predict the label, which was input to the CVAE encoders.

Our approach differs from previous work by explicitly optimizing a VAE architecture with a labeled subspace for classification. We introduce novel elements such as using different KL-divergence loss weights for these subspaces and leveraging the interpretability of VAEs for generating signal-specific rather than population-specific explainable ECG visualizations in the context of classification.

5.2 Methods

5.2.1 Data

The study uses a partially labeled dataset. The unlabeled subset comprises 119,886 raw 10-second, 12-lead ECG signals recorded at a frequency of 500Hz. These signals are obtained from 7,255 patients (71% male, age 64.2 $\pm$ 12.2) between 2010 and 2022 at the Leiden University Medical Center in the Netherlands, who were diagnosed with acute coronary syndrome. The labeled subset set consists of 33,610 labeled ECGs from 2,736

patients within the same cohort. The label, the level of LVF impairment, consists of 4 ordinal categories: normal, mild, moderate, and severe. Each ECG is labeled through visual assessment of echocardiograms conducted within 3 days before or after the ECG recording. A 1-day instead of a 3-day margin is applied when the ECG is obtained within two weeks after cardiac intervention, as the LVF can still change considerably in the first weeks after cardiac intervention. Cases in which a cardiac intervention occurred between ECG and echocardiography are excluded. The relative frequencies of the LVF status were 0.33, 0.56, 0.09 and 0.02 for respectively normal, mild, moderate, and severely impaired LVF. For both data sets, an all-cause mortality indicator is available, at least up to three years after hospital discharge, which was used in survival analysis. The study protocol, identified as nWMODIV2_2022006, received approval from the institutional review board, which waived the requirement to obtain informed consent for the use of ECG data of individual patients.

5.2.2 Data preprocessing

The initial processing of raw ECG signals involves segmentation into individual heartbeats, defined as the 400-ms intervals preceding and following the R-peak, the prominent peak in the ECG denoting ventricular depolarization, see Figure 6.2. All recordings are resting ECGs and the average RR-interval in the dataset is 801ms. Therefore, a 800-ms interval is chosen to strike a balance between selecting too little and thereby missing parts of the heartbeat and selecting too much and including a previous or subsequent beat. This segmentation is achieved through a peak detection method inspired by RPNet, a U-Net-structured CNN with inception blocks. RPNet was previously trained by the authors on manually labeled peak locations, as detailed in Vijayarangan *et al.*, (2020) [146]. The R-peaks are determined in lead I for all leads to prevent misalignment when the peaks are not accurately detected.

The segmented heartbeats can contain noisy or flat signals in one or more of the leads, which is the result of improperly attached electrodes or movement during acquisition. To automatically discard these heartbeats and ensure that only relevant and adequately characterized heartbeats contribute to subsequent analyses and model training, a magnitude and an autocorrelation sieve are applied. The magnitude sieve eliminates heartbeats that show a zero signal by sieving out heartbeats with an average magnitude below 1. The autocorrelation sieve discards signals exhibiting a mean autocorrelation between heartbeats below a predefined threshold. This threshold is set to 0.85, which is determined by visual

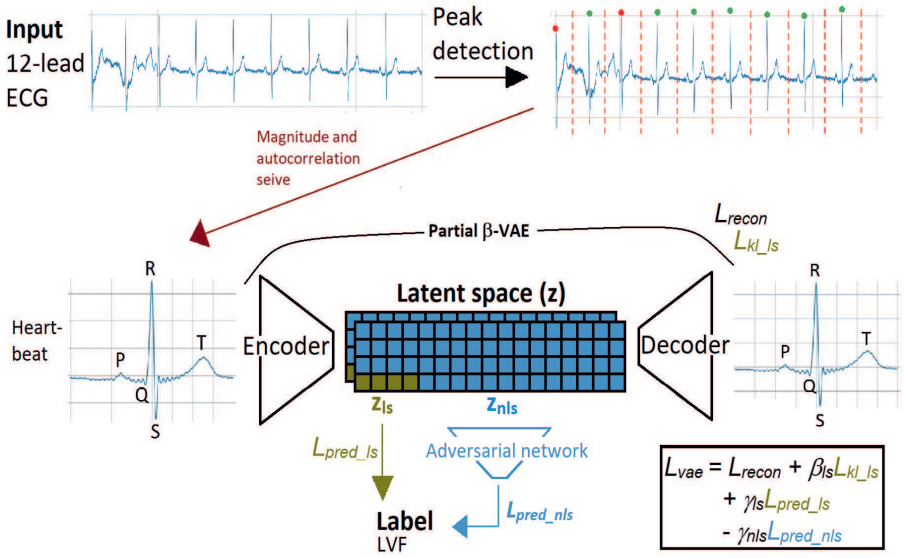


Figure 5.1: **Overview of our Processing and Prediction Framework:** Initially, the 12-lead ECGs are split into individual heartbeats using RPNet[146]. Subsequently, the heartbeats are sieved in two stages: the first based on individual magnitude, identifying low magnitude as indicative of a flat signal, and the second utilizing autocorrelation to assess correlation with other heartbeats from the same signal. The remaining heartbeats (indicated with green dots) are used as input for the VAE. The encoded representation, the latent space z, is divided, comprising a label-specific part, z_{ls} (green part), and non label-specific part, z_{nls} (blue part), i.e. z=[z_{ls}, z_{nls}]. While z_{ls} directly contributes to LVF label prediction with loss $L_{\mathrm{pred\_ls}}$, z_{nls} serves as input for an adversarial MLP also predicting the LVF label, introducing loss $L_{\mathrm{pred\_nls}}$. The decoder's input is sampled from z to produce the reconstructed heartbeat, subsequently compared with the original input heartbeat, resulting in the reconstruction loss L_{recon}.

inspection of the resulting heartbeats. A lower threshold does not remove enough noise signals, whereas a higher threshold removes too many true heartbeats that show naturally occurring variation. ECG signals can potentially show various forms of arrhythmia, which would result in a low autocorrelation between heartbeats. As the aim of this study is to look at the morphology of the heart, these irregular heartbeats are also sieved out. Following the application of these sieves, as a form of data augmentation, a maximum of 5 from the 2-8 retained heartbeats are randomly added to the training set, this number is chosen to prevent over-representation of similar heartbeats from the same ECG signal. Here a heartbeat consists of all 12 leads of the ECG signal and whenever 1 lead shows a loud or flat

signal, the whole heartbeat is discarded, meaning all 12 leads.

5.2.3 Model overview

We introduce an innovative method for the analysis of ECG signals by employing a partial β-VAE, see Figure 6.2. We propose an extension of the β-VAE pipeline to incorporate labeled data, such that classification can be done with the encoder of the VAE, contrary to conditional VAEs. The primary objective of this approach is to promote the disentanglement of label-specific information within the latent representation, which can then be used to improve ECG signal classification. By sampling from a disentangled latent space, it becomes feasible to generate a spectrum of points exhibiting diverse label-specific information while maintaining consistent non label-specific information. Subsequent reconstruction of these sampled points using the VAE decoder results in a variety of ECG signals, which prove valuable for explainability purposes in ECG signal classification. In generative AI, this reconstruction process is also referred to as attribute manipulation, which will here be used to add explainability to the classification process.

To do so, the latent space, denoted z, is divided into two distinct domains: i) The label-specific latent space (z_{ls}) and ii) the non-label-specific latent space (z_{nls}), see Figure 1, with dimensions of L_{ls} and L_{nls} respectively. To extract label-specific information from the ECG signal, the VAE is extended at the bottleneck with a single fully connected layer and a softmax layer. These additional layers are trained to predict the label based on z_{ls}. Consequently, the encoder is tasked with estimating both prior distributions $p(z_{\text{ls}})$ and $p(z_{\text{nls}})$ as well as inferring the label y from the input heartbeat X, modeled as:

$$z_{\text{ls}}, y \sim q(z_{\text{ls}}, y|X) \text{ with } y \sim q(y|z_{\text{ls}}) \text{ and } z_{\text{nls}} \sim q(z_{\text{nls}}|X).$$

The decoders objective, reconstruction of the input X, remains consistent with that of the original VAE and is defined as:

$$X \sim p(X|q(z|X)).$$

The method is designed to be portable to other VAE architectures, but will here be showcased with a relatively basic convolutional VAE. The encoder of the VAE is a CNN with 8 1D convolutional layers followed by 6 2D convolutional layers. This design choice ensures that the leads of the ECG signal undergo mixing only after traversing multiple layers, preserving the integrity of the temporal features of the signal. The decoder is an

exact mirror of the encoder. The latent space at the bottleneck has size $L_{\text{ls}} + L_{\text{nls}}$ and the complete VAE consists of 18.6M parameters in total.

To further enforce the isolation of label-specific information in z_{ls}, the VAE is jointly optimized with an adversarial multilayer perceptron (MLP). This adversarial network is tasked with predicting the label exclusively from z_{nls}, see Figure 1. The negative of this adversarial loss is integrated into the overall VAE loss, discouraging any capture of label-specific information in z_{nls}. Through this dual optimization strategy, label-specific information is explicitly disentangled within the latent space z, providing a comprehensive and effective representation for the analysis of ECG signals.

5.2.4 Training of the Partial β-VAE

Training of the partial β-VAE involves the incorporation of multiple loss terms. These encompass: i) L_{recon}, the mean squared reconstruction error (MSE) quantifying the dissimilarity between the input and output ECG heartbeat, ii)L_{kl_s} the Kullback-Leibler (KL)-divergence loss, $D_{\text{KL}}(P||Q)$ characterizing the disparity between the posterior, $q(z|X)$, and the prior, $p(z)$, which is set to be $\mathcal{N}(0,1)$, iii) a prediction loss for label prediction derived from z_{ls}, and iv) an adversarial prediction loss for label prediction originating from z_{nls}. The first two losses are calculated for the whole dataset and the prediction loss only for the labeled subset.

Consistent with the original work by Higgins et al. [145], the KL-divergence loss is weighted by a β factor. The KL-divergence loss is only calculated for the label specific latent space and is calculated as $D_{\text{KL}}(p(z_{\text{ls}})||q(z_{\text{ls}}|X))$

The labeled subset of the data is annotated with ordinal labels (normal, mild, moderate, and severe impairment), therefore the ranked probability loss function, L_{rps} proposed by Galdran [176] is used here. This function is defined as:

$$L_{\text{rps}} = \frac{1}{nk} \sum_{i}^{N} \sum_{j}^{K} \|\mathbf{P}_{ij} - \mathbf{Y}_{ij}\|_2^2,$$

where K denotes the number of classes, N represents the number of patients, j the one-hot encoded label position, $\mathbf{Y}_i$ signifies the ordinal encoding of the labels ([0,0,0,1], [0,0,1,1], [0,1,1,1], [1,1,1,1]), and $\mathbf{P}_i$ corresponds to the cumulative sum of the output class probabilities. We propose to use a combination of the ranked probability loss and the categorical cross-entropy loss, given as:

$$L_{\text{pred}} = L_{\text{rps}} + L_{\text{CE}}.$$

This choice is informed by optimization experiments which demonstrated its effectiveness. Two prediction losses are used in training. $L_{\text{pred\_ls}}$ is employed for label prediction from z_{ls} and is weighted with a γ_{ls} factor. $L_{\text{pred\_nls}}$ is employed for the adversarial MLP predicting the label from z_{nls} and is weighted with a negative γ_{nls} factor to deter the assimilation of label-specific information in z_{nls}. The MLP is optimized with the same $L_{\text{pred\_nls}}$. This results in the MLP minimizing the prediction loss and the VAE maximizing the prediction loss of the MLP. Since the VAE controls the information that is contained in z_{nls}, maximizing the prediction loss of the MLP will result in minimizing the label specific information in z_{nls}. However, given the dynamic nature of z_{nls} during training and therefore the varying input to the MLP, the MLP is trained for 5 epochs after each epoch of the VAE training phase.

The overall loss function for the VAE is defined as follows:

$$L_{\text{VAE}} = L_{\text{recon}} + \beta_{\text{ls}} D_{\text{KL}}(p(z_{\text{ls}})||q(z_{\text{ls}}|X)) + \gamma_{\text{ls}} L_{\text{pred\_ls}} - \gamma_{\text{nls}} L_{\text{pred\_nls}}.$$

For the MLP, the loss function is expressed as:

$$L_{\text{MLP}} = L_{\text{pred\_nls}}$$

5.2.5 Feature visualization

To visually explore the pertinent features in the ECG signal crucial for LVF prediction, we employ attribute manipulation on a per-signal basis. The application of attribute manipulation in this context differs from the conventional methods employed in data generation. Here it provides a distinctive approach that offers a clear per-signal visualization of the classification rationale adopted by the encoder. The procedure involves encoding each ECG heartbeat into z_{nls} and z_{ls}, and subsequently decoding the encoded signal to obtain the reconstructed signal, denoted as S_{recon}. We assume $L_{\text{ls}} = 1$ here, but other sampling strategies could be adopted when $L_{\text{ls}} > 1$. The above mentioned procedure is applied iteratively to six distinct sets of z_{ls}, covering the mean values for the four LVF groups (S_1, S_2, S_3, S_4), as well as two additional "extreme" z_{ls} values. These extreme values are obtained by sampling the mean of the severely impaired LVF group minus its standard deviation (S_0) and the mean of the normal LVF group plus its standard deviation (S_5). Consequently, a spectrum of reconstructed ECG signals is generated, representing various degrees of LVF impairment. This spectrum can be interpreted as the algorithm's utilization of distinct features within the signal for LVF classification.

5.2.6 Survival analysis

In addition to the primary evaluation metrics, we perform a supplementary analysis to validate the performance of the latent space representation (z_{ls}) by means of survival analysis. The rationale behind this approach lies in understanding that LVF status serves as a significant indicator of cardiac health and exhibits predictive value for mortality. The anticipation is that the performance of z_{ls} in predicting mortality will align closely with the predictive capacity demonstrated by the LVF labels. Survival analysis assesses the time until an event of interest occurs, such as mortality, and compares survival distributions between different groups. The primary survival analysis method employed in this study is the Kaplan-Meier estimator, a non-parametric approach, which provides an estimate of the survival function. The log-rank test is used to compare survival curves between different groups. The survival function, denoted as $S(t)$, represents the probability of survival beyond time t. The evaluation of the survival function is done with the concordance index, which evaluates how well a model ranks the survival times of patients. The Cox proportional hazards model is also applied to assess the impact of z_{ls} on survival outcomes while considering covariates. The model is represented as:

$$h(t) = h_0(t) \exp(\beta_{\text{cox}} z_{\text{ls}}),$$

where $h(t)$ is the hazard function at time t, $h_0(t)$ is the baseline hazard function, β_{cox} represents the hazard ratio for z_{ls} [177].

5.3 Experiments and Results

5.3.1 Experiments

Experimental settings

In all experiments the combined latent space, z, with a dimensionality of L was set to 600, large enough to ensure it was not a limiting factor in reconstruction. Both data sets were combined and divided into a training set (85%) and a test set (15%). The 5-fold cross-validation was performed with the training set with again a ratio of 85:15 between the training and the validation set. The validation set was split into two equal subsets, one used for early stopping and one used for hyper-parameter optimization. All data splits were grouped by patient, which means that no patient in the training set was used in the validation or test set, and stratified by label in the case of labeled data splits. Training was done until convergence, i.e. until the loss on the validation set stopped improving for

30 epochs. To avoid overfitting, balanced sampling, early stopping, regularization with the Adam optimizer with weight decay, and L2 regularization of the fully connected layer were used. To prevent gradient explosion He initialization was used [150], to prevent overflow, the standard deviations of the posterior $q(z|X)$ were restricted to a [-10, 3] interval before the sampling step.

The networks were build and trained in the PyTorch 2.1.0 framework and trained on a Quadro RTX 6000 GPU with CUDA 12.1 [147, 148]. The implementation of our models will be made publicly available via GitHub at `https://github.com/ViktorvdValk/Interpretable_VAE_for_ECG`.

5.3.2 Feature evaluation

The LVF predictions of the models are evaluated with 3 classification metrics. The Area Under the Receiver Operator Characteristic Curve (AUROC) and the F1 score are used for evaluation of the binary split (severe/moderate vs. mild/normal), given the clinical significance of this split and to facilitate comparison with prior studies. Significant difference between AUROC scores was calculated as proposed by DeLong *et al.* (1988) [178]. Another prediction metric, the relative correctness rate, RCR, that takes into account the ordinality of the classes, based on the prediction MSE between the predicted and the ground truth class [179], is used to assess the ordinal classification of the label. The RCR is class balanced and is designed so that the perfect classification results in $\mathrm{RCR} = 1$ and mean classification, ($\hat{y} = \overline{y}$), results in $\mathrm{RCR} = 0$. RCR is defined as follows:

$$\mathrm{RCR} = 1 - \frac{1}{4}\sum_{j=0}^{3}\frac{1}{N_j}\sum_{i\in\mathcal{S}_j}\left(\frac{y_i - \hat{y}_i}{y_i - \overline{y}}\right)^2.$$

Here, $\mathcal{S}_j$ is the set of all patients with LVF label j with $j =$ severe, moderate, mild, normal and $N_j = |\mathcal{S}_j|$. The reconstruction of the heartbeats by the models is evaluated with the mean squared reconstruction error (MSE). For inter-model comparison of the MSE, the paired t-test was used. For 95% confidence interval calculation bootstrapping was used with $n = 1000$. In line with the output class used in L_{rps}, the cumulative sum of the models' output class probabilities was used to predict the final class for RCR calculation.

Impact of the β factor

In this experiment, we investigate the influence of β_{ls}. Notably, given that the pipeline does not incorporate sampling from the posterior $q(z_{\mathrm{nls}}|X)$, it can be argued that structuring this portion of the latent space is unnecessary, particularly if it compromises reconstruction

quality, as observed in prior work by Van der Valk *et al.* (2023) [155]. Consequently, in the proposed algorithm, β_{nls} was set to 0. Furthermore, we examined the impact of β_{ls} on prediction and reconstruction to elucidate the importance of structuring and disentangling the latent space (z_{ls}), which is already subjected to a prediction loss.

Impact of the label-specific split

We investigated the advantages of implementing a label-specific split within the latent space z. Although such a split constrains both reconstruction and prediction capacity, it reduces the susceptibility to overfitting and enhances the interpretability of the model, particularly when L_{ls} is small. L_{ls} plays a crucial role in promoting structure within z_{ls}, thereby facilitating sampling and enhancing the interpretability of the reconstructed signals. The interpretability of the visualization of the features made with attribute manipulation is directly related to L_{ls}, because of the influence of the dimensionality of z_{ls} on the sampling process. Interaction effects between dimensions of z_{ls} when $L_{\mathrm{ls}} > 1$, make attribute manipulation less interpretable.

Ablation experiments

In the ablation experiments we investigate the influence of the prediction loss and the adversarial network. Both extensions to the pipeline promote either direct of indirect disentanglement of the prediction relevant information in the input. Since L is large and thus not a limiting factor, it can be assumed that the addition of the adversarial network forces the model to use the label-specific latent space for signal reconstruction also. This is assessed using a linear regression classifier that aims to predict the label from z_{nls}.

Baseline methods

To give context to the reconstruction, prediction and interpretability values, our method is compared against both principal component analysis (PCA) and the VAE XAI as proposed in Van der Leur *et al.* (2022) [154]. PCA can be considered a general, unsupervised deterministic feature extractor that does not use any form of machine learning. The VAE XAI is a β-VAE that is specifically designed for unsupervised feature extraction from the ECG signal. The features generated with the VAE XAI were shown to contain information of the LVF in a general population of patients who had an ECG taken. In their paper the authors reported an AUC of 0.90 when predicting the LVF from the 12-lead ECG with a latent space of 32. The authors also reported a similar AUC (0.91) for a supervised CNN trained with the LVF labels. Since both PCA and the VAE XAI do not have a classification

Table 5.1: Results for different values of β_{ls}, the KL-divergence weight for z_{ls}, are shown here for both reconstruction (MSE) and prediction (AUROC, F1 and RCR). All metrics show the average (and 95% confidence interval) of 5-fold cross-validation on the validation set. Significant difference, $p < 0.025$, in the AUROC and MSE metrics from β_{ls}=0 for all folds is indicated with an asterisk (*)

β_{ls}	$\mathrm{MSE} \downarrow$	AUROC $\uparrow$	F1 $\uparrow$	$\mathrm{RCR} \uparrow$
0	**2.10** [2.09-2.11]	**0.868** [0.864-0.871]	**0.737** [0.734-0.741]	**0.637** [0.630-644]
0.1	2.96* [2.95-2.97]	0.811* [0.808-0.815]	0.596 [0.592-0.600]	0.498 [0.490-0.505]
0.5	2.37 [2.36-2.38]	0.501* [0.495-0.506]	0.469 [0.468-0.469]	0.275 [0.267-0.283]
1	2.30 [2.29-2.31]	0.556* [0.551-0.556]	0.469 [0.469-0.469]	0.275 [0.266-0.283]

layer at the bottleneck, a linear regression classifier is used to predict the label from z. For PCA the space spanned by the principal components is used as the latent space z with L the number of principal components used for reconstruction and prediction.

5.3.3 Results

Hyperparameter tuning

The prediction and reconstruction tasks have different complexity and data scales. Several adjustments are made to prevent overfitting on the prediction task, which is less complex and is trained on a subset of the data used for reconstruction. γ_{ls} is set to 0 for the first 100 epochs, then gradually increases to 2 in the 400 epochs following. This schedule helps to prevent overfitting on the prediction task while ensuring enough training epochs for the reconstruction task. The value of γ_{ls} is tuned to achieve optimal network prediction performance. γ_{nls} is set to 15 from the beginning, as the adversarial loss does not promote overfitting of the prediction task. Setting γ_{nls} larger than 15 did not improve the capture of any label-specific information in L_{nls}.

Table 5.1 presents the impact of the KL-divergence weight β_{ls} on reconstruction and prediction. Both reconstruction and prediction seem better with small values of β_{ls}. The results indicate that reconstruction and especially prediction are limited by the KL-divergence

Table 5.2: Results for different values of L_{ls} (and thus L_{nls}, given that $L_{\text{ls}} + L_{\text{nls}} = 600$) are shown here for both reconstruction (MSE) and prediction (AUROC, F1, RCR). Setting $L_{\text{ls}} = 600$ essentially means not spitting L, since the whole latent space is then optimized for both reconstruction and prediction. All metrics show the average (and 95% confidence interval) of 5-fold cross-validation on the validation set. Significant difference, $p < 0.025$, in the AUROC and MSE metrics from L_{ls}=1 for all folds is indicated with an asterisk (*)

L_{ls}	$\text{MSE} \downarrow$	AUROC $\uparrow$	F1 $\uparrow$	$\text{RCR} \uparrow$
1	2.10 [2.09-2.11]	0.868 [0.864-0.871]	**0.737** [0.734-0.741]	0.637 [0.631-0.643]
2	2.02 [2.01-2.03]	0.871 [0.868-0.874]	0.726 [0.722-0.730]	0.637 [0.631-0.643]
5	1.99 [1.98-2.00]	0.872 [0.870-0.875]	0.723 [0.720-0.727]	**0.658** [0.652-0.664]
600	**1.87** [1.86-1.88]	**0.873** [0.870-0.876]	0.719 [0.715-0.722]	0.649 [0.643-0.654]

loss. $\beta_{\text{ls}} = 0$ gave both best reconstruction and best prediction results. Therefore, $\beta_{\text{ls}} = 0$ will be used in the rest of the paper.

Table 5.2 summarizes the effect of L_{ls} on reconstruction and prediction. Setting $L_{\text{ls}} = 600$ effectively eliminates the split, as the full latent space is optimized for both reconstruction and prediction. Reducing L_{ls} imposes a constraint on the VAE's reconstruction capacity; however, this effect is minimal, as indicated by the slight reduction in reconstruction quality for $L_{\text{ls}} = 1$ compared with $L_{\text{ls}} > 1$. Importantly, lowering L_{ls} does not compromise prediction performance, suggesting that all label-specific information can be captured in a single latent dimension. Minimizing L_{ls} also improves interpretability (as explained in the Experiments section) and, for these reasons, $L_{\text{ls}} = 1$ was used in all subsequent experiments.

Ablation results

Table 5.3 shows the results of different ablations on the pipeline. Separately, the prediction loss and, to a lesser extent, the addition of the adversarial network contribute to enhancing the prediction quality of the model. When both are used together, the addition of the adversarial network provides negligible benefit to the prediction quality over a network trained

Table 5.3: Results of ablation experiments. The influence of the adversarial network (adv. netw.) and γ are shown here for both reconstruction (MSE) and prediction (AUROC, F1, and RCR). The AUROC for the label prediction with z_{nls} is also shown. All metrics show the average (and 95% confidence interval) of 5-fold cross-validation on the validation set. Significant difference, $p < 0.025$,in the AUROC and MSE metrics from the proposed method for all folds is indicated with an asterisk (*)

Adv. Netw.	γ_{ls}	$\mathrm{MSE} \downarrow$	AUROC $\uparrow$	AUROC (z_{nls}) $\downarrow$	F1 $\uparrow$	$\mathrm{RCR} \uparrow$
Yes	2	2.10 [2.09-2.11]	0.868 [0.864-0.871]	**0.857** [0.854-0.860]	**0.737** [0.734-0.741]	**0.637** [0.631-0.642]
Yes	0	0.96* [0.94-0.98]	0.753* [0.743-0.764]	0.837 [0.829-0.844]	0.668 [0.659-0.677]	0.442 [0.435-0.450]
No	2	1.92 [1.91-1.92]	**0.870** [0.867-0.873]	0.853 [0.849-0.856]	0.720 [0.716-0.724]	0.630 [0.622-0.637]
No	0	**0.82*** [0.82-0.83]	0.535* [0.530-0.540]	0.853 [0.850-0.857]	0.469 [0.469-0.469]	0.274 [0.265-0.282]

solely with the prediction loss. However, less overfitting was observed on the prediction task, when both were used in the pipeline. The benefit of each addition on prediction performance comes at the cost of the reconstruction quality, which is slightly reduced by each term. This degradation occurs because these components impose additional constraints on the latent space z. The prediction quality from z_{nls}, as indicated by AUROC (Adv. Netw.) in Table 5.3, is similar in models with and without the adversarial loss. Preventing the model from capturing label-specific information in z_{nls} appears to be a challenging task in which the adversarial network does not fully succeed.

Benchmark comparison

Table 5.4 shows the comparison of the proposed model with PCA and with the VAE XAI [154] on the test set. The proposed model performs better than the state-of-the-art in LVF prediction, using only one latent dimension to store label specific information, instead of 32 (VAE XAI) or 600 (PCA). This confirms that the proposed model can successfully separate (the majority of) the label specific information and reduce it to 1 dimension. While PCA gives reasonable classification and superior reconstruction results, the visualization of the 600 first principal components is cumbersome and therefore not suitable for meaningful

Table 5.4: Comparison with benchmarks. The proposed method is compared with PCA and the VAE XAI network from [154] for both reconstruction (MSE) and prediction (AUROC, F1 and RCR). The AUROC and F1 scores are shown for the prediction of the normal/mild vs. the moderate/severe groups. All metrics show the average (and 95% confidence interval) of 5-fold cross-validation on the test set. Significant difference, $p < 0.025$, in the AUROC and MSE metrics from the proposed method for all folds indicated with an asterisk (*)

Model	L_{ls} (L)	$\mathrm{MSE} \downarrow$	AUROC $\uparrow$	F1 $\uparrow$	$\mathrm{RCR} \uparrow$
Proposed	1 (600)	2.24 [2.18-2.30]	**0.832** [0.829-0.834]	**0.688** [0.681-0.695]	**0.591** [0.581-0.601]
VAE XAI [154]	32 (32)	32.8* [32.0-33.5]	0.790* [0.786-0.794]	0.640 [0.632-0.647]	0.524 [0.514-0.533]
PCA	600 (600)	**0.356*** [0.354-0.356]	0.815* [0.812-0.818]	0.659 [0.652-0.665]	0.480 [0.470-0.489]

attribute manipulation. On the other hand, the proposed model outperforms the VAE XAI, which would be more suitable for meaningful attribute manipulation, by a large amount in both reconstruction and prediction.

Cox regression

Table 5.5 presents a comparative analysis between two Cox regression models, one using ground truth LVF and the other the ECG-derived LVF prediction. The concordance index and the significance of hazard ratios demonstrate nearly identical performance between the two models. This observation suggests that the ECG-derived LVF captures mortality-related information in patients to a similar extent as the LVF obtained from echocardiograms. Additionally, Figure 5.2 illustrates the predictive value of ECG-derived LVF for mortality outcomes in patients following myocardial infarction.

Evaluation of interpretability

Figure 5.3 shows the histograms of predicted LVF values for the 4 LVF groups. The histograms show separation to a reasonable extent, especially between the mild and the moderate groups. The ground truth LVF is a continuous value, which is discretized in 4 groups in clinical practice. The predicted LVF reintroduces this continuous character, which might be of additional value.

Figure 5.4 presents a visualization example of attribute manipulation of the LVF label using the proposed VAE. Notably, a reconstruction closely tracking the original heartbeat

Table 5.5: Hazard ratios for Cox regression models incorporating both ground truth LVF and ECG-derived LVF, with age included as an additional predictor. All predictors exhibit high significance levels ($p < 0.005$), and the concordance index for survival prediction using ground truth LVF versus ECG-derived LVF predictions demonstrates remarkable similarity. These findings suggest that ECG-derived LVF holds promise as a reliable alternative to echocardiogram-derived ground truth LVF in predicting mortality outcomes

var	hazard ratio	p	concordance ↑
age	1.98 [1.63-2.40]	< 0.005	0.72
ECG-derived LVF	0.59 [0.48-0.71]	< 0.005	
age	2.04 [1.69-2.47]	< 0.005	0.72
ground truth LVF	0.59 [0.48-0.73]	< 0.005	

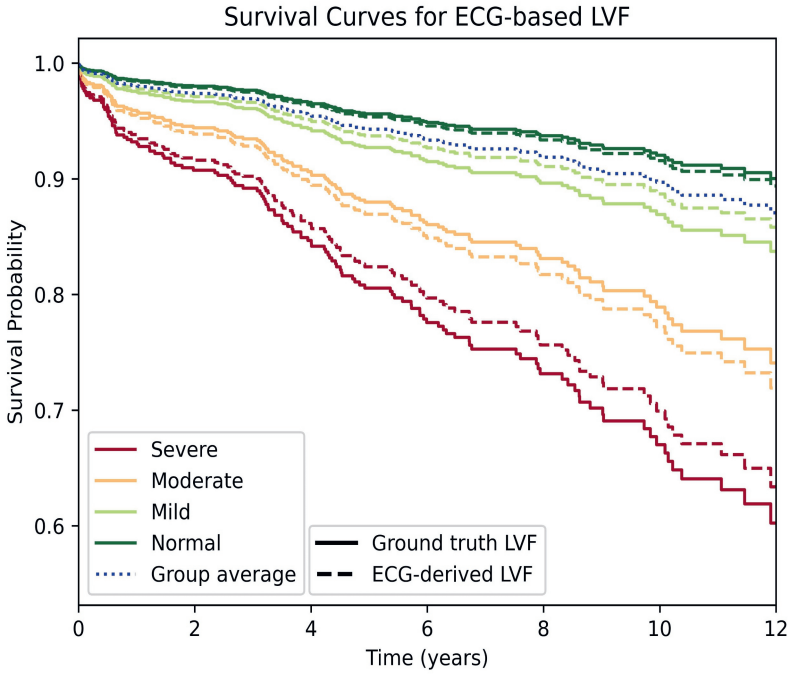


Figure 5.2: Survival curves depicting the stratification of patients into distinct groups based on ECG-derived and ground truth LVF predictions. Clear delineation is observed among the four distinct groups. This underscores the predictive capacity of both ECG-derived and ground truth LVF in delineating survival outcomes, revealing significant differences in mortality risk among the identified groups.

can be seen as a confirmation that the heartbeat is not an out-of-distribution heartbeat, thereby reinforcing the prediction's reliability. The attribute manipulation showcases the

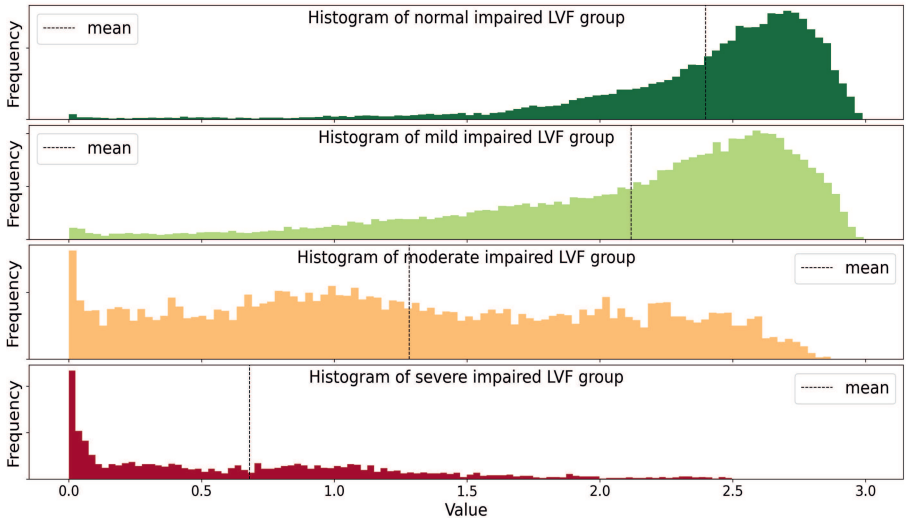


Figure 5.3: The histograms show the distributions of the continuous ECG-derived LVF values for the 4 LVF groups. The distributions show some overlap, which indicates that the prediction is reasonably good but not perfect. The means of the distributions are indicated with a dotted line.

distribution of heartbeats across a spectrum of ECG-derived LVF values, which essentially puts the heartbeat in a population context. Heartbeats shaded in yellow to green represent heartbeats classified as having minimal LVF impairment while retaining non-LVF-related characteristics. Conversely, heartbeats shaded in yellow to dark red indicate greater impairment. The magnitude of the discrepancy between the red and green heartbeats underscores the importance of leads and peaks within the heartbeat for LVF classification. This observation not only elucidates the model's main reason behind the classification decisions, but could also be viewed as potential avenues for desired future improvement. Visual inspection of the ECG spectra by an interventional cardiologist confirmed the relevance of the indicated ECG features. Several patterns known to indicate LVF status where recognized in various attribute manipulated heartbeat distributions, see Figure 5.4. Some of these patterns are routinely used by cardiologists to estimate cardiac function, others were known but not used, since the patterns are deemed too subtle to estimate from a single ECG without context. The model's features are complementary to already known ECG patterns that are associated with reduced LVF, such as LBBB and the presence of Q-waves. In such cases the models' features can help with the quantitative assessment of pathological patterns. For example, the T-wave is an important prognostic factor in myocar-

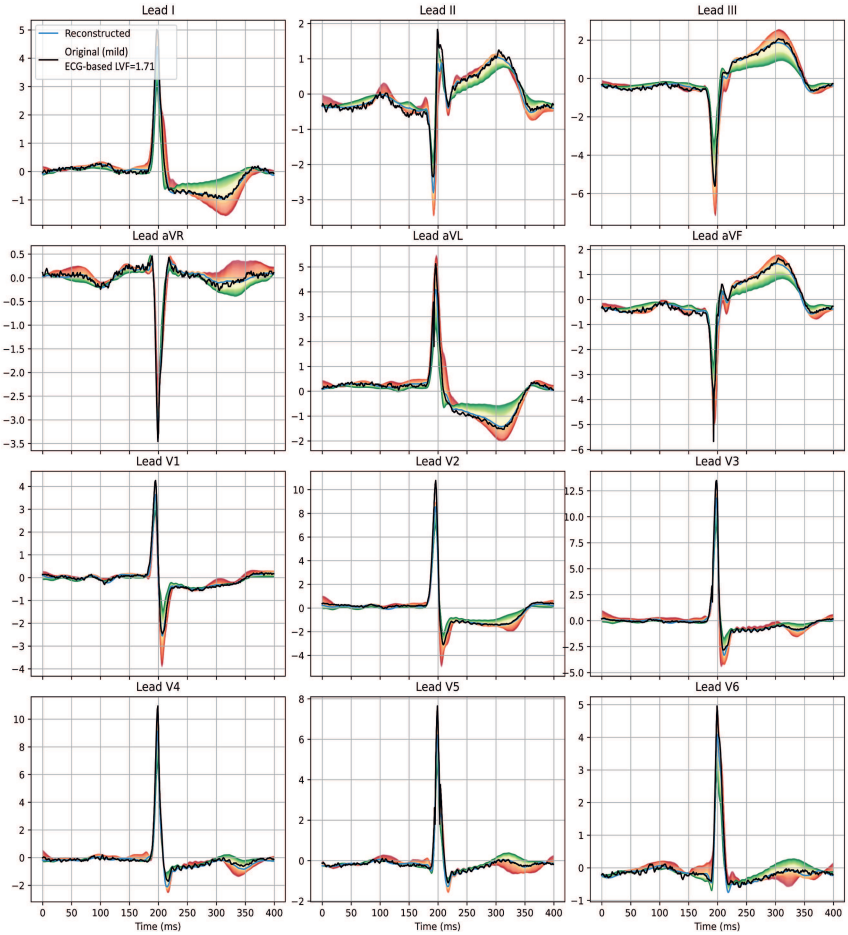
dial infarction, but quantitative assessment of T-wave changes due to myocardial infarction is currently not part of the clinical routine in the non-acute setting. In this case the models output, see Figure 5.4 was considered helpful in providing context.

5.4 Discussion

Explicit disentanglement of the latent space through supervised learning enhances interpretability in classification tasks when combined with attribute manipulation. In this study, we illustrate this approach using a VAE framework, where we i) partition the latent space into a label-specific and non-label-specific part, ii) introduce a classification loss, and iii) leverage an adversarial network into the VAE optimization process. Ablation experiments confirm that these extensions collectively contribute to achieving classification (AUROC$= 0.832$) and reconstruction performance (MSE$= 2.24$) comparable or better than the state-of-the-art, while adding interpretation possibilities. i) Previous studies have shown that incorporating a classification loss encourages the explicit encoding of information relevant for classification in the latent space, thereby enhancing classification accuracy compared to unsupervised methods [155, 168]. ii) Subdividing the latent space does not necessarily improve classification performance. However, it isolates and compresses label-specific information into a dedicated part of the latent space, which can be as small as one dimension (see Table 5.2). This partitioning helps visualize the model's decision-making process through attribute manipulation. It places the signal in a population context, making signal-specific features important for classification easier to interpret (see Figure 5.4). The signal-specific projections used here differ from the population-specific ECG features described by Van der Valk et al. (2023) and Van der Leur et al. (2022) [154, 155]. iii) The inclusion of an adversarial network serves as a regularization mechanism to prevent overfitting on the classification task, as was shown in Table 5.3.

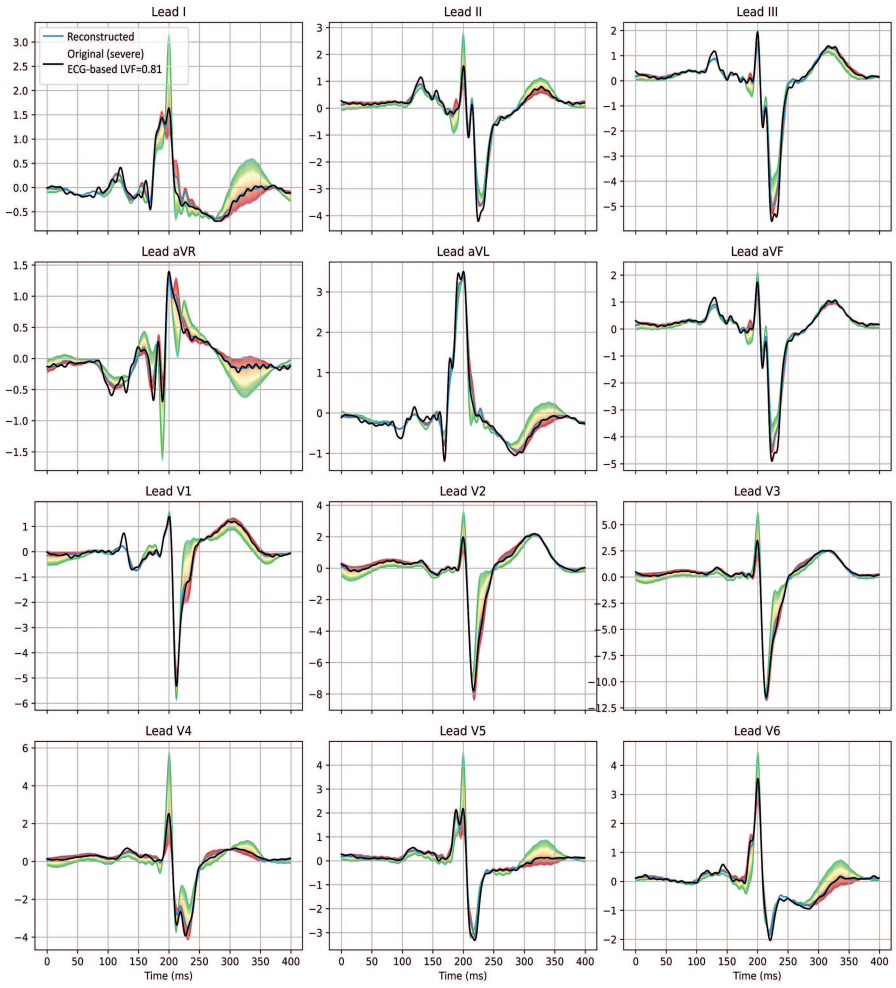
Survival analysis confirmed the effective isolation of LVF-specific information through the ECG-derived LVF, demonstrating performance comparable to the ground truth LVF, as presented in Table 5.5. Although the ECG-based LVF does not serve as a perfect LVF predictor ($\mathrm{AUROC} = 0.832$), it encapsulates all survival-related information present in the ground truth.

The KL-divergence loss in our framework does not provide the same benefits as in conventional VAEs, as evidenced by the negligible impact of the KL-divergence loss weight (β) in our experiments. This can be attributed to two factors: Firstly, for attribute manipulation



(a) Mild LVF impairment

Figure 5.4: Attribute manipulation in 2 heartbeats with mild (a) and severe (b) LVF impairment. The black and blue line respectively show the original and the reconstructed signal. The green to red distributions are the result of attribute manipulation and could be interpreted as the model's rationale to classify the signal with the LVF value shown in the legend. The yellow to green range shows heartbeats that would be classified as having a mildly impaired to normal LVF and the yellow to red range shows heartbeats that could be classified as having a moderate to severely impaired LVF. (a) The ECG shows clear Q-waves, a negative deflection preceding the R-peak, in lead II, III, aVF. The Q-wave pattern is known in clinical cardiology to indicate prior myocardial infarction. The model takes the depth of this peak into account, but also uses the T-wave, the positive deflection after the R-peak.



(b) Severe LVF impairment

Figure 5.4: continued: (b) The ECG shows a pattern known as left bundle branch block (LBBB), indicated by the dominant S wave in lead V1, V2 and V3 and the notched R waves in lead I, AVL V5 and V6. The model shows that not only QRS-characteristics, but also T-wave patterns, are important for prediction of the LVF

we do not sample from the non-label-specific part of the latent space, therefore disentanglement of this part seems unnecessary. Secondly, the prediction loss, already promotes a structure on z_{ls}, as can be seen in Figure 5.3. Namely, the models prediction is a linear combination of z_{ls} (a single fully connected layer without activation). This likely diminishes the benefit of the KL-divergence on the ordering of the label-specific latent space as well.

In summary, our study shows a novel application of VAEs in explainable (ECG) classification. While previous research has shown the potential of VAEs in feature extraction and data augmentation [171, 172, 173, 143, 174], our approach goes beyond by explicitly optimizing the VAE architecture with a labeled subspace for classification. Our method is able to do explainable classification on a per-signal basis, while performing on par with or better than the state-of-the-art in ECG classification. We thereby contribute to the advancement of explainable AI algorithms in the medical context, where transparency and interpretability are essential.

The successful outcomes of our study have important implications for clinical practice. By enhancing the specificity of latent features derived from ECG signals, our method allows a more accurate and reliable assessment of LVF, a critical prognostic determinant in cardiac disease [9]. The ultimate aim of the ECG analysis pipeline as proposed here is integration in clinical practice. The lack of context or a 'normal or healthy' signal to compare against in plain ECG analysis, is one of the reasons why ECG analysis is hard and necessitates extensive expertise. Providing a 'normal or healthy' signal for comparison presents complexities due to the diverse patient characteristics influencing ECG morphology. Therefore, even in patients not affected by disease, the range of patterns considered normal is wide [180, 181, 182]. Our pipeline aims to give such a 'normal or healthy' version of the signal, while preserving patient-specific features. When implemented in ECG-software visuals, as shown in Figure 5.4, can be helpful in providing context to otherwise harder to distinguish ECG characteristics, such as T or Q-wave improvement or decline. Besides, providing context, the visuals also help to assess the quality of the classification and the clinical implication that is attributed to it. Assessment of the difference between the original signal and the reconstruction serves as a quality control measure. A large difference can be used as an indicator of classification inaccuracies, which in some cases are related to signal noise. Clinical integration of our pipeline holds promise for reducing both the time and expertise required for ECG analysis, facilitating patient monitoring, and enabling remote monitoring in home-based settings where remote monitoring with smart devices is becoming increas-

ingly common. If LVF can be determined accurately from ECG measurements, this could, in some cases, reduce the need for echocardiogram acquisition. Moreover, explainable automation of ECG interpretation may allow less specialized healthcare personnel to reliably assess LVF. Accurate LVF prediction from 12-lead ECGs also opens the path toward extending these capabilities to 1-lead ECGs, making home-based monitoring a realistic opportunity for scalable and continuous assessment of cardiac function.

Although our study demonstrates promising results, it is important to acknowledge its limitations and identify directions for future research.

One key limitation lies in the adversarial network's inability to effectively prevent label-specific information from being encoded in z_{nls}. The dimensionality of the latent space (L_{nls}) may facilitate easier adversarial label prediction, undermining the network's ability to suppress label-specific information. In addition, the interaction between L_{nls} and the hyperparameters γ_{ls} and γ_{nls} remains unexplored in this study, but could play a role in mitigating the capture of label-specific information in z_{nls}. Considering the challenges associated with training adversarial networks, future work should investigate the use of extended training schedules to enhance their effectiveness [183].

The joint optimization of the prediction and reconstruction tasks introduced additional challenges, necessitating careful tuning of the hyperparameter γ_{ls} to mitigate overfitting. These challenges stem from the differing complexities and data scales associated with the two tasks. To address this, the reconstruction task was independently trained for several epochs before initiating joint optimization. While this approach was effective, alternative strategies to achieve comparable results were not explored in this study.

Another notable limitation of this study is the imperfect prediction of LVF and the observed overlap in predicted LVF values across true LVF classes, as depicted in Figure 5.3. This overlap can be partially attributed to the labeling process, where ground truth LVF values are rounded to the nearest perceived class, introducing variability near class boundaries. Substituting LVF with left ventricular ejection fraction (LVEF) as the target variable could potentially enhance predictive performance. Unfortunately, concurrent LVEF measurements were unavailable for most of the ECGs analyzed in this study. Previous research has demonstrated that the 12-lead ECG may not encompass all the information required for perfect LVF prediction, which can also explain the overlap [155, 141]. Nonetheless, accurate LVF prediction is critical to ensure the validity of attribute manipulation as illustrated in Figure 5.4. Incorporating additional predictors that provide more detailed information

about LVF status could improve overall predictive accuracy and enhance the robustness of the explainability framework.

In clinical practice, ECG signals are typically preprocessed through methods such as baseline wander correction and de-noising to enhance signal quality and standardization [184]. These steps were not applied in our study, and their inclusion may further improve model performance and clinical applicability. In contrast, some preprocessing procedures used here, such as heartbeat segmentation and exclusion of irregular beats, may restrict applicability in contexts like arrhythmia analysis. Moreover, pathologies where R-peak detection is complicated such as bundle branch blocks or cardiomyopathies may hinder appropriate heartbeat segmentation, which also restricts the applicability of the current pipeline.

The generalizability of our findings to broader patient populations, different clinical settings, and for predicting cardiac parameters beyond LVF remains to be established. Importantly, while this work primarily serves as a proof of concept for interpretable ECG classification, the underlying methodology is not limited to ECG. The approach is applicable to any encoder–decoder architecture, future research could extend it to other cardiac populations or even to entirely different medical domains.

5.5 Conclusion

In conclusion, our study represents a significant step forward in the field of ECG analysis by enhancing the specificity and interpretability of latent features derived from ECG signals. By improving the classification rationale specific to each case, our method has the potential to enhance clinical decision-making and patient care in cardiac disease monitoring.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 860173.

Conflict of Interest Statement

The authors have no relevant conflicts of interest to disclose.

6

Leveraging the wearable 1-lead ECG signal: From cardiac rhythm to cardiac function assessment

Van der Valk, V.O., Atsma, D.E., Scherptong, R.W.C., Staring, M., (2026). Leveraging the wearable 1-lead ECG signal: From cardiac rhythm to cardiac function assessment. (submitted to npj Cardiovascular Health)

Abstract

The electrocardiogram (ECG) is a critical tool in the diagnosis and monitoring of cardiovascular disease. Although traditional 12-lead ECGs offer comprehensive insights into the electrical activity of the heart, they typically require clinical settings and expert interpretation, which limits their accessibility. In contrast, smartwatch 1-lead ECGs can be recorded at home, allowing more frequent and rapid monitoring. This opens opportunities not only for early detection but also for enhancing patient autonomy. This study investigates whether 1-lead ECGs can provide information beyond heart rhythm, specifically whether they can be used to assess left ventricular function (LVF) using explainable deep learning models. Our findings show that LVF can be accurately predicted from 1-lead ECGs (AUC = 0.883), nearly matching the performance of 12-lead ECGs (AUC = 0.897). These results suggest that 1-lead ECGs, when combined with interpretable AI, could support broader clinical applications and empower patients, particularly in resource-limited or remote settings.

6.1 Introduction

Cardiovascular disease remains the leading global cause of death, underscoring the need for accessible and accurate diagnostic tools. The 12-lead electrocardiogram (ECG) is the clinical gold standard for assessing cardiac electrical activity and diagnosing a wide range of heart conditions. However, its application is largely confined to clinical settings due to the need for specialized equipment and trained personnel.

In contrast, the 1-lead ECG offers a portable and cost-effective alternative. Devices such as smartwatches and chest patches enable patients to record ECGs independently, facilitating continuous and remote monitoring. However, the reduced spatial resolution of 1-lead ECGs limits their diagnostic scope, particularly for detecting structural abnormalities. These devices have primarily been optimized for rhythm analysis, especially atrial fibrillation (AF) detection, with several FDA- and EMA-approved systems (e.g., Apple Watch, KardiaMobile) achieving high accuracy in real-world applications [185, 186, 187].

Deep learning has significantly advanced the interpretability and accuracy of ECG-based diagnostics [138, 142]. In the 2017 PhysioNet Challenge, top-performing AF classifiers used convolutional and recurrent neural networks for 1-lead ECG analysis [7, 188]. Since then, explainable AI methods have enabled more transparent and trustworthy models for AF and other arrhythmias [139, 155, 154], supporting broader adoption in clinical practice.

Beyond arrhythmias, recent work has extended 1-lead ECG applications to detecting heart failure [189], chronic kidney disease [190], and age-related cardiac changes [191]. Moreover, predictive modeling using individual leads, such as Lead I or II, has shown potential for detecting myocardial infarction, left ventricular dysfunction, and myocardial injury, although typically with reduced accuracy compared to full 12-lead recordings [192, 193, 194, 195]. These efforts illustrate the growing interest in the use of reduced-lead ECGs for not only rhythm assessments, but also structural and functional cardiac assessments.

Despite these advances, few studies have directly compared wrist-based 1-lead ECGs with full 12-lead ECGs acquired from the same patients. Most prior work has either focused on specific conditions (e.g., arrythmias) or used leads derived from 12-lead datasets, limiting their applicability to real-world wearable recordings. Moreover, while arrhythmia detection has dominated the use of 1-lead ECGs, their potential to identify structural cardiac

changes, such as ventricular dysfunction, remains underexplored, although such abnormalities often leave subtle but detectable signatures in the electrical signal. Attia *et al.* (2022) demonstrated that left ventricular dysfunction (LVD) could be predicted from a 1-lead smartwatch ECG with reasonable accuracy (AUC = 0.885) in a general cardiac patient population [196]. However, to our knowledge, left ventricular function (LVF) assessment from 1-lead smartwatch ECGs has not yet been evaluated with explainable AI, nor in the post–myocardial infarction setting, a population in which early detection of ventricular dysfunction is particularly relevant. The post-myocardial infarction population is homogeneous, since scar tissue is the main cause of abnormalities. The confined heterogeneity in this population is due to the location and severity of myocardial infarction. This is an important aspect that can be leveraged in explainable AI models.

This study conducts a head-to-head comparison of wrist-based 1-lead ECGs, recorded at home as part of the follow-up protocol, and standard 12-lead ECGs, matched by patient and acquisition time, in individuals recovering from myocardial infarction. Focusing on cardiac morphology, we apply deep learning and explainable AI to evaluate whether wearable 1-lead ECGs can move beyond rhythm analysis and provide clinically meaningful insight into structural cardiac abnormalities.

6.2 Methods

6.2.1 Data

The study uses two datasets, a 1-lead ECG dataset and a 12-lead ECG dataset from the same cohort, see Figure 6.1. All data was obtained from patients who were diagnosed with acute coronary syndrome, between 2018 and 2023 and treated at the Leiden University Medical Center in the Netherlands. The 1-lead dataset contains 63.950 raw 30-second, 1-lead ECG signals recorded at a frequency of 300Hz. These signals were recorded by 708 patients (77.3% male, age 59.5 $\pm$ 10.8 years). The 1-lead ECG signals were recorded at home with a Withings ECG Smartwatch as part of a clinical telemonitoring program [197, 198]. The 12-lead dataset contains 77.005 raw 10-second, 12-lead ECG signals recorded at a frequency of 500Hz. These signals are obtained from 6.703 patients (76.3% male, age 60.5 $\pm$ 11.3 years). The study protocol, identified as nWMODIV2_2022006, received approval from the institutional review board, which waived the requirement to obtain informed consent for the use of ECG data of individual pseudonymized patients.

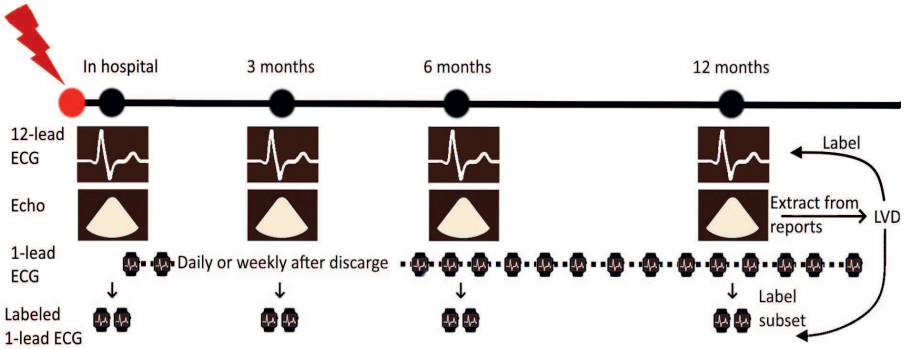


Figure 6.1: **Clinical monitoring and data collection:** Timeline indicates the 12-lead ECG, 1-lead ECG and echocardiogram data collection following the clinical protocol. 12-lead ECGs and echocardiograms are collected during hospital visit as part of the standard monitoring protocol after MI. 1-lead ECGs are collected at home by patients on suggested intervals after discharge. A subset of the 1-lead ECGs, when acquisition overlaps with echocardiography, is labeled with the LVF label extracted from the echocardiography reports.

6.2.2 Data preprocessing

Preprocessing of both 1-lead and 12-lead ECG signals followed the procedure described in Van der Valk *et al.* (2025) [155], see Figure 6.2 for an overview. The raw ECG signals were segmented into individual heartbeats, each defined as an 800 ms interval centered on the R-peak, 400 ms preceding and 400 ms following. R-peak detection was performed using a method inspired by RPNet, a U-Net–structured CNN, previously trained on manually annotated peak locations as described in Vijayarangan *et al.* (2020) [146].

Because the 1-lead recordings were 30s in duration compared to the 10s duration of the 12-lead recordings, the 1-lead data yielded a greater number of segmented beats. Given the low intra-recording variability and to maximize the available training data for the 1-lead model, all valid heartbeats were retained for this modality. Segmented beats containing excessive noise or flatline segments were excluded; for the 12-lead case, exclusion was done if any lead exhibited such artifacts.

6.2.3 Left ventricular function

We selected LVF as a prediction target due to its reflection of structural changes after MI and the presence of well-characterized ECG features. Several of these, such as the first QRS integral, ST integral, ST–T integral, and ST–T area, have shown discrimination in recovering MI patients [199]. The predictive capacity of these features differed per lead

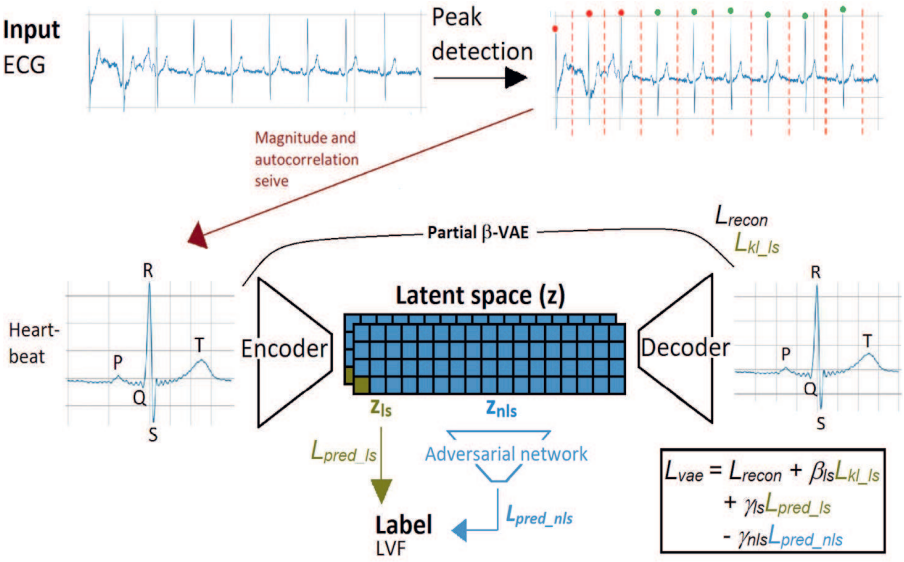


Figure 6.2: **Overview of our Processing and Prediction Framework:** Initially, the 12-lead ECGs are split into individual heartbeats using RPNet[146]. Subsequently, the heartbeats are sieved in two stages: the first based on individual magnitude, identifying low magnitude as indicative of a flat signal, and the second utilizing autocorrelation to assess correlation with other heartbeats from the same signal. The remaining heartbeats (indicated with green dots) are used as input for the VAE. The encoded representation z is derived, comprising z_{nls} and z_{ls}. While z_{ls} directly contributes to LVF label prediction with loss $L_{\mathrm{pred\_ls}}$, z_{nls} serves as input for an adversarial MLP also predicting the LVF label, introducing loss $L_{\mathrm{pred\_nls}}$. The decoder's input is sampled from z to produce the reconstructed heartbeat, subsequently compared with the original input heartbeat, resulting in the reconstruction loss L_{recon}.

and MI location [200]. T-wave morphology changes, including inversion and subsequent reinversion, have also been linked with favorable LVF recovery [201].

The ECG signals were labeled with LVF impairment, consisting of 4 ordinal categories: normal function, mild impairment, moderate impairment, and severe impairment. Labeling was performed using large language model (LLM)–based classification of echocardiogram reports obtained as part of the routine clinical workflow [202]. An echocardiogram report was used as a label if the examination was conducted within seven days before or after the corresponding ECG recording. The reports themselves were based on visual interpretation of the echocardiograms by clinicians.

6.2.4 Model overview and training

LVF prediction from both 1-lead and 12-lead ECGs employed the same model and training strategy as succesfully employed in Van der Valk *et al.* (2025). A partial β-variational autoencoder (VAE) (Figure 6.2) was used for both modalities. The architecture consisted of seven recurrent convolutional blocks in both the encoder and decoder. Four separate model instances were trained: i. a 12-lead model, a model trained and evaluated on the 12-lead ECGs, ii. a lead I model, a model trained and evaluated on lead I of the 12-lead ECGs, iii. a pretrained 1-lead model, a model pretrained on lead I of the 12-lead ECG and finetuned and evaluated on the 1-lead ECGs, and iv. a 1-lead model, a model trained and evaluated on the 1-lead ECGs. The models expect an input of shape 12×400; therefore, for the 1-lead recordings, each 1×400 heartbeat was duplicated along the lead dimension to match the required input shape.

6.2.5 Feature visualization

The model output and interpretability is shown by means of attribute manipulation as explained in Van der Valk *et al.* (2025)[155], see Figure 6.2. Each ECG heartbeat was encoded into z_{nls} and z_{ls}, of dimension 599 and one respectively. Then z was decoded to reconstruct the signal. For attribute manipulation 6 hypothetical signals were created by replacing z_{ls} with six values: four means for the four LVF groups (S_1–S_4) and two extremes (S_0: mean minus one standard deviation of severely impaired group; S_5: mean plus one standard deviation of normal LVF group). These six patient-specific hypothetical signals were used to illustrate how ECG morphology changes across LVF impairment levels and which features the model leverages for classification. These signals create a context for the ECG heartbeat w.r.t. LVF, allowing comparison against the whole database. This facilitates

signal interpretation and potentially enable interpretation by less trained medical personnel.

6.3 Experiments and Results

6.3.1 Experiments

Both datasets were split into a training set (89% of signals) and a holdout test set (11% of signals). The test set was constructed exclusively from the overlapping portion of the datasets, where both patients and acquisition dates (within a seven-day window) were overlapping, to enable a head-to-head comparison. A 5-fold cross-validation was performed in both training sets, maintaining an 85:15 ratio between the training and validation set within each fold. All data splits were grouped by patient and stratified by label when applicable. Training continued until convergence, defined as no improvement in validation loss for 15 consecutive epochs. To mitigate overfitting, we employed balanced sampling, early stopping, L2 regularization on the fully connected layer, and weight decay with the Adam optimizer. To prevent gradient explosion, He initialization was used [150], and to avoid numerical overflow, the standard deviations of the posterior $q(z|X)$ were clipped to the interval [-10,3] prior to sampling. The models were implemented and trained on a 24Gb GPU. Hyperparameter tuning resulted in $\beta_{ls} = 0$, $\beta_{nls} = 0.001$ and $\gamma = 2.0$ for both models. This is in line with the experiments shown by Van der Valk *et al.* (2025) [155].

6.3.2 Feature evaluation

LVF predictions were evaluated using three classification metrics: the Area Under the Receiver Operating Characteristic Curve (AUROC), average precision score (AP) and a heartbeat consistency score (HBC). These metrics were applied to a binary classification task (positive class: severe/moderate impairment vs. negative class: mild impairment/normal function), reflecting clinically relevant groupings and enabling comparison with previous studies. The HBC score was defined as the portion of heartbeats from the same signal that was classified in the same group. The AUROC and AP scores were calculated with the averaged predictions from all heartbeats in a signal.

Heartbeat reconstruction performance was assessed using the mean squared error (MSE) between the ECG input and the output signal. For intermodel comparisons of prediction scores and MSE, paired t-tests were performed. The confidence intervals (95%) for AUC, AP and HBC scores were estimated by bootstrapping with n = 1000 samples of the combined output distribution (5 models). The confidence interval for MSE was estimated

directly with the combined output error distribution.

6.3.3 Results

7.118 1-lead ECG signals and 7.593 12-lead ECG signals were be labeled with LVF. Respectively, 14% and 11% of these were labeled with severe/moderate impaired LVF.

Several data quality issues were identified in the 1-lead recordings, including devices worn on the wrong wrist, leading to signal inversion, device sharing among multiple individuals, and acquisition during non-resting states. These factors introduce substantial noise and may impair model performance. To mitigate this, we applied an autocorrelation-based filter that utilized preceding and subsequent signals from the same patient to exclude low-quality recordings. After filtering, 5,480 1-lead signals comprising 108,364 heartbeats from 460 patients were retained for model training. For the 12-lead recordings, no comparable quality concerns were observed; 7,197 signals and 19,933 heartbeats from 1,856 patients were included. The corresponding test sets consisted of 900 1-lead signals and 559 12-lead signals from 144 patients.

Table 6.1 and Figure 6.3 and 6.4 summarize the performance of the four models for LVF prediction on the test set. As anticipated, the full 12-lead model achieved the highest predictive performance across all metrics. The 1-lead models demonstrated reasonable discriminative ability, with only a modest reduction in performance compared to the 12-lead model. Pretraining the 1-lead model using lead I of the 12-lead ECG improved AUC and AP but did not improve heartbeat consistency (HBC). The lead I model exhibited the lowest AUC yet ranked second HBC after the full 12-lead model. In Figure 6.3 and 6.4 sensitivity at 0.9 and 0.95 are indicated with grey dotted lines. For a sensitivity of 0.9 (modere/severe impaired LVF detected in 90% of cases), specificities are: 0.64, 0.43, 0.71, and 0.65 for respectively the 12-lead, lead I, pretrained 1-lead and 1-lead models; for a sensitivity of 0.95 the respective specificities are: 0.49, 0.50, 0.55, and 0.33. All models showed high intra-recording stability, as reflected by elevated HBC scores: most QRS complexes within a single recording were consistently assigned to the same binary class. This robustness underscores the reliability of signal-level predictions. Figure 6.5 illustrates the AUC when heartbeat-level predictions were aggregated by averaging within each signal.

Evaluation of interpretability

Figure 6.6 and 6.7 show attribute manipulation of a heartbeat of the same patient with the pre-trained 1-lead and the full 12-lead model, respectively. The attribute manipulation can

Table 6.1: Comparison of all models for both reconstruction (MSE) and prediction of LVF (AUROC, AP and HBC). The AUROC, AP, and HBC scores are shown for the prediction of the moderate/severe impaired (positive) vs. normal/mild impaired groups. The AUROC, AP and HBC metrics show the bootstrapped mean (and 95% confidence interval) on the test set.

Model	$\mathrm{MSE}\downarrow$	AUROC ↑	AP ↑	HBC ↑
12-lead	0.027	**0.897**	**0.600**	**0.991**
	[0.010-0.087]	[0.887-0.906]	[0.573-0.626]	[0.857-1.000]
12-lead	0.007	0.805	0.357	0.980
(Lead I)	[0.002-0.036]	[0.792-0.816]	[0.332-0.382]	[0.667-1.000]
1-lead	**0.004**	0.883	0.511	0.959
(pretrained)	[0.001-0.011]	[0.881-0.886]	[0.501-521]	[0.632-1.000]
1-lead	**0.004**	0.863	0.459	0.961
	[0.002-0.012]	[0.859-0.866]	[0.449-0.469]	[0.647-1.000]

be interpreted as the rationale behind the model's prediction. The 12-lead model shows a much more diverse set of features in multiple leads. Lead I of the 12-lead model shows relatively similar output to the 1-lead model.

6.4 Discussion

This study provides a comparison of LVF prediction using matched 1-lead and 12-lead ECG datasets. As expected, the 12-lead model achieved the highest overall performance; however, the performance gap with the 1-lead models was modest, indicating that a substantial portion of LVF relevant information present in the full 12-lead ECG is also captured in a single lead. Model performance was comparable to, and in some instances slightly exceeded, previously reported results in broader cardiac populations [155, 196] (AUCs 0.83 and 0.89, respectively). Since early detection of ventricular dysfunction is particularly relevant in our post-myocardial infarction population, the findings of our study bear significant importance.

Clinical utility and operating points. At a sensitivity of 0.90, specificity scores were 0.64 (full 12-lead), 0.43 (lead I), 0.71 (pretrained 1-lead), and 0.65 (1-lead). These operating points are usable in clinical pathways. The scores show the percentage of healthy patients that should not be reviewed for a test that correctly predicts left ventricular dysfunction with an accuracy of 0.9. They could be used, for example, to trigger manual review or flag patients as likely unchanged/low risk, especially in settings where 12-lead acquisition is not feasible. Notably, averaging heartbeat-level predictions within a recording preferentially improved the 1-lead models, consistent with (i) the larger number of heartbeats available

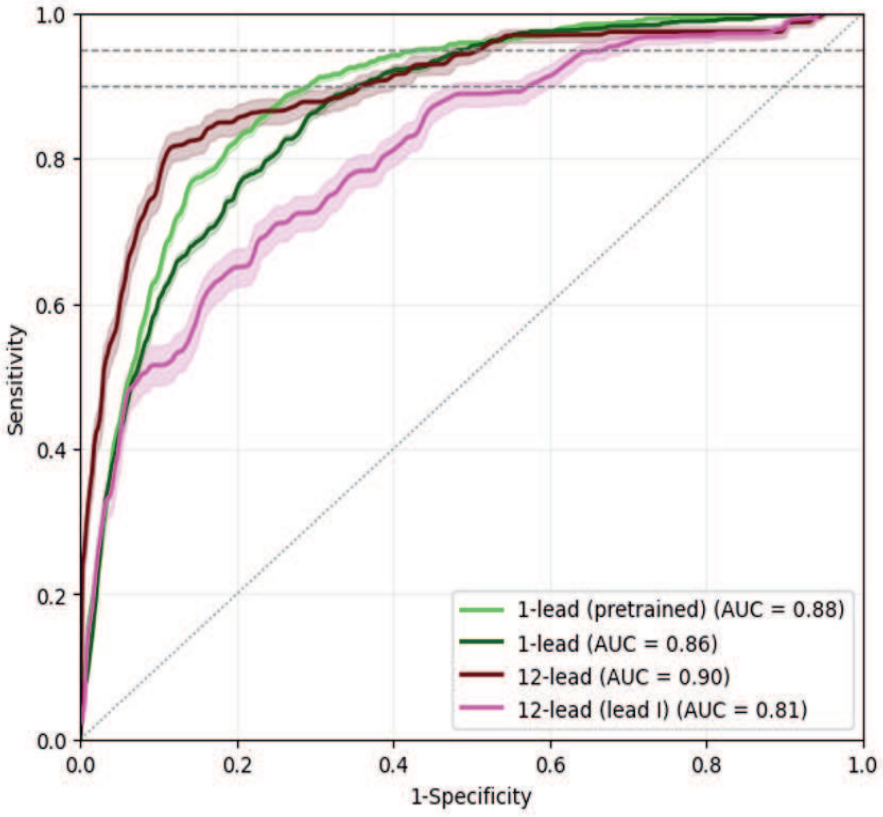


Figure 6.3: ROC analysis including 95% confidence intervals, comparing the four models. The ROC curves with the gray vertical dashed lines indicating recall/sensitivity 0.9 and 0.95

in the longer 1-lead signal (30s vs. 10s) and (ii) noise cancellation through aggregation. Together with the observed high heartbeat consistency, this also supports the benefit of temporal aggregation over multiple recordings.

Explainability as a driver for future research. A distinctive aspect of this work is the use of attribute manipulation to provide a transparent view of model reasoning. The 12-lead model leveraged a broader, multi-lead feature set (e.g. R-peak amplitude, notched R-peaks, T-wave morphology, see figure 6.7), whereas the 1-lead models used features that resembled the 12-lead feature set for lead I, which restricts interpretability and model performance, see figure 6.6. This observation raises important hypotheses for future investigation: (i) whether LVF prediction from a single lead is most effective in disease sub-

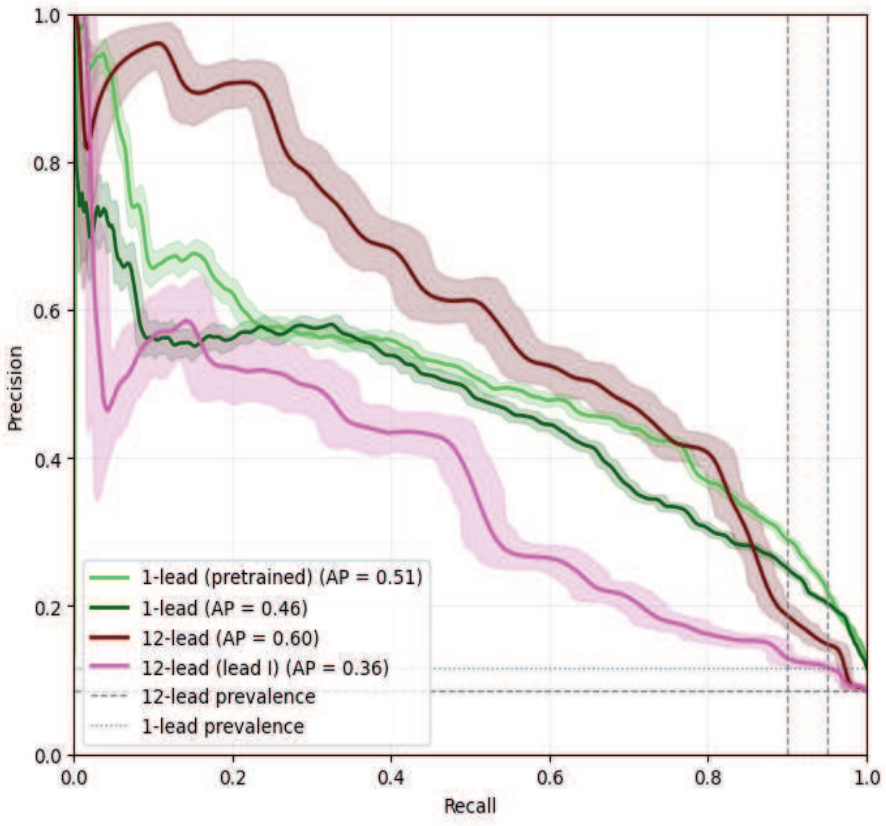


Figure 6.4: PR analysis including 95% confidence intervals, comparing the four models. The PR curves with the gray vertical dashed lines indicating recall/sensitivity 0.9 and 0.95. AP = average precision

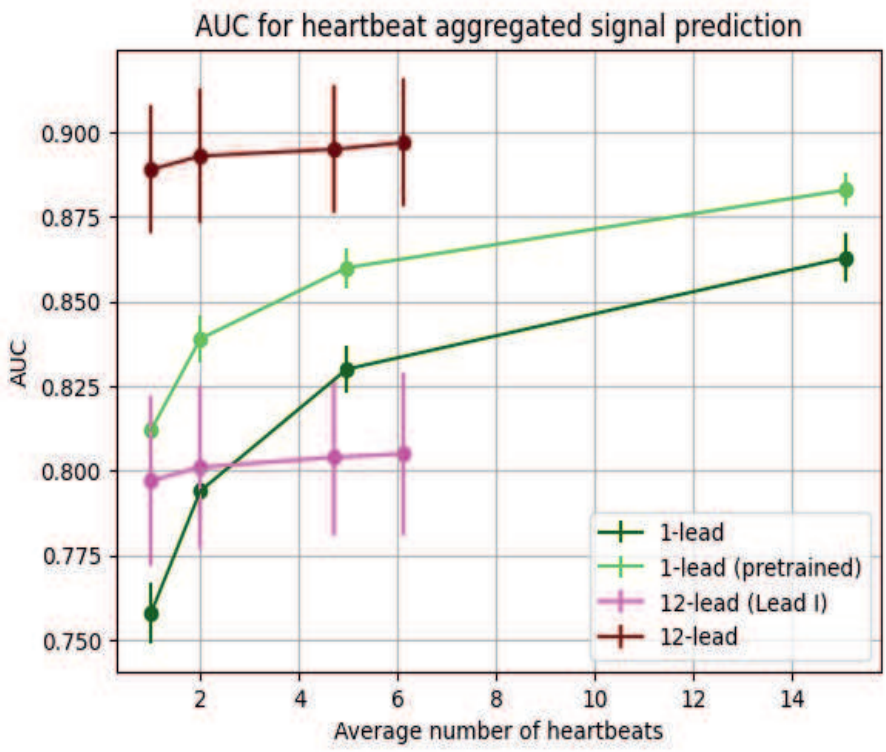


Figure 6.5: AUROC values for intra-signal aggregation of predictions. 1-lead models benefit significantly more from post-processing than 12-lead models.

types characterized by global or lateral abnormalities or lesions, where changes can be observed on lead I (and the 1-lead); (ii) whether multi-lead configurations are essential for more anterior or inferior infarctions; and (iii) how explainability can guide lead selection strategies for simplified ECG acquisition in wearable or resource-limited settings. The observed performance drop for the lead I model compared to the full 12-lead model reinforces the value of complementary spatial information, yet the improvement of the 1-lead model after heartbeat aggregation suggests that temporal averaging may partially compensate for reduced spatial coverage. Future work should validate these hypotheses and explore whether explainability-driven insights can inform adaptive acquisition protocols or subtype-specific prediction models.

Our model uniquely offers explainability that can facilitate clinical interpretation by less specialized personnel. By highlighting which waveform components drive predictions (e.g., T-wave morphology, R-peak amplitude, notched R-peaks), these tools can guide non-

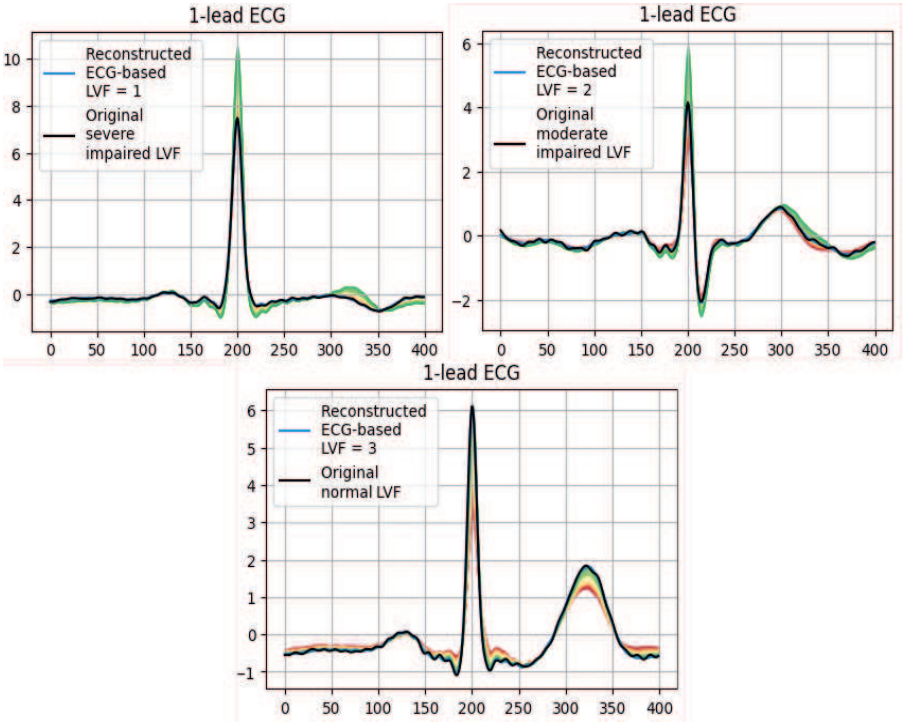


Figure 6.6: Attribute manipulation in three heartbeats from 1-lead ECGs with different LVF impairment. The black line shows the origin heartbeat of the 1-lead ECG, the blue line (largely covered by the black line here) shows the reconstructed heartbeat by the model. The green to red distributions around the reconstructed signal are the result of attribute manipulation and can be interpreted as the reason that the model classified the signal with the LVF value shown in the legend (0=severely impaired, 1=moderately impaired, 2=mildly impaired, 3=normal). The green signal shows the 1-lead signal that would be classified as having a normal LVF. The green signal in all figures shows a R-peak with larger amplitude, more pronounced Q- and S-waves and a T-wave with a slightly larger area. Both the T-wave and the R-peak are important prognostic factors in myocardial infarction. However, quantitative assessment of T-wave and R-peak changes after myocardial infarction is currently not part of the clinical routine in the non-acute setting.

experts in verifying signal plausibility and understanding when predictions are likely reliable versus when additional expert review is warranted. This capability is critical for scaling LVF prediction into resource-limited settings and home-based care, where cardiology expertise may not be immediately available.

Implications for wearables and remote care. While wrist-based 1-lead ECGs have primarily been optimized for rhythm analysis, such as atrial fibrillation detection, this

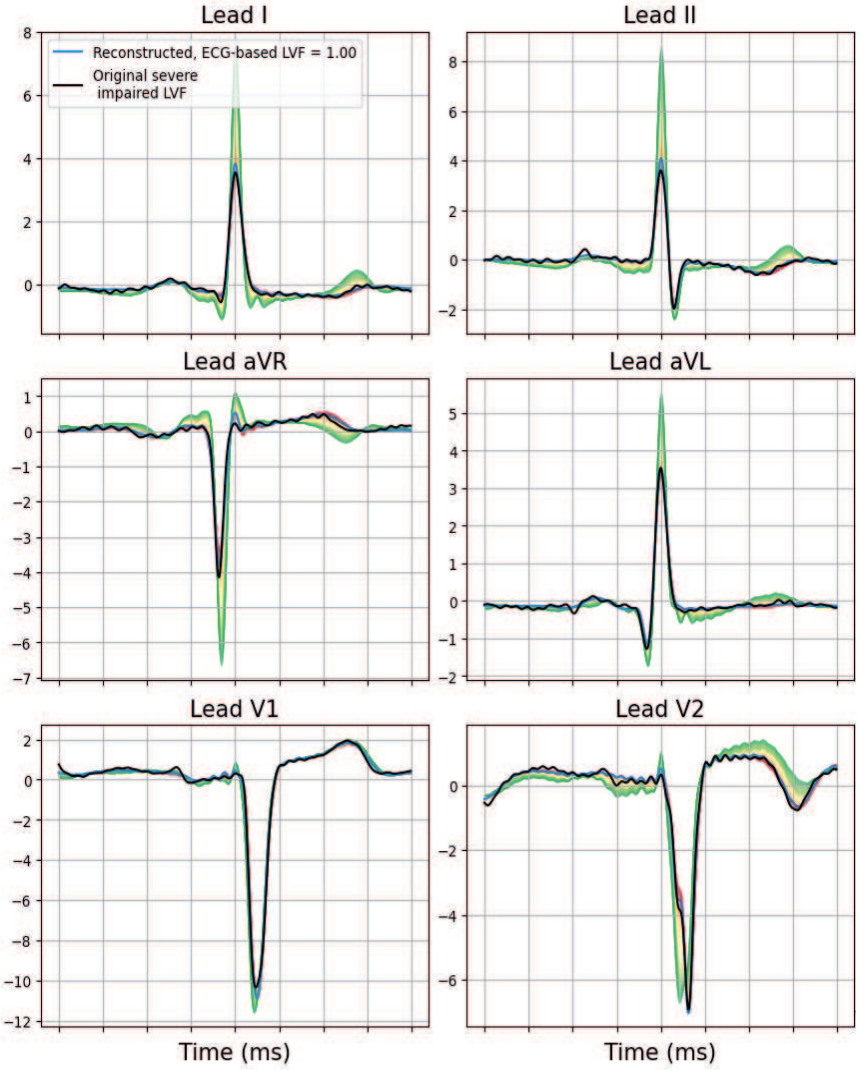


Figure 6.7: Attribute manipulation in a heartbeat from a subset of the leads of a 12-leads ECG of a patient with severe LVF impairment, the patient and visualization method are the same as in the first plot in Figure 6.6, see that caption for detailed explanation. In this case the green signal shows among others a larger R-peak amplitude, a T-wave with larger area and a R-peak without a notch in lead aVR, V2, aVF, V5, and V6. These are all important prognostic factors in myocardial infarction. However, quantitative assessment of these factors after myocardial infarction is currently not part of the clinical routine in the non-acute setting. Prediction to LVF: 0=severely impaired, 1=moderately impaired, 2=mildly impaired, 3=normal.

study demonstrates their potential for structural and functional cardiac assessment. Although 1-lead ECGs are not intented to replace 12-lead ECGs, the combination of reasonable discriminative performance, transparent mechanism, and aggregation-driven stability, supports 1-lead ECGs as a practical tool for screening and potentially longitudinal monitoring. It could be used as a scalable clinical complement for (automated) early warning and for home-based follow-up where high quality 12-lead acquisition is not possible.

Data quality and future work. We addressed 1-lead data quality challenges (device orientation errors, device sharing, non rest acquisition) using an autocorrelation based filter informed by adjacent recordings from the same patient. However, differentiating true pathological inversions from orientation-related inversions remains non-trivial, since several conditions transiently invert peaks or alter large ECG segments. We explored heartbeat autocorrelation, landmark based checks, and learned classifiers with mixed success. Future research directions to address these issues and extend currently methodology include: (i) acquisition protocol refinement to reduce incorrect device orientation and activity artifacts, (ii) robust automated quality control that combines signal-to-signal and even beat-to-beat consistency and anomaly detection, (iii) temporal modeling (e.g. multi visit aggregation) to convert stable beat-level consistency into patient-level trajectories, (iv) model performance analysis in clinical subtypes (global/lateral repolarization abnormalities vs. regional abnormalities) and (v) clinical validation with calibrated, use case specific thresholds (e.g. fixed 0.90 specificity), including health system impact metrics.

Conclusion The 12-lead ECG remains the gold standard for LVF prediction, yet 1-lead models achieve clinically meaningful performance, particularly when predictions are aggregated. Our model was notably designed to enable explainability of the predictions, highlighting a confined single-lead feature set in the 1-lead models versus multi-feature, multi-lead reasoning in the 12-lead model. This provides the mechanistic insight needed to scale these tools to real world applications.

Acknowledgments

Sources of Funding

This project has received funding from the European Institute of Innovation and Technology (EIT) Health under grant agreement 230193 (ASSIST)

Disclosures

None

7

Summary and future work

Summary

This thesis explored the integration of AI, and particularly deep learning, into various stages of the myocardial infarction (MI) clinical workflow, ranging from imaging and physiological monitoring to the extraction of structured data from free-text reports. Across different modalities and tasks, we investigated how AI and, if relevant, explainable AI methods can reduce clinical workload.

In **Chapter 2**, we addressed the growing interest in deep learning–based analysis of invasive coronary angiography (ICA) by conducting a comprehensive systematic review. The field is marked by heterogeneity in task definitions, evaluation metrics, and datasets, which hinders direct comparison and clinical translation. By organizing existing literature into four primary tasks of the clinical workflow: frame selection, vessel segmentation, lesion localization, and lesion assessment, we provided structure to a fragmented research landscape. We further identified methodological trends and common limitations, and proposed clinically relevant future direction and recommendations aimed at facilitating fair benchmarking.

In **Chapter 3**, we investigated how LLMs can automate the extraction of clinically relevant variables from semi-structured ICA and echocardiography reports. By evaluating multiple LLM approaches, prompt engineering and fine-tuning, across different computing environments, we demonstrated that LLM performance in general is strong, but depends on

task complexity, model size and class imbalance. Taking these into account, we concluded that even smaller locally run LLMs have a use-case in extracting simple labels, such as left ventricular function (LVF) or treatment strategy, from cardiac reports. This work enables the automated structuring of large volumes of clinical data, laying the groundwork for scalable cohort creation, model training, and retrospective analysis. The successful classification of echocardiogram reports with LLMs, directly accelerated the labeling process for the ECG classification studies in **Chapter 4, 5 and 6**. LLMs have a direct use case here which speeds up the research workflow, but they could similarly speed up the clinical workflow.

Finally in **Chapter 4, 5 and 6**, we focused on ECG as a widely accessible, non-invasive signal for cardiac assessment. We developed and evaluated explainable AI models based on VAEs, aimed at predicting LVF, a critical post-MI prognostic marker, and mortality. In **Chapter 4**, we improved prediction (AUC=0.853) and explainability by jointly optimizing reconstruction and classification objectives, demonstrating that meaningful cardiac features can be extracted from the 12-lead ECG. In **Chapter 5**, we extended this approach by explicitly disentangling LVF-relevant features in the latent space through supervised adversarial learning. This allowed us not only to maintain predictive performance (AUC=0.868), but also to generate interpretable signal-specific visualizations of how variations in ECG signals relate to LVF, thereby enhancing model transparency and trustworthiness in a clinical setting. In **Chapter 6** we applied the same method to the 1-lead smartwatch ECG and compared LVF prediction from the 1-lead (AUC=0.883) and the 12-lead (AUC=0.897) ECG. This comparison aimed to evaluate the use of the 1-lead ECG as a proxy for both 12-lead ECG-derived LVF and echocardiogram-derived LVF.

Taken together, the studies in this thesis illustrate how deep learning models, when carefully tailored to clinical tasks, can contribute to a more efficient and scalable MI workflow. From automating complex imaging analyses to structuring textual data and deriving interpretable physiological markers, AI offers tangible solutions to alleviate the growing burden on healthcare systems. However, several challenges remain.

Challenges in Cardiac AI

A substantial portion of research in cardiac AI and deep learning lacks rigorous statistical evaluation. Common errors include the use of the term 'significant' without acknowledging its statistical meaning, the failure to perform proper statistical comparisons, and the incorrect application of tests, confidence intervals, or p-values. Peer-reviewed studies frequently

contain many of these mistakes when comparing model performance metrics. Statistical comparisons should be made between distributions of metrics rather than between individual metric values. For aggregated metrics such as AUC, F1 score, or accuracy, there are methods to calculate statistical difference or distributions can be generated through approaches like bootstrap sampling [151]. While cross-validation produces multiple model instances, it is generally intended to assess variability in model performance rather than to create a proper distribution of a metric for formal statistical testing. Incorrect use of statistics can lead to overconfident or incorrect conclusions, which again results in overconfidence and a dilution of the research quality in the field.

Data quality is another important issue in cardiac AI research. Clean dataset collection is essential, yet often challenging. Many variables recorded in the clinic are not collected with algorithm development in mind; rather, they may serve administrative or financial purposes, reflect interobserver variability or bias, depend heavily on local workflows, or are collected by patients who variably adhere to data collection instructions. For example, insurance reimbursement codes at LUMC reflect the expense of procedures more than their clinical importance, resulting in a partial and sometimes misleading representation of the clinical context. Furthermore, in hospital settings, device usage may introduce bias if patients with poorer outcomes are disproportionately imaged on specific devices. Namely, deep learning models can capture device-specific trace marks in images. Similarly, smartwatch-derived data can be compromised when devices are occasionally shared with others without informing healthcare professionals or when data acquisition is rushed due to forgetfulness. These factors not only introduce noise in the dataset, but also complicate the definition of objective ground truths, and introduce bias into AI model development and evaluation. As such, AI models must be critically assessed not just for predictive performance, but also for their relationship to meaningful and unbiased clinical endpoints.

Most work, including ours, does not focus on clinical implementation of algorithms which is an important next step after development and validation. This step is often not taken, since it requires software engineering, integration in available workflows, passing strict regulatory requirements and extensive clinical testing. These steps are usually not the main focus of (PhD) research, but are nonetheless of vital importance for clinical adoption.

Moreover, to fully understand the value of AI in the clinic, a critical question must be asked: *What is the true contribution of AI?* Does it measurably save time, improve diagnostic accuracy, or provide novel insights into disease processes or does it merely out-

perform a benchmark of questionable clinical relevance? More often than not, research is focused on outperforming previous research without relating to the clinical importance of such an improvement. When combined with the aforementioned dilution of research quality in the field, this trend results in a proliferation of seemingly promising studies that do not translate into actionable tools for clinical practice.

Contributions to Cardiac AI

The final three chapters of this thesis demonstrate the potential of both 1-lead and 12-lead ECGs as echocardiogram-derived LVF predictors. Typically, LVF is derived from echocardiograms; automated and explainable ECG-derived LVF predictions would enable a shift in cardiac monitoring from being the exclusive domain of highly trained specialists to also supporting less specialized healthcare personnel in broader clinical and remote settings. Ultimately, LVF predictions from these modalities could serve as early warning signals. Further automation may also facilitate more personalized and adaptive follow-up schedules. For example, patients who demonstrate a decrease in cardiac function based on self-measured 1-lead ECGs could be prompted to have more frequent visits, while patients who show improvement might safely forgo routine appointments. Such an approach could lead to earlier detection of adverse events, reduced disease burden, and fewer unplanned hospital visits, outcomes that are both clinically beneficial and cost-efficient from a healthcare systems perspective.

The explainability methods introduced in this study offer more detailed insights than most other explainable methods, such as the commonly used SHAP or Grad-CAM techniques [203, 115]. Unlike these techniques, which primarily highlight input regions influencing the output, the proposed approach motivates predictions with relative changes in features, such as the shape and size of ECG waveforms, providing a richer interpretation of the models rationale. In addition, it enables projection onto a landscape of ECG data, allowing transitions between signals to be visualized and even labeled as improvements or deterioration. This capability could be particularly valuable for longitudinal monitoring, where multiple patient-specific ECG signals must be contextualized and compared over time. In contrast to GradCAM, which can be applied to any network, the explainability methods introduced in this thesis are less versatile and require an encoder–decoder architecture, which effectively doubles the number of parameters, and rely on both effective disentanglement of the latent space and robust reconstruction performance. On the other

hand, the reconstruction quality itself can serve as an indicator of uncertainty in the explainability output, which could be valuable in, for example, outlier detection.

In Chapter 4, disentanglement is achieved indirectly through the KL-divergence loss, whereas Chapters 5 and 6 complement this with an adversarial network and a split of the latent space, offering a more direct approach. Both strategies fall within the broader field of disentanglement, which is widely applied, among others, in image generation. In principle, alternative or more recent disentanglement models, such as StyleGANs or other (variational) autoencoder-based architectures, could be adapted for the explainability pipelines introduced in Chapters 4–6 [156]. This, as well as the use of larger datasets, could lead to improved disentanglement of the latent space, which, in turn, may result in better predictive performance and more robust explainability. Despite these opportunities for refinement, the results presented in the final chapters are promising. They deserve further validation in clinical environments to determine the true value of estimating cardiac health from both remote and in-hospital ECG recordings. In particular, research that aims to quantify the time and cost-saving potential of automated 1-lead ECG monitoring could provide insight into the clinical utility of wearable devices such as ECG smartwatches. Demonstrating their effectiveness in patient monitoring could support broader adoption and integration into routine care.

Future work in cardiac AI in general should focus not only on improving model architecture and interpretability, but also on the validity and intent of the underlying data and on clinical implementation. Transparency about data origin, quality control, and label reliability is essential. Thereby, collaboration between domain experts and model developers is vital to ensure that AI tools are trained, evaluated, and deployed in a manner that aligns with true clinical needs. Another important direction for future research is the multimodal integration of cardiac data, reflecting the way cardiologists combine patient history and information from multiple modalities when making decisions. In clinical practice, factors such as age and medical background often shape how cardiologists interpret ECG findings. This observation highlights the need for AI models that move beyond single-modality inputs and instead integrate heterogeneous data sources to more closely approximate expert clinical reasoning.

Overall, the work in this thesis contributes to the understanding of how AI can be thoughtfully applied in cardiology, but also underscores the necessity of grounding such applications in rigorous, context-aware clinical and data practices.

Conclusion

The MI care pathway is complex and labor intensive and will therefore become increasingly unsustainable in the face of increasing prevalence of cardiovascular disease [1]. This thesis presents a multifaceted approach to help alleviate this burden through the integration of explainable, efficient, and scalable AI methods across imaging, physiological signals, and clinical text.

We have shown that deep learning can enhance diagnostic accuracy and interpretability in ICA and ECG analysis, and that large language models can unlock the value of unstructured clinical reports, both of which are key enablers for generating large, labeled datasets and supporting data-driven decision-making. By identifying methodological gaps, proposing practical solutions, and demonstrating proof-of-concept implementations, this work contributes to the growing body of research aiming to bridge the gap between AI development and clinical adoption.

Ultimately, the impact of AI in cardiology will depend not only on technical advances but also on careful integration into clinical workflows, transparency of algorithms, and the availability of high-quality data. Continued interdisciplinary collaboration between clinicians, engineers, and data scientists will be crucial to ensure that the full potential of AI is realized in improving patient outcomes in cardiovascular care.

Samenvatting en toekomstig werk

Samenvatting

In deze thesis wordt de integratie van AI, en in het bijzonder deep learning, in verschillende stadia van de klinische workflow van het myocardinfarct (MI) belicht, variërend van beeldvorming en fysiologische monitoring tot het verkrijgen van gestructureerde data uit vrije-tekst rapporten. In verschillende data modaliteiten en voor verschillende taken wordt onderzocht hoe AI en, indien relevant, explainable AI-methoden de klinische werkdruk kunnen verminderen.

In Hoofdstuk 2 gaan we in op de groeiende interesse in deep learning gebaseerde analyse methoden voor invasieve coronaire angiografie (ICA) door een systematische review van het vakgebied te doen. Het veld wordt gekenmerkt door heterogeniteit in taken, evaluatie methoden en datasets, wat directe vergelijking en klinische translatie bemoeilijkt. Door de bestaande literatuur te verdelen in vier primaire taken van de klinische workflow: frameselectie, vatsegmentatie, laesielocalisatie en laesiebeoordeling, hebben we geprobeerd structuur aan te brengen in een gefragmenteerd onderzoekslandschap. We identificeren daarnaast methodologische trends en veelvoorkomende beperkingen, en stellen klinisch relevante toekomstige richtingen en aanbevelingen voor, om eerlijke benchmarking te faciliteren.

In Hoofdstuk 3 onderzoeken we hoe LLM's de extractie van klinisch relevante variabelen uit semi-gestructureerde ICA- en echocardiografie-rapporten kunnen automatiseren. Door meerdere LLM trainingstechnieken, prompt engineering en fine-tuning te evalueren op GPUs van verschillende grootte, tonen we aan dat de nauwkeurigheid van LLM's over het algemeen goed is, maar afhangt van taakcomplexiteit, modelgrootte en klassenverdeling.

Desalniettemin laten we zien dat zelfs kleinere lokaal draaiende LLM's bruikbaar zijn voor de accurate extractie van eenvoudige labels, zoals linkerventrikelfunctie (LVF) of behandelstrategie, uit rapporten van hartpatiënten. Dit werk maakt de geautomatiseerde

structurering van grote hoeveelheden klinische data mogelijk en legt de basis voor schaalbare cohortcreatie, modeltraining en retrospectieve analyse. De succesvolle classificatie van echocardiogramrapporten met LLM's versnelde direct het labelproces voor de ECG-classificatiestudies in Hoofdstuk 4, 5 en 6. LLM's hebben hier een directe toepassing die de onderzoeksworkflow versnelt, maar zouden op vergelijkbare wijze ook de klinische workflow kunnen versnellen.

Tot slot richten we ons in Hoofdstuk 4, 5 en 6 op het ECG als een breed toegankelijke, niet-invasieve databron voor het beoordelen van vaatlijden. We behandelen de ontwikkeling en evaluatie van explainable AI-modellen gebaseerd op VAEs, gericht op het voorspellen van LVF, een cruciale prognostische marker na MI, en mortaliteit.

In Hoofdstuk 4 verbeteren we deze voorspelling (AUC=0.853) en explainability door reconstructie- en classificatiedoelen gezamenlijk te optimaliseren, hiermee tonen we aan dat betekenisvolle kenmerken kunnen worden geëxtraheerd uit het 12-afleidingen ECG. In Hoofdstuk 5 breiden we deze aanpak uit door LVF-specifieke kenmerken expliciet te ontkoppelen in de latente ruimte van de VAE via supervised adversarial learning.

Dit stelt ons niet alleen in staat om de voorspellende waarde te behouden (AUC=0.868), maar ook om interpreteerbare signaal-specifieke visualisaties te genereren van hoe variaties in het ECG signaal zich verhoud tot de LVF. Dit vergroot de transparantie en betrouwbaarheid van het model in een klinische setting. In Hoofdstuk 6 passen we dezelfde methode toe op het 1-afleiding smartwatch ECG en vergelijken we de LVF-voorspelling van het 1-afleiding (AUC=0.883) en het 12-afleidingen (AUC=0.897) ECG. Deze vergelijking is bedoeld om het gebruik van het 1-afleiding ECG als proxy te evalueren voor zowel LVF voorspeld met het 12-afleidingen ECG als direct van het echocardiogram.

De studies in deze thesis tonen aan hoe deep learning-modellen, wanneer zorgvuldig afgestemd op klinische taken, kunnen bijdragen aan een efficiëntere en schaalbare MI-workflow. Van het automatiseren van complexe beeldanalyses tot het structureren van tekst en het afleiden van interpreteerbare fysiologische markers, kan AI handige oplossingen bieden om de toenemende druk op zorgsystemen te verlichten. Er zijn echter nog een groot aantal uitdagingen bij het gebruik van AI in de cardiologie.

Uitdagingen voor het gebruik van AI in de cardiologie

Een aanzienlijk deel van het onderzoek naar cardiale AI en deep learning mist een goed onderbouwde statistische evaluatie. Veelvoorkomende fouten zijn onder meer het gebruik

van de term ‘significant’ zonder recht te doen aan de statistische betekenis ervan, incorrecte statistische vergelijkingen en het verkeerd toepassen van toetsen, betrouwbaarheidsintervallen of p-waarden. Ook gereviewde studies bevatten vaak een van deze fouten bij het vergelijken van modellen. Statistische vergelijkingen moeten worden gemaakt tussen verdelingen van uitkomstmaten in plaats van tussen individuele uitkomstmaten zelf. Voor samengestelde metrics zoals AUC, F1-score of accuracy bestaan er methoden om statistische verschillen te berekenen of verdelingen te genereren via technieken als bootstrap sampling [151]. Hoewel cross-validatie meerdere modelinstanties oplevert, is het doorgaans bedoeld om variabiliteit in modelprestaties te beoordelen en niet om een verdeling van een uitkomstmaat voor statistische toetsing te creëren. Verkeerd gebruik van statistiek kan leiden tot overmatig vertrouwen of verkeerde conclusies, wat resulteert in een verwatering van de onderzoekskwaliteit in het veld.

De kwaliteit van data is een ander belangrijk probleem in cardiaal AI-onderzoek. Het verzamelen van bruikbare datasets is essentieel, maar vaak uitdagend. Veel variabelen die in de kliniek worden geregistreerd, zijn niet verzameld met algoritmeontwikkeling in gedachten; ze kunnen administratieve of financiële doeleinden dienen, interobservervariabiliteit of bias weerspiegelen, sterk afhankelijk zijn van lokale workflows, of worden verzameld door patiënten die meetinstructies niet altijd zorgvuldig volgen. Bijvoorbeeld, declaratiecodes in het LUMC weerspiegelen voornamelijk de kosten van procedures in plaats van hun klinische relevantie, wat resulteert in een gedeeltelijke en soms misleidende weergave van de klinische context. Bovendien kan het gebruik van apparaten in ziekenhuizen bias introduceren als patiënten met slechtere uitkomsten onevenredig vaak met specifieke apparaten worden gescand. Deep learning-modellen kunnen namelijk apparaatspecifieke sporen in beelden herkennen. Ook kan smartwatch data ruis bevatten wanneer smartwatches worden gedeeld met anderen zonder zorgverleners te informeren, of wanneer acquisitie haastig wordt gedaan. Deze factoren zorgen niet alleen voor ruis in de dataset, maar maken sommige metingen ook volledig incorrect, waardoor er bias in de ontwikkeling en evaluatie van AI-modellen sluipt. Daarom moeten AI-modellen kritisch worden beoordeeld, niet alleen op voorspellende uitkomsten, maar ook op hun relatie tot betekenisvolle klinische eindpunten zonder bias.

Het merendeel van de deep learning studies, inclusief het onze, richt zich niet op de klinische implementatie van algoritmen, wat een belangrijke volgende stap is na ontwikkeling en validatie. Deze stap wordt vaak niet gezet, omdat het software-engineering, inte-

gratie in bestaande workflows, het voldoen aan strikte regelgeving en uitgebreide klinische tests vereist. Deze stappen zijn doorgaans niet het hoofddoel van (promotie)onderzoek, maar zijn desalniettemin van cruciaal belang voor klinische adoptie.

Bovendien moet, om de waarde van AI in de kliniek goed te kunnen beoordelen, een kritische vraag worden gesteld: Wat is de werkelijke bijdrage van AI? Bespaart het aantoonbaar tijd, verbetert het de diagnostiek, biedt het nieuwe inzichten in ziekteprocessen, of overtreft het slechts een benchmark van twijfelachtige klinische relevantie? Vaak richt onderzoek zich op het verbeteren van een specifieke uitkomstmaat uit eerder onderzoek zonder de klinische betekenis van een dergelijke verbetering vast te stellen. In combinatie met de eerder genoemde verwatering van de onderzoekskwaliteit in het veld, leidt deze trend tot een proliferatie van ogenschijnlijk veelbelovende studies die niet worden omgezet naar bruikbare tools voor de klinische praktijk.

Bijdragen aan AI in de cardiologie

De laatste drie hoofdstukken van deze thesis tonen het potentieel van zowel het 1-afleiding als het 12-afleidingen ECG als voorspellers van LVF. Normaal wordt LVF bepaald op basis van het echocardiogram. Geautomatiseerde en explainable LVF-voorspellingen op basis van het ECG zouden een verschuiving mogelijk maken van cardiale monitoring als exclusief domein van hoogopgeleide specialisten naar minder gespecialiseerde zorgverleners in de bredere klinische en thuis setting. Uiteindelijk zouden LVF-voorspellingen uit het 1-afleidingen ECG kunnen dienen als early warning. Verdere automatisering kan ook meer gepersonaliseerde en adaptieve follow-up schema's faciliteren. Patiënten die bijvoorbeeld een afname in hartfunctie vertonen op basis van zelfgemeten 1-afleiding ECG's kunnen vaker voor controle worden gevraagd te komen, terwijl patiënten die verbetering tonen een afspraak juist zou kunnen overslaan. Een dergelijke aanpak kan leiden tot snellere detectie van complicaties, een verminderde ziektelast en minder ongeplande ziekenhuisbezoeken, uitkomsten die zowel klinisch gunstig als kostenefficiënt zijn vanuit het perspectief van zorginstituten.

De explainability methoden die in deze studie zijn geïntroduceerd, geven meer gedetailleerde inzichten dan de meeste andere methoden, zoals de veelgebruikte SHAP- of Grad-CAM-technieken [203, 115]. In tegenstelling tot deze technieken, die voornamelijk regio's markeren in de inputdata die de uitkomst beïnvloeden, geeft onze methode de karakteristieken van ECG patronen zoals de vorm en grootte van ECG-segmenten, wat

een betere interpretatie van de redenering van het model biedt. Bovendien maakt het projectie op een landschap van ECG signalen mogelijk, waardoor transities tussen signalen kunnen worden gevisualiseerd en zelfs gelabeld als verbeteringen of verslechteringen.

Deze mogelijkheid kan bijzonder waardevol zijn voor longitudinale monitoring, waarbij meerdere patiëntspecifieke ECG signalen moeten worden gecontextualiseerd en in de tijd vergeleken. In tegenstelling tot GradCAM, dat op elk netwerk kan worden toegepast, zijn de explainability methodes die in deze thesis zijn geïntroduceerd minder veelzijdig en vereisen ze een encoder-decoderarchitectuur, die het aantal parameters effectief verdubbelt. Daarnaast is de ontkoppeling van informatie uit het input signaal in de latente ruimte en de correcte reconstructie van het signaal essentieel, wil het model correcte voorspellingen doen. Anderzijds kan de reconstructiekwaliteit zelf dienen als een indicator van onzekerheid van de voorspelling, wat waardevol kan zijn bij bijvoorbeeld outlierdetectie.

In Hoofdstuk 4 wordt informatie ontkoppeling indirect bereikt via de KL-divergence loss functie, Hoofdstukken 5 en 6 vullen deze loss aan met een adversarial netwerk en een splitsing van de latente ruimte, wat een directere ontkoppeling van informatie geeft. Beide strategieën vallen binnen het bredere veld van disentanglement, dat veel wordt toegepast in onder andere beeldgeneratie. In principe zouden alternatieve of meer recente modellen, zoals StyleGANs of andere (variational) autoencoder-gebaseerde architecturen, kunnen worden aangepast voor de explainability pipelines die in Hoofdstukken 4–6 zijn geïntroduceerd [156]. Dit, evenals het gebruik van grotere datasets, zou kunnen leiden tot verbeterde informatie ontkoppeling in de latente ruimte, wat op zijn beurt kan resulteren in betere voorspellingen en explainability. Ondanks deze mogelijkheden voor verfijning, zijn de resultaten die in de laatste hoofdstukken worden gepresenteerd veelbelovend. Ze verdienen verdere validatie in klinische context om de werkelijke waarde van het voorspellen van de hartfunctie uit zowel remote als inhuis ECG-opnames te bepalen. Onderzoek naar het tijd- en kostenbesparende potentieel van geautomatiseerde 1-afleiding ECG-monitoring kan inzicht geven in de klinische toepasbaarheid van wearables zoals ECG-smartwatches. Het aantonen van hun effectiviteit in patiëntmonitoring kan bredere adoptie en integratie in dagelijkse zorg ondersteunen.

Vervolgonderzoek in cardiale AI moet zich niet alleen richten op het verbeteren van modelarchitectuur en explainability, maar ook op de betrouwbaarheid en bias van verzamelde data en op klinische implementatie. Kwaliteitscontrole van, en transparantie over de herkomst van data is essentieel. Samenwerking tussen domeinexperts en modelon-

twikkelaars is daarbij van vitaal belang om ervoor te zorgen dat AI-tools worden getraind, geëvalueerd en ingezet op een manier die aansluit bij klinische behoeften. Een andere belangrijke richting voor vervolgonderzoek is de multimodale integratie van cardiale data, in navolging op de manier waarop cardiologen patiëntgeschiedenis en informatie uit meerdere modaliteiten combineren bij het nemen van beslissingen. In de klinische praktijk beïnvloeden factoren zoals leeftijd en medische achtergrond vaak hoe cardiologen ECG-bevindingen interpreteren. Dit benadrukt de noodzaak van multimodale AI-modellen, die meerdere databronnen integreren om de klinische redenering van experts beter te benaderen.

Het onderzoek in deze thesis schetst een roadmap voor AI in de cardiologie en benadrukt de noodzaak van betrouwbare data en klinische validatie bij het trainen van AI modellen.

Conclusie

Het MI-zorgpad is complex en arbeidsintensief en zal daarom steeds onhoudbaarder worden gezien de toenemende prevalentie van cardiovasculaire aandoeningen [1]. Deze thesis presenteert hiervoor een veelzijdige aanpak met behulp van schaalbare explainable AI-methoden voor beeldvorming, fysiologische signalen en klinische tekst.

We hebben aangetoond dat deep learning de diagnostische nauwkeurigheid en explainability kan verbeteren in ICA- en ECG-analyse, en dat grote taalmodellen labels uit ongestructureerde klinische rapporten kunnen halen, beide cruciale factoren voor het genereren van grote, gelabelde datasets en het ondersteunen van data-driven besluitvorming.

Door methodologische tekortkomingen te identificeren, praktische oplossingen voor te stellen en proof-of-concept implementaties te demonstreren, draagt dit werk bij aan de groeiende hoeveelheid onderzoek die erop gericht is de kloof tussen AI-ontwikkeling en klinische adoptie te overbruggen.

Uiteindelijk zal de impact van AI in de cardiologie niet alleen afhangen van technische vooruitgang, maar ook van zorgvuldige integratie in klinische workflows, transparantie van algoritmen en de beschikbaarheid van hoogwaardige data. Interdisciplinaire samenwerking tussen clinici, ingenieurs en datawetenschappers zal cruciaal zijn om ervoor te zorgen dat het volledige potentieel van AI wordt benut om patiëntuitkomsten in cardiovasculaire zorg te verbeteren.

8

Appendices

References

[1] M. A. Khan et al., Global epidemiology of ischemic heart disease: Results from the global burden of disease study, Cureus **12** (2020).

[2] I. S. Santos et al., One-year mortality after an acute coronary event and its clinical predictors: the ERICO study, Arquivos brasileiros de cardiologia **105**, 53–64 (2015).

[3] R. A. Byrne et al., 2023 ESC Guidelines for the management of acute coronary syndromes, Eur. Heart J. **44**, 3720–3826 (2023).

[4] C. Vrints et al., 2024 ESC Guidelines for the management of chronic coronary syndromes, Eur. Heart J. **45**, 3415–3537 (2024).

[5] S. Datta et al., Towards a Robust Deep Learning Model for Atrial Fibrillation Detection from a Short Single Lead ECG Recording, Intensive Care Medicine Experimental **11** (2023).

[6] S. Hong et al., Combining Convolutional and Recurrent Neural Networks for Atrial Fibrillation Detection, arXiv preprint arXiv:2110.13985 (2021).

[7] T. Teijeiro, C. A. García, D. Castro, and P. Félix, Arrhythmia Classification from the Abductive Interpretation of Short Single-Lead ECG Records, arXiv preprint arXiv:1711.03892 (2017).

[8] F.-J. Neumann et al., 2018 ESC/EACTS Guidelines on myocardial revascularization, European heart journal **40**, 87–165 (2019).

[9] J. Parker and R. Case, Normal left ventricular function., Circulation **60**, 4–12 (1979).

[10] M. Neves and J. Ševa, An extensive review of tools for manual annotation of documents, Brief. Bioinform. **22**, 146–163 (2021).

[11] H. Zhang et al., Comparison of physician visual assessment with quantitative coronary angiography in assessment of stenosis severity in China, JAMA internal medicine **178**, 239–247 (2018).

[12] R. Herscovici, T. Sedlak, J. Wei, C. J. Pepine, E. Handberg, and C. N. Bairey Merz, Ischemia and no obstructive coronary artery disease (INOCA): what is the risk?, Journal of the American Heart Association **7**, e008868 (2018).

[13] R. Konst, J. Meeder, M. Wittekoek, A. Maas, Y. Appelman, J. Piek, T. van de Hoef, P. Damman, and S. Elias-Smale, Ischaemia with no obstructive coronary arteries, Netherlands Heart Journal **28**, 66–72 (2020).

[14] K. C. Koskinas et al., Current Use of Intracoronary Imaging in Interventional Practice—Results of a European Association of Percutaneous Cardiovascular Interventions (EAPCI) and Japanese Association of Cardiovascular Interventions and Therapeutics (CVIT) Clinical Practice Survey—, Circulation journal **82**, 1360–1368 (2018).

[15] S. Koganti, T. Kotecha, and R. D. Rakhit, Choice of intracoronary imaging: when to use intravascular ultrasound or optical coherence tomography, Interventional Cardiology Review **11**, 11 (2016).

[16] B. K. Nallamothu et al., Comparison of clinical interpretation with visual assessment and quantitative coronary angiography in patients undergoing percutaneous coronary intervention in contemporary practice: the Assessing Angiography (A2) project, Circulation **127**, 1793–1800 (2013).

[17] G. Litjens, F. Ciompi, J. M. Wolterink, B. D. de Vos, T. Leiner, J. Teuwen, and I. Išgum, State-of-the-Art Deep Learning in Cardiovascular Image Analysis, JACC: Cardiovascular Imaging **12**, 1549–1565 (2019).

[18] D. Shen, G. Wu, and H.-I. Suk, Deep learning in medical image analysis, Annual Review of Biomedical Engeneering **19**, 221–48 (2017).

[19] I. Domingues, G. Pereira, P. Martins, H. Duarte, J. Santos, and P. H. Abreu, Using deep learning techniques in medical imaging: a systematic review of applications on CT and PET, Artificial Intelligence Review **53**, 4093–4160 (2020).

[20] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, Deep learning for chest X-ray analysis: A survey, Medical Image Analysis **72**, 102125 (2021).

[21] Y. Zhou, H. Guo, J. Song, Y. Chen, and J. Wang, Review of Vessel Segmentation and Stenosis classification in X-ray Coronary Angiography, 13th International Conference on Wireless Communications and Signal Processing, WCSP 2021 , 1–5 (2021).

[22] M. A. Molenaar, J. L. Selder, J. Nicolas, B. E. Claessen, R. Mehran, J. O. Bescós, M. J. Schuuring, B. J. Bouma, N. J. Verouden, and S. A. Chamuleau, Current state and future perspectives of artificial intelligence for automated coronary angiography imaging analysis in patients with ischemic heart disease, Current cardiology reports **24**, 365–376 (2022).

[23] W. Wu, J. Zhang, H. Xie, Y. Zhao, S. Zhang, and L. Gu, Automatic detection of coronary artery stenosis by convolutional neural network with temporal constraint, Computers in Biology and Medicine **118**, 103657 (2020).

[24] J. H. Moon, D. Y. Lee, W. C. Cha, M. J. Chung, K. S. Lee, B. H. Cho, and J. H. Choi, Automatic stenosis recognition from coronary angiography using convolutional neural networks, Computer Methods and Programs in Biomedicine **198**, 105819 (2021).

[25] C. Ciusdel, A. Turcea, A. Puiu, L. Itu, L. Calmac, E. Weiss, C. Margineanu, E. Badila, M. Berger, T. Redel, T. Passerini, M. Gulsun, and P. Sharma, Deep neural networks for ECG-free cardiac phase and end-diastolic frame detection on coronary angiographies, Computerized Medical Imaging and Graphics **84**, 101749 (2020).

[26] H. Ma, P. Ambrosini, and T. van Walsum, Fast prospective detection of contrast inflow in X-ray angiograms with convolutional neural network and recurrent neural network, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **10435 LNCS**, 453–461 (2017).

[27] C. Cong, Y. Kato, H. D. Vasconcellos, J. Lima, and B. Venkatesh, Automated Stenosis Detection and Classification in X-ray Angiography Using Deep Neural Network, Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019 , 1301–1308 (2019).

[28] R. Zhang, B. Qin, J. Zhao, Y. Zhu, Y. Lv, and S. Ding, Locating X-ray coronary angiogram keyframes via long short-term spatiotemporal attention with image-to-patch contrastive learning, IEEE Transactions on Medical Imaging (2023).

[29] T. J. Jun, J. Kweon, Y. H. Kim, and D. Kim, T-Net: Nested encoder–decoder architecture for the main vessel segmentation in coronary angiography, Neural Networks **128**, 216–233 (2020).

[30] E. Nasr-Esfahani, S. Samavi, N. Karimi, S. M. Soroushmehr, K. Ward, M. H. Jafari, B. Felfeliyan, B. Nallamothu, and K. Najarian, Vessel extraction in X-ray angiograms using deep learning, Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS **2016-Octob**, 643–646 (2016).

[31] E. Nasr-Esfahani, N. Karimi, M. H. Jafari, S. M. Soroushmehr, S. Samavi, B. K. Nallamothu, and K. Najarian, Segmentation of vessels in angiograms using convolutional neural networks, Biomedical Signal Processing and Control **40**, 240–251 (2018).

[32] S. Yang, J. Yang, Y. Wang, Q. Yang, D. Ai, and Y. Wang, Automatic coronary artery segmentation in x-ray angiograms by multiple convolutional neural networks, ACM International Conference Proceeding Series , 31–35 (2018).

[33] Z.-X. Yin and H.-M. Xu, An unsupervised image segmentation algorithm for coronary angiography, BioData Mining **15**, 1–19 (2022).

[34] Z. Xian, X. Wang, S. Yan, D. Yang, J. Chen, and C. Peng, Main Coronary Vessel Segmentation Using Deep Learning in Smart Medical, Mathematical Problems in Engineering **2020** (2020).

[35] R. Hamdi, A. Kerkeni, M. H. Bedoui, and A. Ben Abdallah, *Res-GAN: Residual Generative Adversarial Network for Coronary Artery Segmentation*, volume 13756 LNCS, Springer International Publishing, 2022.

[36] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, Densely Connected Convolutional Networks, (2016).

[37] M. Nobre Menezes, J. L. Silva, B. Silva, T. Rodrigues, C. Guerreiro, J. P. Guedes, M. O. Santos, A. L. Oliveira, and F. J. Pinto, Coronary X-ray angiography segmentation using Artificial Intelligence: a multicentric validation study of a deep learning model, International Journal of Cardiovascular Imaging (2023).

[38] J. Lourenço-Silva, M. N. Menezes, T. Rodrigues, B. Silva, F. J. Pinto, and A. L. Oliveira, Encoder-Decoder Architectures for Clinically Relevant Coronary Artery Segmentation, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **13254 LNBI**, 63–78 (2022).

[39] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2016-December**, 770–778 (2016).

[40] Z. Jiang, C. Ou, Y. Qian, R. Rehan, and A. Yong, Coronary vessel segmentation using multiresolution and multiscale deep learning, Informatics in Medicine Unlocked **24**, 100602 (2021).

[41] K. Iyer, C. P. Najarian, A. A. Fattah, C. J. Arthurs, S. M. Soroushmehr, V. Subban, M. A. Sankardas, R. R. Nadakuditi, B. K. Nallamothu, and C. A. Figueroa, AngioNet: a convolutional neural network for vessel segmentation in X-ray angiography, Scientific Reports **11**, 1–13 (2021).

[42] X. Zhu, Z. Cheng, S. Wang, X. Chen, and G. Lu, Coronary angiography image segmentation based on PSPNet, Computer Methods and Programs in Biomedicine **200**, 105897 (2021).

[43] J. Fan, J. Yang, Y. Wang, S. Yang, D. Ai, Y. Huang, H. Song, A. Hao, and Y. Wang, Multichannel fully convolutional network for coronary artery segmentation in x-ray angiograms, IEEE Access **6**, 44635–44643 (2018).

[44] L. Wang, D. xue Liang, X. lei Yin, J. Qiu, Z. yun Yang, J. hui Xing, J. zeng Dong, and Z. yuan Ma, Coronary Artery Segmentation in Angiographic Videos Using A 3D-2D CE-Net, pages 1–9 (2020).

[45] D. Hao, S. Ding, L. Qiu, Y. Lv, B. Fei, Y. Zhu, and B. Qin, Sequential vessel segmentation via deep channel attention network, Neural Networks **128**, 172–187 (2020).

[46] T. Wan, J. Chen, Z. Zhang, D. Li, and Z. Qin, Automatic vessel segmentation in X-ray angiogram using spatio-temporal fully-convolutional neural network, Biomedical Signal Processing and Control **68**, 102646 (2021).

[47] Y. Meng, Z. Du, C. Zhao, M. Dong, D. Pienta, Z. Xu, and W. Zhou, Automatic extraction of coronary arteries using deep learning in invasive coronary angiograms, Technology and health care : official journal of the European Society for Engineering and Medicine **10.3233/THC-230278.**

[48] S. Song, R. Xiao, T. Lu, and H. Zhu, Coronary Artery Segmentation from X-ray Angiographic Images using Width-aware U-Net, pages 111–114 (2023).

[49] L. Zhao, D. Li, J. Chen, and T. Wan, Automated Coronary Tree Segmentation for X-ray Angiography Sequences Using Fully-convolutional Neural Networks, VCIP 2018 - IEEE International Conference on Visual Communications and Image Processing , 1–4 (2018).

[50] R.-Q. Li, G.-B. Bian, X.-H. Zhou, X. Xie, Z.-L. Ni, and Z. Hou, CAU-net: A Novel Convolutional Neural Network for Coronary Artery Segmentation in Digital Substraction Angiography, in *Neural Information Processing*, edited by H. Yang, K. Pasupa, A. C.-S. Leung, J. T. Kwok, J. H. Chan, and I. King, pages 185–196, Cham, 2020, Springer International Publishing.

[51] C. Zhao, A. Vij, S. Malhotra, J. Tang, H. Tang, D. Pienta, Z. Xu, and W. Zhou, Automatic extraction and stenosis evaluation of coronary arteries in invasive coronary angiograms, Computers in biology and medicine **136**, 104667 (2021).

[52] B. Au, U. Shaham, S. Dhruva, G. Bouras, E. Cristea, A. L. MD, A. Coppi, F. Warner, S.-X. Li, and H. Krumholz, Automated Characterization of Stenosis in Invasive Coronary Angiography Images with Convolutional Neural Networks, (2018).

[53] Y. Gao, L. Zhang, J. Zhao, and Z. Jiang, Improved U-Net with channel and spatial attention for coronary angiography segmentation, 2022 16th ICME International Conference on Complex Medical Engineering (CME) , 123–126 (2023).

[54] P. M. Samuel and T. Veeramalai, VSSC Net: Vessel Specific Skip chain Convolutional Network for blood vessel segmentation, Computer Methods and Programs in Biomedicine **198**, 105769 (2021).

[55] X. Tao, H. Dang, X. Zhou, X. Xu, and D. Xiong, A Lightweight Network for Accurate Coronary Artery Segmentation Using X-Ray Angiograms, Frontiers in Public Health **10** (2022).

[56] R. Hamdi, A. Kerkeni, M. H. Bedoui, and A. Ben Abdallah, CAS-Net: A Novel Coronary Artery Segmentation Neural Network, ESANN 2021 Proceedings - 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning , 177–182 (2021).

[57] K. Jo, J. Kweon, Y. H. Kim, and J. Choi, Segmentation of the Main Vessel of the Left Anterior Descending Artery Using Selective Feature Mapping in Coronary Angiography, IEEE Access **7**, 919–930 (2019).

[58] Y. Zhou, W. Chang, J. Song, H. Guo, J. Wang, and Y. Chen, Semi-supervised Deep Learning of Vessel Segmentation in Coronary Angiography, 2021 6th International Conference on Signal and Image Processing, ICSIP 2021 , 194–198 (2021).

[59] H. He, A. Banerjee, M. Beetz, R. P. Choudhury, and V. Grau, Semi-Supervised Coronary Vessels Segmentation from Invasive Coronary Angiography with Connectivity-Preserving Loss Function, Proceedings - International Symposium on Biomedical Imaging **2022-March**, 1–5 (2022).

[60] H. Zhang, D. Zhang, Z. Gao, and H. Zhang, *Joint Segmentation and Quantification of Main Coronary Vessels Using Dual-Branch Multi-scale Attention Network*, volume 12901 LNCS, Springer International Publishing, 2021.

[61] F. Yu, J. Zhao, Y. Gong, Z. Wang, Y. Li, F. Yang, B. Dong, Q. Li, and L. Zhang, Annotation-Free Cardiac Vessel Segmentation via Knowledge Transfer from Retinal Images, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11765 LNCS**, 714–722 (2019).

[62] B. Ma, S. Liu, Y. Zhi, and Q. Song, Flow Based Self-supervised Pixel Embedding for Image Segmentation, (2019).

[63] S. Yang, J. Kweon, J. H. Roh, J. H. Lee, H. Kang, L. J. Park, D. J. Kim, H. Yang, J. Hur, D. Y. Kang, P. H. Lee, J. M. Ahn, S. J. Kang, D. W. Park, S. W. Lee, Y. H. Kim, C. W. Lee, S. W. Park, and S. J. Park, Deep learning segmentation of major vessels in X-ray coronary angiography, Scientific Reports **9**, 1–11 (2019).

[64] S. Mulay, K. Ram, B. Murugesan, and M. Sivaprakasam, Style Transfer based Coronary Artery Segmentation in X-ray Angiogram, 2021.

[65] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation, arXiv preprint arXiv:1802.06955 (2018).

[66] L. Mou et al., CS2-Net: Deep learning segmentation of curvilinear structures in medical imaging, Medical image analysis **67**, 101874 (2021).

[67] F. Cervantes-Sanchez, I. Cruz-Aceves, A. Hernandez-Aguirre, M. A. Hernandez-Gonzalez, and S. E. Solorio-Meza, Automatic segmentation of coronary arteries in X-ray angiograms using multiscale analysis and artificial neural networks, Applied Sciences (Switzerland) **9** (2019).

[68] I. Cruz-Aceves, F. Cervantes-Sanchez, and M. S. Avila-Garcia, A Novel Multiscale Gaussian-Matched Filter Using Neural Networks for the Segmentation of X-Ray Coronary Angiograms, Journal of Healthcare Engineering **2018** (2018).

[69] L. Busto, J. A. González-Nóvoa, P. Juan-Salvadores, V. Jiménez, A. Íñiguez, and C. Veiga, Using deep learning on X-ray orthogonal coronary angiograms for quantitative coronary analysis, in *Annual Conference on Medical Image Understanding and Analysis*, pages 869–881, Springer, 2022.

[70] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, Springer, 2015.

[71] J. Long, E. Shelhamer, and T. Darrell, Fully Convolutional Networks for Semantic Segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[72] V. Badrinarayanan, A. Kendall, and R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence **39**, 2481–2495 (2017).

[73] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, Pyramid scene parsing network, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[74] R. M. Luque-Baena, I. R. Granados, A. Jimenez-Partinen, E. J. Palomo, and M. Jimenez-Navarro, Stenosis detection in coronary angiography images using deep learning models, 2022 IEEE International Workshop on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering, MetroXRAINE 2022 - Proceedings , 22–27 (2022).

[75] T. Du, X. Liu, H. Zhang, and B. Xu, Real-time Lesion Detection of Cardiac Coronary Artery Using Deep Neural Networks, Proceedings of 2018 6th IEEE International Conference on Network Infrastructure and Digital Content, IC-NIDC 2018 , 150–154 (2018).

[76] X. Liu, T. Du, H. Zhang, and C. Sun, Detection and Classification of Chronic Total Occlusion lesions using Deep Learning, Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS , 828–831 (2019).

[77] T. Du, L. Xie, H. Zhang, X. Liu, X. Wang, D. Chen, Y. Xu, Z. Sun, W. Zhou, L. Song, C. Guan, A. J. Lansky, and B. Xu, Training and validation of a deep learning architecture for the automatic analysis of coronary angiography, EuroIntervention **17**, 32–40 (2021).

[78] V. V. Danilov, K. Y. Klyshnikov, O. M. Gerget, A. G. Kutikhin, V. I. Ganyukov, A. F. Frangi, and E. A. Ovcharenko, Real-time coronary artery stenosis detection based on modern neural networks, Scientific reports **11**, 7582 (2021), L634762455¡br/¿2021-04-19.

[79] S. A. Freitas, F. A. Zeiser, C. A. Da Costa, and G. De O. Ramos, DeepCADD: A Deep Learning Architecture for Automatic Detection of Coronary Artery Disease, Proceedings of the International Joint Conference on Neural Networks **2022-July**, 1–8 (2022).

[80] L. He, Q. Zhou, X. Li, L. Niu, G. Cheng, X. Li, W. Liu, Y. Tong, L. Ma, and L. Zhang, End-to-end video object detection with spatial-temporal transformers, in *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1507–1516, 2021.

[81] T. Han, D. Ai, X. Li, J. Fan, H. Song, Y. Wang, and J. Yang, Coronary artery stenosis detection via proposal-shifted spatial-temporal transformer in X-ray angiography, Computers in Biology and Medicine **153**, 106546 (2023).

[82] Z. Cai and N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[83] S. Ren, K. He, R. Girshick, and J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems **28**, 91–99 (2015).

[84] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, Focal loss for dense object detection, in *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[85] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, Foveabox: Beyound anchor-based object detection, IEEE Transactions on Image Processing **29**, 7389–7398 (2020).

[86] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, End-to-end object detection with transformers, in *European conference on computer vision*, pages 213–229, Springer, 2020.

[87] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, arXiv preprint arXiv:2010.04159 (2020).

[88] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, Flow-guided feature aggregation for video object detection, in *Proceedings of the IEEE international conference on computer vision*, pages 408–417, 2017.

[89] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, Deep feature flow for video recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2349–2358, 2017.

[90] H. Wu, Y. Chen, N. Wang, and Z. Zhang, Sequence level semantics aggregation for video object detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9217–9225, 2019.

[91] T. Gong, K. Chen, X. Wang, Q. Chu, F. Zhu, D. Lin, N. Yu, and H. Feng, Temporal ROI align for video object recognition, in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1442–1450, 2021.

[92] E. Ovalle-Magallanes, J. G. Avina-Cervantes, I. Cruz-Aceves, and J. Ruiz-Pinales, Hybrid classical–quantum Convolutional Neural Network for stenosis detection in X-ray coronary angiography, Expert Systems with Applications **189**, 116112 (2022).

[93] K. Pang, D. Ai, H. Fang, J. Fan, H. Song, and J. Yang, Stenosis-DetNet: Sequence consistency-based stenosis detection for X-ray coronary angiography, Computerized Medical Imaging and Graphics **89**, 101900 (2021).

[94] D. Thanou, O. Senouf, O. Raita, E. Abbé, P. Frossard, F. Aminfar, D. Auberson, N. Dayer, D. Meier, M. Pagnoni, O. Muller, S. Fournier, and T. Mahendiran, Predicting future myocardial infarction from angiographies with deep learning, pages 1–4 (2021).

[95] H. Yabushita, S. Goto, S. Nakamura, H. Oka, M. Nakayama, and S. Goto, Development of novel artificial intelligence to detect the presence of clinically meaningful coronary atherosclerotic stenosis in major branch from coronary angiography video, Journal of Atherosclerosis and Thrombosis **28**, 835–843 (2021).

[96] S. Chen, Y. Tang, X. Shi, H. Zhang, L. Xie, and B. Xu, Convolution Pyramid Network: A Classification Network on Coronary Artery Angiogram Images, Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS **2020-July**, 1186–1189 (2020).

[97] H. Cho, J. G. Lee, S. J. Kang, W. J. Kim, S. Y. Choi, J. Ko, H. S. Min, G. H. Choi, D. Y. Kang, P. H. Lee, J. M. Ahn, D. W. Park, S. W. Lee, Y. H. Kim, C. W. Lee, S. W.

Park, and S. J. Park, Angiography-based machine learning for predicting fractional flow reserve in intermediate coronary artery lesions, Journal of the American Heart Association **8** (2019).

[98] D. Zhang, G. Yang, S. Zhao, Y. Zhang, H. Zhang, and S. Li, Direct Quantification for Coronary Artery Stenosis Using Multiview Learning, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11765 LNCS**, 449–457 (2019).

[99] I.-d. Sievering, D. Nanchen, E. Abbé, O. Muller, and P. Frossard, Anatomy-informed multimodal learning for myocardial infarction prediction, (2022).

[100] T. Mahendiran, D. Thanou, O. Senouf, D. Meier, N. Dayer, F. Aminfar, D. Auberson, O. Raita, P. Frossard, M. Pagnoni, S. Cook, B. De Bruyne, O. Muller, E. Abbé, and S. Fournier, Deep learning-based prediction of future myocardial infarction using invasive coronary angiography: a feasibility study, Open Heart **10**, e002237 (2023).

[101] L. Liu, F. Ding, Y. Shen, S. Tu, J. Yang, Q. Zhao, M. Chu, W. Shen, R. Zhang, M. Zimarino, G. S. Werner, and J. L. Gutiérrez-Chico, Automatic assessment of collateral physiology in chronic total occlusions by means of artificial intelligence, Cardiology Journal (2022).

[102] H. Fang, H. Li, S. Song, K. Pang, D. Ai, J. Fan, H. Song, Y. Yu, and J. Yang, Motion-flow-guided recurrent network for respiratory signal estimation of x-ray angiographic image sequences, Physics in Medicine and Biology **65** (2020).

[103] R. Royer-Rivard, F. Girard, N. Dahdah, and F. Cheriet, End-to-End Deep Learning Model for Cardiac Cycle Synchronization from Multi-View Angiographic Sequences, Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS **2020-July**, 1190–1193 (2020).

[104] I. T. Moon, S.-H. Kim, J. Y. Chin, S. H. Park, C.-H. Yoon, T.-J. Youn, I.-H. Chae, and S.-H. Kang, Accuracy of Artificial Intelligence–Based Automated Quantitative Coronary Angiography Compared to Intravascular Ultrasound: Retrospective Cohort Study, JMIR cardio **7**, e45299 (2023).

[105] J. Haase, J. Escaned, E. M. van Swijndregt, Y. Ozaki, E. Gronenschild, C. J. Slager, and P. W. Serruys, Experimental validation of geometric and densitometric coronary measurements on the new generation cardiovascular angiography analysis system (caas ii), Catheterization and Cardiovascular Diagnosis **30**, 104–114 (1993).

[106] J. J. Fischer, H. Samady, J. A. McPherson, I. J. Sarembock, E. R. Powers, L. W. Gimple, and M. Ragosta, Comparison between visual assessment and quantitative angiography versus fractional flow reserve for native coronary narrowings of moderate severity, American Journal of Cardiology **90**, 210–215 (2002).

[107] K. Masdjedi, L. J. van Zandvoort, M. M. Balbi, R.-J. Nuis, J. Wilschut, R. Diletti, P. P. de Jaegere, F. Zijlstra, N. M. Van Mieghem, and J. Daemen, Validation of novel 3-dimensional quantitative coronary angiography based software to calculate fractional flow reserve post stenting, Catheterization and Cardiovascular Interventions **98**, 671–677 (2021).

[108] A. Erbay, J. Steiner, A. Lauten, U. Landmesser, D. M. Leistner, and B. E. Stähli, Assessment of intermediate coronary lesions by fractional flow reserve and quantitative flow ratio in patients with small-vessel disease, Catheterization and Cardiovascular Interventions **96**, 743–751 (2020).

[109] B. Xu et al., Angiographic quantitative flow ratio-guided coronary intervention (FAVOR III China): a multicentre, randomised, sham-controlled trial, The Lancet **398**, 2149–2159 (2021).

[110] M. R. Patel, E. D. Peterson, D. Dai, J. M. Brennan, R. F. Redberg, H. V. Anderson, R. G. Brindis, and P. S. Douglas, Low diagnostic yield of elective coronary angiography, New England Journal of Medicine **362**, 886–895 (2010).

[111] J. D. Schuijf, M. B. Matheson, M. R. Ostovaneh, A. Arbab-Zadeh, K. F. Kofoed, A. J. H. A. Scholte, M. Dewey, C. Steveson, C. E. Rochitte, K. Yoshioka, C. Cox, M. F. Di Carli, and J. A. C. Lima, Ischemia and No Obstructive Stenosis (INOCA) at CT Angiography , CT Myocardial Perfusion , Invasive Coronary Angiography and SPECT, Radiology **294**, 61–73 (2020).

[112] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, *Object Recognition with Gradient-Based Learning*, pages 319–345, Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[113] S. Hochreiter and J. Schmidhuber, Long Short-term Memory, Neural computation **9**, 1735–1780 (1997).

[114] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks, Communications of the ACM **63**, 139–144 (2020).

[115] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[116] W. Wang, D. Ferrari, G. Haddon-Hill, and V. Curcin, Electronic health records as source of research data, in *Machine Learning for Brain Disorders*, Neuromethods, pages 331–354, Springer US, New York, NY, 2023.

[117] K. Roberts, A. T. Chin, K. Loewy, L. Pompeii, H. Shin, and N. L. Rider, Natural language processing of clinical notes enables early inborn error of immunity risk ascertainment, J. Allergy Clin. Immunol. Glob. **3**, 100224 (2024).

[118] C. A. Lovejoy, A. Arora, V. Buch, and I. Dayan, Key considerations for the use of artificial intelligence in healthcare and clinical research, Future Healthc. J. **9**, 75–78 (2022).

[119] S. C Pereira, A. M. Mendonça, A. Campilho, P. Sousa, and C. Teixeira Lopes, Automated image label extraction from radiology reports - A review, Artif. Intell. Med. **149**, 102814 (2024).

[120] G. L. Nicolosi, Artificial Intelligence in Cardiology: Why so many great promises and expectations, but still a limited clinical impact?, J. Clin. Med. **12** (2023).

[121] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, A Comprehensive Overview of Large Language Models, (2024).

[122] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, page 6000–6010 (2017).

[123] J. Zech, M. Pain, J. Titano, M. Badgeley, J. Schefflein, A. Su, A. Costa, J. Bederson, J. Lehar, and E. K. Oermann, Natural language–based machine learning models for the annotation of clinical radiology reports, Radiology **287**, 570–580 (2018).

[124] B. Meskó, Prompt engineering as an important emerging skill for medical professionals: Tutorial, J. Med. Internet Res. **25**, e50638 (2023).

[125] S. Sivarajkumar, M. Kelley, A. Samolyk-Mazzanti, S. Visweswaran, and Y. Wang, An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study, JMIR Med Inform **12**, e55318 (2024).

[126] R. Ramesh, A. T. Raju, H. V. Reddy, and S. Varma, Fine-tuning large language models for task specific data, pages 1–6 (2024).

[127] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, Multilingual E5 Text Embeddings: A Technical Report, arXiv preprint arXiv:2402.05672 (2024).

[128] S. Bannur, S. Hyland, Q. Liu, F. Perez-Garcia, M. Ilse, D. C. Castro, B. Boecking, H. Sharma, K. Bouzid, A. Thieme, A. Schwaighofer, M. Wetscherek, M. P. Lungren, A. Nori, J. Alvarez-Valle, and O. Oktay, Learning to Exploit Temporal Structure for Biomedical Vision–Language Processing, 2023.

[129] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, and Z. Lu, MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval, Bioinformatics **39**, btad651 (2023).

[130] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, CoRR **abs/1910.13461** (2019).

[131] V. Aryabumi et al., Aya 23: Open Weight Releases to Further Multilingual Progress, 2024.

[132] Z. Chen, A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, A. Sallinen, A. Sakhaeirad, V. Swamy, I. Krawczuk, D. Bayazit, A. Marmet, S. Montariol, M.-A. Hartley, M. Jaggi, and A. Bosselut, MEDITRON-70B: Scaling Medical Pretraining for Large Language Models, 2023.

[133] K. Singhal et al., Large language models encode clinical knowledge, Nature **620**, 172–180 (2023).

[134] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, Well-Read Students Learn Better: On the Importance of Pre-training Compact Models, 2019.

[135] M. J. Timmermans, S. Houterman, E. D. Daeter, P. W. Danse, W. W. Li, E. Lipsic, M. M. Roefs, D. van Veghel, P. R. C. of the Netherlands Heart Registration, and the Cardiothoracic Surgery Registration Committee of the Netherlands Heart Registration, Using real-world data to monitor and improve quality of care in coronary artery disease: results from the Netherlands Heart Registration, Netherlands Heart Journal **30**, 546–556 (2022).

[136] M. M. Balbi, P. Scarparo, M. N. Tovar, K. Masdjedi, J. Daemen, W. Den Dekker, J. Ligthart, K. Witberg, P. Cummins, J. Wilschut, F. Zijlstra, N. M. Van Mieghem, and R. Diletti, Culprit Lesion Detection in Patients Presenting With Non-ST Elevation Acute Coronary Syndrome and Multivessel Disease, Cardiovascular Revascularization Medicine **35**, 110–118 (2022).

[137] DeepSeek-AI et al., DeepSeek-V3 Technical Report, 2025.

[138] G. D. Clifford, C. Liu, B. Moody, H. L. Li-wei, I. Silva, Q. Li, A. Johnson, and R. G. Mark, AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017, in *Computing in Cardiology (CinC)*, pages 1–4, 2017.

[139] E. A. P. Alday et al., Classification of 12-lead ECGs: the PhysioNet/computing in cardiology challenge 2020, Physiological measurement **41**, 124003 (2020).

[140] R. R. van de Leur et al., Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders, European Heart Journal-Digital Health **3**, 390–404 (2022).

[141] T. Van Steenkiste, D. Deschrijver, and T. Dhaene, Interpretable ECG beat embedding using disentangled variational auto-encoders, in *IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pages 373–378, 2019.

[142] J. H. Jang, T. Y. Kim, H. S. Lim, and D. Yoon, Unsupervised feature learning for electrocardiogram data using the convolutional variational autoencoder, PLoS ONE **16**, 1–16 (2021).

[143] V. V. Kuznetsov, V. A. Moskalenko, and N. Y. Zolotykh, Electrocardiogram Generation and Feature Extraction Using a Variational Autoencoder, arXiv , 1–6 (2020).

[144] S. Basu, K. Wagstyl, A. Zandifar, L. Collins, A. Romero, and D. Precup, Early Prediction of Alzheimer's Disease Progression Using Variational Autoencoders, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, edited by D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, pages 205–213, Cham, 2019, Springer International Publishing.

[145] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, in *International Conference on Learning Representations*, 2017.

[146] S. Vijayarangan, R. Vignesh, B. Murugesan, S. Preejith, J. Joseph, and M. Sivaprakasam, RPnet: A Deep Learning approach for robust R Peak detection in noisy ECG, in *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 345–348, 2020.

[147] A. Paszke et al., PyTorch: An Imperative Style, High-Performance Deep Learning Library, in *Advances in Neural Information Processing Systems*, pages 8024–8035, Curran Associates, Inc., 2019.

[148] NVIDIA, P. Vingelmann, and F. H. Fitzek, CUDA, release: 10.2.89, 2020.

[149] F. Pedregosa et al., Scikit-learn: Machine learning in Python, Journal of machine learning research **12**, 2825–2830 (2011).

[150] K. He, X. Zhang, S. Ren, and J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[151] J. A. Hanley and B. J. McNeil, A method of comparing the areas under receiver operating characteristic curves derived from the same cases., Radiology **148**, 839–843 (1983).

[152] L. A. Jeni, J. F. Cohn, and F. De La Torre, Facing imbalanced data–recommendations for the use of performance metrics, in *2013 Humaine association conference on affective computing and intelligent interaction*, pages 245–251, IEEE, 2013.

[153] E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh, Disentangling disentanglement in variational autoencoders, in *International Conference on Machine Learning*, pages 4402–4412, PMLR, 2019.

[154] R. R. van de Leur et al., Inherently explainable deep neural network-based interpretation of electrocardiograms using variational auto-encoders, medRxiv (2022).

[155] V. van der Valk, D. Atsma, R. Scherptong, and M. Staring, Joint Optimization of a beta-VAE for ECG Task-Specific Feature Extraction, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, edited by H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, pages 554–563, Cham, 2023, Springer Nature Switzerland.

[156] T. Karras, S. Laine, and T. Aila, A Style-Based Generator Architecture for Generative Adversarial Networks , IEEE Transactions on Pattern Analysis & Machine Intelligence **43**, 4217–4228 (2021).

[157] D. P. Kingma and M. Welling, Auto-Encoding Variational Bayes, CoRR **abs/1312.6114** (2013).

[158] P. H. Le-Khac, G. Healy, and A. F. Smeaton, Contrastive representation learning: A framework and review, Ieee Access **8**, 193907–193934 (2020).

[159] P. Chormai, J. Herrmann, K.-R. Müller, and G. Montavon, Disentangled explanations of neural network predictions by finding relevant subspaces, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[160] Y. Skandarani, A. Lalande, J. Afilalo, and P.-M. Jodoin, Generative adversarial networks in cardiology, Canadian Journal of Cardiology **38**, 196–203 (2022).

[161] C. T. Wei, M.-E. Hsieh, C.-L. Liu, and V. S. Tseng, Contrastive Heartbeats: Contrastive Learning for Self-Supervised ECG Representation and Phenotyping, in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1126–1130, 2022.

[162] O. Atamny, A. Saguner, R. Abaecherli, and E. Konukoglu, Outlier Detection in ECG, in *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4, IEEE, 2023.

[163] Y. Cho, J.-m. Kwon, K.-H. Kim, J. R. Medina-Inojosa, K.-H. Jeon, S. Cho, S. Y. Lee, J. Park, and B.-H. Oh, Artificial intelligence algorithm for detecting myocardial infarction using six-lead electrocardiography, Scientific reports **10**, 20495 (2020).

[164] Y. M. Ayano, F. Schwenker, B. D. Dufera, and T. G. Debelee, Interpretable machine learning techniques in ECG-based heart disease classification: a systematic review, Diagnostics **13**, 111 (2022).

[165] P. Wagner, T. Mehari, W. Haverkamp, and N. Strodthoff, Explaining deep learning for ecg analysis: Building blocks for auditing and knowledge discovery, Computers in Biology and Medicine **176**, 108525 (2024).

[166] H. Liu, Z. Zhao, X. Chen, R. Yu, and Q. She, Using the VQ-VAE to improve the recognition of abnormalities in short-duration 12-lead electrocardiogram records, Computer Methods and Programs in Biomedicine **196**, 105639 (2020).

[167] A. Pagnoni, K. Liu, and S. Li, Conditional Variational Autoencoder for Neural Machine Translation, arXiv preprint arXiv:1812.04405 (2018).

[168] Q. Zhu and R. Zhang, A classification supervised auto-encoder based on predefined evenly-distributed class centroids, arXiv preprint arXiv:1902.00220 (2019).

[169] A. Pati and A. Lerch, Attribute-based regularization of latent spaces for variational auto-encoders, Neural Computing and Applications **33**, 4429–4444 (2021).

[170] J. Klys, J. Snell, and R. Zemel, Learning latent subspaces in variational autoencoders, Advances in neural information processing systems **31** (2018).

[171] Y. Xia, W. Wang, and K. Wang, ECG signal generation based on conditional generative models, Biomedical Signal Processing and Control **82**, 104587 (2023).

[172] E. Adib, F. Afghah, and J. J. Prevost, Synthetic ecg signal generation using generative neural networks, arXiv preprint arXiv:2112.03268 (2021).

[173] Y. Sang, M. Beetz, and V. Grau, Generation of 12-lead electrocardiogram with subject-specific, image-derived characteristics using a conditional variational autoencoder, in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, IEEE, 2022.

[174] M. Beetz, A. Banerjee, and V. Grau, Multi-domain variational autoencoders for combined modeling of MRI-based biventricular anatomy and ECG-based cardiac electrophysiology, Frontiers in physiology **13**, 886723 (2022).

[175] P. K. Gyawali, B. M. Horacek, J. L. Sapp, and L. Wang, Learning disentangled representation from 12-lead electrograms: application in localizing the origin of Ventricular Tachycardia, arXiv preprint arXiv:1808.01524 (2018).

[176] A. Galdran, Performance Metrics for Probabilistic Ordinal Classifiers, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 357–366, Springer, 2023.

[177] D. R. Cox, Regression models and life-tables, Journal of the Royal Statistical Society: Series B (Methodological) **34**, 187–202 (1972).

[178] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics , 837–845 (1988).

[179] L. Gaudette and N. Japkowicz, Evaluation methods for ordinal classification, in *Advances in Artificial Intelligence: 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009 Kelowna, Canada, May 25-27, 2009 Proceedings 22*, pages 207–210, Springer, 2009.

[180] C. C. Tan, T. M. Hiew, and B. Chia, Right chest electrocardiographic patterns in normal subjects, Chest **97**, 572–575 (1990).

[181] E. Baiocco, D. Contadini, A. Maolo, and M. V. Matassini, Early Repolarization: When Is It a Normal Pattern?, New Concepts in ECG Interpretation , 111–117 (2019).

[182] P. Ahmadi et al., Age and gender differences of basic electrocardiographic values and abnormalities in the general adult population; Tehran Cohort Study, BMC Cardiovascular Disorders **23**, 303 (2023).

[183] M. M. Saad, R. O'Reilly, and M. H. Rehmani, A survey on training challenges in generative adversarial networks for biomedical image analysis, Artificial Intelligence Review **57**, 19 (2024).

[184] M. Blanco-Velasco, B. Weng, and K. E. Barner, ECG signal denoising and baseline wander correction based on the empirical mode decomposition, Computers in biology and medicine **38**, 1–13 (2008).

[185] K. Rajakariar, A. N. Koshy, J. K. Sajeev, S. Nair, L. Roberts, and A. W. Teh, Accuracy of a smartwatch based single-lead electrocardiogram device in detection of atrial fibrillation, Heart **106**, 665–670 (2020).

[186] S. A. Lubitz, A. Z. Faranesh, C. Selvaggi, S. J. Atlas, D. D. McManus, D. E. Singer, S. Pagoto, M. V. McConnell, A. Pantelopoulos, and A. S. Foulkes, Detection of atrial fibrillation in a large population using wearable devices: the Fitbit Heart Study, Circulation **146**, 1415–1424 (2022).

[187] M. P. Turakhia et al., Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple Heart Study, American heart journal **207**, 66–75 (2019).

[188] M. Zabihi, A. B. Rad, A. K. Katsaggelos, S. Kiranyaz, S. Narkilahti, and M. Gabbouj, Detection of Atrial Fibrillation in ECG Signals Using a Hybrid Deep Learning Approach, Computing in Cardiology **44**, 1–4 (2017).

[189] M. Neyazi, J. Bremer, M. Knorr, J. Brederecke, F. Ojeda, S. Blankenberg, and R. Schnabel, Prediction of incident heart failure through deep learning-based analysis of 1-lead ECGs, European Heart Journal **44**, ehad655–900 (2023).

[190] L. Holmstrom et al., Deep learning-based electrocardiographic screening for chronic kidney disease, Communications Medicine **3**, 73 (2023).

[191] G. Ott, Y. Schaubelt, J. M. Lopez Alcaraz, W. Haverkamp, and N. Strodthoff, Using explainable AI to investigate electrocardiogram changes during healthy aging—From expert features to raw signals, Plos one **19**, e0302024 (2024).

[192] B. T. Jin, R. Palleti, S. Shi, A. Y. Ng, J. V. Quinn, P. Rajpurkar, and D. Kim, Transfer learning enables prediction of myocardial injury from continuous single-lead electrocardiography, Journal of the American Medical Informatics Association **29**, 1908–1918 (2022).

[193] C. M. Gibson et al., Evolution of single-lead ECG for STEMI detection using a deep learning approach, International journal of cardiology **346**, 47–52 (2022).

[194] A. Khunte, V. Sangha, E. K. Oikonomou, L. S. Dhingra, A. Aminorroaya, B. J. Mortazavi, A. Coppi, C. A. Brandt, H. M. Krumholz, and R. Khera, Detection of left ventricular systolic dysfunction from single-lead electrocardiography adapted for portable and wearable devices, npj Digital Medicine **6**, 124 (2023).

[195] D. Zhang, S. Yang, X. Yuan, and P. Zhang, Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram, Iscience **24** (2021).

[196] Z. I. Attia et al., Prospective evaluation of smartwatch-enabled detection of left ventricular dysfunction, Nature medicine **28**, 2497–2503 (2022).

[197] R. W. Treskes, M. E. van den Akker-van Marle, L. van Winden, N. van Keulen, E. T. van der Velde, S. Beeres, D. Atsma, and M. J. Schalij, The Box—eHealth in the outpatient clinic follow-up of patients with acute myocardial infarction: cost-utility analysis, Journal of Medical Internet Research **24**, e30236 (2022).

[198] D. Campo et al., Atrial fibrillation detection with an analog smartwatch: prospective clinical study and algorithm validation, JMIR Formative Research **6**, e37280 (2022).

[199] X. Chen, P.-O. Hansson, E. Thunström, Z. Mandalenakis, K. Caidahl, and M. Fu, Incremental changes in QRS duration as predictor for cardiovascular disease: a 21-year follow-up of a randomly selected general population, Scientific Reports **11**, 13652 (2021).

[200] M. M. Kylmälä, T. Konttila, P. Vesterinen, M. Lindholm, H. Väänänen, M. Stenroos, M. S. Nieminen, H. Hänninen, and L. Toivonen, Predicting recovery of myocardial

function by electrocardiography after acute infarction, Annals of Noninvasive Electrocardiology **18**, 230–239 (2013).

[201] H. Yokoyama, H. Tomita, F. Nishizaki, K. Hanada, S. Shibutani, M. Yamada, N. Abe, T. Higuma, T. Osanai, and K. Okumura, Deeply Reinverted T Wave at 14 Days After the Onset of First Anterior Acute Myocardial Infarction Predicts Improved Left Ventricular Function at 6 Months, Clinical Cardiology **38**, 157–163 (2015).

[202] W. van der Loo, V. van der Valk, T. van den Broek, D. Atsma, M. Staring, and R. Scherptong, Large Language Models for Structured Cardiovascular Data Extraction: A Foundation for Scalable Research and Clinical Applications, European Heart Journal-Digital Health , ztaf127 (2025).

[203] S. Lundberg and S.-I. Lee, A Unified Approach to Interpreting Model Predictions, 2017.

Publications

Journal articles

Viktor O. Van der Valk, Wouter Van der Loo, Douwe E. Atsma, Jesper Hjortnaes, Marius Staring, Roderick Scherptong, (2025). The current state of deep learning based invasive coronary angiography image analysis: a systematic review. (in progress)

Wouter P.M. Van der Loo, **Viktor O. Van der Valk**, Douwe E. Atsma, Marius Staring, Roderick W.C. Scherptong, (2025). Large Language Models for Structured Cardiovascular Data Extraction: A Foundation for Scalable Research and Clinical Applications. European Heart Journal-Digital Health.

Viktor O. Van der Valk, Douwe E. Atsma, Roderick W.C. Scherptong, Marius Staring, (2025). Explainable ECG analysis by explicit information disentanglement with VAEs. IEEE Transactions on Bio-Medical Engineering.

Viktor O. Van der Valk, Douwe E. Atsma, Roderick W.C. Scherptong, Marius Staring (2026). Leveraging the wearable 1-lead ECG signal: From cardiac rhythm to cardiac function assessment. (in progress)

Tomás Domingo-Gardeta, José M. Montero-Cabezas, Alfonso Jurado-Román, Manel Sabaté, Jaime Aboal, Adrián Baranchuk, Xavier Carrillo, Sebastián García-Zamora, Hélder Dores, **Viktor O. van der Valk**, Roderick W.C. Scherptong, Joan F. Andrés-Cordón, Pablo Vidal, Daniel Moreno-Martínez, Raquel Toribio-Fernández, José María Lillo-Castellano, Roberto Cruz, François De Guio, Manuel Marina-Breysse, Manuel Martínez-Sellés, (2024). Rationale and design of the artificial intelligence scalable solution for acute myocardial infarction (ASSIST) study, Journal of Electrocardiology, 86.

Alba González-Cebrián, Pablo García-Cristóbal, Fernando Galve, Efe Ilıcak, **Viktor O. Van Der Valk**, Marius Staring, Andrew Webb, Joseba Alonso, (2025). An imageless magnetic resonance framework for fast and cost-effective decision-making, ArXiv preprint, 2505.04550

Giuseppe D'ancona, Mattia Savardi, Mauro Massussi, **Viktor O. Van Der Valk**, Alberto Signoroni, Davide Farina, Hüseyin Ince, Stefano Benussi, Salvatore Curello, Fatih Arslan, (2023). Deep learning to predict long-term mortality from plain chest radiographs in patients referred for suspected angina. European Heart Journal, 44(2).

Giuseppe D'Ancona, Mattia Savardi, Mauro Massussi, **Viktor O. Van Der Valk**, Roderick W. C. Scherptong, Alberto Signoroni, Davide Farina, Monica Murero, Hüseyin Ince, Stefano Benussi, Salvatore Curello, Fatih Arslan, (2024). Deep learning to predict long-term mortality from plain chest X-ray in patients referred for suspected coronary artery disease. Journal of Thoracic Disease, 16(8), 4914-4923.

Conference proceedings

Viktor O. Van der Valk, Douwe E. Atsma, Roderick W.C. Scherptong, Marius Staring, (2023). Joint Optimization of a β-VAE for ECG Task-Specific Feature Extraction. International Conference on Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science.

Viktor O. Van der Valk, Wouter P.M. van der Loo, Douwe E. Atsma, Marius Staring, Roderick W.C. Scherptong, (2025). Towards expert-level AI-based classification of coronary angiography reports: a comparison of LLM approaches. European Society of Cardiology Congress 2025, 83722.

Wouter P.M. van der Loo, **Viktor O. Van der Valk**, Douwe E. Atsma, Marius Staring, Roderick W.C. Scherptong, (2025). Towards expert-level AI-based classification of echocardiography reports: a comparison of LLM approaches. European Society of Cardiology Congress 2025, 83698.

Yicong Tan, Prerak Mody, **Viktor O. van der Valk**, Marius Staring, Jan van Gemert, (2023). Analyzing components of a transformer under different dataset scales in 3D prostate CT segmentation, SPIE Medical Imaging, 12464.

Konstantinos Ntatsis, Niels Dekker, **Viktor O. van der Valk**, Tom Birdsong, Dženan Zukic, Stefan Klein, Marius Staring, Matthew McCormick, (2023). itk-elastix: Medical image registration in Python, Python in Science Conference 2023.

Acknowledgments

First and foremost, I want to thank Marius Staring for supporting and allowing me to work on my PhD in a way that suited me best. Allowing me to work remotely in winter, pushing me to explore education in the form of giving lectures and making exams, helping me find research topics that were closer to my interests, and being the opposite of a micromanager as a supervisor, helped me a lot during my PhD. Along the same lines, I want to thank Boudewijn Lelieveldt, who, as Head of LKEB and as supervisor of my postdoctoral research position, also gave me a lot of freedom and responsibility to work in a way that suited me. I appreciate the effort he took to make sure my PhD could continue after the funding almost overnight stopped, and the help he provided when I was looking for a change in research topic, on roughly 3/4 of my PhD. I also want to thank Roderick Scherptong for introducing me to cardiac research and all the conferences and collaborations that entailed, and for all the scientific and sometimes personal and philosophical discussions we had, in which humor and making fun of each other had a place. I want to thank Douwe Atsma for facilitating and shaping my PhD research in the Cardiology department. I want to thank all members of the LKEB. Michèle and Patrick for providing all the hardware and software help I needed. All PI's and especially Silvia Pintea, for answering all my questions about the academic world, for brainstorming, and for giving much-needed feedback on my work. All (former) PhD students, in particular Qiuyu, Ruochen, Faeze, Efe, Baris, Chang, Li-Hsin, Yunjie, Fredrik and Jenia for the informal discussions about food, culture and life, Chinmay Rao and Alexander Vieth for the collaborations and scientific and informal discussions, and especially Prerak Mody for all of the above and for paving the road of my PhD career. All master students I supervised, for the intellectual discussions and for helping me understand a part of the cardiac workflow. And last but definitely not least, I want to thank Wouter van der Loo for collaborating on multiple papers and presentations in which he provided the cardiac and clinical knowledge I lacked.

Then, outside of the academic world, I want to thank my parents, Joke and Wim, for

their mental support and for always being ready to listen to my PhD plans, insights, and complaints, even if I was not ready to talk about them. And for providing a very warm home, especially while growing up, but still available to return to afterward. I want to thank my sister, Esther, for both providing mental support, joining me on adventures, and designing images that I, even with the help of AI, could not have made. And I want to thank Hilde, Nico and Wies for the support and interest they showed in my PhD research.

I want to thank Nila Overbeek and Sofieke Temmink for listening to and sharing PhD insights, struggles, and complaints over almost weekly walks in the parks.

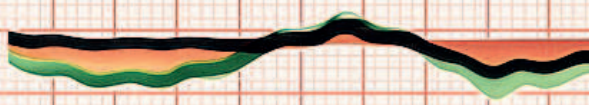
I want to thank Joe, Sergio, Mika, Robert, Jim, Bas, Pieter, Drim, Rinse, and Daniel for the endless jokes, good conversations, and adventures we had over many surftrips, festivals, and dinners, that inspired me, helped me enjoy life, and took my mind off PhD topics. I want to thank Aad, Noud, Joep, Maarten, Ronen, Floortje, Neeltje, Sofie, Roos, Wouter, Jakob and Amy for inspiring me with ideas and ways of living, but also for the support and joy I had during holidays, in bars, and over food. And I want to thank Vera, Jolie, Guusje, Yara, Imme, and Hidde for the personal conversations, adventures, and reflective moments that helped me a lot with my personal development, but also for the endless jokes in between.

I also want to thank Marit Prins for being an enthusiastic supporter of my PhD journey at the start. And along different lines, I want to thank Florence Mooij for always being open to listen to my struggles both in my personal and academic life at the end of my PhD journey. And last, I want to thank the whole Volkskeuken crew for all the cooking sessions, which helped me relax and enjoy after work. In this regard, I want to give a special thanks to Liselotte Bredius and Hans van den Broek for being the two most inspirational entrepreneurs I have met so far. They showed me that there is a different way to do business in a capitalist society.

Curriculum Vitae

Viktor van der Valk was born in Den Haag, the Netherlands, on the very odd date of 19-03-1993. As a true beta scholar, after graduating cum laude from the Vrijzinnig Christelijk Lyceum (VCL) in Scheveningen, he went to study Mechanical Engineering at the University of Delft (TU Delft). After realizing that his true interest was more in the biology of the human mind, he switched to obtain a Bachelor's degree in Neurobiology at the University of Amsterdam (UvA) in 2016. After a year of traveling, a deep dive in Vipassana Meditation and Portuguese and Spanish language courses, he found a way to combine his skill in Mathematics and his passion for the human body and mind in the Master's degree Computational Science (Biology track) at the UvA. This Master degree encompassed courses from deep learning to statistical modeling and systems biology, mostly applied to the field of biology. The final 9 months of the master he did a research internship at De Nederlandse Organisatie voor toegepast-natuurwetenschappelijk onderzoek (TNO) in system dynamic modeling of complex disease (obesity) in which he combined classical AI models with systems biology and modeling. After obtaining this degree cum laude in 2019 and a brief period of freelancing as a python-web developer, he started as a scientific programmer at the Laboratorium voor Klinische en Experimentele Beeldverwerking (LKEB) in 2020 in Leiden. As a scientific programmer, he helped to maintain and extend the (python wrap of the) open-source image registration package Elastix and gained relevant knowledge in the field of image processing. This position slowly transitioned into a PhD in Explainable AI in Cardiology which he started in 2021. His PhD research focused on practical applications of explainable AI in the cardiac workflow, in specific ECG analysis with deep learning and cardiac report classification with LLMs. During his PhD he not only focused on research but also got an introduction in education in the form of supervising Master students, giving lectures, and preparing exams. Towards the end of his PhD he started a position as postdoctoral researcher in Dermatology, in specific: data analysis of the skin and gut microbiome of skin patients. Suffering from a very mild form of urticaria, a skin disease, he

had become fascinated by the microbiome and its influence on, amongst others, the skin. This position was the perfect match to combine genuine interest with skill and personal relevance. Besides a technical researcher, Viktor is a passionate surfer and vegan chef. Two hobbies which he practices after work hours respectively in the north sea or abroad, and at the Volkskeuken in Den Haag.



Leiden University
Medical Center